UNIVERSITÉ TOULOUSE III - PAUL SABATIER

UFR SVT - Sciences de la vie et de la terre -

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ TOULOUSE III

Discipline : Neurosciences

présentée et soutenue

par

Marc J-M. Macé

le 3 mai 2006

Titre :

Représentations visuelles précoces dans la catégorisation rapide de scènes naturelles chez l'homme et le singe.

JURY

Pr. Pier-Giorgio Zanone - LAMPA - U. Paul Sabatier, Toulouse	**Président**
Dr. Sylvie Chokron - LPNC - CNRS, U. Mendès-France, Grenoble	**Rapporteur**
Dr. Pierre-Paul Vidal - LNRS - CNRS, U. Paris 5, Paris	**Rapporteur**
Pr. Rufin Vogels - KUL, Leuven	**Examinateur**
Dr. Catherine Tallon-Baudry - LENA - CNRS, Paris	**Examinateur**
Dr. Guillaume Masson - INCM - CNRS, U. de la Méditerranée, Marseille	**Examinateur**
Dr. Michèle Fabre-Thorpe - CERCO - CNRS, UPS, Toulouse	**Directrice de thèse**

Centre de recherche CERveau et COgnition (CerCo)
UMR 5549 CNRS - Université Paul Sabatier
Faculté de médecine de Rangueil
31062 Toulouse Cedex 9

Remerciements

Je tiens tout d'abord à remercier Michèle Fabre-Thorpe qui a encadré ce travail et m'a guidé depuis mes premiers pas dans le monde de la recherche. Son amitié et son soutien permanents ont été très précieux, tout comme sa disponibilité et … son souci du détail dans les corrections ! Grâce à toi Michèle, je serai bientôt écrivain ! :-)

Merci à Jean Bullier de m'avoir accueilli dans son laboratoire lorsque j'étais en DEA puis en thèse. Nous nous sommes plus rapprochés après ton accident, et je suis très heureux que cette amitié se poursuive.

Merci à Simon Thorpe pour son enthousiasme. Son imagination scientifique débordante est parfois difficile à canaliser, mais qu'est ce qu'on est content d'avoir notre savant fou dans l'équipe !

Merci aux rapporteurs, Sylvie Chokron et Pierre-Paul Vidal et aux examinateurs, Rufin Vogels, Catherine Tallon-Baudry et Guillaume Masson d'avoir accepté de donner de leur temps pour évaluer ce travail. Merci également à Pier-Giorgio Zanone d'avoir accepté de présider ce jury.

Guillaume Rousselet tient une place particulière dans ce travail. Difficile de décrire Guillaume, à la fois exubérant dans la vie et rigoureux en sciences. Nous avons passé des moments formidables pendant mon année de DEA et ma première année de thèse et je dois dire que ça n'a plus été pareil après son départ. Ce n'est pas grave Gibbon, on se retrouvera !

Merci à Rufin VanRullen pour les discussions scientifiques passionnantes et les analyses de données à s'arracher les cheveux : il faut le suivre l'animal !

Merci à tous les étudiants du laboratoire que j'ai rencontré au cours de ces (trop ?) nombreuses années passées au Cerco. Courage, un jour viendra peut-être où le soutien aux jeunes chercheurs ne sera plus seulement un discours de campagne…

Merci à tous les autres habitants du Cerco. C'est grâce à chacun d'eux qu'il y règne une bonne ambiance et un vrai plaisir de "faire de la science".

Merci enfin à Nadège pour m'avoir encouragé et soutenu chaque fois que j'en avais besoin. Sa présence à mes côtés est inestimable et grâce à elle j'ai mieux que des ERP à mettre au centre de ma vie… (du moins, jusqu'à ce que nos enfants soient en âge de porter le bonnet ☺)

Liste de publications

1 – Articles publiés :

Macé M. J-M., Thorpe S. J. & Fabre-Thorpe M. (2005). Rapid categorization of achromatic natural scenes: how robust at very low contrasts? *Eur J Neurosci.,* 21, 2007-2018.

Macé M. J-M., Richard G., Delorme A. & Fabre-Thorpe M. (2005). Rapid categorization of natural scenes in monkeys: Target predictability and processing speed. *Neuroreport*, 16(4), 349-354.

Bacon-Macé N., **Macé M. J-M.**, Fabre-Thorpe M. & Thorpe S. J. (2005). The time course of visual processing: Backward masking and natural scene categorization. *Vis Res*, 45, 1459-1469.

Rousselet G. A., **Macé M. J-M.** & Fabre-Thorpe M. (2004). Spatiotemporal analyses of the N170 for human faces, animal faces and objects in natural scenes. *NeuroReport*, 15(17), 2607-2611.

Rousselet G. A., **Macé M. J-M.** & Fabre-Thorpe M. (2004). Comparing animal and face processing in the context of natural scenes using a fast categorization task. *Neurocomputing*, 58-60, 783-791.

Delorme A., Rousselet G. A., **Macé M. J-M.** & Fabre-Thorpe M. (2004). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Res*, 19(2), 103-113.

Rousselet G. A., **Macé M. J-M.** & Fabre-Thorpe M. (2004). Animals and Humans faces in natural scenes: How specific to human faces is the N170 ERP component? *J Vis,* 4 (1), 13-21. *http://journalofvision.org/4/1/2/.*

Rousselet G. A., **Macé M. J-M.** & Fabre-Thorpe M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *J Vis,* 3 (6), 440-456. *http://journalofvision.org/3/6/5/.*

2 – Articles en révision ou en préparation :

Rousselet G. A., **Macé M. J-M.** & Fabre-Thorpe M. (soumis). Temporal course of ERP in fast object categorization in natural scenes: a story more complicated than expected?

Boucart M., Naili F., Despretz P., Defoort-Delhemme S., **Macé M. J-M.** & Fabre-Thorpe M. (soumis). Implicit but no explicit recognition at very large visual eccentricities.

Macé M. J-M., Bacon-Macé N., Nespoulous J-L. & Fabre-Thorpe M. (en préparation). What's seen first: The animal or the bird?

Macé M. J-M., Richard G., Delorme A. & Fabre-Thorpe M. (en préparation). Monkey and humans can categorize achromatic natural scenes with large variations of luminance and contrast.

Macé M. J-M., Joubert O. & Fabre-Thorpe M. (en préparation). Influence of feature diagnosticity on categorization speed.

3 – Résumés de conférence publiés :

Fabre-Thorpe M., Rousselet G. A., **Macé M. J-M.** & Thorpe S.J. (2006). Teasing apart meaningful from meaningless ERP differences in object categorization: a complicated story. *J Vis*.

Fabre-Thorpe M., Bacon-Macé N., **Macé M. J-M.** & Thorpe S. J. (2005). Coarse to fine processing in natural scene categorisation. *Acta Neurobiologiae Experimentalis*.

Fabre-Thorpe M., **Macé M. J-M.**, Joubert O. (2005). Dog or animal? What comes first in vision? *Perception,* 34 *suppl,* 8.

Macé M. J-M., Joubert O. & Fabre-Thorpe M. (2005). Entry level at the superordinate level in visual categorization. *Proceedings of the 9th International Conference on Cognitive and Neural Systems*.

Rousselet G. A., **Macé M. J-M.** & Fabre-Thorpe M. (2003). Comparing animal and face processing in the context of natural scenes using a fast categorization task. *Proceedings of the 12th Annual Computational Neuroscience Meeting.*

Macé M. J-M., Richard G., Thorpe S. J. & Fabre-Thorpe M. (2003). Category-level Hierarchy: What comes first in Vision? *Acta Neurobiologiae Experimentalis,* 63, C20, 24.

Bacon N., **Macé M. J-M.**, Kirchner H., Fabre-Thorpe M. & Thorpe S. J. (2003). Dynamics of rapid scene categorization: Backward masking and RSVP studies. *Acta Neurobiologiae Experimentalis,* 63, C2, 20.

Macé M. J-M., Fabre-Thorpe M. & Thorpe S. J. (2002). How robust is rapid visual categorization of natural images to large variations of contrast? *J. Cog. Neurosci., Suppl,* A108, 40.

Macé M. J-M., Rousselet G. A., Sternberg C., Fabre-Thorpe M. & Thorpe S. J. (2002). Very early ERP effects in rapid visual categorisation of natural scenes: Distinguishing the role of low-level visual properties and task requirements. *Perception,* 31 *suppl,* 150.

Rousselet G. A., **Macé M. J-M.**, Sternberg C., Fabre-Thorpe M., & Thorpe S. J. (2002). Rapid categorization of faces and animals in upright and inverted natural scenes: no need for mental rotation and evidence for a selective visual streaming of upright faces. *Perception,* 31 *suppl,* 132a.

Thorpe S. J., Bacon N., Rousselet G. A., **Macé M. J-M.**, & Fabre-Thorpe M. (2002). Rapid categorization of natural scenes: feedforward vs. feedback contribution evaluated by backward masking. *Perception,* 31 *suppl,* 132b.

Macé M. J-M. & Fabre-Thorpe M. (2001). Catégorisation visuelle rapide de scènes naturelles chez l'homme et le singe : robustesse des performances aux modifications de contraste et de luminance. *Proceedings du 5ème Colloque des Neurosciences (Toulouse, France).*

Fabre-Thorpe M., **Macé M. J-M.** & Thorpe S. J. (2001). Rapid visual categorisation of grey-scale natural scenes: robustness to large variations in luminance and contrast. *Perception*, 30 *Suppl*, 72b.

Table des matières

Introduction générale

Les quelques pages qui suivent présentent des notions générales sur la vision et plus spécifiquement sur les concepts de catégorisation qui seront abordés dans ce mémoire. La vision est la modalité sensorielle prédominante chez l'homme et chez de nombreux primates, mais l'interprétation des informations visuelles est coûteuse du point du vue computationnel, et chez l'homme plus de la moitié de la surface du cortex est impliquée dans leur traitement. L'architecture fonctionnelle du système visuel des primates est complexe et ce n'est qu'au cours du dernier demi-siècle, avec l'avènement des neurosciences, alliées à la miniaturisation des instruments de mesure et la montée en puissance de l'informatique que la compréhension des mécanismes neuronaux qui sous-tendent la vision a pris son essor. La notion fondamentale de champ récepteur a ainsi émergé dans les années 40-50 avec des travaux portant sur le nerf optique, la rétine et le corps genouillé latéral (Hartline, 1940 ; Kuffler, 1953 ; Barlow, 1953 ; De Valois *et al.*, 1958) et dans les années 60 avec les travaux d'Hubel et Wiesel sur le cortex visuel primaire chez le chat et le singe (Hubel & Wiesel, 1959 ; Hubel & Wiesel, 1968). Les grands principes de fonctionnement du système visuel ont été établis par la suite avec l'analyse des propriétés des neurones des différentes aires visuelles et la compréhension de leur architecture globale. Gross (Gross *et al.*, 1972) a par exemple montré dans les années 70 que de nombreuses cellules du cortex inféro-temporal sont sélectives à des objets complexes (comme des visages). Cette propriété s'interprète comme une signature neuronale de l'une des fonctions les plus fondamentales qu'assure le système visuel : la reconnaissance d'objets. Depuis cette étude pionnière, de nombreux chercheurs ont contribué à caractériser les réponses des neurones dans les différentes aires visuelles (Zeki, 1978 ; Van Essen, 1979 ; Perrett *et al.*, 1982 ; Desimone *et al.*, 1984 ; Tanaka, 1993).

Le système visuel doit avoir la capacité de reconnaître des objets, mais il doit également extraire les propriétés communes entre différents objets pour les regrouper dans des classes. Cette catégorisation est essentielle pour donner un sens aux informations en provenance du monde extérieur et un organisme peut ainsi déterminer de manière pertinente et rapide ce qui est similaire et ce qui est différent. Il s'agit pour le système de traitement de l'information concerné, de gommer des différences pour établir des liens d'équivalence, ou bien au contraire de faire ressortir de petites variations pour marquer des limites. La double exigence de pouvoir discriminer des éléments ou bien de savoir les regrouper s'applique à tous les canaux sensoriels et les systèmes nerveux biologiques sont le plus souvent capable de traiter ces

informations en parallèle. Ces capacités de traitement permettent à l'animal d'extraire les régularités présentes dans l'environnement et sont à la base de son aptitude à calquer sur le monde une représentation d'un haut degré d'abstraction par rapport au signal brut capturé. Cette simplification du monde, en regroupant les objets dans des classes bien définies, permet de manipuler non plus des objets réels mais des concepts associés à ces objets, beaucoup plus souples au regard des opérations mentales qui peuvent leur être appliquées.

Si nous prenons pour exemple la vision, chaque vue d'un objet autour duquel nous nous déplaçons est différente de la précédente. Et pourtant, cet objet nous apparaît comme étant toujours la même entité, comme un tout cohérent qui poursuit son existence indépendamment de l'image sans cesse changeante qu'il projette sur notre rétine. Cette remarquable stabilité des objets perçus possède une base neurobiologique bien établie puisque des neurones invariants aux changements de point de vue ont été trouvés dans le cortex inféro-temporal des macaques (Logothetis & Pauls, 1995 ; Vogels & Orban, 1996 ; Booth & Rolls, 1998). Ce qui est vrai pour les différentes vues d'un objet l'est également pour différents objets d'un même groupe qui peuvent être classés comme faisant partie d'une même catégorie alors qu'ils présentent pourtant des différences intrinsèques. Des enregistrements unitaires chez l'homme ont ainsi montré que certains neurones du lobe temporal médian présentent une grande sélectivité à diverses catégories d'objets donnés (Kreiman *et al.*, 2000 ; Quian Quiroga *et al.*, 2005). Ces auteurs ont enregistré des neurones qui déchargent dès que l'image d'un objet précis ou d'une personne donnée est présentée au sujet. Ils montrent ainsi des exemples de neurones détecteurs de Jeniffer Aniston (quand elle est sans Brad Pitt !), de Halle Berry, du grand opéra de Sydney, mais aussi d'animaux variés (chevaux, araignées, …). La représentation atteinte dans ces cellules est parfois si abstraite qu'il est possible de les activer avec n'importe quelle photographie représentant leur objet préféré et même dans certains cas avec le nom de l'objet écrit en toute lettres ! Ces neurones du lobe temporal médian ne sont pas sous influence exclusivement visuelle, mais les neurones à l'origine des afférences qu'ils reçoivent depuis les aires visuelles présentent probablement une sélectivité à des stimuli visuels similaires.

Bien avant ces considérations neuronales, les plus anciennes approches du problème de la catégorisation datent des grand penseurs grecs. Pour Aristote, les catégories sont immuables et sont définies sur la base de propriétés communes entre certains objets, à la fois nécessaires et suffisantes pour expliquer ou non l'appartenance à un groupe. La pensée d'Aristote a influencé la philosophie occidentale pendant des siècles, mais au Moyen-Age, la question des catégories change quelque peu de sens et l'on ne s'intéresse plus qu'à l'essence des choses (ce qui fait qu'une chaise est une chaise), avec souvent une réponse au problème plus théologique que philosophique. A partir du 18ème siècle, les idées de Descartes font sentir leur influence et

des philosophes comme Wolff puis Kant se séparent d'une conception purement métaphysique des catégories pour leur conférer un caractère plus cognitif. Cette démarche est poursuivie par Wittgenstein (Wittgenstein, 1953) dans la première partie du 20ième siècle et permet d'ancrer la question de la catégorisation dans la psychologique moderne. Le travail de psychologie cognitive de Rosch dans les années 70 (précédé par celui de Shepard dans les années 60 (Shepard & Chang, 1963)), marque une rupture avec l'approche philosophique de la question des catégories et s'oriente résolument vers une démarche expérimentale et scientifique dans laquelle les déductions proviennent de l'analyse de mesures comportementales dans des tâches de catégorisation effectuées par des humains (Rosch, 1973). Rosch et ses collègues ont proposé que les catégories ne sont pas d'arbitraires constructions de l'esprit mais qu'elles sont élaborées en prenant en compte les régularités statistiques du monde et que la mesure du comportement des sujets doit permettre de comprendre comment fonctionnent les processus de catégorisation.

Les catégories peuvent être représentées à la fois sur un plan horizontal : les différentes catégories entre elles (chien, chaise ou voiture), et sur un plan vertical : les différents niveaux d'inclusion observés au sein d'une catégorie (lévrier, chien, animal) (Rosch, 1978). Pour Rosch, les catégories ne s'articulent pas simplement autour de propriétés nécessaires et suffisantes comme dans la conception aristotélicienne, mais autour de prototypes qui peuvent être considérés comme des individus rassemblant un maximum de caractéristiques spécifiques de la catégorie tout en étant les plus différents possible des exemplaires des autres catégories. C'est cette distance relative entre les prototypes des différentes catégories qui détermine leur organisation horizontale. L'organisation verticale au sein des catégories correspond quant à elle à différents niveaux d'abstraction ; niveaux qui ne présentent pas un intérêt égal dans le processus de catégorisation. Rosch a en effet démontré qu'il existe un niveau de catégorisation préféré dans cette structure hiérarchique. C'est lorsqu'on accède aux catégories par ce niveau d'abstraction optimal que les performances comportementales sont les meilleures. Il est appelé le niveau de base (chien, oiseau, chaise, table, voiture, camion, etc…) et c'est le niveau le plus abstrait pour lequel les exemplaires d'une catégorie partagent encore une forme commune (Rosch *et al.*, 1976). Il est considéré comme le niveau le plus "informatif" ; celui qui optimise le "coût cognitif" du traitement par rapport à la quantité d'information obtenue. Les niveaux de catégorisation supérieurs (plus abstraits) sont appelés superordonnés (animal, meuble ou véhicule) et les niveaux inférieurs (plus figuratifs), sont dits subordonnés (labrador, chaise haute ou Porsche).

Des travaux ultérieurs sur les niveaux de catégorisation sont venus compléter cette vue pour expliquer des résultats divergents de ceux établis par Rosch. C'est notamment le cas pour les

objets atypiques qui peuvent être catégorisés aussi rapidement au niveau subordonné qu'au niveau de base. Ainsi, un merle est catégorisé plus vite comme un oiseau que comme un merle, mais une autruche est catégorisée à la même vitesse comme oiseau et comme autruche. Pour expliquer ce résultat, Jolicoeur (Jolicoeur *et al.*, 1984) a proposé la notion de niveau d'entrée de la catégorisation ; niveau dont la position dans la hiérarchie des catégories dépendrait de la typicité d'un objet au sein de cette catégorie. Un objet très atypique peut ainsi avoir un niveau d'entrée inférieur (plus figuratif) par apport au niveau de base. Comme l'ont montré Tanaka et al. (Tanaka & Taylor, 1991), les catégories subordonnées peuvent aussi bénéficier d'un avantage de traitement lorsqu'elles correspondent à des domaines dans lesquels les sujets possèdent une grande expertise. Cette notion d'expertise est pratiquement impossible à quantifier pour les objets qui nous entourent. Nous pouvons simplement constater que nous possédons pour la plupart d'entre nous une grande expertise dans l'analyse des visages humains mais pas dans celle des visages de chauve-souris ou des formes de galaxie (à de rares exceptions près !). Pour étudier finement les effets de l'expertise sur la reconnaissance d'objets, il est donc nécessaire de faire catégoriser aux sujets des objets qu'ils n'ont jamais vu auparavant en mesurant leurs performances au cours de l'apprentissage. C'est ce qu'ont fait Gauthier et al. (Gauthier & Tarr, 1997) avec des marionnettes appelées greebles qu'ils ont créées avec l'idée qu'elles partagent entre elles certaines caractéristiques propres aux visages, comme des propriétés morphologiques communes, un air de famille pour les greebles apparentés, un genre, une identité, etc. Cet outil s'est avéré très puissant pour étudier les mécanismes de la catégorisation des visages et plus généralement les mécanismes de catégorisation impliquant une grande précision, qui se développent en parallèle avec l'expertise.

Il faut noter que l'étude des processus de catégorisation menée par Rosch est intimement liée à l'existence d'ensembles qui se prêtent bien à un découpage taxinomique comme les êtres vivants ou les objets manufacturés. Les critiques de ces travaux ont remis en cause l'universalité de cette organisation verticale des catégories et de nombreux auteurs ont proposé de replacer les idées de Rosch dans un cadre théorique plus étendu en intégrant des principes de catégorisation fondés sur l'organisation temporelle (scripts, scenarii… ; Barsalou & Sewell, 1985), les relations partie/tout (arbre/forêt ; Markman *et al.*, 1980), la contiguïté ou la ressemblance (couleurs et formes ; Lévi-Strauss, 1962, p85) et non plus strictement la relation d'inclusion. Les conclusions de Rosch sur la structure des catégories ne sont pas remises en cause en elle-même, mais il faut garder à l'esprit que le cadre d'étude des catégories dans lequel nous nous plaçons lorsque nous reviendrons sur les travaux de Rosch ne représente qu'une partie des questions liées à la catégorisation et à l'organisation des catégories.

Le retour au premier plan de l'étude de la catégorisation depuis les années 60-70 a amené de nombreuses équipes de recherche à se pencher sur les questions qu'elle suscite et à l'utiliser à la fois comme un objet d'étude et comme un moyen d'investigation du système nerveux. L'équipe de recherche dans laquelle j'ai effectué mon travail de thèse étudie depuis une dizaine d'années les performances du système visuel des humains et des singes (macaques) dans des expériences de catégorisation visuelle de scènes naturelles. Rosch soulignait dès ses premiers travaux l'importance du caractère écologique d'une tâche en critiquant les expériences ayant cours à son époque et dans lesquelles la catégorisation ne portait que sur des ensembles artificiels, construits pour être les plus contrôlés possible et de fait très éloignés de la réalité. Utiliser des stimulis naturels complexes rend bien sûr le contrôle de tous les paramètres difficile et parfois même impossible ; mais se servir des stimuli qui ont façonné les systèmes sensoriels au cours de millions d'années d'évolution pour étudier leurs capacités de catégorisation donne l'assurance qu'il s'agit des éléments pour lesquels le traitement sensoriel est optimal (Tolhurst & Tadmor, 2000). La tâche qui est couramment utilisée dans l'équipe consiste à relâcher un bouton lorsqu'une photographie de scène naturelle flashée sur un écran contient un animal et à maintenir l'appui en l'absence d'animaux. Cette tâche est communément appelée "tâche de catégorisation visuelle ultra-rapide" car elle fait peser de lourdes contraintes temporelles sur le système visuo-moteur pour plusieurs raisons. Tout d'abord, les images ne sont présentées que pendant une durée très courte, 20 à 30 ms ce qui empêche toute exploration oculaire de la scène. Ensuite, les images sont très variées et sont toutes nouvelles pour le sujet qui ne peut utiliser un quelconque apprentissage. Enfin, les sujets sont contraints de répondre le plus vite possible (tout en restant le plus précis possible) afin de rechercher le temps de traitement minimum nécessaire à une telle catégorisation.

Malgré la complexité du problème à résoudre, l'homme est capable de fournir dans de telles conditions environ 94% de réponses correctes pour un temps de réaction (TR) médian compris entre 380 et 440 ms (Figure 1 A) (Thorpe *et al.*, 1996 ; Rousselet *et al.*, 2003). Mais les réponses les plus précoces sont obtenues dès 250 ms. Ce ne sont pas des anticipations car c'est à partir de cette latence que les réponses correctes vers les cibles sont statistiquement plus nombreuses que les réponses incorrectes déclenchées vers des distracteurs (en cas d'anticipations, il y a autant de réponses correctes que d'erreurs puisque cibles et distracteurs sont en proportions identiques dans une série). Le temps de traitement minimal entre l'entrée visuelle et la sortie motrice est donc d'environ 250 ms chez l'homme, ce qui impose de sévères contraintes pour les modèles de reconnaissance d'objets.

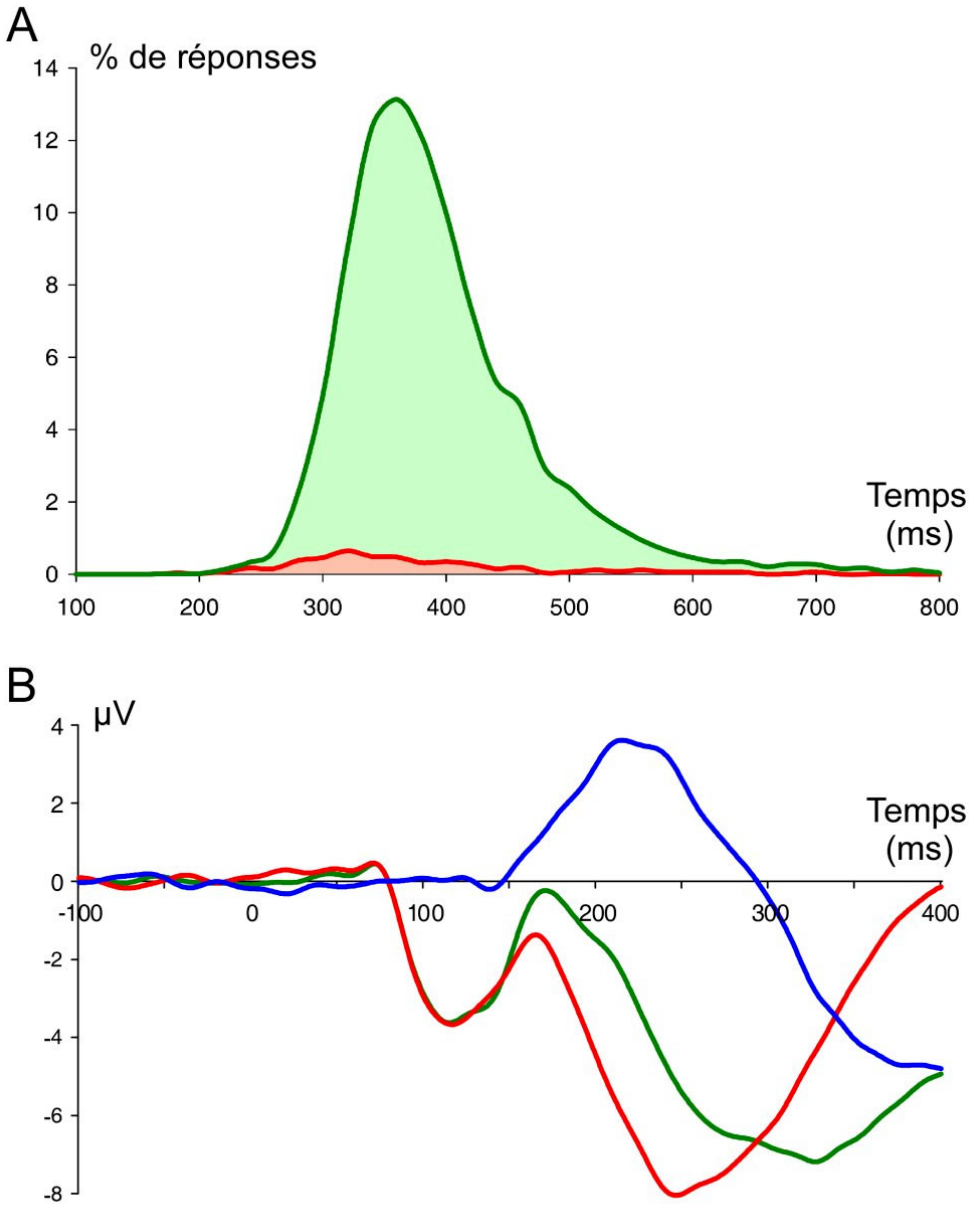


Figure 1 : Performance comportementale et enregistrements électrophysiologiques lors d'une tâche de catégorisation visuelle rapide animal/non animal en go/no-go. A. Distribution en fréquence des temps de réaction en fonction du temps. La courbe verte représente les réponses correctes sur les cibles et la courbe rouge les réponses incorrectes sur les distracteurs. B. Signal EEG enregistré sur l'électrode Fz. Les courbes vertes et rouges correspondent respectivement à la moyenne des potentiel évoqué sur les essais cibles et distracteurs correctement catégorisés. La courbe en bleue représente la différence entre le signal enregistré sur les cibles et les distracteurs. Elle devient significativement différente de la ligne de base vers 150 ms.

Cette étude originale a été répliquée chez le singe : les macaques sont capables d'effectuer la même tâche de catégorisation visuelle rapide que les hommes après un apprentissage de quelques mois. A la fin de l'apprentissage, les animaux atteignent une précision moyenne de 90% sur la première présentation d'images qu'ils n'ont jamais vues auparavant. Les singes sont donc légèrement moins précis que les hommes, mais beaucoup plus rapides avec un TR

médian de 250 ms (Figure 2). Leurs premières réponses correctes apparaissent ainsi dès 180 ms et ces latences très courtes ont été reportées sur un schéma représentant les différentes étapes que doit parcourir l'information visuelle dans le cerveau du singe avant de pouvoir déclencher une réponse motrice (Figure 3). On s'aperçoit sur ce schéma que la durée des traitements à chaque étape doit être limitée à environ 10 ms pour que la main du singe puisse se déplacer dès 180 ms. Ce temps très court pour chacune des étapes du traitement de l'information visuelle constitue une très forte contrainte pour les différents modèles de reconnaissance d'objets, en particulier ceux qui font appel à la synchronisation d'oscillations corticales, relativement lentes à se mettre en place (Singer & Gray, 1995; Kreiter & Singer, 1996), ou à des boucles itératives, consommatrices de temps (Grossberg, 1997 ; Deco & Zihl, 2001 ; Hamker & Worcester, 2002). Cette rapidité remet même en question la notion de codage de l'information en terme de fréquences de décharge (Gautrais & Thorpe, 1998 ; Thorpe, 1990).

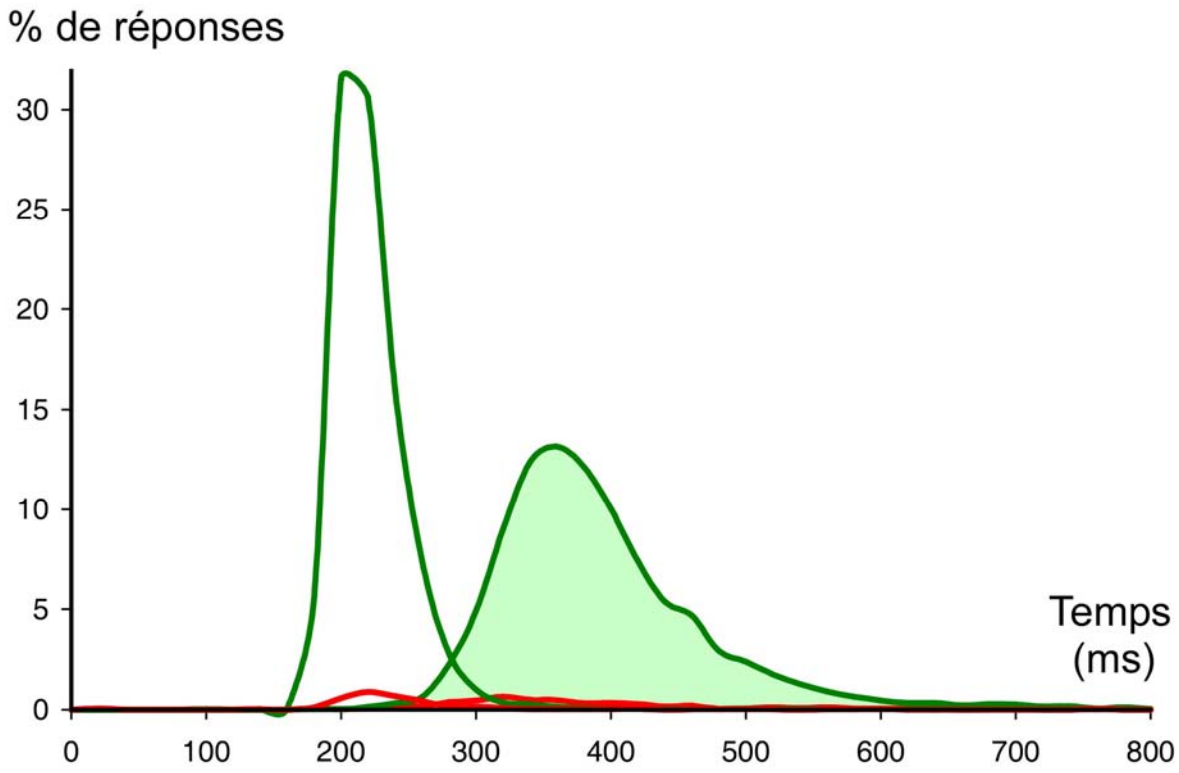


Figure 2 : Distributions comparées des temps de réaction de 14 humains et d'un singe dans une tâche de catégorisation visuelle rapide animal/non animal. Les courbes pleines représentent la performance des hommes et les courbes évidées celle des singes. Les réponses correctes sur les cibles sont en vert et les réponses incorrectes sur les distracteurs sont en rouge.

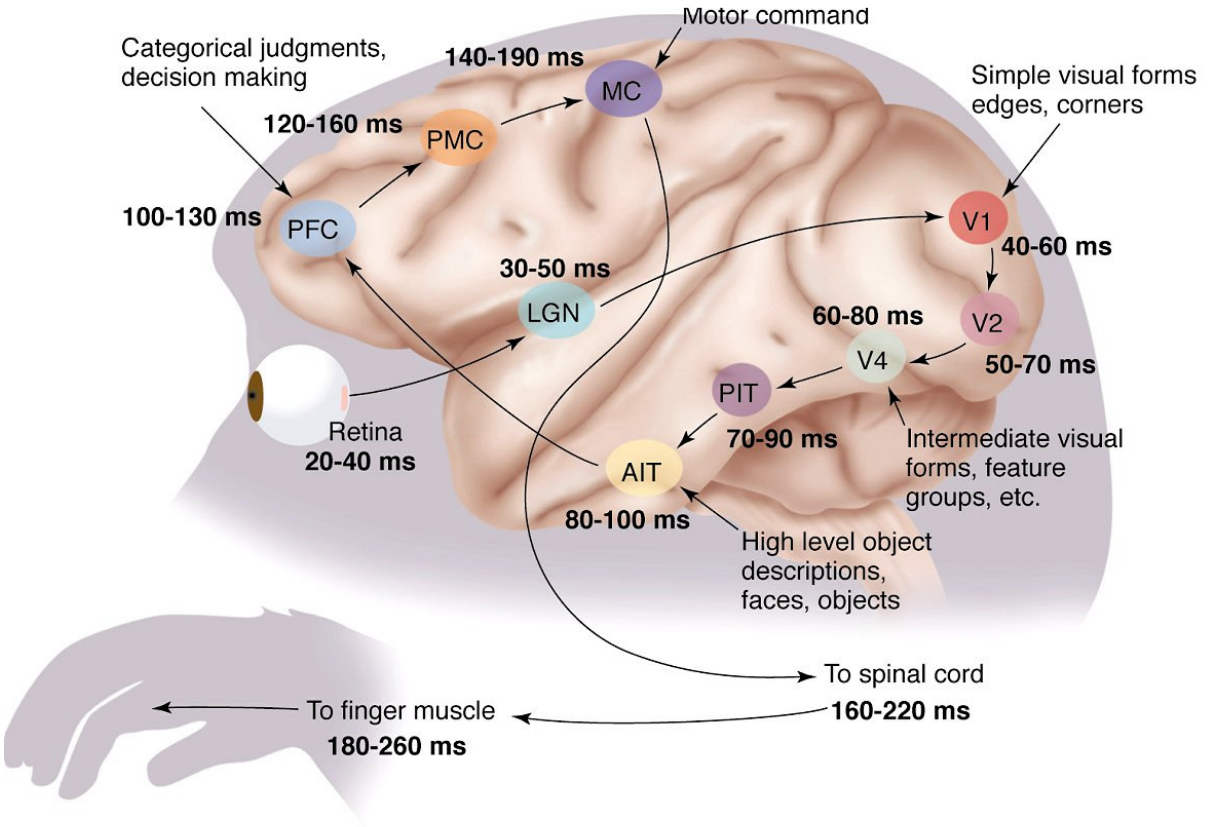


Figure 3 : Les différentes aires de traitement qui interviennent lors d'une tâche de catégorisation rapide ont été figurées sur une représentation schématique d'un cerveau de macaque. Pour chaque étape, le premier et le deuxième nombre correspondent respectivement aux latences minimale et moyenne auxquelles les réponses neuronales dans ces aires ont été enregistrées. Le trajet dans les aires frontales n'est qu'hypothétique et les latences indiquées dans les aires pré-motrices et motrices sont ajustées pour rendre compte d'un TR minimal autour de 180 ms. Reproduit d'après Thorpe et Fabre-Thorpe, Science (2001)

Étudier en parallèle l'animal et l'homme exécutant la même tâche dans les mêmes conditions a pour objet de mesurer jusqu'à quel point le fonctionnement du système visuel de ces deux espèces est comparable, non seulement dans les mécanismes de bas-niveau (De Valois *et al.*, 1974b ; De Valois *et al.*, 1974a ; Ungerleider, 1995 ; Imbert, 2000) mais aussi dans les processus impliqués dans des tâches plus complexes. Une bonne similarité des performances et des variations de ces performances avec la manipulation des paramètres des expériences menées chez l'homme et le singe ouvre la voie à une analyse plus poussée des mécanismes nerveux qui sous-tendent les facultés cognitives étudiées. En effet, cette analyse ne peut être effectuée qu'en ayant recours à des moyens plus invasifs disponible de manière anecdotique chez l'homme. L'architecture cérébrale du macaque est beaucoup mieux connue que celle de l'homme et les outils d'investigation dont nous disposons chez l'animal sont plus puissants. Il est par exemple possible d'enregistrer des neurones de manière unitaire ou multi-unitaire, d'enregistrer des potentiels évoqués à la surface du cortex ou encore de pratiquer des lésions

réversibles pendant que le singe effectue une tâche. Ces moyens d'investigation trop invasifs chez l'homme permettent d'explorer plus en détail les mécanismes cérébraux chez le singe. Pour transposer ces résultats chez l'humain, il est nécessaire (mais pas suffisant) de montrer que les performances entre les deux espèces sont semblables dans un grand nombre de tâches.
Certains auteurs considèrent cependant que les macaques ne sont pas capables de réaliser une véritable catégorisation visuelle parce que regrouper des éléments très différents dans une même catégorie implique un passage par des représentations conceptuelles abstraites. Ils expliquent les performances des animaux dans ces tâches de catégorisation visuelle par une analyse bas-niveau de la scène et des réponses qui s'appuient sur des indices tels que des plages de luminance ou des motifs de couleur (Premack, 1983; Huber & Fagot-Joel, 1999). Mais il a été montré que comme chez l'homme, l'absence d'indices de couleur ne perturbait pas les performances de catégorisation du macaque (Delorme *et al.*, 2000) ; nous verrons également dans ce mémoire dans deux études successives que les singes ne basent pas non plus leur catégorisation sur de simples variations statistiques liées au contraste ou à la luminance (Macé *et al.*, 2005). Ils peuvent également catégoriser des images à différents niveaux en fonction de la tâche (animal/non animal ou oiseau/non oiseau par exemple). De plus, l'analyse des erreurs de catégorisation commises par les hommes et les singes montre un important recouvrement (Delorme *et al.*, 2000 ; Macé *et al.*, 2005) ; un témoignage indirect mais très fort en faveur de l'utilisation des mêmes indices par les hommes et les singes pour déclencher leurs réponses. Enfin, les hommes et les singes sont très rapides dans cette tâche et cette similitude dans la vitesse de traitement des informations visuelles est un argument de plus qui laisse supposer que les processus mis en œuvre dans les deux espèces sont relativement similaires.

Dans leur étude de 1996 chez l'homme (Thorpe *et al.*, 1996), l'enregistrement de l’activité électrique cérébrale a permis de cerner avec une meilleure précision temporelle la durée du traitement sensoriel. L’enregistrement des potentiels évoqués visuels (PEV) permet de s’affranchir de la composante motrice inévitablement prise en compte dans une mesure du temps de réaction. En moyennant séparément les PEV enregistrés sur les essais cibles et les essais distracteurs corrects, ils ont pu déterminer que le temps de traitement nécessaire au système visuel pour catégoriser un objet complexe tel qu'un animal dans une image naturelle est d'environ 150 ms (Figure 1 B). C’est en effet à cette latence que les tracés correspondant aux cibles et aux distracteurs se séparent, reflétant probablement le début de la décision perceptive prise par le système visuel (Fize *et al.*, 2000). Cette différence de traitement est illustrée sur la figure 1 B par la courbe bleue qui représente la différence entre le signal

enregistré sur les cibles (en vert) et le signal enregistré sur les distracteurs (en rouge). Ces données électrophysiologiques, en contraignant les traitements dans une fenêtre temporelle encore plus réduite que celle déterminée par l'étude comportementale seule, suggèrent que le traitement des images doit être massivement parallèle et essentiellement vers l'avant (feed-forward). En effet, les informations visuelles doivent transiter par une dizaine d'étapes de traitement entre la rétine et le cortex inféro-temporal antérieur, étape ultime de la voie ventrale et du traitement visuel. C'est là que s'effectue la mise en correspondance entre les objets perçus et leurs représentations stockées en mémoire. Si un délai de seulement 150 ms chez l'homme suffit pour parcourir toutes ces étapes, c'est probablement parce que l'information de toute l'image est propagée en parallèle et que les traitements n'impliquent pas de boucles itératives coûteuses en temps de calcul.

Nous reviendrons plus longuement dans le deuxième chapitre sur le sens à donner à cette activité différentielle à 150 ms et sur les activités différentielles enregistrées par d'autres équipes à des latences plus précoces (50-120 ms : Seeck *et al.*, 1997; Debruille *et al.*, 1998; Mouchetant-Rostaing *et al.*, 2000) ou plus tardives (200-300 ms : Johnson & Olshausen, 2003) dans diverses tâches de catégorisation visuelle. La meilleure réponse à ces questions a été apportée au moyen d'une double tâche de catégorisation (VanRullen & Thorpe, 2001b) dans laquelle les sujets devaient effectuer deux catégorisations successives. En soustrayant le signal obtenu sur les mêmes images, vues soit comme cibles dans la première tâche, soit comme distracteurs dans la seconde tâche, on voit disparaître les activités différentielles précoces alors que l'activité différentielle à 150 ms est préservée. Ceci indique que les activités différentielles avant 150 ms sont probablement dues à des différences physiques entre les groupes d'images et non aux processus cérébraux liés à la tâche de catégorisation. Nous reviendrons sur cette question à plusieurs reprises dans le mémoire grâce à trois expériences dont les résultats peuvent être utilisés pour confirmer et étendre cette interprétation.

On peut également noter qu'il apparaît autour de 250-300 ms une autre activité différentielle de grande amplitude sur les électrodes centrales et centro-pariétales (C3 et C4 essentiellement, voir annexe A pour la position des électrodes). Ce signal différentiel limité aux électrodes placées au-dessus de territoires impliqués dans la motricité correspond à la différence d'activité motrice entre les réponses effectuées sur les cibles (réponses go, mouvement de la main) et les distracteurs (réponses no-go, la main reste immobile). Une manière élégante de faire ressortir l'aspect moteur de cette activité différentielle consiste à regarder sur une moyenne de sujets droitiers le moment à partir duquel apparaît une différence d'amplitude

entre les hémisphères gauche et droit pour le signal enregistré sur les cibles. Les potentiels évoqués dans l'hémisphère gauche (celui qui commande la main droite, utilisée pour répondre dans la tâche) deviennent plus amples que ceux enregistrés dans l'hémisphère droit dès 200 à 220 ms, ce qui est compatible avec la latence des premières réponses correctes observées à partir de 250 ms chez l'homme.

Ces premiers travaux sur la catégorisation visuelle rapide ont ouvert la voie à toute une série d'études orientées vers la compréhension des mécanismes qui permettent au système visuel de résoudre si rapidement ce problème complexe. Ces études permettent de caractériser l'architecture fonctionnelle du système visuel en déterminant les indices utilisés dans cette tâche et les limites du système. La première étude portant sur la catégorisation visuelle ultra-rapide permettait d'émettre l'hypothèse d'un traitement essentiellement vers l'avant de l'information dans tout le système visuel pour expliquer la vitesse de traitement élevée (Thorpe *et al.*, 1996). La principale implication de cette hypothèse était que le temps de réaction des sujets ne pouvait pas être écourté malgré l'apprentissage des images. C'est effectivement ce qu'a révélé une expérience conçue spécifiquement pour répondre à cette question (Fabre-Thorpe *et al.*, 1998). Le traitement quotidien et répété de 200 images pendant 3 semaines n'a pas permis aux sujets de traiter plus rapidement ces images familières présentées parmi des images nouvelles. Seules les images qui étaient catégorisées à l'origine avec des TR très longs (environ 10% des images) ont bénéficié d'un effet d'apprentissage. L'immense majorité des images utilisées (90%) peuvent donc être traitées grâce à des processus ultra-rapides, la répartition des temps de réaction étant due à la variabilité motrice et attentionnelle qui influence le résultat de chaque essai. Ces résultats constituent un argument fort en faveur de l'idée d'un traitement des informations vers l'avant dans le système visuel. De plus, ils permettent de penser que ce processus, déjà très optimisé, ne peut être amélioré que dans une faible proportion de situations. Il existerait ainsi un nombre incompressible d'étapes de traitement pour effectuer l'analyse d'une image, et ce traitement serait suffisant pour la plupart des images naturelles. Cette conclusion ne s'applique toutefois qu'à la catégorisation d'objets naturels dans des images naturelles et il était important de savoir si ce résultat pouvait être généralisé à la catégorisation d'objets artificiels. En effet, les catégories biologiques pourraient bénéficier d'un statut particulier dans le système visuel, peut être sous la forme de pré-réglages sous contrainte génétique au cours du développement pour faciliter la reconnaissance des animaux. On comprend en effet qu'un animal capable de reconnaître rapidement les proies et les prédateurs dans son environnement puisse bénéficier d'un avantage évolutif. C'est la raison pour laquelle VanRullen (VanRullen & Thorpe, 2001a)

a mené une étude pour savoir si la catégorisation des moyens de transports donne lieu à des performances similaires à celle obtenue pour les animaux. Les moyens de transports présentent le double avantage d'être des objets totalement artificiels pour la plupart d'entre eux, et d'avoir une grande diversité de forme, à l'instar des animaux. Les résultats de cette étude permettent de renforcer l'idée selon laquelle les capacités de catégorisation des objets par le système visuel sont très généralistes puisque les sujets se sont montrés aussi rapides et aussi précis pour catégoriser les moyens de transport que les animaux. La très grande vitesse du système visuel constitue certainement un avantage évolutif important, mais ses capacités sont très générales et s'appliquent probablement à tous types d'objets appris au cours de l'existence, sans se limiter aux objets biologiques ou importants du point de vue de la survie.

Une autre hypothèse mentionnée dans l'étude de 1996 pour expliquer la vitesse du système visuel concerne le parallélisme des traitements. Une image de scène naturelle contient souvent plusieurs objets et l'on peut supposer que la très grande vitesse du système visuel n'est possible qu'au prix d'un traitement de toute l'image "en une seule passe". Mais Rousselet et al. (Rousselet *et al.*, 2002) ont aussi montré que ce parallélisme dans le traitement visuel est encore plus impressionnant. Présenter 2 images en même temps, une dans chaque hémichamp, ne ralentit pas le système visuel et les TR médians et minimaux sont similaires que le système ait à traiter une ou deux images simultanément. Une expérience poussant le défi jusqu'à 4 images présentées simultanément (Rousselet *et al.*, 2004) a montré que 2 images dans le même hémichamp pouvaient être catégorisées aussi bien qu'une image dans chaque hémichamp, mais que 4 images présentées à la fois provoquent une saturation du système de traitement, peut être à cause de conflits survenant dans le cortex préfrontal lors de la prise de décision comme le suggèrent les auteurs dans leur discussion approfondie sur les différences entre activités différentielles frontales et occipitales.

En résumé, l'ensemble de ces travaux montre que le système visuel est capable de reconnaître différentes catégories d'objets complexes, de manière très rapide et très précise, sans utiliser de mouvements oculaires et sans être limité au traitement d'un seul objet ni même d'une seule image. La grande vitesse des opérations cognitives sous-jacentes permet de supposer que le traitement des informations visuelles s'effectue essentiellement vers l'avant et de manière massivement parallèle. D'autre part, nous savons que la catégorisation est pratiquement insensible à la suppression des informations de couleur (Delorme *et al.*, 2000 ; Rousselet *et al.*, 2005) et qu'elle peut être effectuée en grande périphérie (Thorpe, 2001). D'après les connaissances que nous possédons sur l'architecture anatomo-fonctionnelle du système visuel,

les contraintes temporelles mises en évidence dans ces expériences de catégorisation ultra-rapide sont très importantes et les réponses les plus précoces doivent s'appuyer sur les toutes premières informations disponibles.
Ainsi, le fait que ces traitements soient rapides, orientés vers l'avant et parallèles, et qu'il s'appuient sur des indices achromatiques disponibles à la fois au centre du champ visuel et en périphérie laisse penser qu'ils se basent sur les informations magnocellulaires, qui satisfont à ces critères et sont les premières disponibles.

Les travaux qui font l'objet de cette thèse ont eu pour but de rechercher l'importance des informations magnocellulaires dans le traitement rapide de l'information en apportant des indications sur les performances qui peuvent être réalisées sur la base de ces seules informations. Ils ont permis de recueillir également de nouvelles données qui se révèlent être en accord avec les résultats déjà obtenus sur la grande rapidité du système visuel à chaque étape de traitement et sur l'importance de la préactivation des aires visuelles (influences top-down). Enfin, ils se sont intéressés à la notion de représentation visuelle précoce. Plusieurs expériences ont été réalisées pour mieux comprendre la nature et le contenu des représentations construites à partir des toutes premières informations visuelles.
Une approche comparative entre l'homme et le singe est présente tout au long de ce travail afin de vérifier à chaque avancée si les caractéristiques du système visuel de ces deux espèces sont similaires. Le but de cette démarche étant bien sûr de "valider" du mieux possible le modèle macaque dans l'étude de la reconnaissance d'objet chez l'homme.

1 - Quel rôle pour le système magnocellulaire dans la catégorisation visuelle rapide ?

1.1 - Catégorisation visuelle rapide

Nous avons brièvement évoqué dans l'introduction les contraintes que doivent prendre en compte les modèles de la vision pour expliquer les performances du système visuel dans les différentes expériences qui ont été menées jusqu'à présent chez l'homme et l'animal. La première contrainte est d'ordre temporel puisqu'il faut expliquer qu'une trace cérébrale de la catégorisation d'objets complexes apparaisse à une latence de 150 ms. Les études présentées dans ce chapitre vont tenter de renforcer l'idée selon laquelle les premières informations disponibles (magnocellulaires) jouent un rôle considérable dans la catégorisation visuelle ultra-rapide. Deux expériences évoquées dans l'introduction apportent des éléments en faveur de cette hypothèse.

1.1.1 - Catégorisation sans indices de couleur

Pour essayer d'aller plus loin dans la compréhension des mécanismes mis en œuvre dans le système visuel, il est essentiel de déterminer quels sont les attributs les plus importants dans une image pour effectuer une tâche donnée. On peut imaginer manipuler une partie du contenu des images et observer l'impact de cette modification sur les performances et l'activité cérébrale pour savoir si ces informations sont essentielles au fonctionnement du système. C'est ce type de manipulation qu'ont fait Delorme et ses collègues dans une étude parue en 2000 (Delorme *et al.*, 2000). Testés sur des images présentées aléatoirement en couleur ou en niveaux de gris, les sujets (hommes ou singes) se sont montrés aussi rapides (TR minimal) et pratiquement aussi précis dans les deux conditions de présentation. Ainsi de manière surprenante, retirer une importante partie des informations de l'image, qui peuvent sembler très utiles pour interpréter la scène ou lever des ambiguïtés, ne ralentit pas la vitesse du traitement.

Il existe une hypothèse assez simple pour expliquer ce résultat, qui fait appel aux caractéristiques respectives des deux systèmes principaux qui transmettent les informations de la rétine au cortex visuel. Au niveau de V1, les informations magnocellulaires ont 10 à 20 ms d'avance sur les informations parvocellulaires, ce qui peut s'avérer un avantage primordial lors de processus de traitement très rapides. Si l'on fait l'hypothèse que le système visuel doit tirer

partie des toutes premières informations disponibles pour traiter aussi rapidement les images, la catégorisation pourrait avantageusement s'appuyer sur les informations transmises par le système magnocellulaire. Étant donné que les informations qu'il transporte sont achromatiques, la suppression de la couleur ne ralentirait pas la catégorisation des images.
Il faut cependant noter que le système parvocellulaire encode à la fois des contrastes de couleur et des contrastes de luminance. Le résultat précédent sur les images en noir et blanc et en couleur pourrait donc être expliqué sans faire appel au système magnocellulaire si les contrastes de luminance sont encodés aussi rapidement que les contrastes de couleur dans le système parvocellulaire. Cette question n'est pas entièrement tranchée, mais certains travaux laissent penser qu'il pourrait effectivement exister un avantage d'environ 5 ms pour les contrastes de luminance par rapport aux contrastes chromatiques au sein du système parvocellulaire (Benardete & Kaplan, 1997 ; Benardete & Kaplan, 1999). L'absence d'effet de la suppression des informations de couleur sur la catégorisation ne constitue donc pas une preuve formelle de l'implication du système magnocellulaire dans la catégorisation visuelle rapide, mais elle permet néanmoins de proposer cette hypothèse comme meilleure candidate en explication du phénomène.

1.1.2 - Catégorisation en périphérie

Dans les expériences présentées dans l'introduction, les images sont flashées pendant seulement 20 ms au centre de l'écran. Les sujets n'ont pas le temps d'effectuer de saccades d'exploration oculaire et leur regard est dirigé toujours au centre de l'image puisqu'ils doivent fixer à chaque essai une croix située au milieu de l'écran. Dans l'expérience sur le parallélisme des traitements, les images sont légèrement excentrées et il est intéressant de comparer la performance des sujets lorsqu'une image est présentée quelques degrés à droite ou à gauche du point de fixation et lorsqu'elle est présentée au centre, comme dans les expériences précédentes. On constate que la baisse de précision due à la présentation en périphérie est très réduite puisqu'elle est inférieure à 4% (90,4 contre 94-95%) bien que la tâche du sujet soit compliquée par l'alternance aléatoire des conditions à 1 et 2 images (Rousselet *et al.*, 2002). Ce résultat montre que ni les mouvements oculaires exploratoires, ni même une présentation au centre du champ visuel ne sont nécessaires pour qu'une image soit correctement catégorisée. Le fait que les sujets puissent maintenir une performance élevée lorsque les images sont présentées en périphérie du champ visuel a d'importantes implications sur le rôle de l'attention, la finesse de la représentation et les informations nécessaires (les détails accessibles) dans la catégorisation visuelle. Une expérience a été réalisée pour tester directement l'effet de l'excentricité sur la catégorisation en mesurant la robustesse du système

visuel à une présentation des images en périphérie. Grâce à un dispositif composé de 3 projecteurs dans une pièce spécialement équipée, les images pouvaient être présentées jusqu'à de très grandes excentricités. Les résultats montrent que les humains peuvent effectuer la tâche de catégorisation animal/non animal dans des images naturelles pratiquement jusqu'aux limites du champ visuel (Thorpe, 2001). Lorsque les images (39° x 26°) sont affichées à 70° d'excentricité, les sujets sont toujours capables de répondre au-dessus du niveau de la chance (60% correct ; Figure 1), même s'il ne sont pas certains de leurs réponses et ne peuvent pas indiquer de quel animal il s'agit. Lorsque les images apparaissent au centre de l'écran, les sujets conservent une performance élevée (93,3%), ce qui montre qu'il est tout a fait possible de réaliser la tâche avec une attention distribuée sur l'ensemble du champ visuel sans baisse de précision. Ainsi, une bonne acuité et une attention focalisée ne sont pas nécessaires pour que le système visuel puisque effectuer une catégorisation d'objets complexes.

Comme ceux de l'expérience précédente, ces résultats constituent un nouvel indice en faveur d'un rôle du système magnocellulaire dans la reconnaissance d'objet. Le fait qu'il soit possible de catégoriser des images en périphérie avec une précision relativement élevée alors que l'acuité y est médiocre laisse penser que la faible résolution spatiale du système magnocellulaire n'est pas rédhibitoire pour que les informations qu'il transporte soient utilisées dans des processus de catégorisation visuelle.

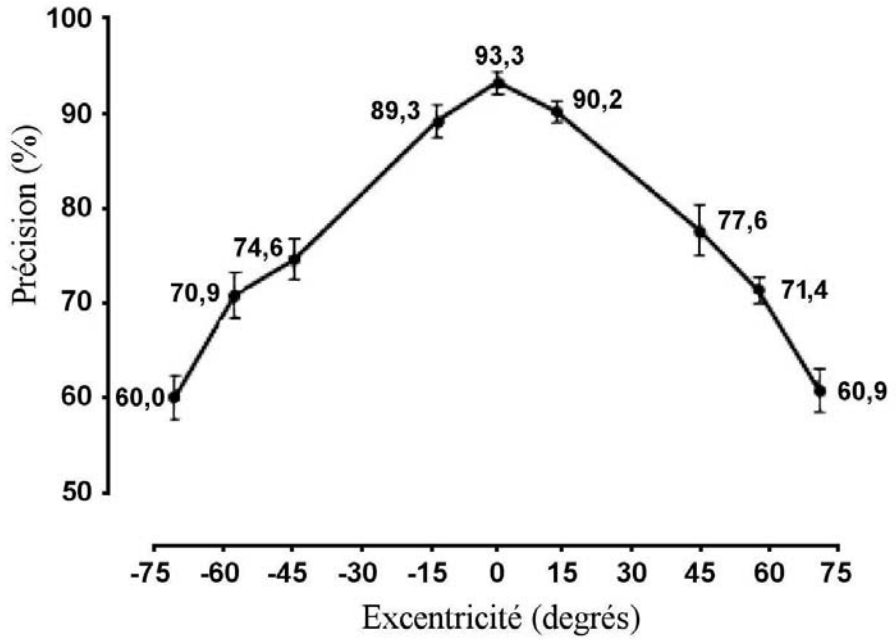


Figure 1 : Précision de 10 sujets humains dans une tâche de catégorisation visuelle d'images naturelles en fonction de l'excentricité des images présentées. Les performances en présentation centrale restent pratiquement au même niveau que lorsque l'image est toujours présentée au centre de l'écran. A 75° d'excentricité, la précision des sujets reste encore au-dessus du niveau de la chance, bien qu'ils ne puissent pas donner de description de la cible. Reproduit d'après Thorpe et al., EJN (2001).

1.1.3 - Implications pour la catégorisation visuelle

Toutes ces données concernant la vitesse de traitement, l'absence d'effet de la suppression de la couleur et la robustesse de la catégorisation pour des présentations en périphérie ont amené à penser que le système magnocellulaire pourrait être le principal système impliqué dans les toutes premières réponses fournies par le sujet. L'idée n'est pas totalement nouvelle et Sherman a proposé (Sherman, 1985) que le rôle du système parvocellulaire pourrait être de compléter et détailler une représentation grossière de la scène visuelle établie en premier lieu sur la base des information magnocellulaires. C'est une hypothèse qui est à contre-courant de l'idée dichotomique très répandue selon laquelle le système magnocellulaire, identifié à la voie dorsale, ne servirait qu'à détecter des mouvements et le système parvocellulaire, identifié à la voie ventrale, à reconnaître des objets (Livingstone & Hubel, 1987 ; Livingstone & Hubel, 1988). Cette intervention du système magnocellulaire a déjà été proposée pour expliquer des performances très élevées dans une tâche de reconnaissance de chiffres à faible contraste (Strasburger & Rentschler, 1996), mais il faut considérer que les chiffres forment une catégorie finie dans laquelle un sujet n'a que 10 possibilités à considérer par opposition à la catégorie "animal" choisie dans les travaux précédemment présentés, qui est beaucoup plus ouverte.

1.2 - Architecture générale du système visuel

La connectivité à l'intérieur de la rétine entre les différents types de cellules qui la compose est très complexe, et c'est après un premier traitement dans ce réseau rétinien que les informations visuelles quittent l'œil par le nerf optique. Elles se projettent en grande majorité vers le cortex visuel primaire (V1) à l'arrière du cerveau après un relais thalamique dans le corps genouillé latéral (CGL). Signalons qu'une faible partie des informations visuelles emprunte des trajets sous-corticaux, notamment vers le colliculus supérieur et le pulvinar et servent entre autre à attirer l'attention vers des éléments importants présents dans le champ visuel et à guider l'action (Grieve *et al.*, 2000). En ce qui concerne le trajet cortical, après V1, les informations visuelles sont transmises vers l'aire V2, étape à partir de laquelle deux grandes voies se séparent anatomiquement et fonctionnellement (Mishkin *et al.*, 1983 ; Van Essen & Zeki, 1978). La voie dorsale (V3, MT, MST, LIP…) est impliquée dans le codage du mouvement et dans la représentation spatiale des objets (Pohl, 1973) tandis que la voie ventrale (V4, TEO, TE…) est impliquée dans la reconnaissance des objets et plus généralement dans la représentation de la forme et de la couleur (Cowey & Gross, 1970 ; Desimone *et al.*, 1985). Nous nous intéressons donc ici principalement à la voie ventrale.

Dans la rétine ou le CGL, les neurones que l'on enregistre présentent une réponse optimale lorsqu'ils sont stimulés par un point blanc sur un fond noir ou inversement. Dans le cortex visuel primaire, les champs récepteurs sont déjà plus complexes et une grande partie des neurones ne répondent plus à des points isolés mais à des barres orientées dans une direction précise. Cette complexité croissante dans la sélectivité des neurones, qui va de pair avec une augmentation de la taille des champs récepteurs, se poursuit à travers les différentes étapes de traitement des voies visuelles. On peut par exemple enregistrer dans l'aire V2, juste après V1, des neurones sélectifs à des intersections de droites ou à des contours illusoires et dans l'aire V4 les premiers neurones sélectifs à la couleur, à des formes simples ou des textures. Les étapes suivantes de la voie ventrale sont situées dans le cortex inféro-temporal où ont été enregistrés chez le singe et l'homme des neurones qui déchargent pour des catégories d'objets (Gross *et al.*, 1972 ; Perrett *et al.*, 1982 ; Baylis & Rolls, 1987 ; Quian Quiroga *et al.*, 2005). Le travail minutieux de Tanaka (Tanaka, 2003) a permis de mettre en évidence l'existence d'un gradient dans les propriétés des réponses neuronales au sein même du cortex inféro-temporal. On distingue ainsi la partie postérieure du cortex inféro-temporal (TEO), dans lequel des cellules répondent à des dessins au trait ou des représentations relativement simples et la partie antérieure de ce cortex (TE) où les neurones sont sélectifs à des représentations plus complexes des objets (scènes naturelles par exemple) (Figure 2).

La région antérieure du cortex inféro-temporal est l'ultime aire "purement" visuelle de la voie ventrale. Les étapes de traitement ultérieures sont moins bien connues, mais consistent pour une bonne part en une mise en relation des informations visuelles avec les éléments en provenance d'autres modalités sensorielles ou de représentations mnésiques, comme dans le cortex périrhinal avec lequel le cortex inféro-temporal partage de nombreuses connexions (Suzuki & Amaral, 1994).

Le cortex périrhinal fait partie du lobe temporal médian qui contient également le cortex enthorhinal et les cortex parahipocampiques. Alors que le cortex inféro-temporal reçoit ses afférences principalement du système visuel, le cortex périrhinal est fortement connecté aux autres aires corticales du lobe temporal médian, à l'amygdale et aux cortex auditifs et somatosensoriels (Suzuki & Amaral, 1994). Le rôle du cortex périrhinal est donc fondamentalement intégrateur et multimodal (Murray & Richmond, 2001) et de nombreuses études ont démontré son importance dans des tâches visuelles d'apprentissage et de mémoire (Mumby & Pinel, 1994 ; Buckley *et al.*, 1997 ; Buffalo *et al.*, 1999). Le cortex inféro-temporal antérieur et le cortex périrhinal possèdent des connexions en direction du néocortex (Van Hoesen *et al.*, 1981 ; Lavenex *et al.*, 2002), qui serait impliqué dans la prise de décision dans la tâche pour activer les centres prémoteurs et moteurs et déclencher la réponse motrice.

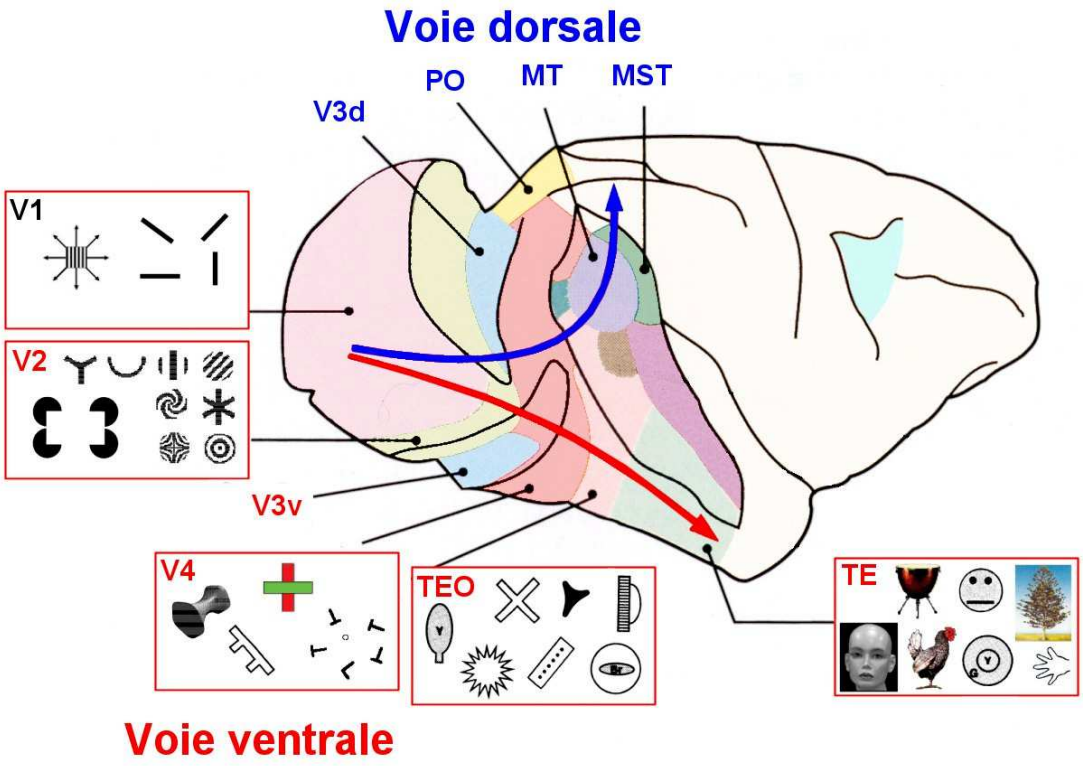


Figure 2 : Représentation schématique de l'architecture du système visuel chez le macaque.
Les informations visuelles sont traitées en V1, puis en V2 et partent ensuite vers la voie dorsale (en bleu : V3, PO, MT, MST) ou la voie ventrale (en rouge : V3, V4, TO, TEO). Les différentes aires visuelles sont représentées en couleur, libellées et accompagnées d'exemples de stimuli pour lesquels les neurones qu'elles contiennent présentent une sélectivité.
Reproduit d'après Bullier, Brain Res (2001) et modifié par Nadège Bacon-Macé.

1.3 - Flux magno- et parvo-cellulaires dans les voies visuelles

1.3.1 - Connexions anatomiques

Outre la dichotomie voie ventrale / voie dorsale du système visuel, il existe une division entre trois systèmes parallèles qui ne véhiculent pas les mêmes types d'informations depuis la rétine. La voie la moins importante en terme de connexions est appelée koniocellulaire ; elle est la moins rapide des trois (Irvin *et al.*, 1986) et véhicule essentiellement des informations sur les contrastes de couleur jaune/bleu (Martin *et al.*, 1997). Nous ne présenterons pas plus en détail cette voie encore peu explorée. Les voies magno- et parvo-cellulaires, tout comme la voie koniocellulaire, prennent leur origine dans la rétine. On distingue ainsi des cellules ganglionnaires de type magnocellulaires (cellules "parasol") et des cellules ganglionnaires de

type parvocellulaire (cellules "midget"). Les cellules parasol représentent environ 10% des cellules ganglionnaires de la rétine, contre 80% pour les midgets (Silveira & Perry, 1991). Les cellules magnocellulaires ont une distribution relativement uniforme sur la rétine contrairement aux cellules parvocellulaires qui sont beaucoup plus concentrées aux abords de la fovéa (Dacey & Petersen, 1992). Il est important de déterminer les propriétés de réponse de ces neurones ganglionnaires car elles déterminent la nature des informations transmises par les systèmes magno- et parvo-cellulaires. L'activité des cellules parasol est par exemple indépendante de la longueur d'onde utilisée et ces cellules possèdent un champ récepteur plus étendu que celui des cellules midget (Shapley & Perry, 1986). L'absence de sélectivité à la couleur et la largeur du champ récepteur des cellules parasol font que le système magnocellulaire est achromatique et qu'il véhicule des informations portant essentiellement sur les basses fréquences spatiales. Il transmet donc une image relativement grossière de la scène visuelle, dépourvue de détails et en noir et blanc. Il est en revanche très sensible aux variations de contraste et sa fréquence de fusion temporelle est élevée (Kaplan & Shapley, 1986 ; Tootell *et al.*, 1988). Le système parvocellulaire véhicule quant à lui les informations chromatiques et les hautes fréquences spatiales présentes dans l'image. Il fournit donc une description détaillée du monde environnant, mais il est peu sensible au contraste (Derrington & Lennie, 1984). Les axones de toutes les cellules ganglionnaires forment le nerf optique qui transmet les informations visuelles jusqu'au corps genouillé latéral dans le thalamus. Chez l'homme, le CGL possède 6 couches bien distinctes, alternativement ipsi- et contra-latérales et organisées de manière rétinotopique (Schiller & Malpeli, 1978 ; Nelson & LeVay, 1985). Les 4 couches supérieures reçoivent des projections parvocellulaires et les 2 couches inférieures des projections magnocellulaires ; les intercouches recevant des afférences koniocellulaires (Malpeli & Baker, 1975 ; Wilson *et al.*, 1976; Irvin *et al.*, 1986) (Figure 3).

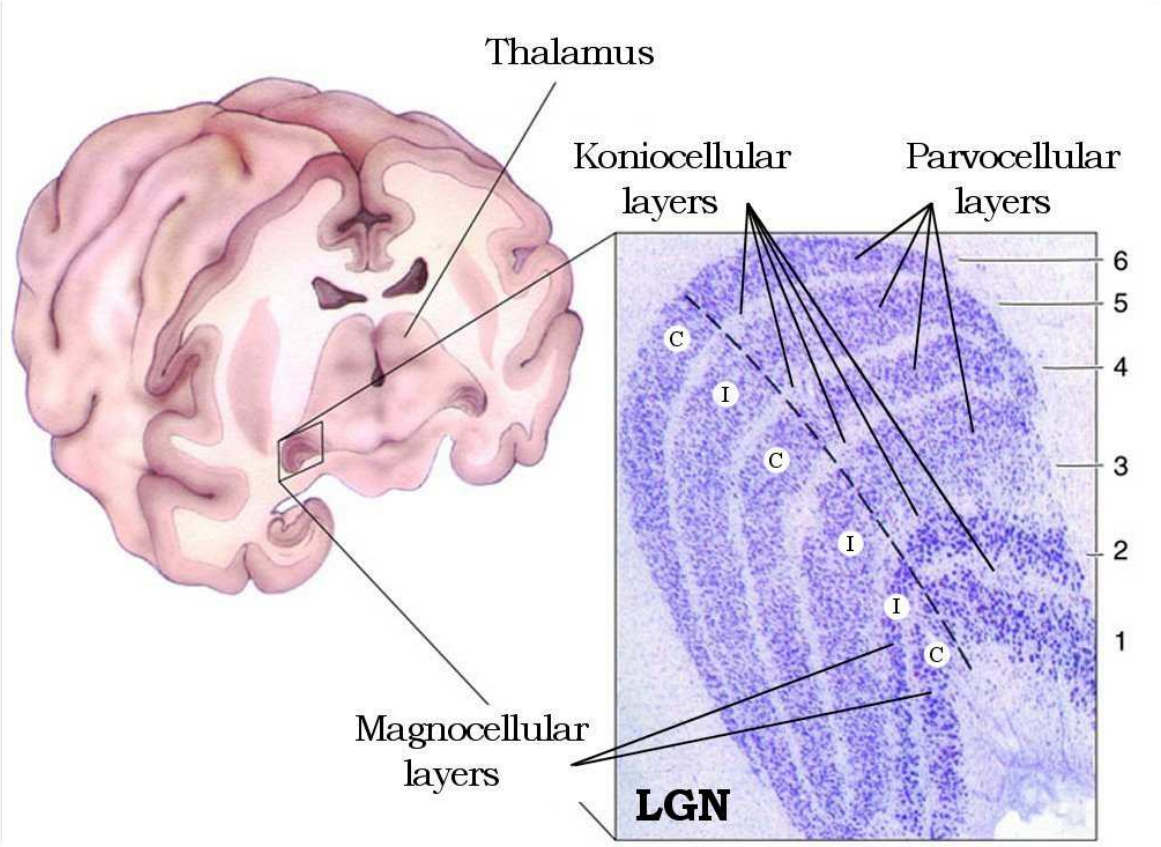


Figure 3 : Organisation des couches du corps genouillé latéral. Les couches 1 et 2 reçoivent des projections magnocellulaires, les couches 4 à 6 des projections parvocellulaires et les intercouches des projections koniocellulaires. Les lettres I et C indiquent si la couche reçoit des afférences Ipsi ou Contra-latérales. Reproduit d'après Hubel (1985).

Les radiations optiques qui partent des différentes couches du CGL pour se projeter sur le cortex visuel primaire conservent une bonne rétinotopie. Les afférences magnocellulaires trouvent leur terminaisons majoritairement dans la couches 4Cα et les afférences parvocellulaires dans les couches 4A et 4Cβ (Figure 4) (Hendrickson *et al.*, 1978). Une coloration du cortex visuel à la cytochrome oxydase révèle des structures plus foncées dans les couches 2 et 3 appelées blobs et interblobs en V1 (Horton & Hubel, 1981) et bandes pales/fines/épaisses en V2 (Livingstone & Hubel, 1982). Les bandes épaisses de V2 reçoivent très majoritairement des informations magnocellulaires par l'intermédiaire de la couche 4B en V1 et se projettent à leur tour dans la voie dorsale sur l'aire MT (middle temporal area). En conséquence, il est généralement admis que ce sont les informations magnocellulaires qui sont très majoritairement utilisées pour la détection des mouvements et la localisation dans l'espace (Maunsell *et al.*, 1990 ; Kessels *et al.*, 1999). Les neurones situés dans les blobs et les interblobs reçoivent un mélange de projections parvocellulaires et magnocellulaires. Les blobs, par l'intermédiaire de connexions avec les couches 4A et 4B qui reçoivent elles-même

des projections des couches 4Cβ (parvocellulaire) et 4Cα (magnocellulaire) et les interblobs directement par les couches 4Cα et β (Fitzpatrick *et al.*, 1985 ; Munk *et al.*, 1995). Les bandes fines et les bandes pales de V2 reçoivent leurs entrée respectivement des blobs et des interblobs de V1 et se projettent toutes les deux dans la voie ventrale vers V4 (DeYoe & Van Essen, 1985 ; Nakamura *et al.*, 1993). Le système parvocellulaire, qui se projette presque exclusivement dans la voie ventrale, est donc très impliqué dans la reconnaissance d'objets (Desimone *et al.*, 1984, Zeki & Shipp, 1988), mais il a été montré, en accord avec les données anatomiques et contrairement à la simplification couramment utilisée : voie ventrale = parvocellulaire et voie dorsale = magnocellulaire, que les systèmes parvocellulaire et magnocellulaire contribuent de manière pratiquement équivalente à l'activité dans la voie ventrale (Ferrera *et al.*, 1992 ; Nealey & Maunsell, 1994 ; Allison *et al.*, 2000), sans que le rôle précis du système magnocellulaire n'y soit connu.

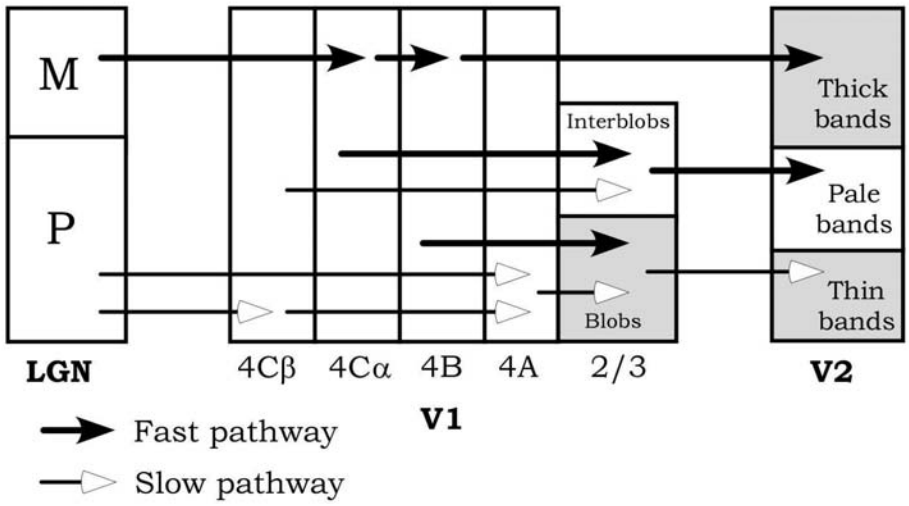


Figure 4 : Détail des connexions corticales pour les systèmes magnocellulaire et parvocellulaire en V1. Les informations magnocellulaires et parvocellulaires se mélangent dans les blobs et les interblobs, mais les projections des blobs vers les bandes fines sont cependant 20 ms plus lentes que les projections des interblobs vers les bandes pales ou que les projections magnocellulaires directes vers les bandes épaisses. Reproduit d'après Munk et al., PNAS (1995).

1.3.2 - Caractéristiques physiologiques

L'une des différences les plus importantes entre les systèmes magno- et parvo-cellulaires concerne la vitesse de transfert de l'information : la voie parvocellulaire est en effet moins rapide que la voie magnocellulaire dans laquelle les corps cellulaires sont plus volumineux et les axones ont un diamètre plus important. Le décalage entre les premières décharges enregistrées pour chaque système dans le LGN et dans V1 atteint entre 10 et 20 ms (Maunsell

& Gibson, 1992 ; Nowak *et al.*, 1995), ce qui peut constituer une différence importante lorsque les contraintes temporelles sont très fortes. Dans l'aire V2, ce sont les neurones des bandes épaisses qui déchargent le plus rapidement, autour de 40 ms (moyenne : 70 ms). Viennent ensuite les bandes pâles autour de 50 ms (moyenne : 71 ms) et enfin les bandes fines autour de 60 ms (moyenne : 92 ms). Les neurones des bandes épaisses de V2, qui reçoivent des informations uniquement magnocellulaires, déchargent donc en moyenne pratiquement en même temps que ceux des bandes pâles, mais 20 ms avant ceux des bandes fines, alors que ces deux dernières structures reçoivent un mélange d'afférences magnocellulaire et parvocellulaire depuis V1 (Munk *et al.*, 1995). En V4, les premières réponses sont observées autour de 75 ms (Schmolesky *et al.*, 1998), mais la latence moyenne est supérieure à 100 ms, la déviation standard dans les latences de décharge étant particulièrement importante (104 ±23 ms). Une telle distribution dans les temps de réponse des neurones provient probablement de l'arrivée progressive des informations magnocellulaires puis parvocellulaires dans cette aire. Ainsi, du côté de la voie dorsale, dans laquelle les afférences parvocellulaires sont extrêmement réduites, les aires MT et MST (Medial Superior Temporal area) présentent à la fois des latences de décharges très courtes autour de 70 ms et une distribution des latences beaucoup plus réduite qu'en V4.

Ainsi, bien que la ségrégation des voies magno- et parvo-cellulaires ne soit que partiellement conservée du point de vue anatomique dans les blobs et les interblobs, puis dans leurs projections sur V2, le mélange des informations n'est probablement pas complet avec une ségrégation temporelle de 10 à 20 ms qui pourraient persister jusqu'en V4.

1.3.3 - Modèles de traitement rapide de l'information visuelle

L'étude de la latence de réponse des neurones dans les différentes aires des voies ventrales et dorsales permet de comprendre comment le flux d'information traverse le système visuel et de proposer des hypothèses sur la manière dont le "décodage" de l'information s'effectue. Plusieurs modèles cherchent ainsi à tirer parti des différences de latences qui existent entre les voies dorsale et ventrale ou les systèmes magnocellulaire et parvocellulaire. L'idée est que la ségrégation des voies et des systèmes peut être utilisée pour effectuer un premier traitement rapide de l'information avant même de commencer des calculs plus lourds sur la totalité des données. Ce pré-traitement "intelligent" a été proposé par Bullier (Bullier, 2001) dans un modèle qui s'appuie sur la brièveté des latences de traitement de l'information dans la voie dorsale pour que le résultat d'un traitement rapide de la scène visuelle le long de cette voie soit propagé suffisamment rapidement par des connexions en retour vers les aires primaires

pour améliorer l'efficacité des traitements sur le flux d'informations parvocellulaires dans la voie ventrale (Figure 5). Vidyasagar (Vidyasagar *et al.*, 2002) propose un système apparenté lorsqu'il décrit un modèle de fonctionnement du système visuel dans lequel la voie dorsale est utilisée pour guider intelligemment l'attention dans la voie ventrale (Figure 6). Notre propre modèle, relativement proche de celui de J. Bullier, utilise également la ségrégation temporelle entre systèmes parvocellulaire et magnocellulaire, mais **au sein même de la voie ventrale**. Le traitement en une première passe rapide des informations grossières fournies par le système magnocellulaire serait utilisé pour orienter de manière pertinente les traitements subséquents (Figure 7). Munk propose également que l'avance qu'ont les informations magnocellulaires dans la voie ventrale puisse être mise à profit pour orienter les traitement sur le flux d'informations parvocellulaire grâce à des connexions en retour (Munk *et al.*, 1995). Un idée très similaire est développée par Nakamura qui voit dans les rares connexions qui court-circuitent des aires visuelles (V1 vers V4 ou V2 vers TEO) une autre manière de prendre suffisamment d'avance pour interagir par des connexions en retour rapides avec les informations en cours de traitement dans les aires précédentes (Nakamura *et al.*, 1993). Enfin un dernier modèle du même type est celui développé par Bar (Bar, 2004), dans lequel le FEF (Frontal Eye Field) sert à pré-traiter la scène visuelle pour en extraire des cartes de saillance perceptive et la région parahipoccapique à fournir des informations sur le contexte de la scène au cortex inféro-temporal pour limiter l'activation aux seules représentations des objets pertinents dans la scène (Figure 8).

Les différents modèles que nous avons présenté ci-dessus sont tous compatibles avec des traitements de type "coarse to fine" dans lesquels une première idée globale de la scène visuelle est reconstruite avant que les détails n'y soient implémentés comme cela a été proposé d'un point de vue théorique (Schyns & Oliva, 1994), puis retrouvé expérimentalement (Sugase *et al.*, 1999 ; Matsumoto *et al.*, 2005).

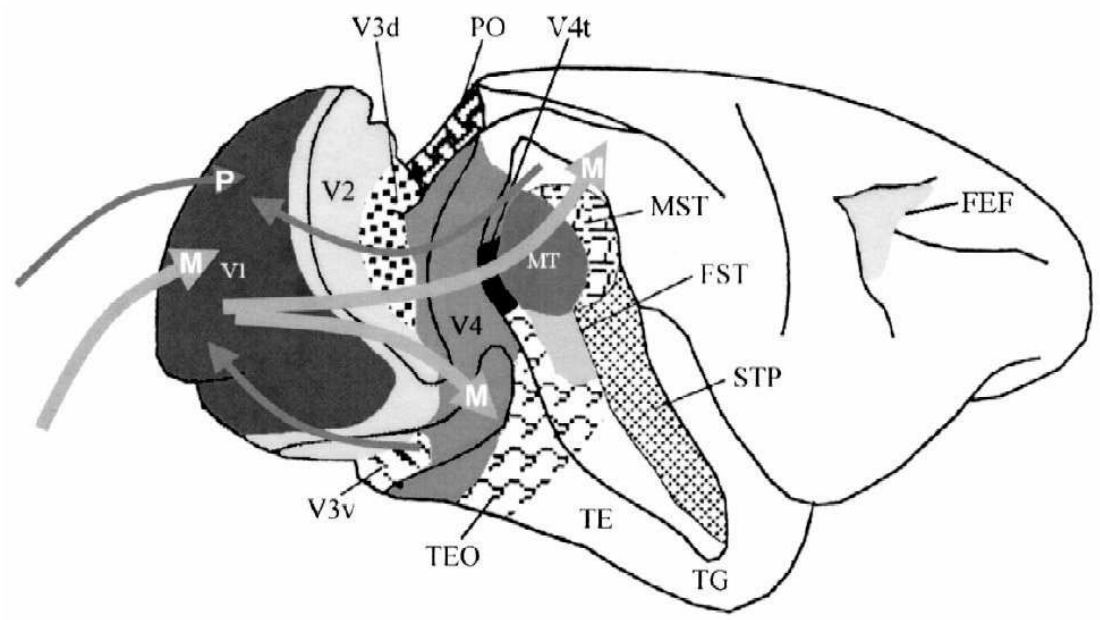


Figure 5 : Les informations magnocellulaires se propagent très vite dans la voie dorsale et les projections en retour vers les aires visuelles primaires arrivent pratiquement en même temps que les informations parvocellulaires. Cette coïncidence temporelle pourrait permettre à un système visuel d'utiliser la voie dorsale comme un pré-traitement de la scène visuelle utile pour les traitements de la voie ventrale.
Reproduit d'après Bullier, Brain Res (2001).

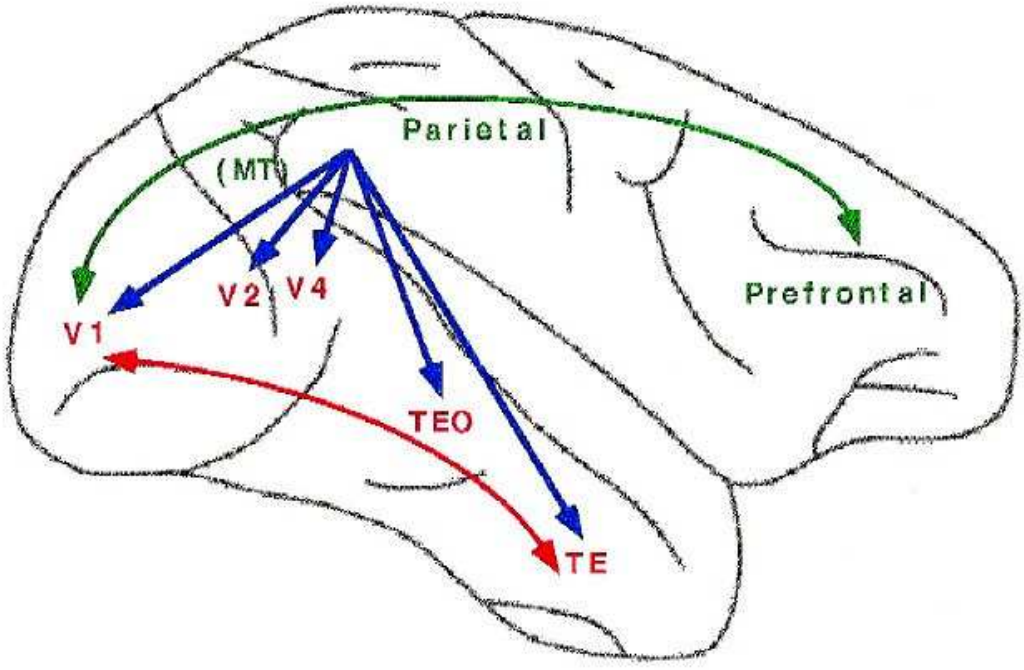


Figure 6 : Dans le modèle proposé par Vidyasagar, les informations magnocellulaires qui activent très rapidement toute la voie dorsale sont utilisées pour détecter les zones les plus intéressantes du champ visuel et orienter les traitements effectuées dans la voie ventrale sur les informations parvocellulaires grâce à des modulations attentionnelles. Reproduit d'après Vidyasagar, Brain Res (1999).

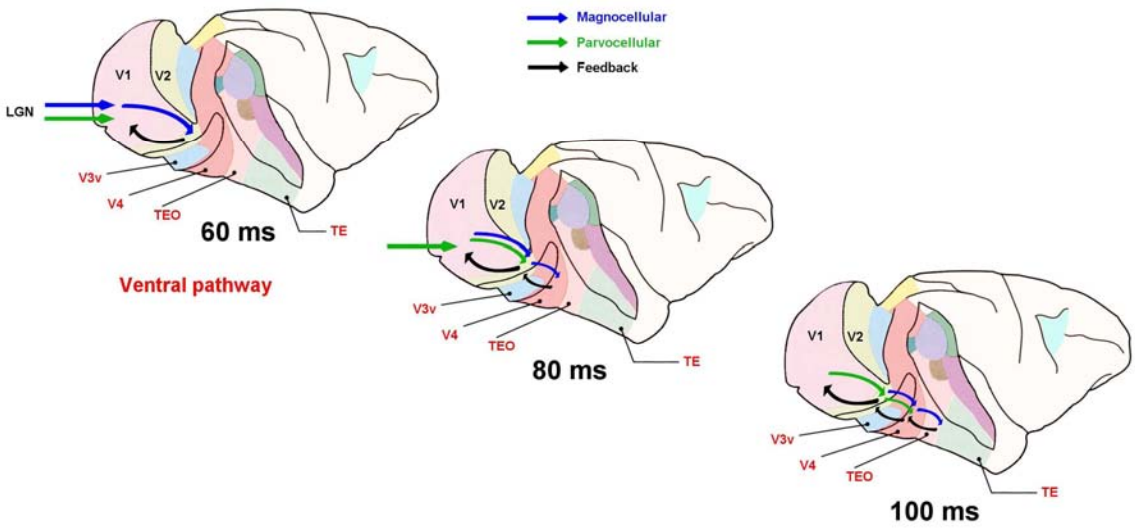


Figure 7 : Même si les systèmes magnocellulaires et parvocellulaires convergent en V2 avant de se projeter sur V4, il est toujours possible de distinguer ces 2 systèmes dans les premières étapes de la voie ventrale grâce à leurs différences importantes en terme de décours temporel. Ainsi, les 20 ms d'avance que possède le système magnocellulaire dans le LGN et dans V1 seraient partiellement conservées dans V2. Ces informations magnocellulaires qui donnent une description relativement grossière de la scène pourraient venir influencer les traitements en cours dans cette même voie ventrale sur les informations parvocellulaires.

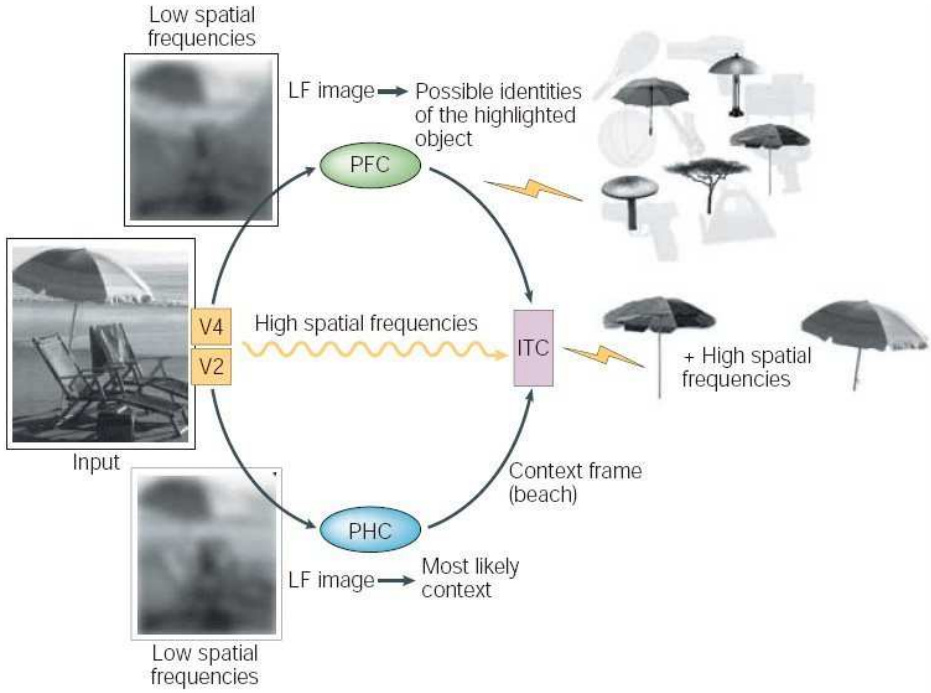


Figure 8 : On retrouve dans ce modèle l'idée de pré-traitements effectués rapidement sur une quantité limitée d'informations pour faciliter les traitements ultérieurs lorsque le reste de l'information parvient dans le système. Ici, le rôle principal échoirait au FEF qui servirait à établir très rapidement des cartes de saillance perceptive pour guider les traitements et au cortex parahippocampique qui aiderait le cortex inféro-temporal dans sa reconnaissance d'objets en restreignant la recherche aux représentations les plus pertinentes grâce à des indices provenant du contexte de l'objet. Reproduit d'après Bar, Nature Rev Neurosci (2004).

Dans les modèles présentés ci-dessus, les informations magnocellulaires jouent à chaque fois un rôle important dans la reconnaissance des objets. Une manière de préciser cette importance consiste à essayer de déterminer ce que le système visuel peut réaliser sur la base des seules informations magnocellulaires.

1.4 - Catégorisation ultra-rapide : robustesse aux variations de contraste

Nous avons vus que les systèmes magno- et parvo-cellulaire présentent de nombreuses différence en termes de sensibilité à la couleur, de vitesse de conduction et de sensibilité aux fréquences spatiales. Nous avons également mentionné que la sensibilité aux contrastes de luminance du système magnocellulaire est bien plus élevée que celle du système parvocellulaire. Les cellules du système magnocellulaire continuent de décharger pour des contrastes inférieurs à 2-3% alors que les réponses des cellules ganglionnaires du système parvocellulaire cessent en dessous de 10% de contraste (Enroth-Cugell & Robson, 1966 ; Shapley *et al.*, 1981 ; Derrington & Lennie, 1984). Ainsi, lorsque le contraste est fortement diminué dans des images en niveaux de gris, la perception ne repose probablement plus que sur des informations provenant du système magnocellulaire. Si l'hypothèse que nous proposons d'une catégorisation basée sur les informations magnocellulaires est correcte, les sujets devraient pouvoir catégoriser avec de bonnes performances des images dont le contraste est réduit à des niveaux inférieurs au seuil de réponse du système parvocellulaire.

1.4.1 - Expériences chez l'homme et le singe : article n°1

Nous avons exploré les performances de catégorisation rapide de l'homme lorsque des images sont présentées dans des conditions extrêmes de contraste. L'article qui décrit cette expérience a été publié en 2005 dans "*European Journal of Neurosciences"*. La même expérience a été reproduite chez le singe et les résultats obtenus font l'objet d'un article en préparation.

Résumé des deux publications : "Rapid categorization of achromatic natural scenes: how robust at very low contrasts?" et "Monkey and humans can categorize achromatic natural scenes with large variations of luminance and contrast"

Les sujets (hommes et singes) devaient effectuer une tâche de catégorisation animal/non animal. Les images étaient présentées en niveau de gris dans 8 conditions de contraste différentes. Le contraste initial de l'image en N&B était divisé par 4, 8, 10... et jusqu'à 32 tout en conservant la même luminance moyenne. Les conditions dans lesquelles le contraste était

divisé par dix ou plus présentent des niveaux de contrastes en dessous du seuil de réponse du système parvocellulaire et seul le système magnocellulaire procure alors des informations sur les variations locales de luminance encore présentes dans les images. En fait il s'agit même ici d'une franche sur-estimation, le seuil de 10% de contraste pour les réponses parvocellulaires ayant été déterminé à l'aide d'échiquiers ou de réseaux en noir et blanc (Kaplan & Shapley, 1986 ; Shapley & Perry, 1986). Dans une image naturelle, il est très rare de trouver côte à côte les pixels ayant les luminances maximale et minimale de l'image. Au contraire, une évaluation du contraste local dans les images naturelles montre que le contraste local le plus fréquent est 0% et qu'une proportion pratiquement nulle de pixels possèdent un contraste local supérieur à 10% lorsque le contraste est divisé par 8. Les performances des sujets reposent donc principalement sur les informations du système magnocellulaire dès la condition N/8. Le temps de réaction, la précision et, dans le cas des sujets humains, l'activité cérébrale étaient mesurés pendant l'expérience pour pouvoir comparer entre elles les performances dans les différentes conditions de contraste.

Les sujets travaillaient sur des séquences de stimuli présentant des images dans toutes les conditions de contraste. La principale différence entre l'expérience chez l'homme et le singe concerne la proportion de conditions de contraste difficiles dans une série donnée, qui a été diminuée pour tester les singes afin de préserver un bon niveau de motivation (la condition où le contraste est divisé par 14 est remplacée par une condition où le contraste est divisé par 2 et la fréquence des conditions les plus difficiles est réduite de 1/8 à 1/25 afin d'atteindre environ 80% de réponses correctes -donc récompensées- dans une série). Le nombre d'images est également réduit (600 contre 1728) pour limiter le nombre total d'essais à 30 000 par singe (chaque image est en effet vue au moins 2 fois dans chacune des condition (2*600*25). L'expérience a pour objectif de mesurer l'évolution de la performance en fonction du contraste et de comparer les effets de réductions de contraste chez l'homme et le singe. Les résultats obtenus chez l'homme dans ce type de tâches de catégorisation ont jusqu'à présent toujours été reproduits chez le singe ; cette validation étant bien sûr très importante avant de mener des études plus invasives chez le singe pour étudier à une échelle plus fine les processus mis en œuvre lors de cette tâche.

Résultats résumés :

Chez les deux espèces, la précision diminue régulièrement avec la baisse de contraste dans les images : elle est élevée quand le contraste des images n'est pas modifié (88% chez les hommes et 95% chez les singes), mais elle n'atteint le niveau chance (49-52%), et ceci pour

les deux espèces, que dans la condition la plus extrême quand le contraste est divisé par 32 (Figure 9 (attention, la courbe des humains est différente de celle de l'article, voir légende)). Le résultat principal est que la précision reste élevée dans les conditions pour lesquelles le contraste est divisé par 8, 10 ou 12 (de 80 à 63% correct), alors que le système magnocellulaire est pratiquement le seul activé. La précision pour la condition dans laquelle le contraste est divisé par 16 est également au-dessus du niveau chance avec 56% de réponses correctes chez les hommes et 63% chez les singes.

Le temps de réaction minimal (voir table 1 de l'article n°1) augmente régulièrement avec la diminution de contraste, reflétant, au moins en partie, le fait que les informations visuelles sont disponibles de plus en plus tardivement avec la diminution du contraste et du rapport signal/bruit, en accord avec l'augmentation des latences dans le système visuel lorsque le contraste diminue (voir Albrecht *et al.*, 2002 pour une revue)

En ce qui concerne les EEG enregistrés chez l'homme, le fait le plus marquant concerne la réduction de l'amplitude de l'activité différentielle entre les cibles et les distracteurs avec la diminution de contraste. Cette amplitude est fortement corrélée avec la précision des sujets dans les différentes conditions. A noter qu'il n'y a plus d'activité différentielle détectable quand la précision des sujets chute au niveau de la chance (condition N/32).

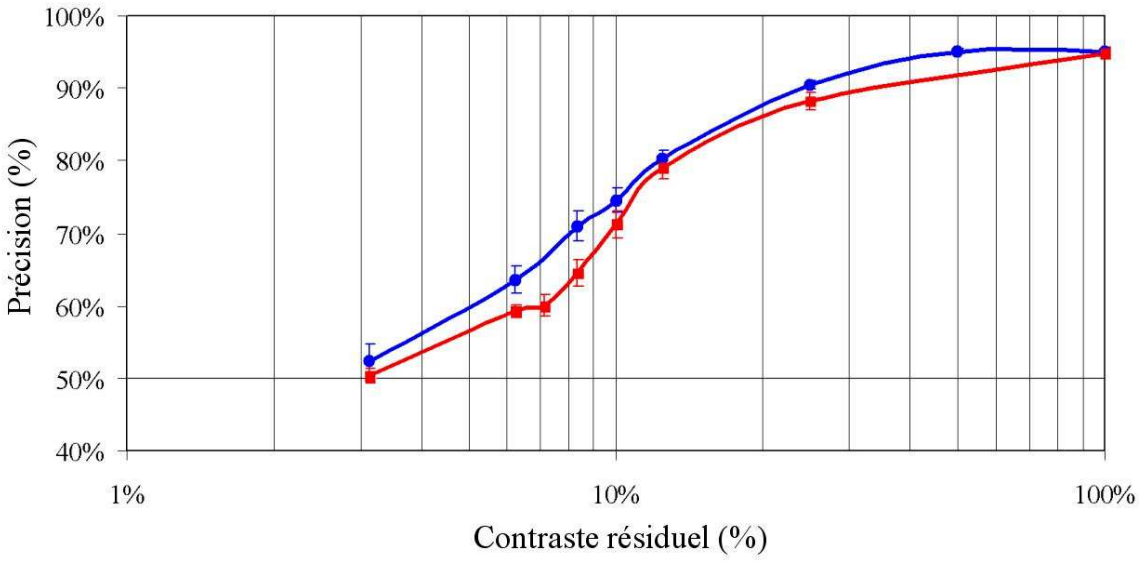


Figure 9 : Évolution de la précision en fonction du contraste résiduel pour les hommes (en rouge) et les singes (en bleu). Pour une meilleure comparaison des performances des deux espèces, les résultats des deux singes ont été moyennés et les données des humains se limitent à la performance des seuls sujets experts (habitués à effectuer la tâche de catégorisation animal/non animal, soit 16 sujets parmi les 24) sur le même groupe d'images que les singes. La courbe de performance des humains est donc différente de celle de l'article qui rapporte la précision pour l'ensemble des sujets sur l'ensemble des images.

Discussion :

La tâche de catégorisation visuelle rapide est donc extrêmement robuste aux variations de contraste, ce qui donne tout d'abord un nouvel argument pour affirmer que la catégorisation visuelle rapide peut s'effectuer sur la seule base des informations magnocellulaires. Cette robustesse au contraste est aussi un bon argument pour penser que le système visuel ne s'appuie pas simplement sur la statistique des zones de contraste ou des contrastes locaux pour effectuer la tâche puisqu'il n'est que très peu sensible aux manipulations qui les affectent.

La grande similarité des résultats entre les singes et les hommes dans cette expérience vient une fois de plus confirmer que les systèmes visuels de ces deux espèces leur permettent d'atteindre des performances similaires malgré une divergence évolutive de plusieurs dizaines de millions d'années. Ces résultats suggèrent que le système visuel des humains et des macaques pourraient partager des principes de fonctionnement similaires.

Les informations magnocellulaires impliquées dans les processus sous-jacents peuvent appartenir au contingent de la voie dorsale ou à celui de la voie ventrale. Mais comme il semble difficile d'exclure la participation de la voie ventrale dans des tâches de reconnaissance d'objet, les informations magnocellulaires au sein de la voie ventrale pourraient donc avoir un important rôle à jouer ; un rôle plus précoce que celui qu'assureraient les informations parvocellulaires. Lors du fonctionnement normal du système visuel, les informations magnocellulaires pourraient être utilisées pour orienter les traitements sur le flux d'informations parvocellulaires grâce aux 10 à 20 ms d'avance dont elles disposent. La représentation grossière de la scène visuelle construite à partir de ces informations magnocellulaires précoces pourrait être très souvent suffisante pour déclencher une réponse correcte dans la tâche de catégorisation utilisée ici, sans que l'analyse plus fine et chromatique apportée par les informations parvocellulaires ne soit nécessaire. Il nous faudra regarder dans une prochaine expérience si une tâche de catégorisation dans laquelle des distinctions plus fines qu'animal/non animal doivent être faites, par exemple pour des catégorisations de type chien/non chien et oiseau/non oiseau comme nous le verrons dans le 3ème chapitre sont possibles à réaliser sur la base des seules informations magnocellulaires. Le niveau de catégorisation à partir duquel cette tâche deviendra impossible nous donnera des indications précieuses sur la finesse des représentations créées grâce aux informations magnocellulaires.

1.4.2 - Les activités différentielles précoces et le contraste

Nous avons souligné dans l'introduction que l'utilisation de deux tâches différentes de catégorisation avaient permis à VanRullen et al. de montrer que les activités différentielles

présentes avant 150 ms sont -au moins en grande partie- dues à des différences physiques entre les groupes de stimuli. Les données électrophysiologiques de l'expérience de catégorisation d'images à différents niveaux de contraste présentée ci-dessus peuvent être analysées d'un autre point de vue que celui exposé dans l'article pour porter un éclairage différent sur l'interprétation des activités différentielles précoces. Le signal enregistré sur les cibles et les distracteurs pour lesquels le contraste est normal présente un profil caractéristique avec une différence précoce entre les deux signaux qui apparaît à une latence courte : entre 100 et 130 ms (Figure 10). On retrouve ensuite l'activité différentielle autour de 150 ms corrélée à la tâche effectuée par le sujet et discutée en détail dans l'article. Il est intéressant de remarquer que l'activité différentielle précoce n'est enregistrée que pour la condition où le contraste est maximal. En effet, la condition dans laquelle le contraste est divisé par 4, ainsi que toutes les autres conditions, ne présentent pas ce profil. La disparition de cette activité différentielle précoce est peut-être à rechercher dans la diminution de la saillance perceptive des détails d'une image avec la diminution de contraste, ce qui a pour conséquence de réduire les différences statistiques entre images cibles et distracteurs. Ainsi, lorsque les différences physiques sont réduites (ici, par une réduction du contraste), les activités différentielles précoces disparaissent. Une réduction plus graduelle du contraste aurait probablement permis d'observer plusieurs stades dans la disparition de cette activité différentielle précoce.

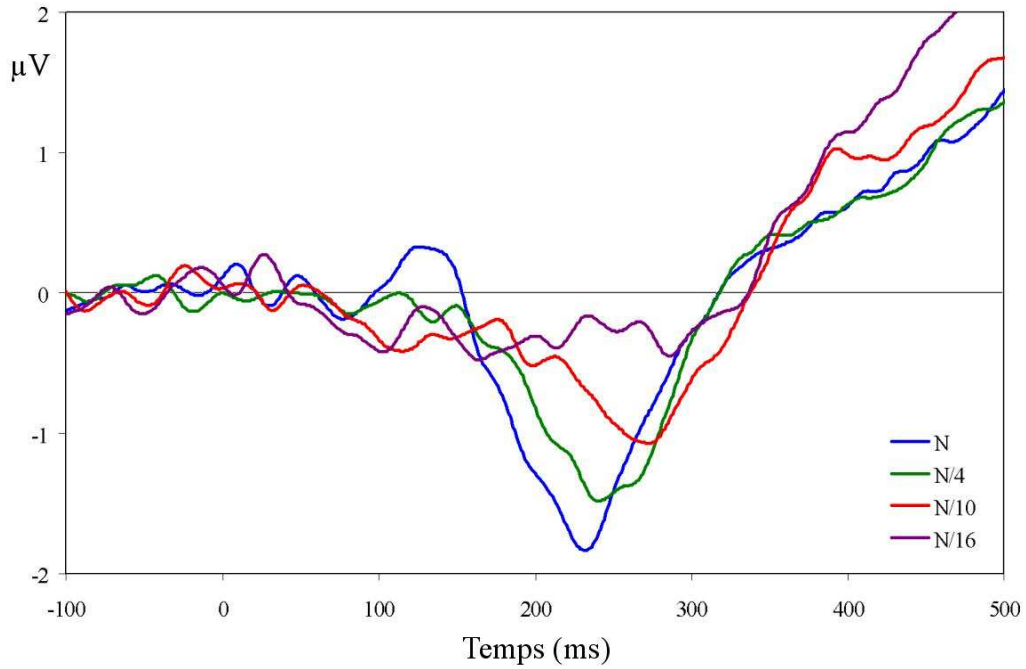


Figure 10 : Signal différentiel (cibles - distracteurs) enregistré pour différentes conditions de contraste. Seul le signal enregistré pour la condition dans laquelle les images ont un contraste normal présente une activité différentielle précoce entre 100 et 150 ms (N, en bleu). Dans les autres conditions, la réduction du contraste entraîne la disparition de cette activité différentielle précoce (N/4, N/8 et N/16 en vert, rouge et violet).

Article n°1

Eur J Neurosci, **21**, 2007-2018

Rapid categorisation of achromatic natural scenes: how robust at very low contrasts?

Marc J-M. Macé, Simon J. Thorpe & Michèle Fabre-Thorpe

European Journal of Neuroscience, Vol. 21, pp. 2007–2018, 2005

Rapid categorization of achromatic natural scenes: how robust at very low contrasts?

Marc J.-M. Macé, Simon J. Thorpe and Michèle Fabre-Thorpe
Centre de Recherche Cerveau et Cognition (UMR 5549, CNRS-UPS), Faculté de Médecine de Rangueil, 133, route de Narbonne, 31062 Toulouse, France

Keywords: animal categorization, EEG, human performance, low contrast, magnocellular, natural images

Abstract

The human visual system is remarkably good at categorizing objects even in challenging visual conditions. Here we specifically assessed the robustness of the visual system in the face of large contrast variations in a high-level categorization task using natural images. Human subjects performed a go/no-go animal/nonanimal categorization task with briefly flashed grey level images. Performance was analysed for a large range of contrast conditions randomly presented to the subjects and varying from normal to 3% of initial contrast. Accuracy was very robust and subjects were performing well above chance level ($\approx$ 70% correct) with only 10–12% of initial contrast. Accuracy decreased with contrast reduction but reached chance level only in the most extreme condition (3% of initial contrast). Conversely, the maximal increase in mean reaction time was $\approx$60 ms (at 8% of initial contrast); it then remained stable with further contrast reductions. Associated ERPs recorded on correct target and distractor trials showed a clear differential effect whose amplitude and peak latency were correlated respectively with task accuracy and mean reaction times. These data show the strong robustness of the visual system in object categorization at very low contrast. They suggest that magnocellular information could play a role in ventral stream visual functions such as object recognition. Performance may rely on early object representations which lack the details provided subsequently by the parvocellular system but contain enough information to reach decision in the categorization task.

Introduction

There is a huge literature concerning the sensitivity of the visual system as a function of contrast, but the vast majority of these studies have involved electrophysiological or behavioural responses to relatively simple visual stimuli such as static, moving or flickering gratings and bars (De Valois *et al*., 1974; Kaplan & Shapley, 1982; Schiller *et al*., 1990; Sclar *et al*., 1990; Shapley, 1990). A few studies have looked at particular visual tasks such as letter and figure recognition, conjunction search or reading (Legge *et al*., 1987; Strasburger *et al*., 1991; Strasburger & Rentschler, 1996; Nasanen *et al*., 2001; Cheng *et al*., 2004). Complex objects and human faces at different contrasts were used in two studies. In the first one, the authors used a limited set of hand drawings at four different contrasts and showed that accuracy performance remained high above 10% contrast. Moreover, the response along the ventral stream brain areas became increasingly contrast-invariant (Avidan *et al*., 2002). The second study used a simple detection task and contrast was at most divided by two, with a marginal effect on reaction time (Lewis & Edmonds, 2003). To our knowledge, no other study involved high-level object recognition and scene processing as a function of contrast. This is unfortunate given that object recognition in natural scenes is one of the most important functions of the visual system.

Under normal visual conditions, human beings can be extremely fast in extracting the meaning of natural visual scenes (Potter, 1976; Intraub, 1981; Keysers *et al*., 2001). In a go/no-go categorization task in which subjects have to determine whether or not a photograph of a natural scene contains a target object (e.g. an animal or a means of transport) they are able to score $\approx$94% correct, with early motor responses appearing before 300 ms (Thorpe *et al*., 1996; VanRullen & Thorpe, 2001a). However, in everyday life, visual conditions are often far from being optimal; at dusk or dawn, for example, luminance and contrast can be very low and conditions might not allow the processing of colours. When faced with such challenging everyday conditions our visual system still appears very efficient. To what extent is high-level scene categorization possible when the contrast of the image is severely reduced, as in the case of a natural phenomenon such as fog?

Low contrast and luminance prevent information about colour from being used efficiently, and it might be thought that the absence of colour would have a major impact on performance. Indeed, there have been a number of studies showing that colour can have an early important role for high-level visual tasks (Gegenfurtner & Rieger, 2000; Delorme *et al*., 2004), but previous studies from our laboratory have also demonstrated that removing of colour information in rapid visual categorization tasks has remarkably little effect (Delorme *et al*., 2000). Specifically, the influence of colour cues on the onset of correct 'go' responses towards targets is not visible before 400 ms, at which point more than 50% of the responses have already been produced. An achromatic object representation can thus be sufficient to trigger an adequate motor response. Because the achromatic magnocellular information reaches V1 $\approx$20 ms before the chromatic parvocellular information (Maunsell & Gibson, 1992; Nowak *et al*., 1995; Schmolesky *et al*., 1998), this result had led us to propose that the magnocellular achromatic pathway could have a crucial role to play in

Correspondence: Dr Marc Macé, as above.
E-mail: marc.mace@cerco.ups-tlse.fr

Received 6 September 2004, revised 31 January 2005, accepted 2 February 2005

doi:10.1111/j.1460-9568.2005.04029.x

early object processing. However, in such cases this representation is presumably very coarse. Because magnocellular ganglion cells in the macaque retina are eight times less densely packed than parvocellular cells (Silveira & Perry, 1991), with more convergence from photoreceptors (Dacey & Brace, 1992; Dacey & Petersen, 1992; Sun, 2001), magnocellular spatial resolution is relatively poor. Nonetheless, such coarse representations might be sufficient for some forms of object categorization.

The present experiment was specifically designed to determine the robustness of human performance in an animal vs. nonanimal rapid visual categorization task using achromatic natural images and large reductions of contrast.

In addition, we can use the different contrast sensitivities of the different visual pathways to address another question. In the cat's retina, parvocellular (X) cells stop responding below 10% contrast whereas magnocellular (Y) cells can still fire at residual contrasts of 2–3% (Enroth-Cugell & Robson, 1966). Similar results have been found in the macaque retina (Kaplan & Shapley, 1986) and in the lateral geniculate nucleus (Shapley *et al*., 1981; Derrington & Lennie, 1984). Thus, the present experiment could also provide clues about a possible role of magnocellular pathways in object vision at very low contrasts.

Materials and methods

Subjects

Twenty-four subjects (12 males and 12 females) aged 22–52 years (mean 30) performed the experiment. All participants had normal or corrected-to-normal vision. They volunteered for the study and gave their written informed consent. The study conformed to the Code of Ethics of the World Medical Association. Reaction times and accuracy were recorded as well as brain electrical activity using a 32-channel electrocap and a Synamps system.

Go/no-go rapid visual categorization task

The methods were similar to those used in a number of previous studies (e.g. Fabre-Thorpe *et al*., 1998; Delorme *et al*., 2000). Subjects were seated ≈40–50 cm in front of a tactile computer screen in a dimly lighted room. They had to place their fingers on a response pad (a plate with photodiodes) to trigger image presentation. An image was then flashed at the centre of the screen for only 28 ms to prevent ocular exploration. The subjects were verbally instructed to perform a go/no-go animal/nonanimal visual categorization task as quickly and as accurately as possible. When a photograph that contained a target was flashed, subjects had to lift their hand and touch the screen in < 1 s (go response). The reaction time was measured between the onset of the visual stimulus and the finger lift from the response pad. When the trial was a distractor, subjects had to keep their finger(s) on the button (no-go response). To avoid behavioural anticipations, the interstimulus interval time was randomly selected between 1.6 and 2 s (mean 1.8 s). Subjects were given online feedback of results: correct responses, both go and no-go, were indicated by a brief sound.

Stimuli

For the present experiment, 1728 grey-level photographs of natural scenes were used in eight different contrast conditions (13824 stimuli). All images came from a large commercial database (Corel photo library) and were chosen specifically to be as varied as possible (see Fig. 1) with one or more animals of many different kinds and sizes as target images: mammals, fish, reptiles and birds. Distractors were also highly varied with landscapes, trees, flowers, objects of all kinds, human constructions and cars. Subjects had no clues concerning the next photograph and when it contained a target they had no information concerning the viewpoint, the size, the number, the location and the possible occlusion of the target(s).

Images resolution was 384 × 256 pixels and the 17-inch tactile screen was set at a resolution of 800 × 600 pixels. The apparent size of the pictures was 15 × 10° and most images (75%) were horizontal. The 1728 images in 16 million colours were converted to 256 grey levels using Corel photo CD lab software, and then processed using Adobe Photoshop to generate seven other exemplars of each image in which the normal original contrast (N) of the photograph was divided by 4, 8, 10, 12, 14, 16 and 32 (N, N/4, N/8, N/10, N/12, N/14, N/16, N/32). This contrast reduction was done with mean luminance of the image kept constant and corresponds to a division of the standard deviation of the pixel luminance values. Each subject saw each image in a single contrast condition over 18 blocks of 96 images (1728 trials). All contrast conditions for an image were counterbalanced across the group of subjects ($n = 24$) so that any given image was seen at each contrast condition by three different subjects. Contrast conditions and targets and distractors for a given contrast condition were all equiprobable in each testing block and subjects were instructed to try to respond on about half of the trials in each testing condition. Prior to testing, all subjects performed a 50-trial training session using a different set of photographs.

Image statistics

If we consider that the original normal contrast of the image is at 100% contrast, the N/4, N/8, N/10, N/12, N/14, N/16, N/32 stimuli obtained with contrast reduction have, respectively, residual contrast levels of 25, 12.5, 10, 8.3, 7.1, 6.2, and 3.1%.

This residual contrast is a strong overestimation of the overall local contrasts of the test photographs. Classically, contrast studies have used regular sine wave gratings or checkerboard patterns but this type of artificial stimulus is very different from natural images. Local contrasts of natural scenes hardly ever reflect the optimal 100% Michelson contrast that could be achieved with a checkerboard stimulus, as pixels with maximum and minimum values are virtually never placed next to each other. We analysed the local contrast distribution of the images by calculating, for each pixel value converted in luminance intensity of the screen (in candelas per square meter), the mean and maximum absolute Michelson contrast values with the eight surrounding pixels (Fig. 2A and B). Relatively to simpler psychophysical stimuli, maximal local contrasts in natural images seldom reach 100%; nearly 90% of the photographs had < 3% of their maximum local contrast values > 90% Michelson contrast.

Two contrast values, namely 10% and 3%, are of special interest as they correspond to the maximal contrast sensitivity usually attributed to, respectively, the parvocellular and the magnocellular pathways. In the original images, only 41% of the mean pixel-based contrast values were > 10% threshold. This proportion was strongly reduced to 5.9, 0.82 and 0.26 in the N/4, N/8 and N/10 conditions, a proportion that dropped to 0.02 for N/16. For the hardest conditions, N/14, N/16 and N/32, only 4.2, 3.12 and 0.26% of the mean local contrasts were > 3%. Considering the optimal 256 grey level values that were used for the normal condition, subjects could only rely on a maximum of 32 consecutive grey levels for N/8, 25 for N/10, 18 for N/14 and 16 for N/16. These local contrast statistics on the image set show that the visual system had to deal only with local contrasts < 10% for all contrast conditions below N/8 or N/10. The distributions of local contrasts obtained with our set of images were similar to those from

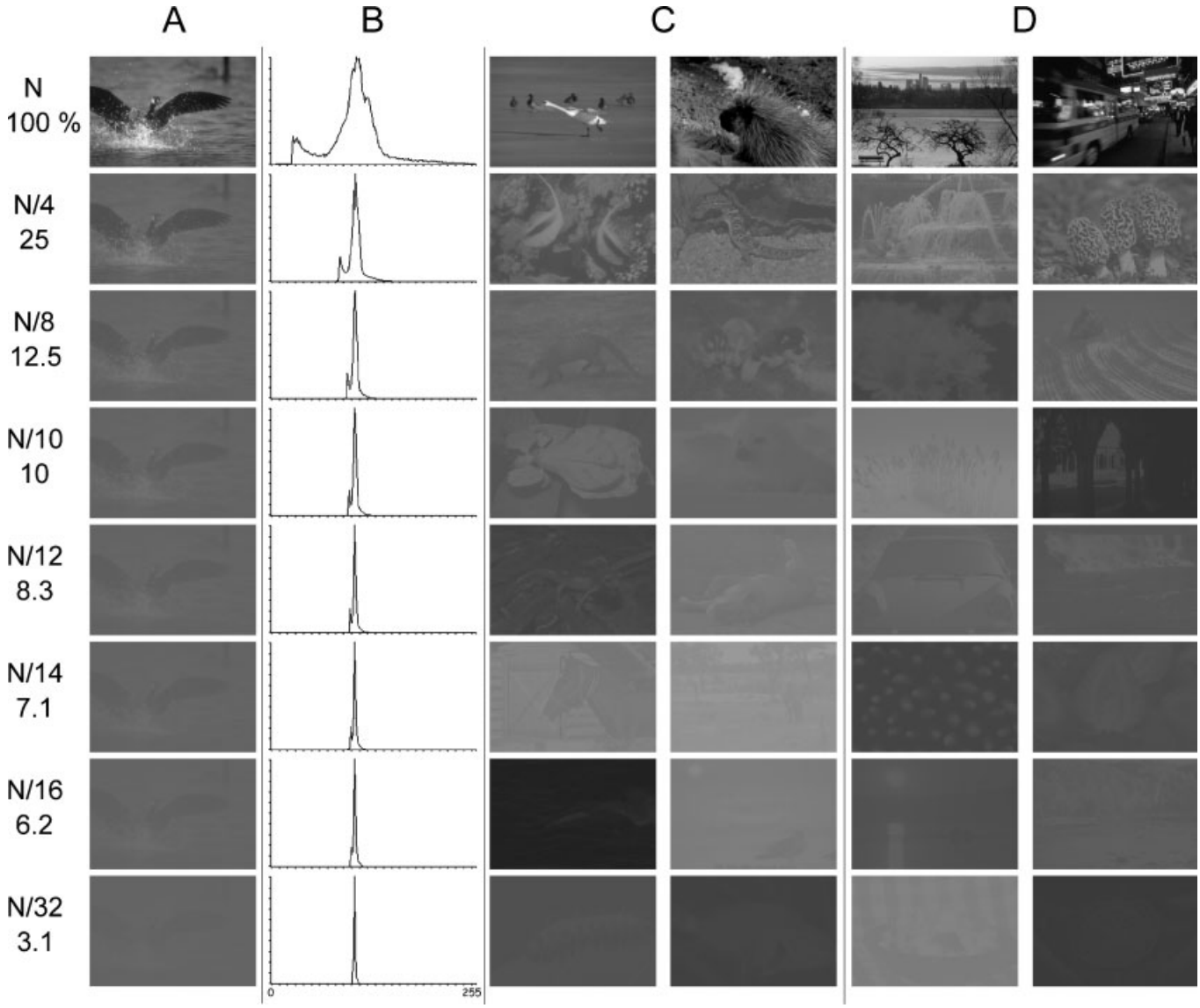


FIG. 1. Examples of stimuli for the eight contrast conditions. From N to N/32, the residual contrast calculated from an initial image considered at 100% contrast in the N condition is indicated below. (A) The same target image is shown in the eight different contrast conditions and (B) the associated distribution of pixel luminance corresponding to the various grey levels (0–255) is shown. Note that the distributions are centred on the same mean luminance value. (C and D) Various examples of target (C) and distractor (D) images for each of the eight contrast conditions are shown.

Ruderman (1994), Brady & Field (2000) and Tadmor & Tolhurst (2000) with small discrepancies which could be explained by differences in image sets.

Evoked-potential recording and analysis

Brain electrical activity was recorded from 32 electrodes mounted in an elastic cap in accordance with the 10–20 system and completed by additional occipital electrodes connected to a Synamps amplifier system (Neuroscan Inc., El Paso, TX, USA). The ground electrode was placed along the midline, ahead of Fz. Impedances were kept < 5 kΩ. The signal was sampled at 1000 Hz and low-pass filtered at 100 Hz with a notch filter at 50 Hz. Potentials were on-line referenced relative to electrode Cz and average re-referenced off-line. Baseline correction was performed using the 100-ms prestimulus interval. Two artefact rejections were applied over the −100 ms to +400 ms time period, the first on frontal electrodes FP1 and FP2 with a criterion of −50 to +50 μV to reject trials with eye movements, and the second on parietal electrodes Oz and Pz with a criterion of −30 to +30 μV to remove trials with excessive alpha rhythms. Only correct trials were averaged. Statistical tests were performed on the original data and the electroencephalogram (EEG) signal shown on figures is low-passed at 30 Hz. Event-related potentials (ERPs) were computed separately for correct target trials and correct nontarget trials and a differential activity was calculated by subtracting the distractor signal from the target signal. This 'differential cerebral activity' was calculated to focus on the differences between the two kinds of trials. It has been shown to reflect successively three different stages of processing. Whereas its early phase, starting ≈ 75 ms post stimulus onset, appears linked to low-level differences between image sets, and its late phase (after 250 ms) to the motor response on target trials, the intermediate phase in the 150–250 ms time window develops in relation with task performance (VanRullen & Thorpe, 2001b; Rousselet *et al*., 2004) and was the focus of the present experiment. As the differential activity

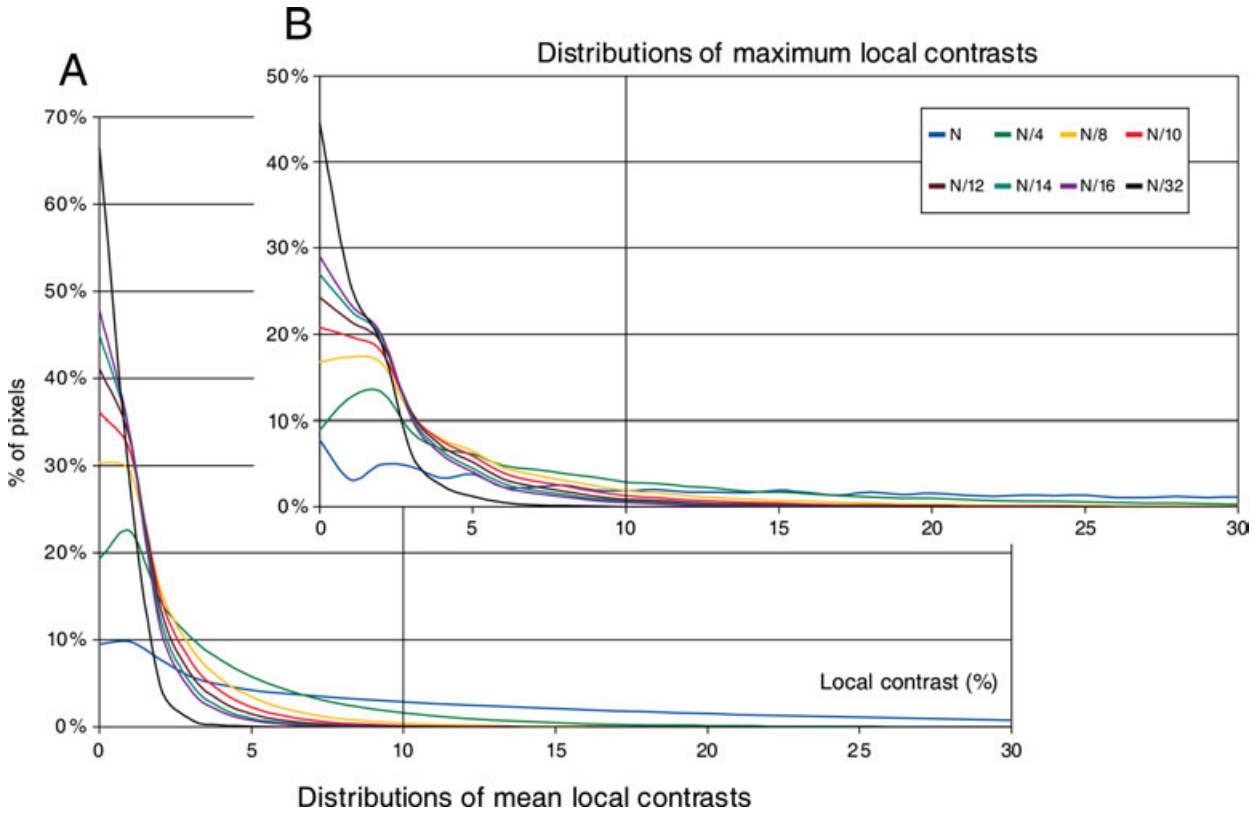


FIG. 2. Distribution of local contrasts (bin size 1%) in all the stimuli and for each condition from N to N/32. The percentage of pixels is plotted in relation to the percentage of (A) mean or (B) maximum Michelson local contrast [Michelson contrast: (Lmax − Lmin)/(Lmax + Lmin)]. The vertical line at 10% corresponds to the contrast commonly given as the parvocellular contrast sensitivity threshold measured in the LGN.

was delayed in extreme contrast conditions, its peak amplitude and peak latency were measured in a much larger time window, 220–320 ms, which included the task-related differential activity in all contrast conditions. To look for the onset of this differential activity a 130–320-ms time window was considered. Following Rugg *et al.* (1995), the onset value of this differential activity is evaluated by applying paired *t*-tests every ms at each scalp location. Normally, the *t*-tests values have to result in probabilities <0.01 for at least 15 consecutive bins; however, in the present study, because of low signal-to-noise ratio in extreme contrast conditions, an estimation of the onset value is given using a significant *t*-test value <0.05 for 10 consecutive steps.

Results

Behaviour

Accuracy

We evaluated behavioural performance in terms of accuracy and reaction time for each condition. A χ^2 test between correct and incorrect responses determined whether accuracy was above chance level, set at 50% as targets and distractors were equally likely. For the 100% contrast condition, the mean accuracy was $>88\%$, a score that is slightly below the accuracy obtained previously (Delorme *et al.*, 2000) with grey level photographs categorized among coloured photographs (93% correct) and closer to the value (91.4% correct) obtained in a challenging categorization experiment when grey level images were followed by a strong mask after 100 ms (Bacon-Mace *et al.*, 2005). As expected, we observed a significant accuracy decrease with contrast reduction (Fig. 3A and Table 1). Compared to the N condition, accuracy dropped by 7% in the N/4 condition where subjects scored 81.2% correct. Each contrast reduction induced a statistically significant drop in accuracy relatively to the preceding contrast condition (Fig. 3A). However, for intermediate conditions (N/8, N/10 and N/12), accuracy remained at a good level (72.4, 67.4 and 62.7% correct) even though the visual system was faced with images where virtually all the mean local contrast values were $<10\%$. Even at more extreme conditions N/14 and N/16, accuracy, although very poor, was still above chance (N/16: 56%, $\chi^2 = 93.982$, d.f. $= 1$, $P < 0.001$). In fact, chance level was not reached until the hardest task condition in which contrast was divided by 32 (49% correct).

Reducing contrast did not affect all images equally. In particular, it appears that the amount of local image contrast is important. When the 864 target images were classified into three equal groups of 288 images, according to their average value of local contrasts (using either maximum or mean local contrasts), there was no difference in accuracy between the three groups when the contrast was normal (condition N). However, with reduced contrast, there was a clear accuracy advantage for the group with the largest amount of local contrast. The maximal accuracy bias was observed in the N/12 condition with 17% more correct responses for the photographs with highest local contrasts.

Subjects were instructed to try and keep responding on $\approx 50\%$ of the trials in all contrast conditions. Overall, they succeeded well because they responded on 51.0% of trials. However, the response rate depended on the condition with a bias towards not responding at low contrasts (below N/12) and a tendency to over-respond at higher

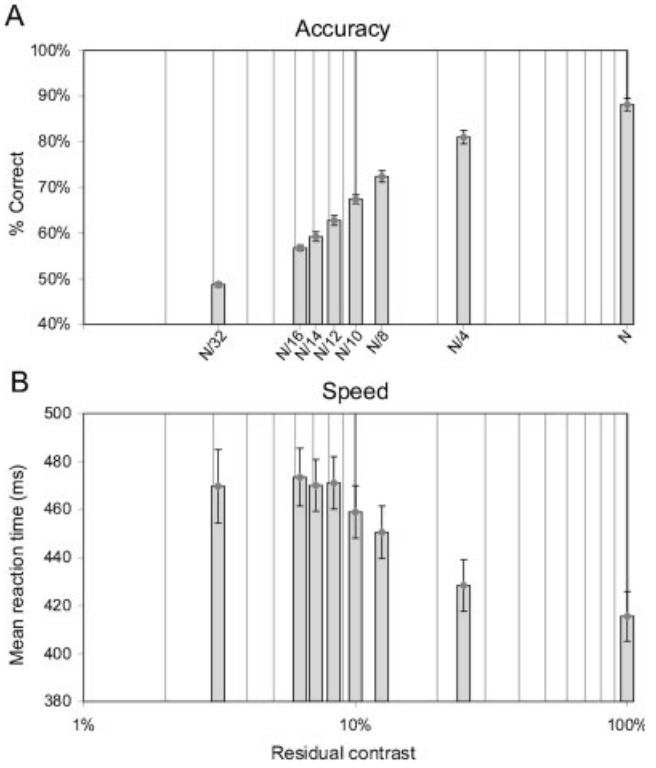


FIG. 3. Average (A) accuracy and (B) speed of performance illustrated by the mean reaction time for the group of 24 subjects and across all contrast conditions as indicated on graph A (from N to N/32). The horizontal axis represents the residual contrast computed with the N condition at 100% contrast and expressed on a logarithmic scale. Error bars are ± SEM. Note that the accuracy takes into account correct responses on target and distractor trials whereas RT values are only obtained with correct go responses on target trials.

contrasts. Interestingly, despite these variations, the false alarm rate remained remarkably constant across all contrast reduced conditions (mean 18%, range 15–19.7% of the total trials). As a consequence, the main effect of reducing contrast was on the proportion of correct hits which decreased from 47% for condition N to 14.1% of the total number of trials at N/32 (see Table 1).

Speed

Mean and median reaction times were also affected by the reduction of contrast but this effect was limited to the first few conditions only. From the N condition [mean reaction time (RT) 416 ms] and up to the N/12 condition (mean RT 473 ms), the increase in mean RT was progressive to reach a maximum of 57 ms (Fig. 3B). Each contrast reduction induced a statistically significant increase (see Table 1). There was virtually no more increase when contrast was further reduced (N/14, N/16 and N/32 compared to N/12). This plateau could suggest that maximal processing of the available information had been done so that any further delay was unable to provide more evidence for decision making.

This increase in reaction time can be seen in the RT distributions computed for each contrast condition (Fig. 4A). The shape of the distribution compared to the 100% contrast condition was nearly unaffected for N/4. As contrast decreased, the distribution was more and more flattened and shifted towards longer latencies. All responses were affected including the earliest ones. As subjects were explicitly required to produce their responses as fast as possible, these early responses are of great interest and can set the minimum input–output processing time. This minimum RT can be defined as the latency at which correct go responses start to statistically outnumber incorrect ones. Table 1 clearly shows that this minimum latency regularly increased with contrast reduction. Calculated on the cumulative number of go responses, it increased from 280 ms for the original N condition to 410 ms for N/16, suggesting that the minimum amount of information used to trigger the earliest responses is available later and later when contrast is reduced.

Electrophysiology

Event-related potentials

The visual and cognitive processing of target and distractor images can be reflected in the electrical activity recorded while the subjects are performing the task. It is assumed that early ERP components are heavily dependant on the physical characteristics of the stimuli and that more and more cognitive processing is reflected by components with longer latencies (Halgren *et al.*, 1994; Foxe & Simpson, 2002; Liu *et al.*, 2002; Proverbio *et al.*, 2002). The effect of contrast reduction was present on the earliest recorded ERP components with the occipital P1 wave being significantly delayed by ≈ 20 ms when

TABLE 1. Accuracy and speed of performance in each of the testing conditions from N to N/32

	N	N/4	N/8	N/10	N/12	N/14	N/16	N/32
Accuracy (%)								
Overall	88.1 ± 6.6	81.0 ± 7.3	72.4 ± 6.0	67.4 ± 4.6	62.7 ± 5.3	59.2 ± 5.2	56.7 ± 2.9	48.7 ± 2.5
Correct go	47.8 ± 1.8	46.0 ± 4.0	40.8 ± 5.0	37.1 ± 6.1	32.3 ± 6.7	28.7 ± 5.5	25.2 ± 8.4	14.1 ± 4.0
Correct no-go	40.3 ± 7.4	35.0 ± 5.4	31.5 ± 4.8	30.3 ± 5.5	30.4 ± 6.0	30.5 ± 5.8	31.5 ± 9.1	34.6 ± 4.5
Go-response rate	57.5 ± 7.1	61.0 ± 8.0	59.3 ± 7.5	56.9 ± 8.6	51.9 ± 10.3	48.2 ± 11.5	43.7 ± 10.9	29.5 ± 17.3
RT (ms)								
Mean	416 ± 50	430 ± 53	452 ± 54	462 ± 54	473 ± 54	472 ± 53	476 ± 59	476 ± 75
Median	407 ± 52	417 ± 53	441 ± 54	450 ± 54	459 ± 52	458 ± 56	459 ± 63	464 ± 73
Minimum RT (ms)								
10-ms bin	280	310	330	370	410	NS	NS	NS
Cumul. 10-ms bin	280	300	320	340	350	370	410	NS

Accuracy and RT values are ± SD. The average overall accuracy is given for the group of 24 subjects together with the relative accuracy on target (go responses) and distractors (no-go responses). As subjects were instructed to respond on half of the trials, the response rate obtained in each testing condition is also indicated. The mean and median RT (in ms) are averages of the 24 individual mean and median RTs. The minimum RT was calculated for each condition with all subjects pooled together. It was computed for noncumulated and cumulated data over 10-ms time bins. A minimum of three consecutive significant χ^2 tests at $P < 0.01$ was required to be confident that the performance was over chance level.

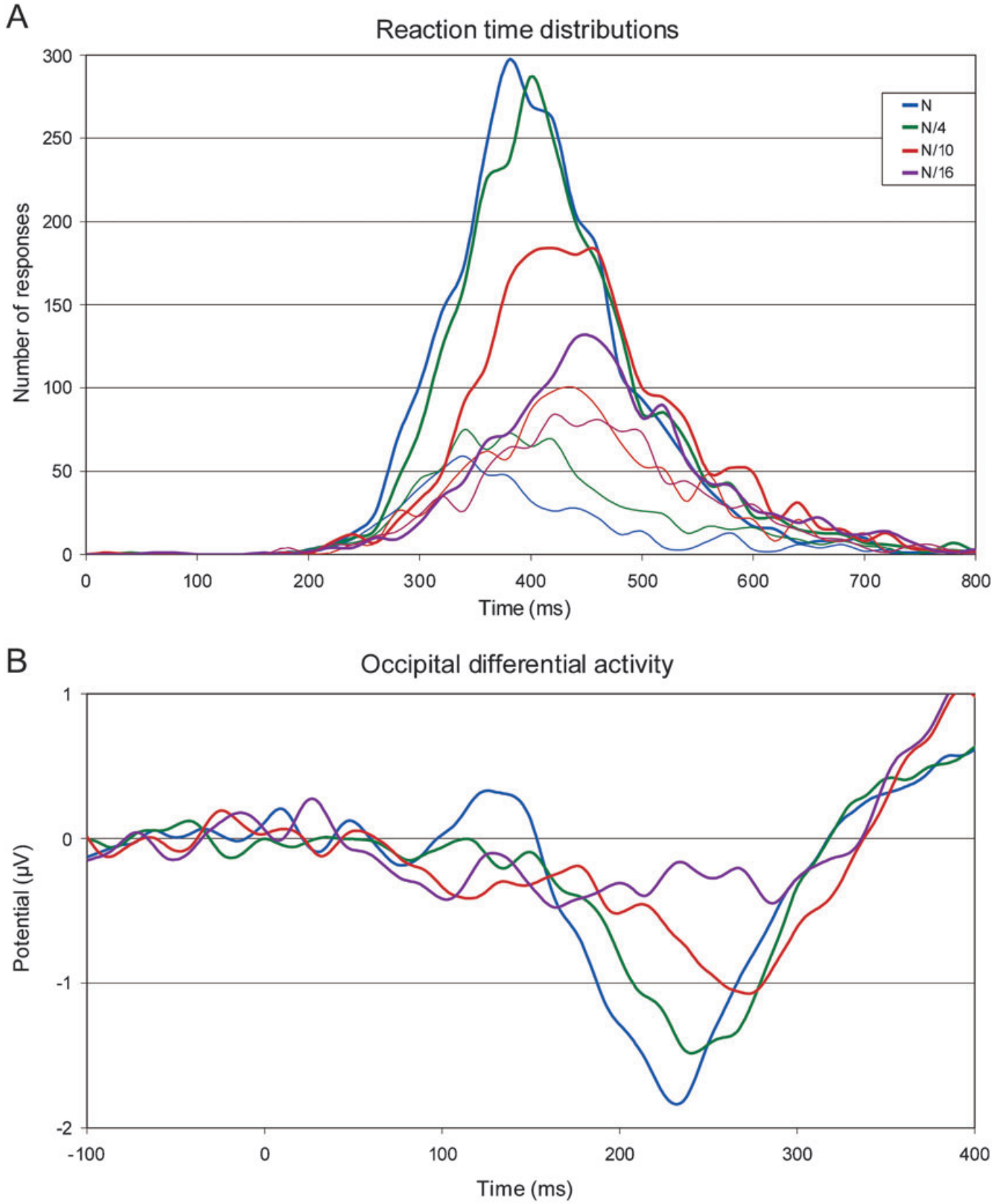


FIG. 4. (A) Reaction time (RT) distributions for the 24 subjects in four of the eight contrast conditions: N (blue), N/4 (green), N/10 (red) and N/16 (purple); time bins are 10 ms. RT distributions of correct go responses on targets are shown in thick lines and RT distributions of false alarms on distractors in thin lines. (B) Differential activity between target and distractor ERPs are averaged from seven occipital electrodes (O1, O2, PO7, PO8, O9, O10 and Oz) for the same four contrast conditions.

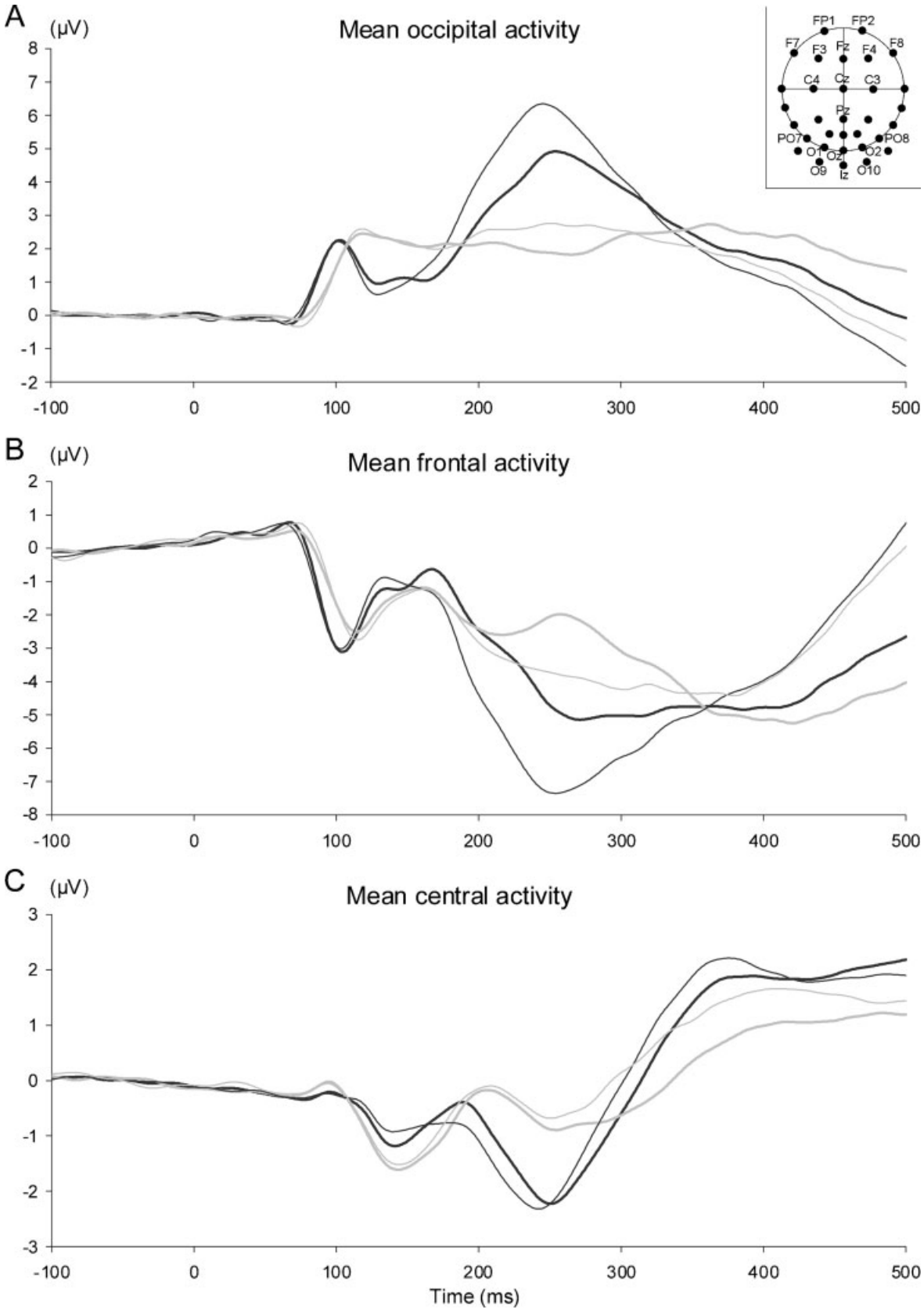


FIG. 5. Mean ERP signal in two contrast conditions, N and N/8; N condition in black, N/8 condition in grey. Targets, thick traces; distractors, thin traces. (A) Average signal from seven occipital electrodes (O1, O2, PO7, PO8, O9, O10 and Oz). (B) Average from five central electrodes (C3, C4, P3, P4 and Pz). (C) Average from seven frontal electrodes (FP1, FP2, F3, F4, F7, F8, Fz). The location in the 10–20 system of all cited electrodes is shown in the top right-hand corner.

TABLE 2. Average onset latency, peak latency and peak amplitude of the differential activity (DA) obtained by subtracting the grand average signal obtained on correct distractor trials from those obtained on correct target trials (24 subjects)

	N	N/4	N/8	N/10	N/12	N/14	N/16
Seven occipital electrodes							
DA onset (ms)	166	171	206	208	221	230	257
Peak latency (ms)	231	240	263	269	270	282	292
Peak amplitude (μV)	−1.94	−1.65	−1.15	−1.26	−0.71	−0.59	−0.75
Seven frontal electrodes							
DA onset (ms)	162	184	217	211	260	227	247
Peak latency (ms)	240	260	263	283	281	281	302
Peak amplitude (μV)	3.11	2.66	2.16	2.20	1.12	1.14	1.45

The grand averages were computed from seven occipital electrodes (O1, O2, PO7, PO8, O9, O10 and Oz) and from seven frontal electrodes (FP1, FP2, F3, F4, F7, F8 and Fz) for the seven contrast conditions from N to N/16.

contrast was divided by 4. However, no further delay was seen with enhanced contrast reductions (Fig. 5). Many studies have shown that with contrast reductions, information flow through the ventral visual system is slowed down due to longer integration times [retina, Shapley & Victor, 1978; lateral geniculate nucleus (LGN), Kaplan *et al*., 1987; Hartveit & Heggelund, 1992; Maunsell *et al*., 1999; V1, Albrecht & Hamilton, 1982; Lupp *et al*., 1976; Maunsell & Gibson, 1992; see also Albrecht *et al*., 2002, for a review]. With contrast reduction, the increased P1 peak latency could partly reflect the increase in neuronal firing latencies in the visual pathway although no scaling effect was observed with increasing contrast reductions.

Differential activities

ERPs were analysed separately for target and distractor correct trials. Target ERP and distractor ERP grand averages were computed for the whole group of subjects. The differential activity was always calculated by subtracting the signal recorded on distractors from the signal recorded on targets. Studies that aimed at localizing the brain generators involved showed that >90% of the occipital and frontal differential activities could be explained by two dipoles located ventrally and laterally in the extrastriate cortex (Rousselet *et al*., 2002; Delorme *et al*., 2004). In the present experiment, statistically significant differential activity could be observed on occipital sites in all contrast conditions with the exception of the N/32 condition in which subjects performed at chance level.

The early differential activity which has been shown to reflect physical differences between the image sets (VanRullen & Thorpe, 2001b), can be observed on occipital electrodes in the highest contrast condition at ≈100 ms. With increasing contrast reductions, this early differential activity disappears progressively as low level differences between target and distractor images become less prominent.

On the other hand, the large differential activity building up after 250–300 ms corresponds to the differential motor activation between correct go and no-go responses. In the present study the effect related to motor activation was also evaluated by comparing left and right EEG signals in right-handed subjects. Whereas there was no asymmetry between left and right occipital and frontal recorded signals, an important lateralization effect was seen when comparing ERPs recorded on central electrodes C3 and C4. This motor activation developed over the left hemisphere at a latency that was never earlier than 250 ms across all contrast conditions. Such left–right asymmetry limited to central electrodes and developing at longer latencies than the task-related signal shows that, despite its large amplitude, the motor activation cannot contaminate the categorization-related activation. This has also been clearly stated by others (Antal *et al*., 2000 and Johnson & Olshausen, 2003), who showed that the sign of the 150–250 ms differential activity remained unchanged after an inversion of the motor response (i.e. no-go on previous targets and go on previous distractors).

Now, focusing on the categorization-related differential activity that appears in normal contrast conditions in the 150–250 ms window after stimulus onset, an effect of contrast reduction could be seen across all task conditions, both on the latency of the differential activity and on the amplitude and latency of its peak (cf Table 2). Concerning the latency from which the differential activity develops on occipital sites (averaged on seven occipital sites: O1, O2, PO7, PO8, O9, O10 and Oz), there was a pronounced increase with contrast reduction from 166 ms in the N condition to 257 ms for N/16 (see Table 2 and Fig. 4B). The delayed onset of the differential activity was associated with a significant reduction in its amplitude. At occipital sites, the peak amplitude was reduced by more than a half between conditions N and N/16 (Fig. 4B). Finally, this drop in amplitude was also associated with an increase in the latency at which it peaks, from 231 ms to 292 ms. Similar results were observed on frontal sites (Table 2).

Correlations between behaviour and electrophysiological recordings

In the original study (Thorpe *et al*., 1996) it was proposed that the differential activation between go and no-go trials could reflect inhibitory mechanisms on no-go trials. Indeed, the lack of correlation found between the onset latency of the differential effect and the behavioural reaction times -recently confirmed (Johnson & Olshausen, 2003) was consistent with such a hypothesis. However, generators for this differential activity were subsequently found in the extrastriate visual areas. Moreover, we recently showed (Rousselet *et al*., 2004) that ERPs associated with missed target trials were similar to ERPs on distractor trials whereas a differential activity could clearly be seen between ERPs on false alarms and ERPs on distractor trials. It is reasonable to imagine that a behavioural response can be triggered once a sufficient number of neurons tuned to animal features are recruited (correctly or erroneously) by the visual stimulation. For a discussion about the possible origins of this differential activity, see Rousselet *et al*. (2004).

It is thus of great interest to look for correlations between behaviour and the various features of the recorded task-related differential activity.

First, across the different contrast conditions, the decrease in accuracy was highly correlated with the decrease in the peak amplitude of the differential activity. In the case of the occipital and the frontal grand averages, the Pearson R^2 correlation indexes were, respectively, 0.93 and 0.88 (Fig. 6A).

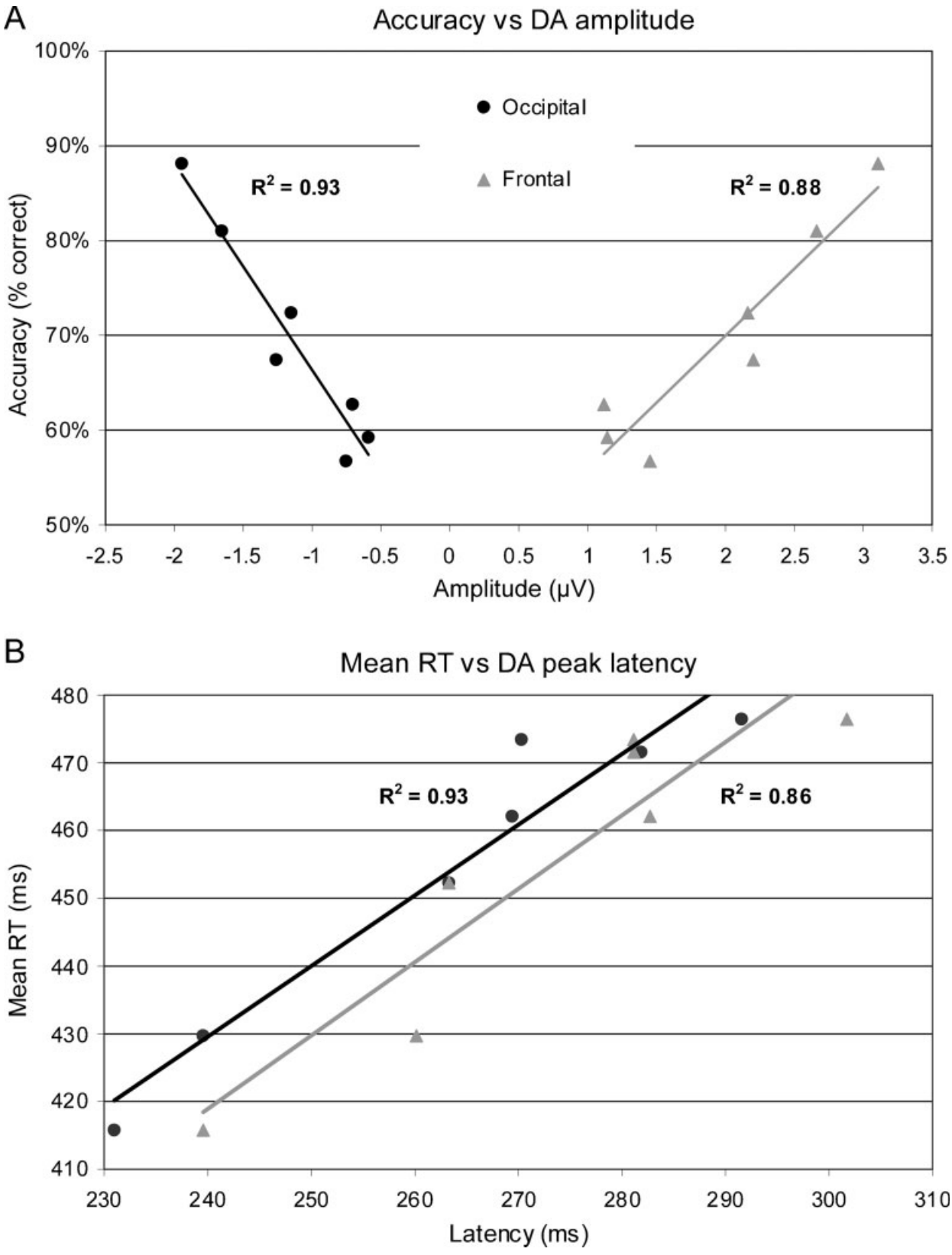


FIG. 6. Correlations between behavioural results and the ERP differential activity (DA) data on occipital (●) and on frontal (▲) electrodes. Mean values were averaged on seven occipital electrodes (O1, O2, PO7, PO8, O9, O10 and Oz) and on seven frontal electrodes (FP1, FP2, F3, F4, F7, F8, Fz). (A) Correlation between the behavioural accuracy in each of the seven contrast conditions and the peak amplitude of the ERP differential activity (expressed in µV). (B) Correlation between the mean reaction time (RT) and the latency of the DA peak (both in ms). R^2 indexes are Pearson's correlation coefficients. Note that the RT values are obtained with correct go responses on target trials whereas the DA signal reflects the difference between ERPs recorded on correct target and distractor trials.

Second, a high correlation (0.93 for occipital electrodes and 0.86 for frontal ones) was also found between the peak latency of the differential activity and the mean reaction time observed in all task conditions (Fig. 6B); the N/32 condition was excluded as no differential activity could be observed and subjects responded at chance level. It is worth noting that the regression curves of occipital and frontal signal correlations are virtually parallel and separated by ≈20 ms, a delay which could reflect the intervention of a second mechanism, presumably more frontal but nevertheless time-linked to the occipital activation.

Discussion

The main result of this study concerns the high robustness of the human visual object recognition system under extreme conditions of stimulus contrast. Subjects still score above chance level with achromatic natural photographs in which only 6–7% of the original contrast is left. In such degraded images, they have to base their responses on a very limited amount of information. In the original N condition, where 256 grey levels were available, 90% of the images used > 200 grey levels but only 3% used the full range of grey levels. When contrast was decreased, the image sharpness dropped dramatically as the number of grey levels was very limited (≤ 25 in the N/10 condition).

The results reported here are the first to specifically address the question of the human visual system efficiency at low contrast with natural image stimuli. In a recent study, Avidan and collaborators presented line drawings of complex objects and faces at different contrasts (Avidan *et al*., 2002) and reported a drop in performance below 10% contrast. Using fMRI, they also showed increasing contrast invariance from V1 to the lateral occipital complex (LOC) in the ventral visual pathway. They stressed the fact that contrast invariance is higher in areas in which neuronal activity is related to complex object representations. Such impressive invariance in high level visual areas to large modifications of contrast has also been stressed in monkeys (Rolls & Baylis, 1986) with natural stimuli such as photographs of faces. Other series of studies have mainly used digits and letters. Strasburger and collaborators (Strasburger *et al*., 1991; Strasburger & Rentschler, 1996) investigated the accuracy of human subjects in a high-level visual categorization task where the contrast of the stimuli was reduced. They found impressive performance levels at low contrast but only when the task was performed centrally, as performance dropped rapidly with eccentricity. The effect of contrast reduction has also been studied in cognitive visual tasks such as visual search for an uppercase character among digits (Nasanen *et al*., 2001), or reading (Legge *et al*., 1987). In these tasks, contrast reduction had a large effect on speed and it also increased the number of eye fixations necessary to perform the task as low contrast impairs peripheral vision.

Unlike the present study where we used a complex object recognition task in which the targets can have a wide range of unpredictable forms and sizes, the results using letters and digits were obtained with a limited number of simple form elements and we could have expected here a dramatic effect of contrast reduction. Although we indeed observed an accuracy decrease, this decrease was very progressive and contrast had to be divided by 32 before subjects reached chance level. Together with the decrease in accuracy, there was a progressive increase in mean reaction time that reached a plateau at N/12. Finally we also observed an increase in the minimum processing time, which could well reflect the fact that subjects need more and more time to gather information about the image features before they had accumulated enough to trigger their behavioural response. This idea can be linked to the model of information accumulation proposed by Schall (2001). Furthermore, this idea of information accumulation is also supported by the results of a masking study using the same sort of go/no-go animal categorization task that showed how performance drops off progressively as the stimulus–mask interval is decreased (Bacon-Macé *et al*., 2005).

In our data, we could relate this accumulation of information to the peak amplitude and latency of the EEG differential activity associated with the task. With contrast reduction we observed a decrease in its amplitude, which may reflect the fact that less and less evidence is available to discriminate targets from distractors (Rousselet *et al*., 2004). This amplitude was indeed highly correlated with the subject's accuracy. Such correlations are interesting as they suggest a direct relation between brain activity and performance level. This differential EEG activity between target and distractor trials is barely visible when subjects performed very poorly (56% correct) in condition N/16, and totally disappears in condition N/32 in which subjects responded at chance level. The slope of the differential activity, less and less steep across the different contrast conditions, may also reflect the speed at which information about the visual scene accumulates over time.

These electrophysiological observations can be discussed in relation to neuronal responses to different contrast conditions. Contrast is a very critical factor for both the strength and the latency of neuronal responses: firing rate is decreased and response onset is greatly delayed when contrast is reduced (Albrecht *et al*., 2002). In the cat retina and LGN, reducing contrast for sinusoidal gratings from 40% to detection threshold leads to a 15–25-ms increase in onset latency (Sestokas & Lehmkuhle, 1986; Sestokas *et al*., 1987). In the striate cortex, a decrease in contrast from 100% to 5–10% generally induces a latency increase of 30–50 ms (Carandini & Heeger, 1994; Albrecht, 1995; Gawne *et al*., 1996; Reich *et al*., 2001). While the effects of contrast on latency up to V1 are relatively modest, they are much more dramatic at higher levels of the visual system. Only one study has reported a value for the increase in onset latency for neurons in the superior temporal sulcus and the infero-temporal cortex (IT) with decreasing contrast (Oram *et al*., 2002). This study used grey level drawings or photographs of various objects at different contrast and showed an increase in latency of up to 150 ms when the contrast is reduced from 100% to 6% (see also Xiao *et al*., 2001). This value is in the same range as the 130 ms increase in minimum processing time observed here on behavioural reaction time between 100% and 6.25% contrast (corresponding to N and N/16; see Table 1). The similarity between the results reported in the present experiment and illustrated in Fig. 4 and the neuronal response curves illustrated in Fig. 6 of the Oram *et al*. (2002) paper is especially striking and argues in favour of a strong relationship between neuronal responses in higher order visual areas, differential EEG activity and psychophysical results. On the other hand this processing delay is not as marked for the mean behavioural reaction times (60 ms) or for the peak latency of the EEG differential activity (61 ms). These two parameters are tightly correlated across all contrast conditions, providing further evidence for a relation between the accumulation of information and the behavioural responses. These latencies increase less than the minimum reaction time as they might reflect the average time needed to process the total amount of available information, an amount that is more and more limited when contrast is reduced.

A role for the magnocellular pathway in early processing for object recognition?

In the rapid animal/nonanimal categorization task used here, subjects could still perform largely above chance level with very low

luminance contrast photographs which might only activate magnocellular cells. Commonly, 10% contrast is considered the minimum contrast value to activate parvocellular retinal cells. The proportion of pixels over this 10% contrast threshold drops below 1% from condition N/8 (see image statistics). Thus, the activation of the parvocellular system might be very low from the N/8 condition with the task being performed in the near-absence of parvocellular inputs at more extreme contrast conditions. Some behavioural (Merigan & Eskin, 1986; Schiller *et al*., 1990) and electrophysiological (Blasdel & Fitzpatrick, 1984; Hubel & Livingstone, 1990) studies have argued for a high parvocellular sensitivity which could result from 'probability summation' in V1 (Watson, 1992). Nevertheless, there is evidence that the significant contrast sensitivity advantage of the magnocellular ganglion cells cannot be totally suppressed by cortical integration (Kaplan *et al*., 1990; for a review see Vidyasagar *et al*., 2002).

Most experiments performed to dissociate the role of the ventral visual system from the dorsal visual system have concentrated on tasks that typically rely on visual features processed by the parvocellular pathways. Even though Sherman (1985) proposed that the parvocellular system could provide high acuity capacity to a coarse magnocellularly driven form of vision, only a few studies have taken into account this possibility (Kruger *et al*., 1988; Strasburger & Rentschler, 1996; Bullier, 2001). A recent study Sugase *et al*. (1999) showed a biphasic response of IT neurons to faces with a first phasic component related to face recognition and a second late tonic component related to finer computations about facial characteristics (such as its expression). Some authors have proposed an influence of the dorsal magnocellular stream over the ventral pathway (Bullier, 2001; Vidyasagar, 1999). However, as magnocellular projections might account for as much as half of the information in the ventral pathway (Ferrera *et al*., 1992; Nealey & Maunsell, 1994), such interactions could also take place within the ventral stream itself (Sherman, 1985; Nakamura *et al*., 1993). The rapid preprocessing of magnocellular inputs could thus guide, in an intelligent way, the detailed visual processing of the slower parvocellular information.

In the present study, the latency of the earliest correct go-responses that appear at $\approx$280 ms set a severe constraint on the input–output processing time. In such a short delay, early motor responses to visual scenes should mainly be based on the coarse processing of the first wave of magnocellular visual information (VanRullen & Thorpe, 2002; Thorpe & Fabre-Thorpe, 2001).

Overall, the robustness of the categorization performance at very low contrast suggests that early object representations underlying behavioural performance in our rapid categorization task are very coarse. They could rely on magnocellular visual information and be subsequently refined by parvocellular inputs. Such coarse transient representations might only be unveiled in tasks using severe time constraints or forced-choice responses.

Acknowledgements

This work was supported by the Integrative and Computational Neuroscience ACI program of the CNRS. Financial support was provided to M.J-M.M. by a PhD grant from the French government. Many thanks to Nadège M. Bacon-Macé for improvements in the manuscript. The authors declare that they have no competing financial interests.

Abbreviations

EEG, electroencephalogram; ERP, event-related potential; IT, infero-temporal cortex; LGN, lateral geniculate nucleus; RT, reaction time.

References

Albrecht, D.G. (1995) Visual cortex neurons in monkey and cat: effect of contrast on the spatial and temporal phase transfer functions. *Vis. Neurosci.*, **12**, 1191–1210.

Albrecht, D.G., Geisler, W.S., Frazor, R.A. & Crane, A.M. (2002) Visual cortex neurons of monkeys and cats: temporal dynamics of the contrast response function. *J. Neurophysiol.*, **88**, 888–913.

Albrecht, D.G. & Hamilton, D.B. (1982) Striate cortex of monkey and cat: contrast response function. *J. Neurophysiol.*, **48**, 217–237.

Antal, A., Keri, S., Kovacs, G., Janka, Z. & Benedek, G. (2000) Early and late components of visual categorization: an event-related potential study. *Brain Res. Cogn. Brain Res.*, **9**, 117–119.

Avidan, G., Harel, M., Hendler, T., Ben-Bashat, D., Zohary, E. & Malach, R. (2002) Contrast sensitivity in human visual areas and its relationship to object recognition. *J. Neurophysiol.*, **87**, 3102–3116.

Bacon-Macé, N., Mace, M.J., Fabre-Thorpe, M. & Thorpe, S. (2005) The time course of visual processing: backward masking and natural scene categorization. *Vision Res.*, **45**, 1459–1469.

Blasdel, G.G. & Fitzpatrick, D. (1984) Physiological organization of layer 4 in macaque striate cortex. *J. Neurosci.*, **4**, 880–895.

Brady, N. & Field, D.J. (2000) Local contrast in natural images: normalisation and coding efficiency. *Perception*, **29**, 1041–1055.

Bullier, J. (2001) Integrated model of visual processing. *Brain Res. Brain Res. Rev.*, **36**, 96–107.

Carandini, M. & Heeger, D.J. (1994) Summation and division by neurons in primate visual cortex. *Science*, **264**, 1333–1336.

Cheng, A., Eysel, U.T. & Vidyasagar, T.R. (2004) The role of the magnocellular pathway in serial deployment of visual attention. *Eur. J. Neurosci.*, **20**, 2188–2192.

Dacey, D.M. & Brace, S. (1992) A coupled network for parasol but not midget ganglion cells in the primate retina. *Vis. Neurosci.*, **9**, 279–290.

Dacey, D.M. & Petersen, M.R. (1992) Dendritic field size and morphology of midget and parasol ganglion cells of the human retina. *Proc. Natl Acad. Sci. USA*, **89**, 9666–9670.

De Valois, R.L., Morgan, H. & Snodderly, D.M. (1974) Psychophysical studies of monkey vision. 3. Spatial luminance contrast sensitivity tests of macaque and human observers. *Vision Res.*, **14**, 75–81.

Delorme, A., Richard, G. & Fabre-Thorpe, M. (2000) Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. *Vision Res.*, **40**, 2187–2200.

Delorme, A., Rousselet, G.A., Mace, M.J. & Fabre-Thorpe, M. (2004) Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Brain Res. Cogn. Brain Res.*, **19**, 103–113.

Derrington, A.M. & Lennie, P. (1984) Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *J. Physiol. (Lond.)*, **357**, 219–240.

Enroth-Cugell, C. & Robson, J.G. (1966) The contrast sensitivity of retinal ganglion cells of the cat. *J. Physiol. (Lond.)*, **187**, 517–552.

Fabre-Thorpe, M., Richard, G. & Thorpe, S.J. (1998) Rapid categorization of natural images by rhesus monkeys. *Neuroreport*, **9**, 303–308.

Ferrera, V.P., Nealey, T.A. & Maunsell, J.H. (1992) Mixed parvocellular and magnocellular geniculate signals in visual area V4. *Nature*, **358**, 756–761.

Foxe, J.J. & Simpson, G.V. (2002) Flow of activation from V1 to frontal cortex in humansA framework for defining 'early' visual processing. *Exp. Brain Res.*, **142**, 139–150.

Gawne, T.J., Kjaer, T.W. & Richmond, B.J. (1996) Latency: another potential code for feature binding in striate cortex. *J. Neurophysiol.*, **76**, 1356–1360.

Gegenfurtner, K.R. & Rieger, J. (2000) Sensory and cognitive contributions of color to the recognition of natural scenes. *Curr. Biol.*, **10**, 805–808.

Halgren, E., Baudena, P., Heit, G., Clarke, J.M., Marinkovic, K., Chauvel, P. & Clarke, M. (1994) Spatio-temporal stages in face and word processing. 2. Depth-recorded potentials in the human frontal and Rolandic cortices. *J. Physiol. (Paris)*, **88**, 51–80.

Hartveit, E. & Heggelund, P. (1992) The effect of contrast on the visual response of lagged and nonlagged cells in the cat lateral geniculate nucleus. *Vis. Neurosci.*, **9**, 515–525.

Hubel, D.H. & Livingstone, M.S. (1990) Color and contrast sensitivity in the lateral geniculate body and primary visual cortex of the macaque monkey. *J. Neurosci.*, **10**, 2223–2237.

Intraub, H. (1981) Identification and processng of briefly glimpsed visual scenes. In Fisher, D.F., Monty, R.A., & Senders, J.W. (eds), *Eye Movements: Cognition and Visual Perception*. Erlbaum, Hillsdale, pp. 181–190.

Johnson, J.S. & Olshausen, B.A. (2003) Timecourse of neural signatures of object recognition. *J. Vis.*, **3**, 499–512.

Kaplan, E., Lee, B.B. & Shapley, R.M. (1990) New views of primate retinal function. In Osborne, N. & Chader, J. (eds), *Progress in Retinal Research*. Pergamon Press, Oxford, pp. 273–336.
Kaplan, E., Purpura, K. & Shapley, R.M. (1987) Contrast affects the transmission of visual information through the mammalian lateral geniculate nucleus. *J. Physiol. (Lond.)*, **391**, 267–288.
Kaplan, E. & Shapley, R.M. (1982) X and Y cells in the lateral geniculate nucleus of macaque monkeys. *J. Physiol. (Lond.)*, **330**, 125–143.
Kaplan, E. & Shapley, R.M. (1986) The primate retina contains two types of ganglion cells, with high and low contrast sensitivity. *Proc. Natl Acad. Sci. USA*, **83**, 2755–2757.
Keysers, C., Xiao, D.K., Foldiak, P. & Perrett, D.I. (2001) The speed of sight. *J. Cogn. Neurosci.*, **13**, 90–101.
Kruger, K., Donicht, M., Muller-Kusdian, G., Kiefer, W. & Berlucchi, G. (1988) Lesion of areas 17/18/19: effects on the cat's performance in a binary detection task. *Exp. Brain Res.*, **72**, 510–516.
Legge, G.E., Rubin, G.S. & Luebker, A. (1987) Psychophysics of reading – V. The role of contrast in normal vision. *Vision Res.*, **27**, 1165–1177.
Lewis, M.B. & Edmonds, A.J. (2003) Face detection: mapping human performance. *Perception*, **32**, 903–920.
Liu, J., Harris, A. & Kanwisher, N. (2002) Stages of processing in face perception: an MEG study. *Nat. Neurosci.*, **5**, 910–916.
Lupp, U., Hauske, G. & Wolf, W. (1976) Perceptual latencies to sinusoidal gratings. *Vision Res.*, **16**, 969–972.
Maunsell, J.H., Ghose, G.M., Assad, J.A., McAdams, C.J., Boudreau, C.E. & Noerager, B.D. (1999) Visual response latencies of magnocellular and parvocellular LGN neurons in macaque monkeys. *Vis. Neurosci.*, **16**, 1–14.
Maunsell, J.H. & Gibson, J.R. (1992) Visual response latencies in striate cortex of the macaque monkey. *J. Neurophysiol.*, **68**, 1332–1344.
Merigan, W.H. & Eskin, T.A. (1986) Spatio-temporal vision of macaques with severe loss of P beta retinal ganglion cells. *Vision Res.*, **26**, 1751–1761.
Nakamura, H., Gattass, R., Desimone, R. & Ungerleider, L.G. (1993) The modular organization of projections from areas V1 and V2 to areas V4 and TEO in macaques. *J. Neurosci.*, **13**, 3681–3691.
Nasanen, R., Ojanpaa, H. & Kojo, I. (2001) Effect of stimulus contrast on performance and eye movements in visual search. *Vision Res.*, **41**, 1817–1824.
Nealey, T.A. & Maunsell, J.H. (1994) Magnocellular and parvocellular contributions to the responses of neurons in macaque striate cortex. *J. Neurosci.*, **14**, 2069–2079.
Nowak, L.G., Munk, M.H., Girard, P. & Bullier, J. (1995) Visual latencies in areas V1 and V2 of the macaque monkey. *Vis. Neurosci.*, **12**, 371–384.
Oram, M.W., Xiao, D., Dritschel, B. & Payne, K.R. (2002) The temporal resolution of neural codes: does response latency have a unique role? *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **357**, 987–1001.
Potter, M.C. (1976) Short-term conceptual memory for pictures. *J. Exp. Psychol [Hum. Learn.]*, **2**, 509–522.
Proverbio, A.M., Esposito, P. & Zani, A. (2002) Early involvement of the temporal area in attentional selection of grating orientation: an ERP study. *Brain Res. Cogn. Brain Res.*, **13**, 139–151.
Reich, D.S., Mechler, F. & Victor, J.D. (2001) Temporal coding of contrast in primary visual cortex: when, what, and why. *J. Neurophysiol.*, **85**, 1039–1050.
Rolls, E.T. & Baylis, G.C. (1986) Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Exp. Brain Res.*, **65**, 38–48.
Rousselet, G.A., Fabre-Thorpe, M. & Thorpe, S.J. (2002) Parallel processing in high-level categorization of natural images. *Nat. Neurosci.*, **5**, 629–630.
Rousselet, G.A., Thorpe, S.J. & Fabre-Thorpe, M. (2004) Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vision Res.*, **44**, 877–894.
Ruderman, D.L. (1994) Statistics of natural images. *Network: Computation Neural Systems*, **5**, 517–548.
Rugg, M.D., Doyle, M.C. & Wells, T.J. (1995) Word and nonword repetition within-modality and across-modality: an event-related potential study. *J. Cogn. Neurosci.*, **7**, 209–227.
Schall, J.D. (2001) Neural basis of deciding, choosing and acting. *Nat. Rev. Neurosci.*, **2**, 33–42.
Schiller, P.H., Logothetis, N.K. & Charles, E.R. (1990) Role of the color-opponent and broad-band channels in vision. *Vis. Neurosci.*, **5**, 321–346.
Schmolesky, M.T., Wang, Y., Hanes, D.P., Thompson, K.G., Leutgeb, S., Schall, J.D. & Leventhal, A.G. (1998) Signal timing across the macaque visual system. *J. Neurophysiol.*, **79**, 3272–3278.
Sclar, G., Maunsell, J.H. & Lennie, P. (1990) Coding of image contrast in central visual pathways of the macaque monkey. *Vision Res.*, **30**, 1–10.
Sestokas, A.K. & Lehmkuhle, S. (1986) Visual response latency of X- and Y-cells in the dorsal lateral geniculate nucleus of the cat. *Vision Res.*, **26**, 1041–1054.
Sestokas, A.K., Lehmkuhle, S. & Kratz, K.E. (1987) Visual latency of ganglion X- and Y-cells: a comparison with geniculate X- and Y-cells. *Vision Res.*, **27**, 1399–1408.
Shapley, R. (1990) Visual sensitivity and parallel retinocortical channels. *Annu. Rev. Psychol*, **41**, 635–658.
Shapley, R., Kaplan, E. & Soodak, R. (1981) Spatial summation and contrast sensitivity of X and Y cells in the lateral geniculate nucleus of the macaque. *Nature*, **292**, 543–545.
Shapley, R.M. & Victor, J.D. (1978) The effect of contrast on the transfer properties of cat retinal ganglion cells. *J. Physiol. (Lond.)*, **285**, 275–298.
Sherman, S.M. (1985) Functional organization of the W-, X- and Y-cell pathways in the cat: a review and hypothesis. *Prog. Psychobiol. Physiol. Psychol.*, **11**, 233–314.
Silveira, L.C. & Perry, V.H. (1991) The topography of magnocellular projecting ganglion cells (M-ganglion cells) in the primate retina. *Neuroscience*, **40**, 217–237.
Strasburger, H., Harvey, L.O. Jr & Rentschler, I. (1991) Contrast thresholds for identification of numeric characters in direct and eccentric view. *Percept. Psychophys*, **49**, 495–508.
Strasburger, H. & Rentschler, I. (1996) Contrast-dependent dissociation of visual recognition and detection fields. *Eur. J. Neurosci.*, **8**, 1787–1791.
Sugase, Y., Yamane, S., Ueno, S. & Kawano, K. (1999) Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, **400**, 869–873.
Sun, H. (2001) Rod–cone interactions assessed in inferred magnocellular and parvocellular postreceptoral pathways. *J. Vision*, **1**, 42–54.
Tadmor, Y. & Tolhurst, D.J. (2000) Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes. *Vision Res.*, **40**, 3145–3157.
Thorpe, S.J. & Fabre-Thorpe, M. (2001) Neuroscience. Seeking categories in the brain. *Science*, **291**, 260–263.
Thorpe, S., Fize, D. & Marlot, C. (1996) Speed of processing in the human visual system. *Nature*, **381**, 520–522.
VanRullen, R. & Thorpe, S.J. (2001a) Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, **30**, 655–668.
VanRullen, R. & Thorpe, S.J. (2001b) The time course of visual processing: from early perception to decision-making. *J. Cogn. Neurosci.*, **13**, 454–461.
VanRullen, R. & Thorpe, S.J. (2002) Surfing a spike wave down the ventral stream. *Vision Res.*, **42**, 2593–2615.
Vidyasagar, T.R. (1999) A neuronal model of attentional spotlight: parietal guiding the temporal. *Brain Res. Brain Res. Rev.*, **30**, 66–76.
Vidyasagar, T.R., Kulikowski, J.J., Lipnicki, D.M. & Dreher, B. (2002) Convergence of parvocellular and magnocellular information channels in the primary visual cortex of the macaque. *Eur. J. Neurosci.*, **16**, 945–956.
Watson, A.B. (1992) Transfer of contrast sensitivity in linear visual networks. *Vis. Neurosci.*, **8**, 65–76.
Xiao, D.K., Edwards, R.H., Bowman, E.M. & Oram, M.W. (2001) The influence of stimulus contrast on response latency and response strength of neurones in the superior temporal sulcus of the macaque monkey. *Soc. Neurosci. Abstr.*, **23**, 450.

1.5 - Catégorisation ultra-rapide : robustesse aux variations de luminance

Nous avons vu dans les deux études précédentes que les humains et les singes ne s'appuient pas sur la valeur absolue des contrastes locaux pour effectuer la tâche de catégorisation, sinon leur performance chuterait très vite dès que le contraste des images est diminué. Existe-t-il d'autres indices bas niveau que les sujets pourraient utiliser pour maintenir de bonnes performances dans cette tâche, sans avoir à faire intervenir une représentation élaborée de la scène visuelle ? Les humains, comme les singes, pourraient par exemple utiliser une analyse statistique des plages de luminance et de leur distribution spatiale dans les images pour effectuer la tâche. Ce type de stratégie peut paraître à première vue complexe à mettre en œuvre étant donné la diversité des images à catégoriser dans notre tâche. Elle a cependant été montrée chez le pigeon, qui est capable d'utiliser la texture d'un visage ou des gradients de luminance pour catégoriser le genre des visages humains (Troje *et al.*, 1999 ; Huber *et al.*, 2000). Pour vérifier cette hypothèse, nous avons mené une expérience conjointement chez l'homme et le singe dans laquelle la luminance des images était manipulée.

1.5.1 - Expériences chez l'homme et le singe

Dans ces expériences, les hommes et les singes sont toujours testés sur le même dispositif pour que les performances soient directement comparables. Les résultats de ces expériences feront l'objet d'un article (actuellement en préparation), comparant les performances des hommes et des singes face à des stimuli dont la luminance a été manipulée. Ces données n'étant pas encore publiées, nous les présenterons en détail dans ce mémoire pour insister sur les similitudes et les différences entre les deux espèces.

MATERIEL ET METHODES :

Les conditions de test sont similaires à celles utilisées pour les autres tâches. Les sujets (hommes ou singes) sont assis face à un écran tactile dans une pièce très faiblement éclairée. Une image est flashée au centre d'un écran noir pendant 28 ms. Pour déclencher la série d'images, le sujet doit placer sa main juste sous l'écran, sur une plaque équipée de photodiodes. La tâche du sujet est de relever sa main le plus vite possible pour aller toucher l'écran tactile lorsque l'image flashée contient une cible (réponse go) et de conserver sa main sur la touche dans le cas contraire (réponse no-go). Une limite de 1s est imposée pour toucher l'écran. Le délai entre deux images est aléatoire et compris entre 1,6 et 2 secondes (durée moyenne 1,8 s).

Les sujets effectuaient une catégorisation animal/non animal (An/nAn) et les réponses correctes (go et no-go) étaient signalées par un son. Pour les singes, ces réponses correctes étaient de plus récompensées par une dose de sirop de fruit et les réponses incorrectes étaient "punies" par le réaffichage de l'image pendant une durée de 3 secondes. Ceci permet l'exploration de l'image mal catégorisée et impose un délai d'attente au singe avant d'accéder à l'essai suivant ; donc à la récompense suivante.

Les hommes effectuaient 1800 essais en une séance de 18 séries séparées par des pauses pour une durée totale d'expérience d'environ 2 heures. Les singes travaillaient 5 jours par semaine et effectuaient entre 300 et 1500 essais par jour, selon leur motivation. Ils disposaient d'eau à volonté à la fin de chaque session quotidienne (avant d'être remis en cage sans accès à l'eau) et pendant le week-end.

Stimuli :

Nous avons utilisé 600 photographies provenant d'une grande banque d'images (commercialisée par Corel). Les images sont choisies pour être les plus variées possible, avec des vues en gros plan ou éloignées, des animaux présentés sous des angles atypiques, des images contenant plusieurs animaux ou des animaux seulement partiellement visibles. Cibles et distracteurs sont équiprobables au sein d'une série. Les cibles incluent des poissons, des reptiles, des mammifères et des oiseaux photographiés dans leur environnement naturel. Les distracteurs sont des paysages, des arbres, des fleurs, des objets, des monuments, des voitures… Lors de l'exécution de la tâche, le sujet ne dispose d'aucune information concernant le contenu de l'image à venir. Au sein d'une image, la cible peut être variable en nombre et en position ; elle peut être partiellement occultée par un objet ou présentée en partie (pattes, bec…).

La résolution des images est de 256x384 pixels et celle de l'écran tactile de 17" est réglée sur 800x600. Les ¾ des images sont horizontales. Selon la distance du sujet à l'écran (50 cm pour les humains, 35 cm pour les singes), la taille angulaire de l'image est de 13,0°x8,9° pour les hommes et 18,5°x12,7° pour les singes. L'image est affichée pendant 3 trames à 90 Hz, ce qui correspond à une durée de présentation de 28 ms.

Modification des images :

Les images couleur ont été converties en 256 niveaux de gris allant du noir (0) au blanc (255). Les images subissent ensuite une modification du contraste et de la luminance.

6 conditions de présentation sont utilisées :

- N : Image normale
- N/2 : Contraste divisé par 2 (sans changer la luminance moyenne de l'image)
- +24 ou +48 : Image N/2 avec luminance augmentée de 24 ou 48 (sur l'échelle 0-255)
- -24 ou -48 : Image N/2 avec luminance diminuée de 24 ou 48 (sur l'échelle 0-255)

Les 600 images, présentées dans 6 conditions différentes, donnent un total de 3600 stimuli. Utiliser la condition N/2 (plutôt que N) comme base pour la modification de la luminance permet de conserver une plus grande quantité d'information car la proportion de pixels qui vont saturer dans les valeurs extrêmes est moins importante (voir Figure 11).

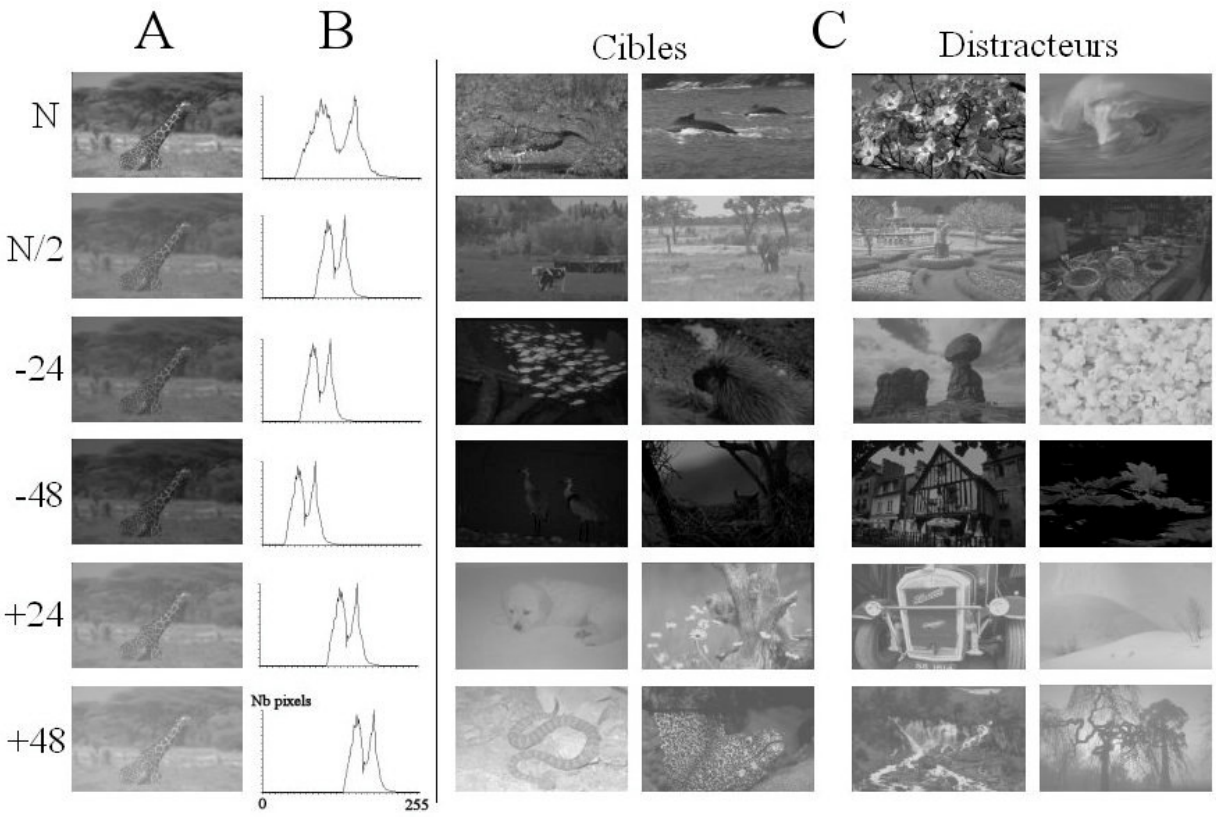


Figure 11 : A : Même image vue dans toutes les conditions de présentation. B : Histogramme de distribution des valeurs de luminance des pixels de l'image en A. Lorsqu'on divise le contraste par 2, l'écart de la luminance d'un pixel donné par rapport à la luminance moyenne de l'image est divisé par deux. Cela a pour effet de resserrer l'histogramme de distribution des luminances de tous les pixels d'une image autour de sa luminance moyenne (N/2) et de limiter la proportion de pixels saturés à 0 ou 255 dans les conditions -48 et +48. C : Exemple de stimuli utilisés dans la tâche animal/non animal.

Sujets :

18 sujets humains (âge moyen : 27 ans, min : 19 , max : 49 ; 9 femmes) ont donné leur consentement pour participer à cette expérience.

Deux macaques rhésus de 6 ans (un mâle -Ry- et une femelle -Eu-) travaillent sur la tâche An/nAn depuis 2-3 années. Ils ont déjà participé à certaines des séries expérimentales décrites en introduction. Comme pour les sujets humains, leur capacité de catégorisation n'a pas été perturbée lorsque les indices de couleur étaient supprimés (ce qui permet d'utiliser ici des images en noir et blanc pour s'affranchir des questions de contrastes de couleur) et l'utilisation d'images familières n'a entraîné qu'une légère augmentation en précision sans effet sur les temps de réaction rapides (ce qui valide également la répétition des images vues pour un sujet donné dans le présent travail).

Seuls les indices comportementaux (précision et vitesse) ont été analysés chez le singe. Chez l'homme, l'approche comportementale a été associée à l'enregistrement des potentiels évoqués.

Protocole expérimental :

Hommes :

Les sujets naïfs effectuent un entraînement sur 50 images en noir et blanc sans modification de contraste ou de luminance avant de commencer l'expérience. Chaque sujet catégorise au cours du test l'ensemble des 600 images dans 3 conditions au hasard parmi les 6. Il effectue donc 1800 essais, en 18 séries de 100 images, équilibrées en cibles et distracteurs et en conditions de luminance. Chacune des 6 conditions est équiprobable, chaque sujet catégorise donc 300 images (150 cibles et 150 distracteurs) dans chacune des conditions.

Singes :

Chaque singe a vu l'ensemble des 600 images dans les 6 conditions de présentation. Les 3600 images à catégoriser ont été réparties en 12 séries de 300 images (équilibrées en cibles et distracteurs ainsi qu'en conditions de luminance). Lors d'une séance quotidienne le singe travaille à volonté sur une séquence de 300 stimuli différents. Les 50 premiers essais ne sont pas pris en compte pour éliminer la variabilité intense du début d'expérience (stress, fébrilité, présence de l'expérimentateur...). L'analyse est effectuée de l'essai 51 à l'essai 350, l'animal travaillant ensuite à volonté. Les séances au cours desquelles l'animal effectuait moins de 650 essais n'étaient pas comptabilisées afin que l'état de motivation de l'animal reste comparable d'une série à l'autre. De plus une séance de 650 essais permet une double analyse sur chaque présentation d'image, ce qui donne une mesure de la constance des performances.

Avant de débuter les séries expérimentales, chaque singe a été repris en test pendant 3 à 4 semaines pour catégoriser des images en couleur puis en niveaux de gris afin de retrouver des performances stables en catégorisation et de déterminer une ligne de base contrôle pour ses performances face à des stimuli achromatiques. Les images utilisées pendant ces 3-4 semaines étaient différentes de celles utilisées pendant les séries test.

Évaluation de la performance, tests statistiques :

Comportement :

La performance des sujets est évaluée à la fois en termes de vitesse, de précision et d'échange précision/vitesse. On considère que le sujet a répondu dès qu'il relâche le bouton et on enregistre la latence de cette réponse (temps de réaction). Cette réponse est correcte sur une cible et incorrecte sur un distracteur. Le toucher d'écran n'est qu'un indice complémentaire pour apprécier une éventuelle correction d'erreur (qui ne sera pas considérée même si chez le singe, jusqu'à 30% des réponses go incorrectes sont "corrigées" avant de toucher l'écran). A l'inverse, pour qu'une réponse de type no-go soit enregistrée, il faut que le sujet maintienne sa main sur le bouton durant au moins une seconde après l'apparition de l'image.

La **précision** est appréciée par le pourcentage de réponses correctes (go et no-go). La différence de précision entre 2 conditions est évaluée à l'aide d'un Khi^2 sur l'ensemble des réponses correctes et incorrectes. Des comparaisons appariées (Test de Wilcoxon) sont également effectuées pour permettre d'analyser la constance de l'effet observé sur le groupe d'individus.

Vitesse : Pour chaque condition donnée on calcule la distribution des TR pour les réponses correctes et les réponses incorrectes par pas de temps de 10 ms. La comparaison des distributions de TR dans deux conditions données est effectuée à l'aide de tests de Mann-Whitney (distributions ne passant pas le test de normalité). Pour tous les sujets du groupe, les TR médians obtenus dans deux conditions différentes sont comparés par des tests appariés (Wilcoxon).

Précision/vitesse : Pour analyser ces performances en tenant compte de l'échange précision/vitesse (loi de Fitts, 1954), on calcule un *d'* "dynamique" spécifique à notre tâche qui prend en compte les TR sur les erreurs et les réponses correctes en fonction du temps (détails en annexe B).

Potentiels évoqués :

L'EEG de scalp était enregistré chez les sujets humains à l'aide d'un bonnet de 32 électrodes (voir en annexe A leur répartition) relié à un Neuroscan afin d'analyser les potentiels évoqués associés à la tâche. Après rejet des essais parasités par des mouvements oculaires ou des ondes alpha, les potentiels évoqués sont moyennés séparément pour les cibles et les distracteurs ; seuls les essais corrects sont pris en considération. Les moyennes sont filtrées par sujet avant d'effectuer la soustraction entre essais cibles et essais distracteurs.

Les grandes moyennes calculées sur l'ensemble des sujets fournissent un intervalle de confiance pour chaque tracé. Ces courbes de différences permettent d'évaluer la latence à partir de laquelle le signal enregistré sur les cibles commence à diverger de celui enregistré sur les distracteurs. La latence de cette activité différentielle est estimée en retenant le pas de temps à partir duquel 15 valeurs consécutives de t-test (soit une durée de 15 ms) sont au-dessus du seuil de significativité (Rugg *et al.*, 1995).

RESULTATS :

L'analyse des résultats a montré une grande robustesse des performances aux modifications de luminance, à la fois chez le singe et chez l'homme.

Comportement :

La performance est analysée en précision et en vitesse chez l'homme et le singe.

Précision :

Chez l'homme : Chaque sujet a effectué 300 essais dans chaque condition ; un total de 5400 essais a donc été effectué dans chaque condition par le groupe de 18 sujets. La précision moyenne du groupe est de 97,0% dans la condition normale, elle est de 96,6% pour les conditions N/2, N/2-24 et N/2+24 ; une différence qui n'est pas significative[1].

Les deux conditions extrêmes de luminance entraînent une dégradation de la précision. La précision moyenne est de 95,7% (N/2-48) et de 95,1% (N/2+48). Comparées à chacune des 4 conditions précédentes ces dégradations sont toujours significatives[2]. Cette significativité apparaît également lorsque l'on compare au niveau du groupe la précision de chacun des sujets dans deux conditions données par un test apparié (exemple : comparaison conditions N

[1] khi^2 calculé sur l'ensemble des essais corrects et incorrects, ns

[2] khi^2, p est respectivement <0,02 et <0,0001 avec toutes les autres conditions

et N/2-48[3]). Il faut pourtant noter que si cette dégradation est significative, elle reste très limitée puisqu'elle n'est inférieure que de 2% à la précision enregistrée sur images normales ! (Figure 12A)

Chez le singe : La précision des deux singes (EU et RY) est évaluée comme pour l'homme, sur 300 essais par condition. Elle atteint 95,8% de réponses correctes en condition N, et chute significativement à 92,6% correct lorsque le contraste est divisé par 2[4]. Lorsque cette réduction de contraste est associée à une diminution de la luminance (-24 et -48), la précision du singe reste similaire à la condition N/2[5] ; en revanche, lorsqu'elle est associée à une augmentation de la luminance une nouvelle chute de précision est observée, à +24 le pourcentage de réponses correctes est de 91,0% et à +48 de 89,3%[6]. (Figure 12B)

Vitesse de réponse :

Pour évaluer la vitesse de réponse chez l'homme, on comparera d'abord les distributions des TR sur les réponses correctes dans les diverses conditions de présentation des images, et on évaluera la constance des effets observés par des tests appariés sur l'ensemble des résultats individuels des sujets du groupe.

Chez l'homme : Le temps de réaction médian pour les 18 sujets humains est de 385 ms dans la condition normale de présentation. Par rapport à cette condition, on observe une augmentation de ce TR pour toutes les autres conditions de présentation (TR médian : N/2 : 394 ms, +24 : 397 ms, -24 : 395 ms +48 : 408 ms et -48 : 407 ms) (Figure 12A). Cette augmentation est toujours significative[7]. Les conditions extrêmes sont celles pour lesquelles la dégradation est la plus marquée. Ces résultats sont robustes et sont observés individuellement chez la majorité des sujets comme le montre un test de Wilcoxon[8].

L'augmentation des TR est d'environ 10 ms dans les 3 conditions de perturbation moyenne de l'image présentée (N/2, +24 et -24). La comparaison des distributions de TR entre ces trois conditions n'est jamais significative[9]. L'augmentation des TR est d'environ 20 ms dans les 2 conditions extrêmes. Il faut noter que les perturbations induites par une augmentation ou une

[3] Wilcoxon : *ddl* = 17, p =0,0043
[4] khi^2, entre les conditions N et N/2, p<0.0009
[5] khi^2 : ns
[6] khi^2 : entre N/2 et +48, p<0,007
[7] Mann-Withney : p < 0,001 dans tous les cas
[8] Wilcoxon : *ddl* = 17, p < 0,001 pour toutes les comparaisons avec la condition N
[9] Mann-Whitney : ns

diminution de la luminance sont pratiquement symétriques. La comparaison des distributions des TR dans ces deux conditions extrêmes ne fait pas apparaître de différence[10]. (Figure 12A)

Chez le singe : Les temps de réaction des singes sont en général beaucoup plus courts que ceux des hommes. Ils sont affectés de façon comparable en fonction des conditions de présentation : ils changent peu pour les conditions de perturbation moyenne N, N/2, -24 et +24 et sont augmentés de manière plus importante pour les modifications extrêmes de luminance.

Étant donné le nombre réduit de singes ayant effectué la tâche de catégorisation, il n'est pas possible de faire des tests appariés sur les données entre les différentes conditions. Seuls des tests concernant les distributions des TR entre les conditions peuvent être utilisés. Les TR médians, minimum et maximum sont enregistrés respectivement dans la condition N (254 ms) et dans les 2 conditions extrêmes -48 (259 ms), +48 (258 ms). Ce TR médian est très proche entre les conditions N, N/2, –24 et + 24 (compris entre 254 ms et 256 ms) et la distribution des réponses entre ces conditions n'est pas significativement différente[11] (Figure 12B). Par rapport à la condition N, l'augmentation de TR enregistrée dans les 2 conditions extrêmes –48 et +48 est faible mais significative [12] (Figure 12B).

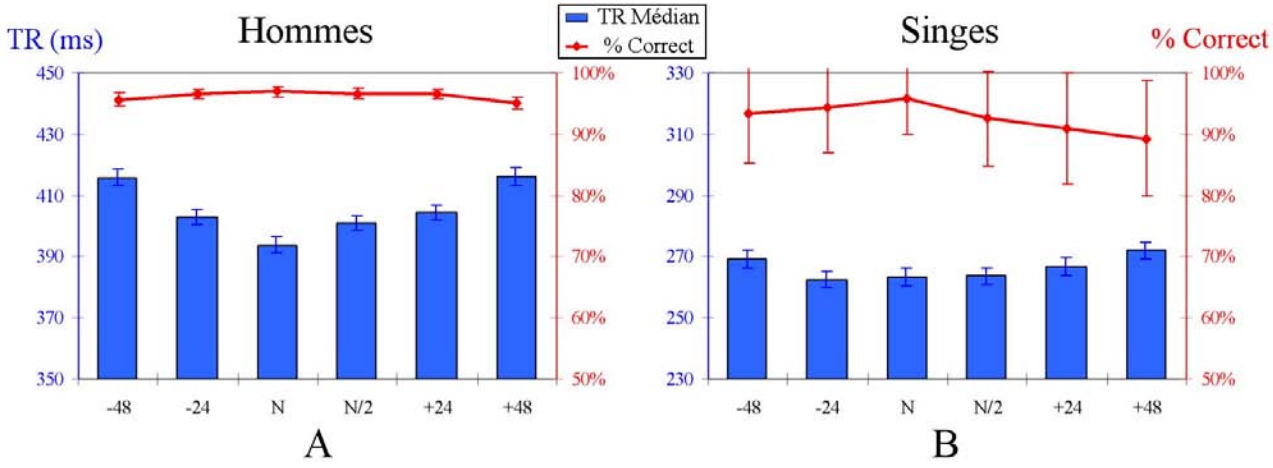


Figure 12 : Les TR médians et la précision sont respectivement représentés dans chacune des conditions de présentation par les histogrammes et les courbes. On observe que la performance est maximale pour les conditions N et N/2 et qu'aux valeurs extrêmes, les TR augmentent légèrement et la précision diminue. Il est à noter que les effets induits sont faibles, même lorsqu'ils sont significatifs. En A (hommes) et en B (singes), les barres verticales représentent les intervalles de confiance à la moyenne à 95%.

[10] Mann-Whitney : ns
[11] Mann-Whitney : ns pour toutes ces conditions prises deux à deux
[12] Mann-Whitney : p respectivement < 0,024 et < 0,004

Échange précision/vitesse :

Dans une telle tâche, il est indispensable de considérer la corrélation qui existe entre vitesse et précision (loi de Fitts). Nous avons donc calculé le d' pour chaque pas de temps et chaque condition, ce qui nous permet de visualiser l'évolution de la performance au cours du temps (Figure 13). L'interprétation du d' donne deux indications principales : la précision finale indiquée par la hauteur du plateau et la latence d'apparition des premières réponses correctes.

Chez l'homme : Les résultats sont clairs, les plateaux atteints dans chacune des conditions illustrent bien le groupe de 4 conditions (N, N/2, -24, +24) très proches, la condition normale permettant d'atteindre la plus grande précision. Dans les 2 conditions extrêmes, le plateau final est moins élevé illustrant une précision moindre.

De façon particulièrement intéressante, on peut remarquer que ces courbes se superposent à leur origine jusqu'à des TR autour de 360 ms. Rappelons que le TR médian est de 385 ms et que 360 ms après l'apparition du stimulus, les sujets ont déjà fourni plus de 17% de leurs réponses "go" correctes.

Pour chacune des conditions, les réponses rapides sont donc quantitativement similaires, ce qui est important pour mieux caractériser les traitements à la base de la catégorisation ultra-rapide. (Figure 13)

Chez le singe : La première caractéristique surprenante des courbes de d' des singes est leur pente très raide, quelle que soit la condition considérée. Cette forte pente est le reflet d'une très grande précision dans la catégorisation, dès les premières réponses. On retrouve en partie cette caractéristique dans le départ très brutal des histogrammes de TR des singes.

Les plateaux atteints pour les différentes conditions restent bien séparés les uns des autres, avec un d' minimum pour les conditions dans lesquelles la luminance est augmentée. L'origine des courbes de d' pour les différentes conditions de présentation des images apparaît légèrement décalées dans le temps : toutes divergent à un pas de temps compris entre 190 et 210 ms. (Figure 13)

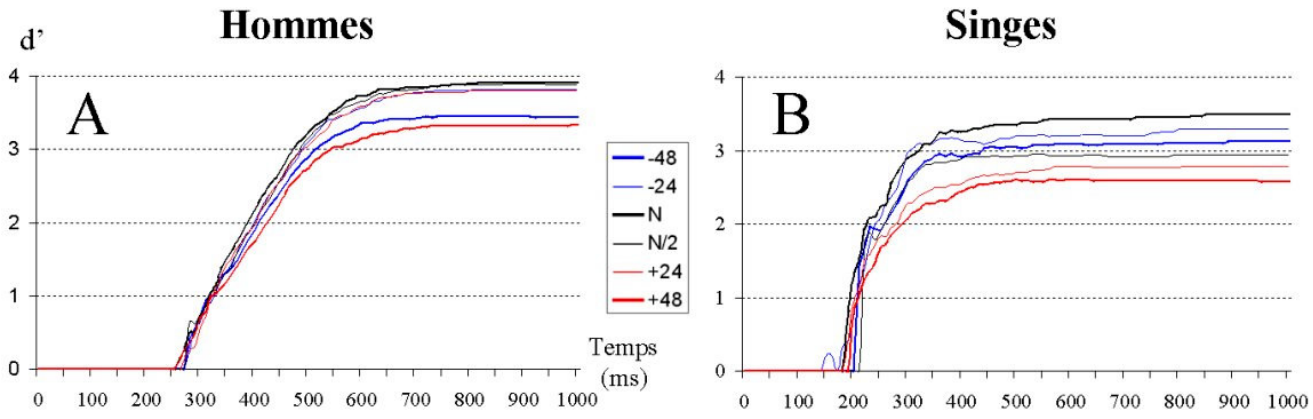


Figure 13 : Évaluation de la performance en fonction du temps : courbes de *d'* cumulés pour les hommes (A) et les singes (B) dans chacune des conditions de luminance.

En résumé :

La précision n'est que peu affectée par les variations de contraste et de luminance imposées aux images naturelles. Par rapport à la condition normale, diviser le contraste par 2 n'entraîne pas de perturbation chez l'homme mais induit une chute de 3% de la précision du singe. Dans les 2 conditions extrêmes de présentation, la précision est toujours affectée et c'est la condition de luminance maximale qui induit le déficit maximal. Toutefois, ce déficit est peu sévère chez l'homme, la précision ne chute que de 2%. Il est plus accentué chez le singe avec une diminution de 6,5%. On observe le tableau inverse en ce qui concerne le TR avec une augmentation des TR beaucoup plus marquée (10-20 ms) chez l'homme que chez le singe (5 ms). Ces différences entre les hommes et les singes pourraient refléter des différences de stratégie, les singes privilégiant plutôt la vitesse et les hommes la précision. Ces perturbations restent malgré tout très modérées lorsque l'on considère la grande dégradation des images et leur courte durée d'affichage (28 ms).

Électrophysiologie :

L'activité différentielle liée à la tâche apparaît autour de 170 ms ; elle est liée au statut "cible" ou "distracteur" occupé par l'image dans la tâche. Un test statistique nous permet de déterminer la latence de cette activité différentielle dans les diverses conditions de présentation. C'est dans la condition N qu'elle apparaît le plus précocement, vers 170 ms, puis dans la condition -48 à 178 ms et enfin dans la condition +48 à 183 ms. Enfin une dernière onde différentielle apparaît tardivement vers 300 ms, elle est liée à la réponse motrice des sujets sur les cibles et se retrouve principalement sur les électrodes centrales.

Dans les expériences menées précédemment dans l'équipe, l'activité différentielle liée à la tâche a le plus souvent été rapportée à des latences proches de 150 ms (Thorpe *et al.*, 1996 ; Fabre-Thorpe *et al.*, 2001 ; VanRullen & Thorpe, 2001 ; Rousselet *et al.*, 2002 ; Johnson & Olshausen, 2003 ; Bacon-Macé *et al.*, 2005 ; Fize *et al.*, 2005), mais à également été trouvée à 170 ms (Delorme *et al.*, 2004 ; Rousselet *et al.*, 2004 ; Macé *et al.*, 2005) sans qu'une véritable explication n'ait été proposée. Nous pensons que quand la visibilité du stimulus est maximale (image en couleur, non modifiée) et que la tâche est simple (par exemple animal/non animal, sans changements de catégorisation), la préparation du sujet est maximale et l'activité différentielle apparaît autour de 150 ms. En revanche, lorsque la tâche devient plus complexe, parce que l'on introduit des conditions dans lesquelles les images sont difficiles à catégoriser (luminance, contraste) ou que la catégorie cible change d'une série sur l'autre, la préactivation des ensembles de cellules permettant de catégoriser les cibles pourrait être moins efficace et entraîner un délai dans la latence à partir de laquelle le signal EEG diffère.

L'activité différentielle qui se manifeste aux alentours de 170 ms a été analysée dans toutes les conditions de luminance des images. La figure 14 illustre la courbe différentielle obtenue dans la condition N et dans les 2 conditions de luminance les plus extrêmes (+48 et -48). On peut remarquer que quelle que soit la condition de luminance, les courbes suivent la même pente et atteignent la même amplitude. En revanche, on constate un décalage dans la latence de ces courbes. C'est dans la condition où les performances comportementales des sujets sont les plus affectées en précision et en vitesse que l'activité différentielle apparaît avec la latence la plus longue. Une durée de traitement additionnelle de 8-13 ms par rapport à la condition normale semble nécessaire dans les conditions de luminance extrême.

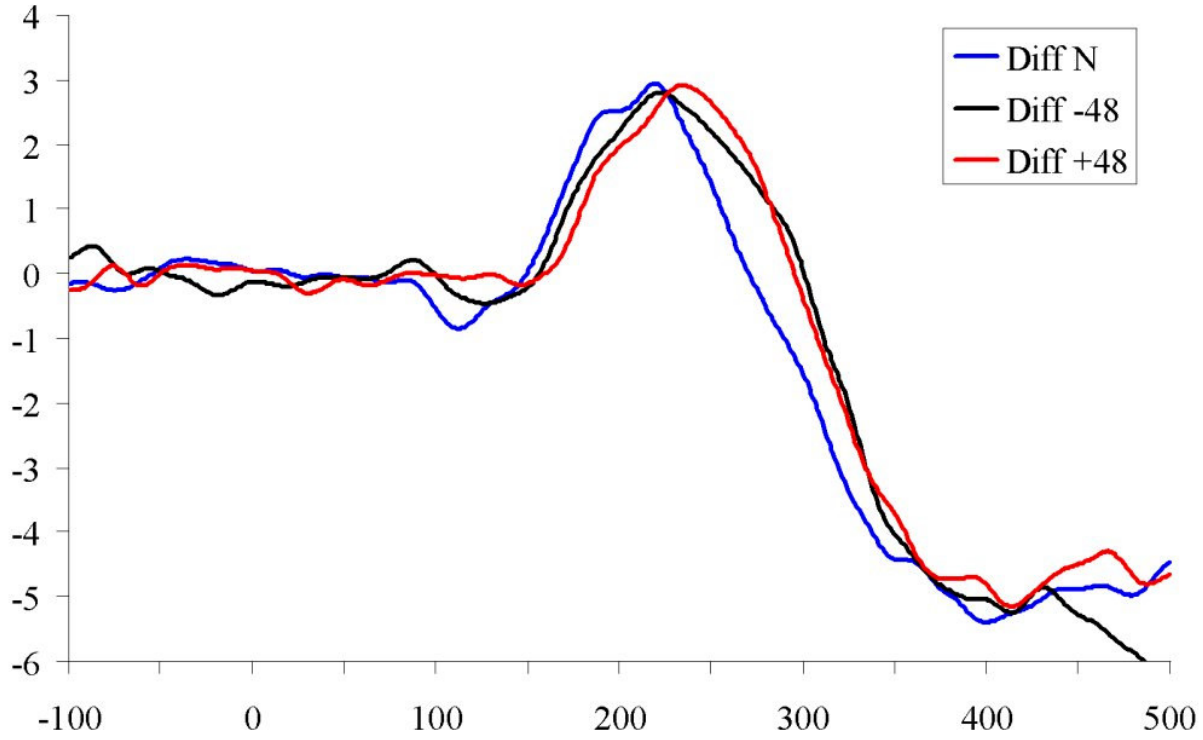


Figure 14 : Les courbes représentent les activités différentielles sur les images normales (bleu), les images pour lesquelles la luminance est augmentée de 48 (rouge) ou diminuée de 48 (noir). On observe très nettement une augmentation de la latence à partir de laquelle l'activité différentielle devient différente de la ligne de base entre les conditions normales, luminance réduite et luminance augmentée.

DISCUSSION :

Pour des images en noir et blanc dont le contraste a été divisé par 2 et la luminance augmentée ou diminuée d'environ 50%, la précision de l'homme dans une tâche de catégorisation ne chute que de 2% et la latence de ses réponses n'augmente que d'environ 20 ms. L'impact de cette modification des images est légèrement plus important pour les singes sur la précision avec une baisse d'environ 6-7% mais beaucoup plus réduite sur les TR qui n'augmentent que de 5 ms.

Notre étude comparative fait ressortir l'importance accordée par le singe à la vitesse de réaction. Comme dans les études précédentes, les singes macaques se sont révélés plus rapides que les hommes dans cette tâche. Dans les conditions de présentation pour lesquelles luminance et contraste sont altérés, la dégradation de la performance est essentiellement enregistrée sur la précision pour les singes. Chez l'homme, elle atteint le TR avant d'affecter la précision. Ces différences peuvent facilement être interprétées par un compromis précision/vitesse, les hommes privilégiant la précision et les singes la vitesse.

Pour les deux espèces, les faibles dégradations des performances malgré les fortes modifications des images semblent confirmer que les réponses produites ne reposent pas sur

une analyse statistique simple des caractéristiques bas niveau des images pour effectuer la tâche de catégorisation visuelle complexe qui est proposée.

1.5.2 - Les activités différentielles précoces et la luminance

Comme pour l'expérience dans laquelle le contraste des images était réduit, l'expérience dans laquelle la luminance des images est manipulée apporte une contribution intéressante à la question des activités différentielles précoces. A nouveau, il n'est possible d'observer une activité différentielle précoce que dans la condition où les images sont présentées avec un contraste et une luminance normale. L'activité différentielle précoce est ainsi quasi-inexistante dès que le contraste est divisé par deux, ce qui confirme les résultats obtenus dans l'étude précédente. Mais la manipulation de la luminance des images offre un autre outil intéressant pour étudier ces phénomènes précoces. On peut par exemple soustraire les distracteurs les plus foncés des cibles les plus claires, ou bien les distracteurs les plus clairs des cibles les plus foncées. Le sens de la soustraction change du point de vue des propriétés de luminance globale (clair - sombre ou sombre - clair) mais pas du point de vue de la catégorisation (cible - distracteur). Si l'on suppose que les activités différentielles entre 80 et 140 ms sont principalement induites par le traitement des cibles et des distracteurs en tant que tels dans la tâche, elles devraient présenter le même profil entre les deux courbes. En revanche, si les potentiels évoqués entre 80 et 140 ms sont plus liés aux différences entre les caractéristiques physiques des images, nous devrions observer une activité différentielle précoce opposée entre les deux courbes. La figure 15 montre bien que c'est la deuxième hypothèse qui est la bonne, puisque les deux différentielles sont parfaitement opposées pour toute la période avant 150 ms. Les activités différentielles que l'on peut observer autour de 80 à 140 ms entre les cibles et les distracteurs sont donc dues en grande partie (sinon totalement) à des différences physiques entre les images et dans le cas présent, à des différences de luminance, indifféremment de leur statut de cible ou de distracteur. Les activités différentielles après 150 ms évoluent bien dans le même sens, ce qui signifie qu'elles sont liées au traitement de l'image en tant que cible ou distracteur (même si l'on voit que les différences physiques très importantes entre les groupes d'images ont un impact sur le décours temporel du signal EEG jusqu'à environ 300 ms).

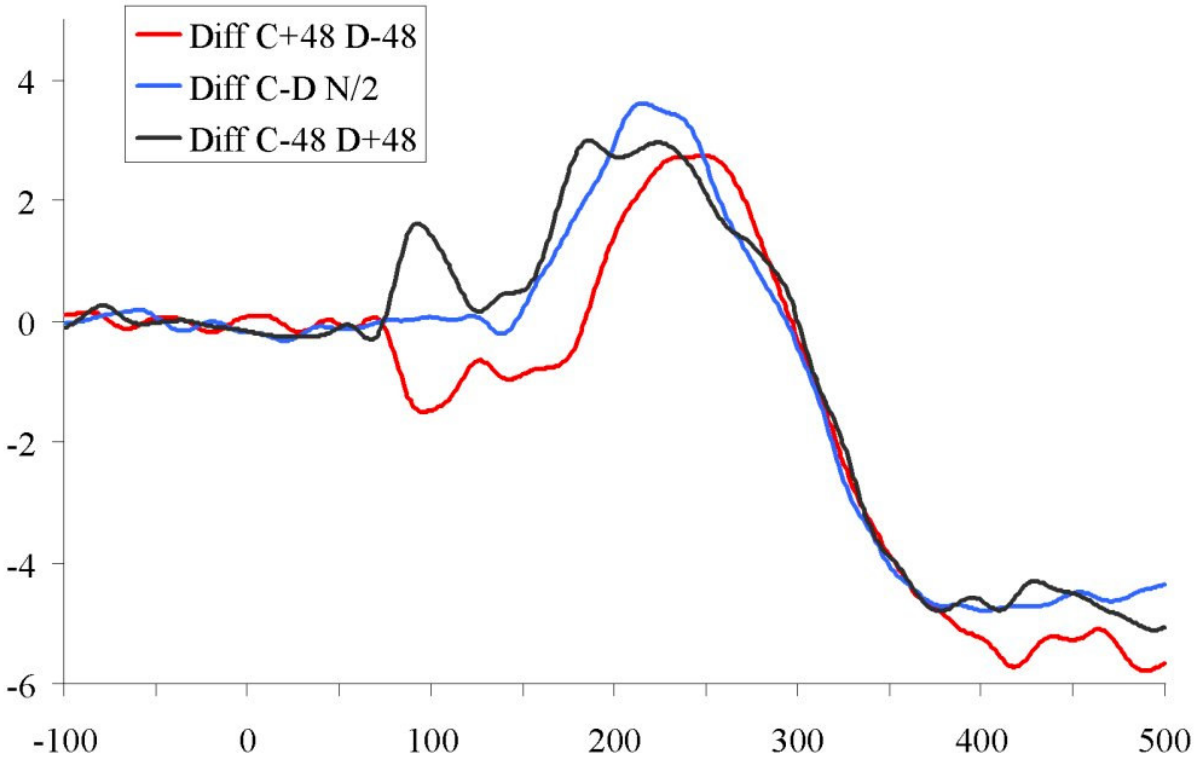


Figure 15 : La courbe en bleu clair représente l'activité différentielle sur les images pour lesquelles le contraste est divisé par deux sans changer la luminance. La courbe en rouge représente l'activité différentielle calculée entre les cibles les plus éclaircies et les distracteurs les plus assombris alors que la courbe noire représente l'activité différentielle pour les cibles les plus assombries moins les distracteurs les plus éclaircis. On peut noter que les activités différentielles précoces sont inversées, en accord avec l'inversion de luminance dans les images et sans suivre le sens de la soustraction cible - distracteur. L'activité différentielle qui apparaît après 150 ms se développe dans le même sens pour les trois courbes : elle est donc plus liée aux processus de catégorisation eux-mêmes qu'au traitement des caractéristiques physiques des images.

2 - Dynamique des premiers traitements visuels

La vitesse avec laquelle les sujets humains effectuent la tâche de catégorisation ainsi que les latences observées sur les enregistrements EEG sont difficiles à concilier avec les contraintes qui pèsent sur le système visuel. En effet, la fréquence de décharge maximale des neurones est relativement faible (100-200 Hz), les vitesses de conduction cortico-corticales sont peu élevées (environ 1 m/s (Nowak *et al.*, 1997) et même encore inférieures au sein d'une aire cortical donnée (Bringuier *et al.*, 1999)) et le nombre d'étapes à franchir pour parvenir dans des aires où des neurones sélectifs à des catégories d'objets ont été enregistrés est important (une dizaine en comptant les connexions intra-aires). Tous ces éléments rassemblés posent de grandes difficultés aux modèles de la vision proposés ces dernières décennies. Un certain nombre d'hypothèses ont été émises pour expliquer l'efficacité du système visuel :

- un traitement massivement parallèle de l'information sur l'ensemble du champ visuel (Thorpe, 2001 ; Rousselet *et al.*, 2004)

- un traitement asynchrone des informations qui donne la priorité aux éléments les plus saillants de l'image (VanRullen, 2003)

- des voies rapides pour transférer les informations sans passer nécessairement par toutes les étapes (Bullier & Kennedy, 1983 ; Yukie & Iwai, 1981 ; Nakamura *et al.*, 1993)

- enfin une architecture de traitement de l'information qui pourrait tirer partie des vitesses de conduction différentes entre les deux principaux systèmes qui véhiculent les informations visuelles (voir à ce sujet les modèles proposés au chapitre 1).

La vitesse de décodage des informations visuelles est cruciale pour la survie de l'individu et l'architecture du système visuel est probablement très optimisée pour que les informations y soient propagées et traitées rapidement. Dans leur expérience menée en 1996, Thorpe et al., montraient qu'une différence significative existait entre le signal EEG enregistré sur les cibles et les distracteurs dès 150 ms. Dans ce chapitre, nous explorerons l'aspect temporel du traitement de l'information visuelle en comparant ces 150 ms avec les latences des neurones enregistrés chez le singe et l'homme et avec les différents résultats de la littérature obtenus en EEG et MEG. Nous étudierons également les variations de latence de cette activité différentielle au cours de différentes tâches de catégorisation en manipulant la difficulté de la tâche et le niveau de préactivation du système visuel. Enfin, dans une expérience de masquage dans laquelle le temps d'intégration des images est limité, il nous sera possible d'explorer plus en détail la dynamique des processus visuels mis en jeu au cours de cette période de 150 ms.

2.1 - Latences de réponses dans le système visuel

2.1.1 - Enregistrements cellulaires et EEG

Des enregistrements unitaires chez le macaque montrent que les neurones du cortex inféro-temporal s'activent à partir de 70 à 100 ms en réponse à la présentation de réseaux ou d'images d'objets comme des visages ou des arbres (Richmond *et al.*, 1983 ; Perrett *et al.*, 1992 ; Nowak & Bullier, 1997 ; Schmolesky *et al.*, 1998 ; Vogels, 1999). Chez l'homme, les enregistrement intracrâniens montrent une activation du cortex inféro-temporal autour de 100 ms (Halgren *et al.*, 1994b ; Allison *et al.*, 1999) et du lobe frontal vers 150 ms (Halgren *et al.*, 1994a) lors de la présentation de visages.

Moins difficiles à mettre en œuvre que les enregistrements intracrâniens chez l'homme, l'EEG et la MEG permettent également d'explorer la dynamique du fonctionnement du cerveau en temps réel. Cependant, la source de l'activité électrique à l'origine du signal enregistré à la surface du scalp est difficile à reconstruire précisément et la localisation spatiale du (ou des) générateur(s) de l'activité électrique est donc assez limitée à partir du signal EEG/MEG seul. De bien meilleurs résultats sont obtenus en couplant ces enregistrements EEG ou MEG (128 canaux ou plus) à des données anatomiques ou fonctionnelles obtenues par IRM pour contraindre le positionnement des sources à une région limitée de l'espace cortical (Halgren *et al.*, 2000 ; Di Russo *et al.*, 2001 ; Vanni *et al.*, 2004 ; Di Russo *et al.*, 2005). L'activité la plus précoce observée sur les potentiels évoqués en EEG/MEG au moyen d'un stimulus visuel de type pattern-reversal est appelée C1. Cette composante apparaît vers 50 à 60 ms et correspond vraisemblablement dans sa première phase à l'activation de V1, puis de V1 et V2 autour de son pic, aux alentours de 70-80 ms. La deuxième composante, présente autour de 100 ms est une onde positive appelée P1. Elle est interprétée comme provenant de diverses aires de la voie ventrale et de la voie dorsale (V1, V2, V4, MT et MST et peut être également le cortex inféro-temporal). La composante suivante, N1, est négative. Elle débute autour de 120-130 ms et atteint son maximum entre 140 et 170 ms. Elle semblerait résulter de l'activation de la quasi-totalité des aires visuelles, mais également d'aires impliquées dans la mémoire (lobe temporal), la catégorisation et la prise de décision (lobe frontal) (Clark & Hillyard, 1996 ; Foxe & Simpson, 2002).

2.1.2 - Les activités différentielles avant 150 ms

Il est cependant difficile d'établir avec certitude les contreparties fonctionnelles de ces composantes électriques et il existe une controverse à la fois sur l'identification des premiers signaux cérébraux qui divergent entre les cibles et les distracteurs dans une tâche de catégorisation et sur le sens qu'il convient de leur donner. Thorpe et al., ont proposé en 1996 que le système visuel avait suffisamment traité l'information en 150 ms pour déterminer si une image contient un animal. L'activité différentielle présente entre les cibles et les distracteurs autour de 150 ms, même si elle est précoce, n'est cependant pas le premier signal différentiel qui peut être mis en évidence dans les enregistrements cérébraux au cours d'une tâche de catégorisation d'images. Certaines études rapportent ainsi des effets très précoces dans des tâches de reconnaissance des visages, dès 30 à 60 ms (Seeck *et al.*, 1997 ; Braeutigam *et al.*, 2001). Nous avons déjà rassemblé un faisceau d'arguments montrant que ces latences, qui ne sont compatibles qu'avec une activation des toutes premières aires de la vision (LGN, V1, et peut être V2 : Di Russo *et al.*, 2001 ; Vanni *et al.*, 2004) ne correspondent probablement qu'à des effets liés à des différences d'encodage lors de la répétition d'un stimulus, comme ceux observés dans une expérience de George et al. (George *et al.*, 1997), et non à des traitements avancés de reconnaissance d'identité. Ces expériences mettent en jeu plusieurs fois successivement les mêmes circuits nerveux lorsqu'un même stimulus ou un même type de stimuli est présenté et cette réactivation à des délais relativement courts peut fort bien donner lieu à des réponses neuronales qui évoluent entre le début et la fin de la série. L'effet observé dans ce type de tâche n'est donc probablement pas lié à une catégorisation proprement dite mais aux modifications physiologiques à court ou moyen terme causées par la répétition du stimulus lui-même. VanRullen a également souligné dans sa thèse que l'étude de Seeck et al. contient une erreur de protocole : la principale source du signal différentiel enregistré à 50 ms peut provenir d'un déséquilibre dans la fréquence de présentation d'une image cible ou distracteur affichée 200 ms avant le stimulus (VanRullen, 2000).

Dans des expériences de catégorisation de visages, de mains et de formes abstraites, Mouchetant-Rostaing et al., (Mouchetant-Rostaing *et al.*, 2000a ; Mouchetant-Rostaing *et al.*, 2000b) rapportent des effets de catégorisation grossière à des latences comprises entre 50 et 80 ms. Dans l'une des expériences, les sujets doivent détecter les visages portant des lunettes dans des séries de visages d'hommes, de femmes ou des deux mélangés en égale proportion. Bien que leur tâche ne porte pas sur la reconnaissance du genre des visages, une différence précoce apparaît à 50 ms dans le signal EEG entre les séries qui ne contiennent que des

hommes ou que des femmes et celles qui contiennent à la fois des hommes et des femmes. Les auteurs proposent que ces différences autour de 50-80 ms reflètent une catégorisation implicite et grossière du genre des visages. Il est cependant possible d'expliquer autrement l'apparition de cette activité différentielle, en considérant que la présentation successive de 100 visages d'hommes puis de 100 visages de femmes pourrait induire des effets d'habituation dans les premières aires de la vision, contrairement à la condition dans laquelle les visages d'hommes et de femmes sont alternés. Cet effet extrêmement précoce trouvé pour les visages, a également été répliqué avec des photographies de mains, alors qu'il n'existe probablement pas de structures spécialisées dans le traitement du genre des mains ! D'ailleurs les activités différentielles de grande amplitude présentes autour de 180 ms dans la tâche de catégorisation du genre des visages n'apparaissent que plus tardivement, vers 200-220 ms, dans la tâche de catégorisation du genre des mains. Il semblerait que dans ces expériences, ce sont ces activités différentielles "tardives" (180 ms pour les visages et 200-220 ms pour les mains), et non les activités différentielles précoces à 50 ms qui reflètent les processus de catégorisation.

D'autres études mettent en évidence un signal différentiel à une latence plus plausible de 80 à 120 ms et l'interprètent comme le témoin d'une catégorisation précoce (catégorisation de visages humains Liu *et al.*, 2002), ou encore d'une reconnaissance de la familiarité d'un visage (Debruille *et al.*, 1998). Dans l'étude de Debruille, dans laquelle les sujets devaient catégoriser des visages de personnes célèbres selon qu'ils leurs étaient familiers ou non, une différence significative apparaissait dans le signal EEG entre 76 et 130 ms. Les auteurs concluent à une reconnaissance visuelle très précoce de la familiarité des visages (et même de l'identité) malgré des réponses motrices n'apparaissant qu'environ 700 ms plus tard ! Le premier problème que l'on peut soulever dans cette étude concerne les différences physiques entre les images. Même si les images de visages connus et inconnus proviennent des mêmes fonds documentaires (les visages inconnus sont des personnes célèbres dans d'autres pays que celui où a lieu l'expérience), l'absence de contrôle statistique des propriétés des images (distribution des contrastes locaux ou des fréquences spatiales par exemple) ne permet pas d'exclure un effet des caractéristiques physiques entre les deux groupes d'images pour expliquer l'apparition d'activités différentielles précoces de faible amplitude et limitées à trois électrodes. Le second problème touche à la nature de la tâche elle-même. Il ne s'agit pas ici d'une tâche de catégorisation proprement dite mais d'une tâche portant sur la reconnaissance de la familiarité. Or on sait qu'il existe dans les cortex enthorhinal, inféro-temporal et périrhinal des neurones spécialisés dans la détection de la familiarité et de la nouveauté qui peuvent décharger à des latences très courtes sans toutefois avoir opéré une quelconque

catégorisation de l'objet détecté (Xiang & Brown, 1998). Les activités différentielles très précoces enregistrées dans cette tâche peuvent donc refléter des différences physiques entre les images ou des réponses neuronales relatives à la familiarité des stimuli.
Dans l'étude de Liu et al., un grand soin est apporté pour tenter de contrôler les propriétés de bas niveau des images de visages et de maisons que les sujets doivent catégoriser. Malgré une égalisation du spectre de puissance des cibles et des distracteurs et l'ajout de bruit aléatoire dans leur phase, les images restent néanmoins physiquement différentes puisque c'est justement sur ces faibles différences qui subsistent que les sujets vont s'appuyer pour effectuer la tâche. Ici encore, il n'est pas possible d'exclure un effet des différences de bas niveau entre les groupes d'images, même si elles ont été réduites, pour expliquer l'apparition d'activités différentielles autour de 100 ms.

Citons enfin une étude en MEG d'Halgren et al., qui mettent en évidence des effets à 110 ms dans une tâche de catégorisation de visages d'humains et d'animaux mais les relient clairement à des phénomènes perceptuels (Halgren *et al.*, 2000). Dans cet article, les auteurs concluent que la catégorisation des visages proprement dite n'intervient pas avant 160 ms.

Après cette revue critique d'un certain nombre d'études électrophysiologiques montrant des effets précoces liés à la catégorisation, il nous apparaît important de préciser que nous ne sommes pas fondamentalement opposé à ce qu'il puisse exister des phénomènes de haut niveau à des latences inférieures à 150 ms, mais aucune des études présentées ci-dessus n'utilise un protocole permettant de s'affranchir totalement des différences de bas niveau entre les images. Nous montrerons dans le 3ème chapitre de ce mémoire qu'il peut effectivement exister des processus liés à la catégorisation de visages à des latences d'environ 120-130 ms, mais qu'ils ne sont pas forcément décelables sur des enregistrements EEG dans lesquels les différences physiques entre les cibles et les distracteurs sont éliminées (Rousselet *et al.*, Submitted).

Afin de s'affranchir totalement à la fois des effets de série et des effets dus aux caractéristiques physiques de images, VanRullen et al. (VanRullen & Thorpe, 2001b) ont utilisé un protocole astucieux dans lequel ils enregistrent le signal EEG sur un même groupe d'images présentées successivement comme cibles puis comme distracteurs (et inversement) dans deux tâches de catégorisation différentes. Ils ont ainsi montré qu'une telle procédure abolit totalement les signaux différentiels précoces entre 80 et 120 ms tout en conservant l'activité différentielle principale qui apparaît vers 150 ms (Figure 1). Les différences

précoces seraient donc dues à des différences statistiques entre les groupes d'images qui se traduiraient à leur tour par des différences dans le signal EEG correspondant à l'encodage physique des images dans le système visuel. Nous présentons à plusieurs reprises dans ce mémoire des données complémentaires provenant de trois expériences qui vont dans le même sens que cette première étude et tendent à montrer la non-spécificité de ces activités différentielles précoces du point de vue de la catégorisation. Il s'agit, comme nous l'avons vu au chapitre 1, des expériences sur la luminance et le contraste des images qui permettent d'agir directement sur l'amplitude des activités différentielles précoces en manipulant les propriétés physiques des images ou encore d'une tâche de catégorisation animal/non animal alternée avec une tâche de catégorisation de visages humains qui sera présentée au 3ème chapitre.

2.1.3 - Les activités différentielles après 150 ms

A l'inverse, certains auteurs considèrent que l'activité différentielle à 150 ms est elle-même un artefact dû à des différences entre les images utilisées ou que les neurones qui répondent à cette latence ne le font pas de manière spécifique à la catégorie recherchée Johnson & Olshausen, 2003. Ces auteurs s'appuient principalement sur le fait qu'une analyse essai par essai révèle qu'il n'existe pas de corrélation entre la latence des temps de réaction et la latence de l'activité différentielle à 150 ms. Ils associent donc les processus de catégorisation proprement dit à un signal plus tardif (autour de 300 ms), fortement corrélé avec le temps de réaction. Mais leur argument tient difficilement après l'étude de DiCarlo et al. (DiCarlo & Maunsell, 2005) qui a montré que la latence de décharge des neurones enregistrés dans le cortex inféro-temporal est toujours fortement corrélée à l'apparition du stimulus et faiblement corrélée aux réponses motrices, alors que l'importance de cette aire est clairement reconnue dans la reconnaissance d'objets et la catégorisation visuelle (Desimone *et al.*, 1984 ; Gross, 1992 ; Vogels, 1999 ; Tanaka, 2003) et qu'elle semble très impliquée dans la génération de l'activité différentielle à 150 ms (Delorme *et al.*, 2004 ; Fize *et al.*, 2005). De plus, nous savons que les premières réponses comportementales se situent autour de 250 ms. En tenant compte du délai entre l'activation du cortex moteur et la détection du mouvement (20-30 ms pour le délai cortex moteur - motoneurones de la main (Hess *et al.*, 1987) et encore 10-15 ms pour que la main se déplace), nous disposons d'arguments très forts pour affirmer que l'activation des cellules liées à la catégorisation doit intervenir bien avant 300 ms !!!

2.1.4 - Encore et toujours 150 ms !

De nombreuses expériences ont confirmé en revanche que l'activité qui apparaît à 150 ms est robuste et constitue probablement un bon marqueur des premiers traitements cognitifs engagés dans la tâche en cours (catégorisation de visages : Schendan *et al.*, 1998 ; catégorisation d'animaux dans des scènes naturelles : Antal *et al.*, 2000 ; catégorisation de visages d'hommes et d'animaux : Halgren *et al.*, 2000 ; catégorisation de véhicules et d'animaux : VanRullen & Thorpe, 2001b). Il est d'ailleurs étonnant d'observer la relative homogénéité des latences recueillies au moyen de protocoles expérimentaux pourtant très différents (go/no-go ou choix forcé, photographies ou dessins…). A titre d'exemple, Van Rullen et al., (VanRullen & Thorpe, 2001a) ont réalisé une expérience dans laquelle les sujets devaient catégoriser des véhicules ou des animaux. Bien que les véhicules ne soient pas des objets biologiques, les sujets avaient des performances comportementales similaires sur les deux catégories et les activités différentielles enregistrées au cours de ces tâches étaient similaires.

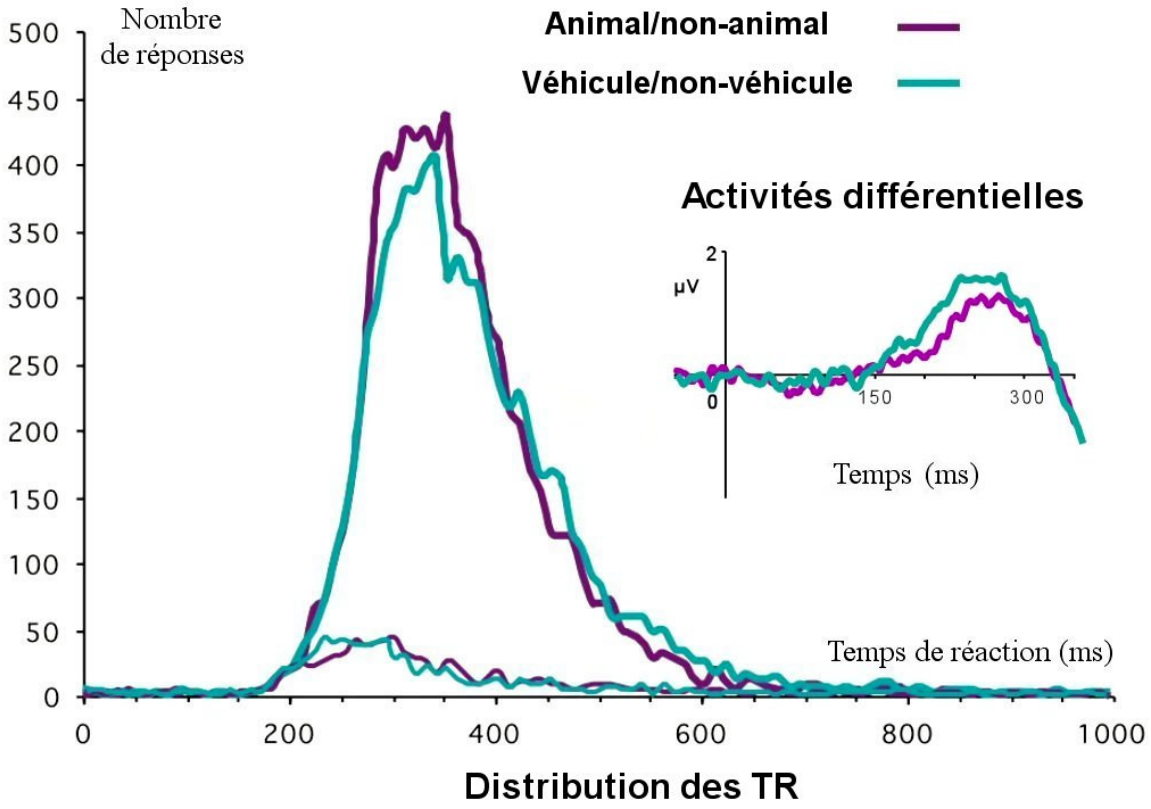


Figure 1 : Distribution des temps de réaction et activités différentielles dans deux tâches de catégorisation alternées animal/non animal et véhicule/non véhicule. Le comportement et l'électrophysiologie montrent qu'il existe une grande similarité dans le décours temporel du traitement visuel de ces deux catégories d'objets. Reproduit d'après VanRullen et al., J Cogn Neurosci (2001).

Étant donné que nous sommes principalement intéressés par les contraintes temporelles qui peuvent peser sur le système visuel, notre but va maintenant être de trouver des tâches dans lesquelles l'activité différentielle apparaît avant 150 ms.

Les tâches de catégorisation dont nous avons parlé jusqu'à maintenant sont relativement complexes, puisque le système visuel doit reconnaître parmi un grand nombre d'attributs ceux qui sont pertinents, c'est à dire "diagnostiques" de la catégorie cible afin de les traiter de façon préférentielle. De plus, l'aspect exact des cibles est imprédictible puisque les images sont toujours nouvelles et que l'ensemble des photographies non-cibles est très vaste et varié. Si l'on veut essayer de diminuer la latence de l'activité différentielle, il nous faut nous tourner vers des tâches plus simples ou des tâches dans lesquelles il est possible de pré-activer plus avant les représentations de la cible.

2.2 - Pré-activation du système visuel…

2.2.1 - En simplifiant les cibles et les distracteurs

Effectuer la tâche de catégorisation ultra-rapide sur des objets biologiques ou artificiels donne lieu à des latences d'activité différentielle similaires. Mais les images utilisées dans ces expériences sont complexes et nous présentons ici une autre expérience dans laquelle les sujets devaient effectuer une catégorisation de formes très simples : rond/carré. Lorsque des sujets doivent catégoriser des cercles parmi des carrés, leur précision est augmentée de 6% et leur temps de réaction moyen diminué de 50 ms par rapport à une tâche de catégorisation animal/non animal (Aubertin *et al.*, 1999). La simplification de la tâche permet donc d'accélérer le processus de décision grâce à une réduction de la complexité et de l'ambiguïté des stimuli, mais cette amélioration reste modeste. Il est probable que les influences descendantes depuis les aires de "haut-niveau" vers les aires visuelles (top-down) soient plus efficaces dans le cas d'une tâche simple comme la détection de cercles que dans le cas d'une catégorisation d'animaux dont la diversité est très importante. Cependant, ces hypothèses ne concernent que les temps de réaction moyens. En effet aucune amélioration des performance n'a lieu du côté des temps de réaction les plus courts. De plus, la latence de l'activité différentielle, vers 150 ms, est encore une fois superposée entre les tâches de catégorisation carré/rond et animal/non animal (Aubertin *et al.*, 1999, Figure 2) ; seule l'amplitude des deux courbes diffère après 200 ms.

Simplifier l'aspect des catégories cibles dans cette tâche de catégorisation ne permet donc pas de faire varier la latence de l'activité différentielle. Pour influencer la latence de cet événement, il est peut être nécessaire de modifier légèrement les conditions de la tâche, par exemple en pré-activant le système visuel grâce à un apprentissage portant sur un nombre réduit d'images.

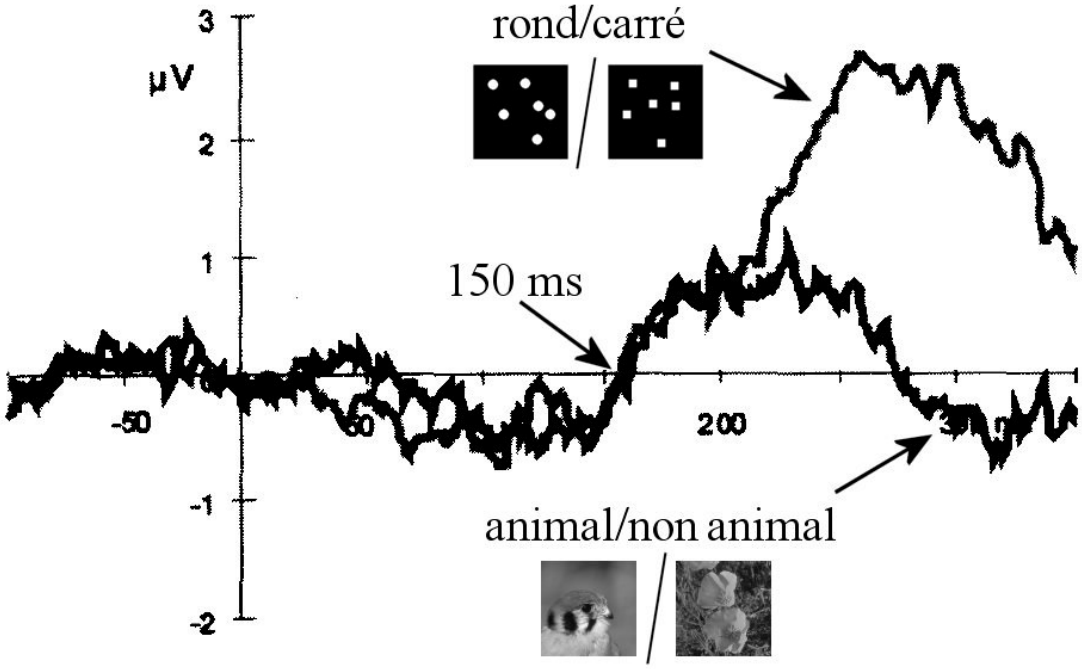


Figure 2 : Activités différentielles comparées entre une tâche de catégorisation rond/carré et une tâche de catégorisation animal/non animal. Reproduit d'après Aubertin et al., C R Acad Sci III (1999)

2.2.2 - En faisant intervenir l'apprentissage

Si l'activité différentielle correspond à l'activation des représentations liées aux objets présents dans la scène, il devrait être possible de diminuer sa latence en pré-activant plus fortement les aires corticales en charge du traitement des objets présents dans la scène. Le moyen le plus simple de préactiver le système visuel consiste à présenter plusieurs fois les images afin que le système visuel des sujets puisse améliorer sa performance spécifiquement sur ces cibles. C'est ce qui a été réalisé dans l'étude que nous décrivions dans l'introduction de ce mémoire, dans laquelle les sujets devaient catégoriser une série de 200 images (animaux et distracteurs) chaque jour pendant trois semaines. A l'issue de ces trois semaines d'apprentissage, un test consistait à catégoriser ces 200 images familières parmi 200 images nouvelles. Nous avons mentionné que les sujets n'étaient pas plus rapides pour catégoriser les images familières que les images nouvelles puisque le TR minimal était similaire pour les 2 groupes d'images. La

figure 3 nous montre que les enregistrement EEG effectués au cours de l'expérience vont dans le même sens que le comportement puisque la latence de l'activité différentielle est similaire pour les deux groupes d'images : 150 ms. Seule l'amplitude de cette activité différentielle est différente, les images familières donnant lieu à une différentielle plus ample.

Ces résultats sont surprenants et montrent que le processus de catégorisation d'images naturelles en jeu dans cette tâche est remarquablement optimisé. Il n'est pas possible de l'accélérer ou de court-circuiter certaines étapes pour aller plus vite, quand bien même les cibles sont parfaitement connues.

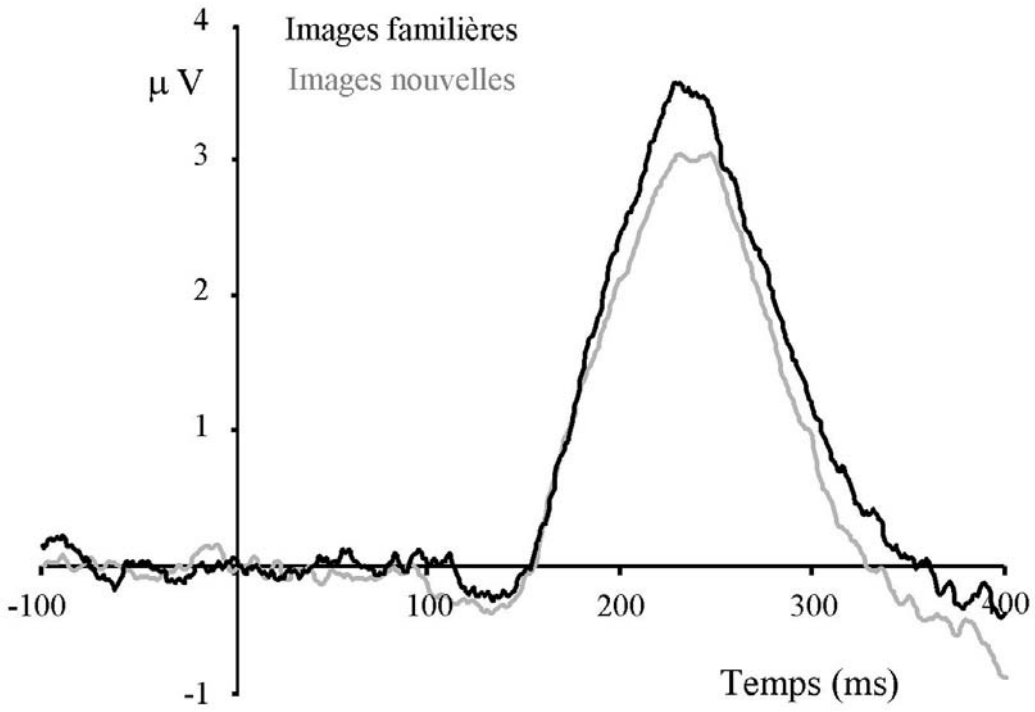


Figure 3 : 14 sujets devaient effectuer une tâche de catégorisation animal/non animal sur un ensemble de 200 images familières (catégorisées tous les jours pendant trois semaines) et de 200 nouvelles images mélangées. La courbe représente les grandes moyennes des activités différentielles (cibles-distracteurs) enregistrées sur les images familières (en noir) et nouvelles (en gris). Reproduit d'après Fabre-Thorpe et al., J Cogn Neurosci (2001).

Si l'on prend pour hypothèse que le processus de catégorisation résulte de l'interaction entre la préactivation du système visuel pour favoriser les représentations de la catégorie cible et la reconstruction le long des voies visuelles du stimulus présenté, il est intéressant de comparer les latences de réponse que nous obtenons dans la tâche de catégorisation avec celles que nous observerions dans des tâches où la préactivation peut être maximale comme pour une reconnaissance (réponse à l'apparition d'un stimulus donné parmi d'autres) ou une simple détection (réponse à l'apparition d'une image quelconque). Les différences de latences entre

les tâches de détection simple, de reconnaissance et de catégorisation nous indiqueront si des étapes de traitement peuvent être évitées lorsque la tâche se simplifie ou si l'optimisation du système visuel est telle que seul un gain de temps très faible est possible. Ces expériences peuvent nous donner accès successivement au "coût temporel" de la catégorisation et de la reconnaissance en les comparant entre elles.

2.2.3 - En maximisant les influences descendantes : articles n°2 et 3

Dans les deux articles qui suivent, nous comparons les performances obtenues sur des tâches de catégorisation et de reconnaissance dans des expériences menées chez l'homme et chez le macaque. Nous proposons une brève synthèse des données de ces deux expériences pour mettre en parallèle les résultats obtenus chez les deux espèces.

Résumé des deux publications : "Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes" et "Rapid categorization of natural scenes in monkeys: target predictability and processing speed" :

Bien que similaires dans leur principe, les expériences menées chez les hommes et les singes présentaient quelques différences dans leur protocole pour s'adapter aux particularités des singes. Les humains (14 sujets), devaient effectuer 20 séries de 100 essais et les singes (3 sujets), 20 séries de 150 essais en alternant à chaque série entre la tâche de catégorisation et la tâche de reconnaissance. La tâche de reconnaissance commençait par la présentation répétée de l'image cible avant le début de la série. Celle-ci était présentée 5 secondes au sujet, puis flashée 5 fois, le tout répété trois fois. Ainsi, les sujets pouvaient prendre leur temps pour bien analyser l'image et choisir implicitement des détails visuels précis pour reconnaître cette cible le plus rapidement possible. A la différence des singes, les humains ne catégorisaient que des images nouvelles et leurs cibles dans la tâche de reconnaissance pouvaient être aussi bien des images d'animaux que des distracteurs.

Résultats résumés :

Les humains, comme les singes, ont obtenu des performances plus élevées dans la tâche de reconnaissance par rapport à la tâche de catégorisation. La précision des humains était supérieure de 5,6% (93,1 vs 98,7%) et celle des singes de 3,9% (92,4 vs 96,3%). Cette augmentation de précision était également accompagnée d'une diminution des temps de réaction. Le TR médian diminuait de 63 ms (400 vs 337 ms) chez les humains et de 19 ms chez les singes (263 vs 244 ms). Ainsi, la performance globale des sujets était bien supérieure

dans la tâche de reconnaissance par rapport à la tâche de catégorisation. De plus, ce ne sont pas seulement les TR moyens ou médians qui étaient diminués mais l'ensemble de la distribution des TR qui était déplacée vers des latences plus courtes puisque les TR minimaux diminuaient également de 40 ms chez les hommes (260 vs 220 ms : article n°2, Figure 2) et de 20 ms chez les singes (180 vs 160 : article n°3, Figure 2). Les comparaisons de données montrent que les latences chez l'homme sont toujours plus tardives que chez le macaque, probablement parce que la taille de notre cerveau a pour conséquence d'augmenter la longueur moyenne de nos connexions corticales par rapport à celles du singe (Thorpe & Fabre-Thorpe, 2001 ; Macé *et al.*, 2005).

L'observation des erreurs commises par les hommes et les singes lors de la tâche de reconnaissance permet de supposer qu'il existe de grandes similitudes dans la manière de traiter les images chez les humains et les macaques. En effet, outre les quelques erreurs en commun que commettaient les deux espèces, une importante fraction des erreurs commises par les singes comme par les hommes pouvait s'expliquer par des réponses anticipées sur des images partageant des caractéristiques physiques de bas niveau avec la cible .

Chez les humains, les enregistrements EEG ont permis de montrer que la tâche de reconnaissance fait appel à des processus cérébraux ayant un décours temporel plus rapide que celui de la tâche de catégorisation. L'activité différentielle commençait 35 ms plus tôt dans la tâche de reconnaissance par rapport à celle enregistrée dans la tâche de catégorisation (Figure 4).

Discussion :

La tâche de catégorisation visuelle que nous utilisons dans nos expériences est déjà réalisée de façon très rapide, et nous avons vu précédemment qu'il n'est pas possible d'accélérer les traitements sous-jacents lorsque les images utilisées sont familières. Dans les expériences présentées ici, nous montrons qu'il est possible de réduire les temps de traitement en simplifiant la tâche du sujet pour qu'il n'ait plus à effectuer qu'une simple reconnaissance de cible en utilisant des influences descendantes optimisées. Dans le cas de la catégorisation, les sujets doivent extraire les informations d'une image afin de déterminer si elle contient ou non un animal. La variabilité des animaux est très grande et les sujets n'ont aucune information à priori sur le type d'animal, son emplacement, l'échelle à laquelle il va apparaître etc… Ces traitements sont donc relativement complexes par rapport à ceux mis en jeu dans la tâche de reconnaissance où l'information à trouver dans l'image est parfaitement connue, sans la moindre variabilité d'une cible à la suivante. Malgré ces différences très importantes en termes de complexité de la tâche, le TR minimal des hommes n'augmente que de 40 ms et

celui des singes de 20 ms. De plus la latence de l'activité différentielle dans la tâche de catégorisation n'est augmentée que de 35 ms par rapport à celle de la tâche de reconnaissance. On peut noter que l'augmentation du TR moyen est beaucoup plus marquée que celle du TR minimal, probablement parce que la simplification de la tâche à un effet important sur la durée nécessaire à la prise de décision des sujets et donc sur la variabilité de la distribution des réponses (Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004). La proportion de réponses avec des TR longs est bien plus élevée dans la tâche de catégorisation, ce qui augmente le TR moyen. Il est très probable que la vitesse de prise de décision ne soit pas le seul facteur qui donne un avantage à la reconnaissance et qu'étant donné la prédictibilité de la cible, il soit possible de gagner du temps à chaque étape lors des traitements visuels. Cette idée est appuyée par les résultats électrophysiologiques qui montrent une activité différentielle dans la tâche de reconnaissance environ 135 ms après l'apparition du stimulus (Figure 4). Cette activité différentielle est localisée majoritairement dans des structures du cortex inféro-temporal (Delorme *et al.*, 2004), tout comme celle enregistrée autour de 150 ms dans la tâche de catégorisation animal/non animal (Fize *et al.*, 2000, Halgren *et al.*, 2000). Il existe donc une corrélation entre la diminution du TR minimal de 40 ms dans la tâche de reconnaissance et une apparition plus précoce (de 35 ms) de l'activité différentielle dans cette zone du cerveau.

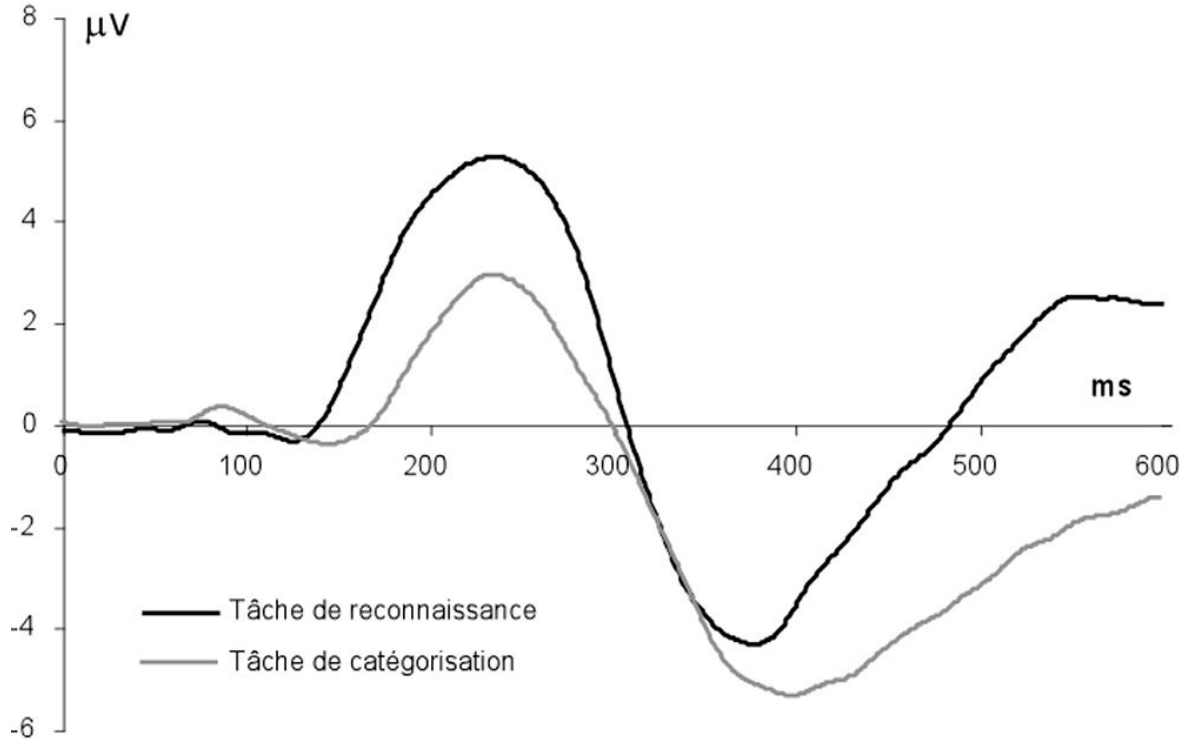


Figure 4 : Activités différentielles enregistrées dans la tâche de reconnaissance et la tâche de catégorisation. Reproduit d'après Delorme et al., Cogn Brain Res (2004).

Article n°2

Brain Res Cogn Brain Res, **19**, 103-113

Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes

Arnaud Delorme, Guillaume A.Rousselet, **Marc J-M. Macé**

& Michèle Fabre-Thorpe

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Cognitive Brain Research 19 (2004) 103–113
COGNITIVE BRAIN RESEARCH
www.elsevier.com/locate/cogbrainres

Research report

Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes

Arnaud Delorme[1], Guillaume A. Rousselet[2], Marc J.-M. Macé, Michèle Fabre-Thorpe*

Centre de Recherche Cerveau et Cognition (UMR 5549, CNRS-UPS), 133 route de Narbonne, 31062, Toulouse Cedex, France

Accepted 19 November 2003

Abstract

The influence of task requirements on the fast visual processing of natural scenes was studied in 14 human subjects performing in alternation an ''animal'' categorization task and a single-photograph recognition task. Target photographs were randomly mixed with non-target images and flashed for only 20 ms. Subjects had to respond to targets within 1 s. Processing time for image-recognition was 30–40 ms shorter than for the categorization task, both for the fastest behavioral responses and for the latency at which event related potentials evoked by target and non-target stimuli started to diverge. The faster processing in image-recognition is shown to be due to the use of low-level cues, but source analysis produced evidence that, regardless of the task, the dipoles accounting for the differential activity had the same localization and orientation in the occipito-temporal cortex. We suggest that both tasks involve the same visual pathway and the same decisional brain area but because of the total predictability of the target in the image recognition task, the first wave of bottom-up feed-forward information is speeded up by top-down influences that might originate in the prefrontal cortex and preset lower levels of the visual pathway to the known target features.

Theme: Neural basis of behaviour
Topic: Cognition

Keywords: Natural scenes; Categorization; Image recognition; Top-down influences; Early Visual Processing; Decision-making; Differential ERPs

1. Introduction

Spotting a specific object among others is an every day task that appears trivial but raises a number of questions concerning the underlying visual processing. In visual search tasks, subjects are asked to look for a pre-specified target embedded in distractor arrays. Typically, for low-level features, ERP studies suggest that a visual decision can be made in about 150 ms [1,21,34]. This latency increases when targets are defined by a conjunction of characteristics such as form and color [18], although pop out has been reported for some specific conjunction of low-level features [7,21,28,38]. Surprisingly, 150 ms has also been reported to be the minimal processing time to differentiate between different classes of natural images. Using a superordinate categorization task in which human subjects had to respond when a natural image that they had never seen before contained an animal, Thorpe et al. [36] showed that visual evoked potentials recorded on correct target trials differed sharply from those recorded on correct distractor trials at about 150 ms after stimulus onset. This differential brain activity has been found at the same latency with non-biological relevant categories of objects such as ''means of transport'' and has been shown to be related to ''visual decision making'' rather than physical differences between photographs belonging to different categories [40]. This speed of processing could well be seen for any well-learned object-category [32]. In such categorization tasks, very different objects have to be grouped together (i.e. a snake and a flock of sheep) and performance cannot rely on the analysis of a single low-level cue or even on a single

* Corresponding author. Tel.: +33-5-62-17-28-07; fax: +33-5-62-17-28-09.
E-mail address: Michele.fabre-thorpe@cerco.ups-tlse.fr (M. Fabre-Thorpe).
[1] Present address: Computational Neurobiology Laboratory, Salk Institute, 10010 N. Torrey Pines Road, San Diego, CA 92037, USA. Tel.: +1-858-458-1927x15; fax: +1-858-587-0417. arno@salk.edu (A. Delorme), *URL:* http://www.cnl.salk.edu/~arno.
[2] Present address: Department of Psychology, McMaster University, 1280 Main St. W., Hamilton, ON, Canada L854K1.

0926-6410/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogbrainres.2003.11.010

conjunction of low-level cues. When considering this very short delay together with the anatomy and physiology of the visual system, it was argued that such severe temporal time constraints imply that the underlying processing probably relies on feed-forward mechanisms during a first wave of visual information [35,36].

It thus seems that high-level search tasks such as looking for an animal in a natural scene might be performed as fast as the simplest pop-out search tasks. To explain speed of processing in visual search tasks, emphasis had been put on the target saliency, and on the number of diagnostic stimulus features [33]. However, increasing stimulus diagnosticity in the animal categorization task of natural images by using highly familiar photographs failed to induce a decrease of the minimal processing time: subjects could categorize novel images as fast as images on which they had been extensively trained [8].

Thus, the fast visual processing mode that underlies rapid categorization cannot be speeded up when top-down presetting of the visual system is optimized with experience. However, it is a difficult experimental issue to determine the relative importance of bottom-up and top-down processes. To investigate further how top-down knowledge related to task requirements could influence the visual analysis of natural images, we tested human subjects in a task in which they were assigned a given photograph as target and had to detect this single target-photograph among a variety of different non-target stimuli. Being fully briefed about the target should allow subjects to maximize the use of top-down influences and to rely only on a limited number of low-level cues specific to the target-image.

In the present experiment, we studied the fast processing of natural images in human subjects performing in alternation the superordinate ‘‘animal/non-animal’’ categorization task and a single-photograph recognition task. Along with behavioral performance, analysis involved associated ERPs and localization of brain sources to investigate the neural dynamics of early information processing. Since both tasks used the same natural images as stimuli and required the same motor response, any processing differences should be related to task requirements.

2. Methods

2.1. Stimuli

All stimuli used in the two tasks were photographs of natural scenes (Corel CD-ROM library). In each group, images were chosen to be as varied as possible (Fig. 1). Subjects were tested on blocks of 100 stimuli including 50% targets and 50% distractors. In the categorization task 1000 photographs were used (50% distractors and 50% targets) and each of them was seen only once by each subject. The target-photographs included pictures of mammals, birds, fish, arthropods, and reptiles. There was no a priori information about the size, position or number of targets in the photograph. There was also a wide range of non-target images, with outdoor and indoor scenes, natural landscapes or city scenes, pictures of food, fruits, vegetables, trees and flowers. . .

In the recognition task, as in the categorization task, targets and non-targets were equiprobable in each block of 100 images so that the target-photograph assigned to a given block was seen 50 times among 50 varied non-target photographs that did not contain an animal. Each of the 14 subjects was tested with 15 targets (a total of 210 targets) and the same 750 non-target stimuli. In the 210 photographs used as targets, 140 (10 images per subject) contained an animal and were thus similar to the target photographs used in the categorization task. They had been categorized by human subjects in a previous study [8] and were known to offer different levels of difficulty. The remaining 70 (five images per subject) did not contain any animal and were thus homogenous with the non-targets used in both tasks.

2.2. Task and protocol

Fourteen human subjects (seven women and seven men, mean age 26 ranging from 22 to 46), with normal or corrected to normal vision volunteered for this study. Participants sat in a dimly lit room at 110 cm from a color computer screen piloted from a PC computer. They were required to start a block of 100 images by pressing a touch-sensitive button. A small fixation point ($<.1°$ of visual angle) appeared in the middle of the black screen. Then, an 8-bit color vertical photograph (256 pixels wide by 384 pixels high which roughly correspond to $4.5 \times 6.5°$ of visual angle) was flashed for 20 ms using a programmable graphic board (VSG 2.1, Cambridge Research Systems). The short presentation time prevented any exploratory eye movement. The stimulus onset asynchrony (i.e. time between the onset of one image and the onset of the next image in a series) was random between 1800 and 2200 ms.

Subjects had to give a go/no-go response: releasing the button as quickly and accurately as possible when they saw a target-image but keeping their finger(s) on the button on non-target trials. They were given a maximum of 1000 ms to

Fig. 1. Targets and associated errors in the recognition task. Target-images used in the recognition task are illustrated on a green background. The figures show the high variety of the animal images used in the 10 testing blocks (images a, b, c, e, f, i, j, k, l, m, n, o, q, v) in which animals are sometimes hardly visible (e, i, j, v) and the non-animal images used in the five control blocks (images d, g, h, p, r, s, t, u, w, x). On the right of each target-image is shown the non-target photograph(s) that induced a false alarm. Errors can clearly be related to global orientation (a, c, d, g, h. . .), color (e, i, j, l. . .), color patches in specific locations (n, t. . .), object identity or semantics (p, s, x. . .), spatial layout of the scene (b, e, f, k, n, v. . .) or any combination. The figures below each error indicate the reaction time of the incorrect go response. Similar natural images were used in the categorization task.


Target
Error 1
Target
Error 1
Error 2
Target
Error 1
a
260
i
247
207
q
259
b
207
j
355
244
r
279
c
454
k
294
314
s
304
d
414
l
242
247
t
278
e
457
m
279
250
u
296
f
336
n
320
931
v
217
g
202
o
325
243
w
324
h
372
p
332
222
x
344

respond, after which delay any response was considered as a no-go response.

On two different days, subjects were tested on 10 categorization blocks and 10 recognition blocks, alternating between the two tasks within a session while their associated EEG was recorded. In the animal categorization task, subjects had to respond whenever the picture contained an animal. In the target-image recognition task, a given animal image was assigned as the target for the following block of 100 images. The five image-recognition control blocks using images that did not contain an animal were inserted at regular intervals.

For the image-recognition task, each testing block was preceded by a learning phase during which the subject was presented with the target-photograph which was both repeatedly flashed for 20 ms (similar to the testing conditions) and presented for 1000 ms to allow ocular exploration (3*5 flashes intermixed with two long—1000 ms—presentations). Participants were instructed to carefully inspect and memorize the target-image in order to respond to it in the following sequence of images as fast and as precisely as possible. The testing block started immediately after the learning phase.

2.3. Evoked-potential recordings and analysis

Electric brain potentials were recorded from 32 electrodes mounted on an elastic cap (Electro-cap International). Data acquisition was made at 1000 Hz using a SynAmps recording system (Neuroscan) coupled with a PC computer. The analog low-pass filter was set at 500 Hz and the default SynAmps analog 50-Hz notch filter was used. Impedances were kept below 5 kΩ. Potentials were recorded with respect to common reference Cz, then average re-referenced. Potentials on each trial were baseline corrected using the signal during the 100 ms that preceded the onset of the stimulus. Trials were checked for artifacts and discarded using a $[-50; +50\ \mu V]$ criterion over the interval $[-100; +400$ ms] at frontal electrodes for eye movements and a $[-30; +30\ \mu V]$ criterion on the period $[-100; +100$ ms] at parietal electrodes to discard alpha brain waves. Only correct trials were considered for ERP averages. The waveforms were low-pass filtered at 35 Hz for use in graphics. Inter-subject two-tailed statistical *t*-tests (13 *df*) were performed on unfiltered ERPs for each electrode to evaluate the latency at which target ERPs diverged from non-target ERPs. This differential activity onset was defined as the time from which 15 consecutive values were statistically different to compensate for multiple comparisons. We computed significance for all electrodes but focused on two groups: frontal electrodes (10–20 system nomenclature: Fz, FP1, FP2, F3, F4, F7, F8) and occipital electrodes (10–20 system nomenclature: O1 and O2 with the addition of Oz, I, O1′, O2′, PO9, PO10, PO9′, PO10′) where the differential activity reached the highest amplitude. The additional occipital electrodes have the following spherical coordinates (theta/phi): Oz = 92/−90, I = 115/−90, O1′ = −92/54, O2′ = 92/−54, PO9 = −115/54, PO10 = 115/−54, PO9′ = −115/72, PO10′ = 115/−72.

2.4. Source localization

The source analysis was performed using a four-shell ellipsoidal model and using Brain Electrical Source Analysis (BESA, version 99). Because of temporal muscle contraction, the two most temporal electrodes were too noisy and were discarded from the analysis. All other electrodes were used to localize the equivalent dipoles. Grand-average waveforms were low-pass filtered at 35 Hz before analysis. Pairs of dipoles were placed in a central position, given a spatial symmetry constraint, then fitted in location and orientation for a particular time window (simplex algorithm).

3. Results

The aim of this study was to compare the visual processing of a natural image when the task required the representation of a high-level object category such as ‘‘animal’’ or when it could be performed using short-term memory of low-level cue(s). Behavior and ERPs were recorded and analyzed in all subjects.

3.1. Behavioral results: recognition vs. categorization

The analysis of behavioral performance included accuracy, speed of response and a study of the non-target images that incorrectly induced a go-response.

3.1.1. Accuracy

Although extremely good in both tasks (93.1% correct in the categorization task; 98.7% correct in the recognition task) accuracy was significantly better in the recognition

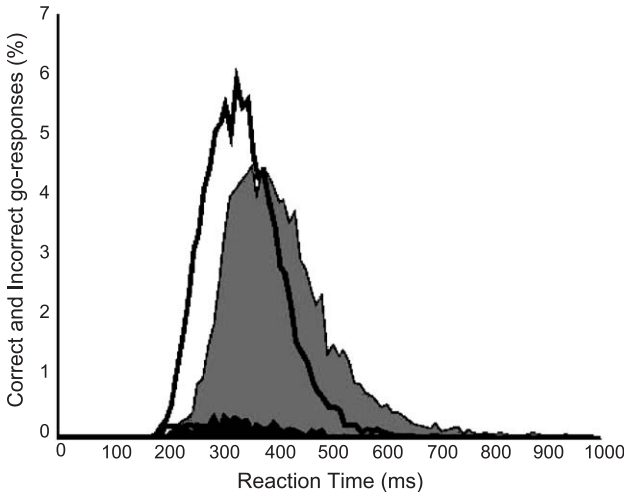


Fig. 2. Overall reaction time distribution of go-responses in both the animal categorization task (gray traces and shaded distribution) and the recognition task (black lines). The top two traces are for correct go responses towards targets, the bottom two traces are for false alarms induced by non-target stimuli.

task (two-tailed χ^2: $df=1$, $p<0.0001$), an effect that was found to be significant at $p<0.05$ for each individual subject. An accuracy bias was found in both tasks, but whereas this bias was in favor of correct no-go responses in the categorization task, it was in favor of correct go responses in the recognition task. Thus, subjects were slightly better at ignoring distractors than responding to animal-targets in the categorization task (93.9% vs. 92.4%; two-tailed χ^2: $df=1$, $p<0.0001$), whereas they were more accurate at detecting the target-image in the recognition task than at ignoring non-target images (99.7% vs. 97.5%; two-tailed χ^2: $df=1$, $p<0.0001$). This result provides an argument for the use of different strategies in the two tasks that will be discussed later.

3.1.2. Reaction time (RT)

As illustrated in Fig. 2, reaction times were significantly faster for the recognition task (median RT: 337 ms) than for the categorization task (median RT: 400 ms; two-tailed Mann–Whitney *U*-test: $p<0.0001$). For individual subjects this difference was always significant ($p<0.01$).

Processing speed can be measured using median RT or mean RT, but these values do not reflect all aspects of processing speed. One very useful value is the minimal processing time needed to complete the tasks. The average slower speed in the categorization task could be due to some difficult photographs that need longer processing time [8]. Thus, although the average processing time could be shorter in the recognition task, the minimal processing time might be similar in both tasks. As in our experimental protocol targets and non-targets were equiprobable in both tasks, we defined the minimal processing time (Fig. 2) as the first time bin for which correct hits to targets started to significantly outnumber false alarms to non-targets. Responses triggered with shorter latency but with no bias towards correct go-responses were presumably anticipations initiated before stimulus processing was completed. Using 10-ms time bins, this "minimal processing time" was found significant at 220 ms (two-tailed χ^2: $df=1$, $p<0.0001$) in the recognition task and at 260 ms in the categorization task (two-tailed χ^2: $df=1$, $p=0.0007$). The minimal processing time to reach decision was thus shortened by about 40 ms in the recognition task relatively to the categorization task. However, this shortening of RT latencies can be seen in Fig. 2 as a shift of the entire RT distribution of the recognition task toward shorter latencies, from the earliest to the latest behavioral responses.

3.1.3. Control set

The results obtained in the recognition task with the control sets (that used non-animal target pictures) show again the better accuracy and the shorter processing time associated with tasks that only require image recognition (Fig. 3). Subjects scored 98.3% correct, with a median RT for correct go-responses at 348 ms. These scores are slightly below the performance level observed when the one-image target contained an animal (respectively 98.7% and 337 ms), a result that could be due to higher similarities with the distractors, but the minimal processing time was found at exactly the same latency (220 ms) in both cases ($p<0.0001$, χ^2 test evaluated over every 10 ms time bin).

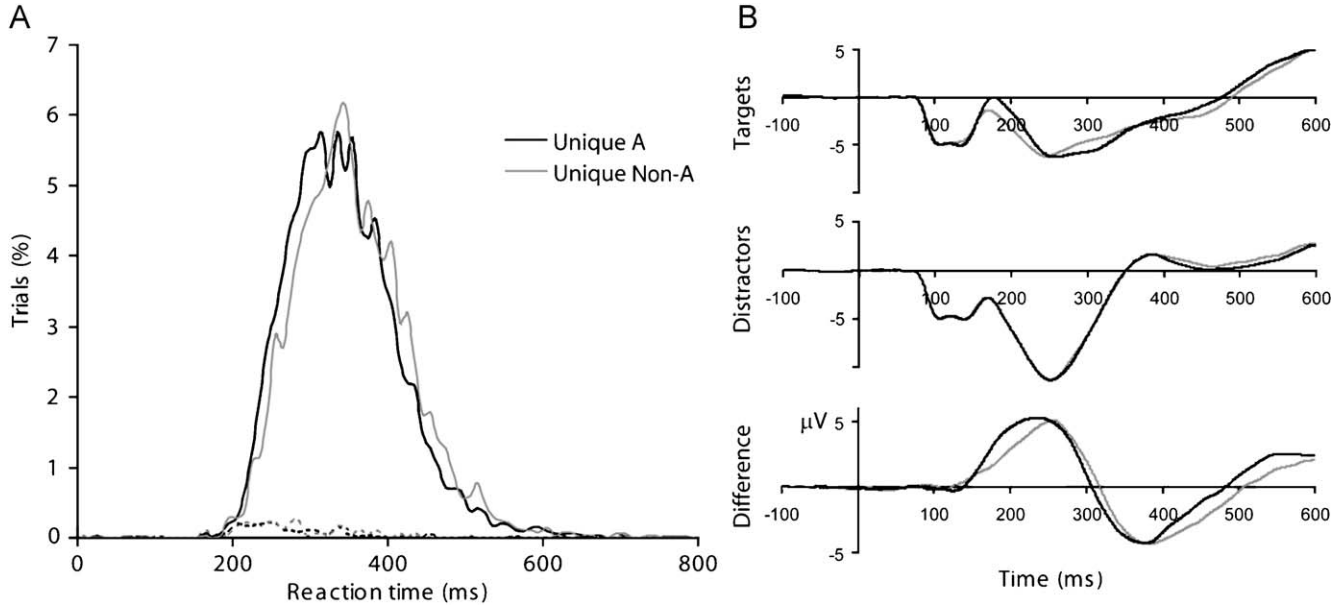


Fig. 3. Overall results from the 14 subjects on the two different target-photograph sets in the recognition task. (A) Histogram of reaction time for the condition where pictures containing animals had to be recognized (Unique A) and for the condition where the target pictures did not contain animals (Unique Non-A). (B) The differences between the frontal (FZ) ERPs recorded on correct target trials (upper curves) and on non-target trials (middle curves) are plotted (lower curves). In A and B: data are plotted in black for the animal set and in gray for the non-animal set.

3.1.4. Errors

A question that needs to be raised concerns the kind of errors that are produced in both tasks. In the categorization task, false alarms on distractors were slightly less common than target misses, and so far it has rarely been possible to objectively determine the reasons for these errors. In contrast, the errors produced in the recognition task were often seen with non-target images that share some obvious low-level properties with the memorized target image. These features (Fig. 1) appear to be related to coarse orientation of objects, prevailing color, patches of color(s) in a given location, context or object identity, spatial layout or complexity of the scene... When performing the recognition task, subjects were thus relying on low-level visual cue(s) that could differ from one memorized target to another.

3.2. Event-related potentials

ERPs were considered separately for correct target and correct distractor trials (Fig. 4). Using both individual data and grand average ERPs, the differential brain activity between the two types of trial was assessed in the two tasks by subtracting the average ERP on correct distractor trials from the average ERP on correct target trials. It is commonly assumed that the averaged electrical responses recorded from the scalp result from stimulus-evoked brain events and that the amplitude and latency of the various components of this evoked response reflect the most relevant features of the brain processing dynamics. Recently it has been shown [23] that these deflections might be generated by partial stimulus-induced phase resetting of multiple electroencephalographic processes. However, by using the difference between the two ERPs, no assumption is made about the relevance of the different ERP components, since the question that is addressed concerns only the differences in the cerebral processing of targets and distractors. The onset latency of this differential activity—which might correspond to the minimal visual processing time to differentiate a target from a distractor—was assessed using a two-tailed paired *t*-test performed for each 1 ms time bin and for each electrode (see Methods).

As reported in previous studies using this categorization task, a positive differential activity was clearly seen on frontal electrodes [8,36]. On occipital sites, a mirror differential activity of inverse polarity was observed [10]. The results are illustrated in Fig. 4 and show that ERPs to targets and non-targets superimposed very well until about 170 ms at which point they diverged abruptly (two-tailed paired *t*-test: $df = 13$, $p < 0.02$; occipital: 169 ms; frontal: 179 ms).

In the recognition task, the ERPs on correct target trials were computed separately for the two different sets of target-images (animal and control non-animal sets) and for their associated non-target images (Fig. 3B). The grand average ERPs computed on all the non-targets superimposed perfectly (Fig. 3B, middle traces) showing that there was no bias in the high variety of distractors used with the two different target sets. On the other hand, ERPs averaged separately on correct trials for the two target sets showed some differences (Fig. 3B, upper traces). The onset latency of the differential ERPs (Fig. 3B, lower traces) was found at 135 ms in the animal picture recognition task (two-tailed paired *t*-test: $df = 13$, $p < 0.02$; occipital: 135 ms; frontal: 148 ms), a latency virtually identical to the one found in the non-animal picture recognition task (two tailed paired *t*-test: $df = 13$, $p < 0.02$; occipital: 134 ms; frontal: 145 ms). Although the onsets were similar for these two sets of recognition targets, they diverged shortly after, the amplitude of the differential ERP increasing with a steeper slope with animal pictures targets. However, in the two sets of target-images, the computed differential activities reached similar amplitudes (on FZ electrode, animal pictures: 5.5 μV; non-animal pictures: 5.1 μV); but, the peak amplitude was observed earlier with animal images (233 ms) than with the set of non-animal images (255 ms). These differences at the ERP level might reflect the higher diagnosticity of animal images among non-animal images compared to the recognition of non-animal images among similar pictures.

Thus, in the picture recognition task, a clear differential activity was also observed at all sites but its onset was seen around 140 ms, much earlier than in the categorization task regardless of whether the images contained an animal or not. Consistent with this result, the difference between the two tasks also reached significance at about 140 ms (two-tailed paired *t*-test: $df = 13$, $p < 0.02$; occipital 141 ms; frontal 158 ms). Thus, the differential activity between target and non-target trials developed much earlier and reached a much higher amplitude in the recognition task than in the categorization task (5.3 vs. 2.9 μV for electrode Fz). Moreover, the peak of amplitude was observed at similar latencies in both tasks when pictures contained an animal (animal categorization: 234 ms, image recognition: 235 ms).

In both tasks the differential ERP between animal-target and non-target ERPs also showed an early small deflection that reached significance at about the same latency in the categorization task (two-tailed paired *t*-test: $df = 13$, $p < 0.02$; first occipital electrode: 98 ms; first frontal electrode 120 ms) and in the recognition task (two tailed paired *t*-test: $df = 13$, $p < 0.05$; occipital: 100ms; frontal: 112 ms). This small deflection does not appear with non-animal target images in the recognition task (Fig. 3B, lower traces) and might thus be linked to statistical differences in physical properties of different subsets of images as documented recently [40].

3.3. Source localization and activation dynamics

For both tasks we used an ellipsoidal source model in the software BESA to analyze the dipole source localization of the differential ERP waveforms and the time course of their activities (Fig. 5). Despite the strong constraints imposed on the model (large time window of 80 ms and only 2 dipoles that were required to be symmetrically positioned), residual

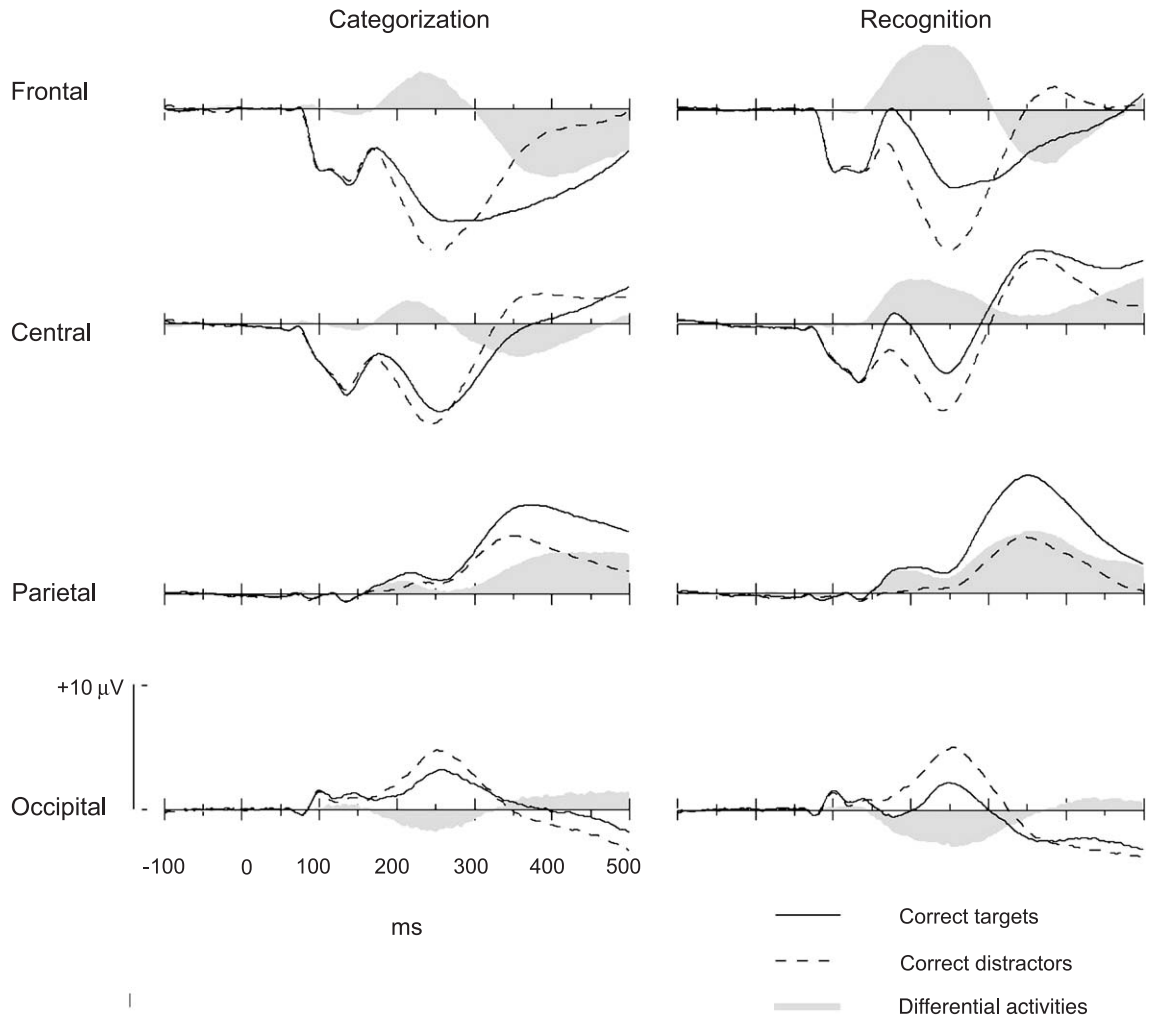


Fig. 4. Grand average differential ERP activity. Average ERPs for all subjects in the categorization task (left column) and in the recognition task (right column) at different scalp locations: frontal, central, parietal and occipital sites corresponding respectively to the midline electrodes Fz, Cz, Pz and Oz. Average ERP on correct target trials (black line), average ERP on correct distractor trials (dashed lines), differential activity between correct target and distractor trials (shaded area). Note that the latency of the differential activity is always shorter in the recognition task.

variance was kept under 4% for both tasks (residual variance: 3.9% in the categorization task and 2.2% in the recognition task), as already found in other studies using the categorization task [10]. Models using shorter and different time windows produced dipole localization that could not be distinguished from those illustrated in Fig. 5. Thus, most of the difference between ERPs to target and non-targets can be explained by a single bilaterally activated brain area located ventrally and laterally in the occipital lobe, in a region that probably corresponds to extra-striate visual cortex. The localization and orientation of the dipoles were similar for the two tasks, the most obvious difference between the observed scalp signals being the time-course of the differential activity that started earlier in the recognition task.

In the recognition task, the two sets of target-images were analyzed separately and were found to be associated with non-distinguishable dipoles that accounted in both cases for about 98% of the differential ERP waveforms. The only difference was seen in the temporal dynamics of activation of both pairs of dipoles that were associated with a stronger activity increase from 150 ms onwards with the set of animal targets, reaching earlier its maximal amplitude.

4. Discussion

The results of the present study show that the processing time of natural scenes by the human visual system depends on task instructions. When subjects are required to recognize a given target-image, they can rely on a variety of low-level cues, a hypothesis supported by the high similarity between the target and the non-target scenes that induced response errors. Consequently, the subjects were faster and more accurate in this natural scene recognition task than when they categorized the same type of natural images on the basis of the presence of an animal, a task that presumably requires access to more abstract representations. The

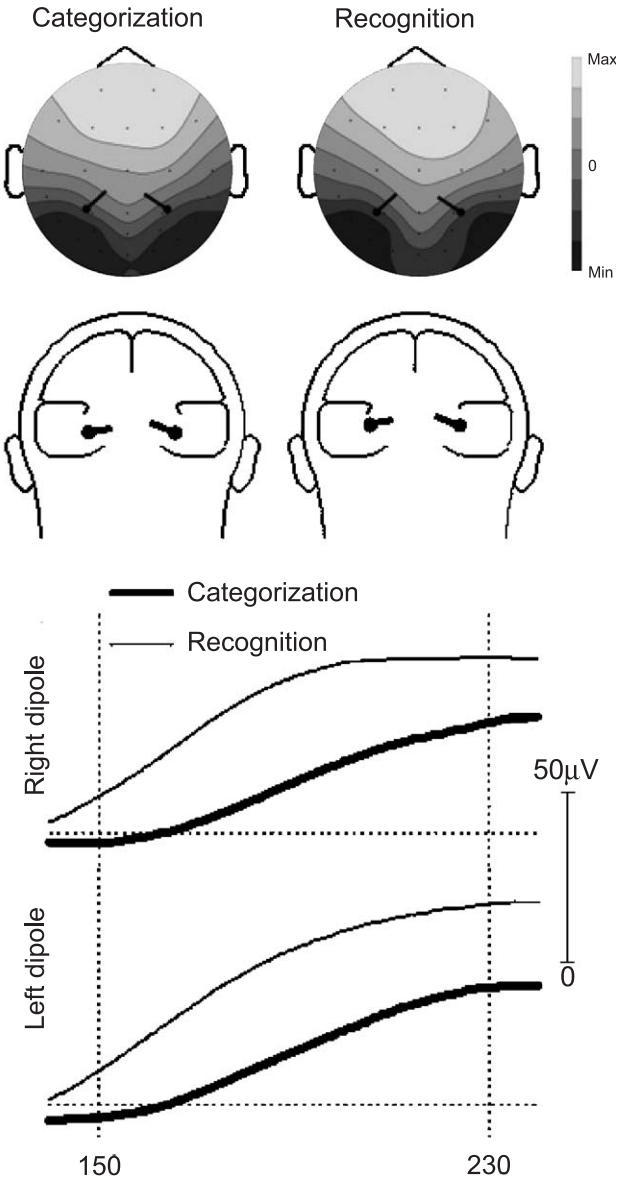


Fig. 5. Cartography of the differential activity between the ERP waveforms of target and non-target data trials and localization of the electrical sources that accounted for this difference. For both tasks, the categorization task and the recognition task, a bilateral source accounted for more than 96 % of the differential ERP waveforms. Top: Gray-level scalp maps illustrate the averaged differential potential at 230 ms. Superimposed on these maps, the localization of the sources was virtually the same in both tasks. The location of the dipoles is also shown on frontal views. Bottom: the temporal dynamics of the left and right electrical sources show that activation starts earlier and reaches a higher amplitude in the recognition task than in the categorization task.

results also provide some evidence that regardless of the visual analysis required in either task, the perceptual decision is made in the same brain structure and the visual information probably processed along the same visual pathway.

The visual processing required for recognizing a given target-image is done in a delay that is about 30–40 ms shorter than the visual analysis required for detecting an animal in the same image. This delay is observed for both the latency at which the earliest behavioral responses are produced and the onset latency of the differential cerebral activity (used as an index of the perceptual decision). It increases to 60 ms when considering the median reaction time, reflecting the fact that the variation in response latencies is larger in the animal categorization task than in the image-recognition task (Fig. 2) because of a larger difficulty range in the categorization task.

One could argue that the main difference between the two tasks is due to a novel vs. familiarity effect. Whereas the categorization task is exclusively performed with previously unseen images (trial unique presentations), the target-image recognition task involves the repetitive visual processing of a recently memorized photograph (i.e. ‘‘familiar’’) among non-target images that have never been seen before (i.e. ‘‘novel’’). Indeed, it has been shown using event-related fMRI, that the activity of brain areas that are thought to be involved in scene categorization (extrastriate visual cortex, inferotemporal cortex and prefrontal cortex) is modulated by stimulus repetition in subjects performing a rapid classification of pictures [4]. However, in the ‘‘animal’’ categorization task, we have recently shown that extensive experience with a given set of natural scenes did not result in faster behavioral responses than with completely novel images nor reduced the latency of the differential ERPs [8]. In agreement with other ERP studies using words, faces and other visual stimuli [12,22,31,39], familiarity effects were not seen until about 300–360 ms post-stimulus and thus could not account for speeding up the visual processing in the recognition task used here.

Various interpretations could account for our results. As target-image recognition task relies on detection of low-level cues, one possibility is that the faster analysis could simply result from the by-pass of higher processing stages that would only be necessary to reach a decision in the superordinate ‘‘animal’’ categorization task, when access to abstract representations is specifically required. In the recognition task, the perceptual decision could be made in brain structures considered as lower in the hierarchy of visual processing but in which low-level features would be already fully analyzed and accessible. Decisions could be made in area V4 or even in the primary visual cortex V1 as suggested by Barbur et al. [2]. Alternatively, we would like to argue that visual information is analyzed along the same brain pathway [16] but that the higher target predictability in the image-recognition task allows faster processing of the pertinent cues using top-down connections to preset neuronal assemblies at various levels of the visual pathway.

The main result supporting this alternative view is the location of the dipoles accounting for 96% and more of the differential activity recorded in both tasks. Even though the 32-recording-site set-up and the ability of the BESA software to specify accurately the ‘‘absolute’’ location of the brain activity may be questioned, the fact that, regardless of the task, the dipoles were found at very similar positions and

orientations in the brain appears difficult to explain if the underlying brain areas were not the same. In both tasks, the perceptual decision could therefore involve the same cerebral structures, most probably the occipito-temporal visual areas involved in object recognition. The location has been confirmed using the same categorization task with an event-related fMRI study [9], and found to be close to areas such as the fusiform gyrus involved in the recognition of various stimuli such as faces, objects or animals [5,14,20]. In correlation with the differential activity that develops 30–40 ms earlier in the target-image task, the main difference between our two tasks was found in the temporal dynamics and amplitude of the dipole activation (Fig. 5) that developed earlier and reached higher amplitude in the image-recognition task.

In preceding studies using the animal categorization task, we have already argued that the short latency at which the scalp differential activity starts to develop imposes such a high temporal constraint that the perceptual decision presumably relies essentially on feed-forward processing [8,35,36]. We postulated that information from the retina had to reach the primary visual cortex, area V1 (via the thalamus), and was subject to further processing in areas V2 and V4 before reaching the high-level brain areas involved in object recognition. These various processing steps are likely to be just as essential in the target-image recognition. Thus the most likely interpretation still relies on a faster visual processing of these images because of total target predictability.

In both tasks, speed of bottom-up processing would depend upon the tuning of neuronal populations along the visual pathways and thus on stimulus diagnosticity. Such bias has been shown for spatial frequencies [29], suggesting that a given scene might be flexibly encoded and perceived at the scale that optimizes information for the on-going task. Automatic target priming has been shown for color and spatial position in pop-out tasks [24,25] and has been attributed to temporary representations that could be updated on the basis of task demands. Saccade latency can be shortened by 30 ms and more, an effect linked to diagnosticity since it builds up with target color repetition [26]. In our tasks, we would expect top-down influences to bias bottom-up visual processing more heavily and more precisely in the recognition task than in the categorization task. The recognition of a target scene might be achieved using a carefully chosen low-level feature or a simple combination of characteristics (a blob of a given color or orientation for example). Compatibility would be maximal in this task because every target-image would activate all preset neuronal populations. Moreover, as the specific location of this feature in the image is also known, focalized spatial attention could be allocated at the exact location of the screen where the cue is going to appear when the target is flashed; a view that is supported by our analysis of the images that induced false alarms. In contrast, in the categorization task, the subject needs to process evenly the whole natural scene: the location of the target-animal in the photograph is unknown and although many features (an eye, a paw, a tail, a beak, a wing...) are diagnostic of the presence of an animal, none of them is necessary to classify an image as a target. Thus the presetting of the visual system cannot be as highly specific as in the recognition task and could not rely on the same features. Indeed, whereas color appears as an important diagnostic feature in the image-recognition task, we have shown that the fast responses in the "animal" categorization task do not rely on color cues [6]. A strong modulation of color processing could be due to top-down influences from high-level predictions about color-specific features [19].

Among the brain structures that might heavily influence the visual pathway through descending connections depending on behavioral requirements is the prefrontal cortex [3,27]. In a categorization task, the firing of prefrontal neurons reflects category membership rather than simple processing of the physical characteristics of the stimuli [11]. In the target-image recognition task, the activity in the frontal cortex is probably very similar to that recorded in a delayed matching to sample task with elevated activity during delay periods [13,15]. Moreover, prefrontal neurons can also convey information about both the physical characteristics of a stimulus and its location [30], a combination of cues used in the target-image recognition task. Thus, in the target-image recognition task, prefrontal activity could very precisely modulate the neuronal activity along the visual pathway [17] to optimize, for each memorized target, the processing of the selected pertinent cues.

Whereas total predictability speeds up visual processing, we showed using a control set of target images that presetting does not have the same strength for all natural scenes. Scenes with animals were, on average, recognized faster than scenes without animals. Certainly some features might be more salient in animal photographs presented among non-animal photographs, whereas the control set of non-animal images presented among other non-animal pictures could lack this diagnostic advantage. Another possible explanation may lie in the performance, in alternation, of the animal categorization task and the image recognition. Subject might have difficulty in inhibiting totally the presetting of neuronal populations tuned to animal features.

Another point that needs stressing is the fact that, in our preceding studies, the onset of the differential activity was found at about 150 ms for the categorization, whereas in the present study it was found about 20–30 ms later. Image size or presentation cannot account for this increased onset latency. On the other hand, this difference could be explained by the switching between two different tasks that required different presettings of the visual system as it has also been seen in another experimental protocol using two different interleaved tasks (manuscript in preparation). It might be that, had we used a blocked procedure in which subject would have completed all the testing series of one task before completing the second task we would have ended with even shorter differential activities.

Regardless of the task, we suggest that natural images are processed along the same visual circuit and that a perceptual decision is made in the same brain area but that the processing speed of bottom-up information is highly dependent upon the subject expectancy and the strength of top-down influences. However, we evaluated the temporal cost of the higher-level visual computations needed to perform the superordinate ''animal'' categorization task at about 30–40 ms. This temporal cost appears low when considering the discrepancy in task requirements. The answer might be in the level of complexity of the most informative features for classification. Fast super-ordinate categorization might rely on diagnostic features of intermediate complexity [37], accessible with coarse visual information rather than on fully integrated high-level object representations.

Acknowledgements

This work was supported by the CNRS and fellowships from the French government. Experimental procedures with human subjects were authorized by the local ethical committee (CCPPRB No. 9614003).

References

[1] L. Anllo-Vento, S.J. Luck, S.A. Hillyard, Spatio-temporal dynamics of attention to color: evidence from human electrophysiology, Hum. Brain Mapp. 6 (1998) 216–238.
[2] J.L. Barbur, J. Wolf, P. Lennie, Visual processing levels revealed by response latencies to changes in different visual attributes, Proc. R. Soc. Lond., B Biol. Sci. 265 (1998) 2321–2325.
[3] F. Barcelo, S. Suwazono, R.T. Knight, Prefrontal modulation of visual processing in humans, Nat. Neurosci. 3 (2000) 399–403.
[4] R.L. Buckner, J. Goodman, M. Burock, M. Rotte, W. Koutstaal, D. Schacter, B. Rosen, A.M. Dale, Functional–anatomic correlates of object priming in humans revealed by rapid presentation event-related fMRI, Neuron 20 (1998) 285–296.
[5] L.L. Chao, J.V. Haxby, A. Martin, Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects, Nat. Neurosci. 2 (1999) 913–919.
[6] A. Delorme, G. Richard, M. Fabre-Thorpe, Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans, Vision Res. 40 (2000) 2187–2200.
[7] J.T. Enns, R.A. Rensink, Influence of scene-based properties on visual search, Science 247 (1990) 721–723.
[8] M. Fabre-Thorpe, A. Delorme, C. Marlot, S.J. Thorpe, A limit to the speed of processing in Ultra-Rapid Visual Categorization of novel natural scenes, J. Cogn. Neurosci. 13 (2001) 171–180.
[9] D. Fize, K. Boulanouar, Y. Chatel, J.P. Ranjeva, M. Fabre-Thorpe, S.J. Thorpe, Brain areas involved in rapid categorization of natural images: an event-related fMRI study, NeuroImage 11 (2000) 634–643.
[10] D. Fize, M. Fabre-Thorpe, G. Richard, B. Doyon, Thorpe, S.J., Rapid categorisation of foveal and extrafoveal natural images: Associated ERPs and effect of lateralisation (Submitted for publication).
[11] D.J. Freedman, M. Riesenhuber, T. Poggio, E.K. Miller, Categorical representation of visual stimuli in the primate prefrontal cortex, Science 291 (2001) 312–316.
[12] D. Friedman, Cognitive event-related potential components during continuous recognition memory for pictures, Psychophysiology 27 (1990) 136–148.
[13] J.M. Fuster, G.E. Alexander, Neuron activity related to short-term memory, Science 173 (1971) 652–654.
[14] I. Gauthier, P. Skudlarski, J.C. Gore, A.W. Anderson, Expertise for cars and birds recruits brain areas involved in face recognition, Nat. Neurosci. 3 (2000) 191–197.
[15] P.S. Goldman-Rakic, Cellular basis of working memory, Neuron 14 (1995) 477–485.
[16] K. Grill-Spector, N. Kanwisher, Different recognition tasks activate a common set of object processing areas in the human brain, Abstr. Soc. Neurosci. (2000) 686.6.
[17] I. Hasegawa, Y. Miyashita, Categorizing the world: expert neurons look into key features, Nat. Neurosci. 5 (2002) 90–91.
[18] S.A. Hillyard, L. Anllo-Vento, Event-related brain potentials in the study of visual selective attention, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 781–787.
[19] J.M. Hopf, E. Vogel, G. Woodman, H.J. Heinze, S.J. Luck, Localizing visual discrimination processes in time and space, J. Neurophysiol. 88 (2002) 2088–2095.
[20] N. Kanwisher, J. McDermott, M.M. Chun, The fusiforme face area: a module in human extrastriate cortex specialized for face perception, J. Neurosci. 17 (1997) 4302–4311.
[21] F. Karayanidis, P.T. Michie, Evidence of visual processing negativity with attention to orientation and color in central space, Electroencephalogr. Clin. Neurophysiol. 103 (1997) 282–297.
[22] T. Liu, L.A. Cooper, The influence of task requirements on priming in object decision and matching, Mem. Cogn. 29 (2001) 874–882.
[23] S. Makeig, M. Westerfield, T.P. Jung, S. Enghoff, J. Townsend, E. Sejnowski, T.J. Sejnowski, Dynamic brain sources of visual evoked responses, Science 295 (2002) 690–694.
[24] V. Maljkovic, K. Nakayama, Priming of pop-out: I. Role of features, Mem. Cogn. 22 (1994) 657–672.
[25] V. Maljkovic, K. Nakayama, Priming of pop-out: II. The role of position, Percept. Psychophys. 58 (1996) 977–991.
[26] R.M. McPeek, V. Maljkovic, K. Nakayama, Saccades require focal attention and are facilitated by a short-term memory system, Vision Res. 39 (1999) 1555–1566.
[27] E.K. Miller, J.D. Cohen, An integrative theory of prefrontal cortex function, Annu. Rev. Neurosci. 24 (2001) 167–202.
[28] K. Nakayama, G.H. Silverman, Serial and parallel processing of visual feature conjunctions, Nature 320 (1986) 264–265.
[29] A. Oliva, P.G. Schyns, Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli, Cogn. Psychol. 34 (1997) 72–107.
[30] G. Rainer, W.F. Asaad, E.K. Miller, Memory fields of neurons in the primate prefrontal cortex, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 15008–15013.
[31] M.D. Rugg, M. Soardi, M.C. Doyle, Modulation of event-related potentials by the repetition of drawings of novel objects, Brain. Res. Cogn. Brain. Res. 3 (1995) 17–24.
[32] H.E. Schendan, G. Ganis, M. Kutas, Neurophysiological evidence for visual perceptual categorization of words and faces within 150 ms, Psychophysiology 35 (1998) 240–251.
[33] P.G. Schyns, Diagnostic recognition: task constraints, object information, and their interactions, in: M.J. Tarr, H.H. Bülthoff (Eds.), Object Recognition in Man, Monkey, and Machine, Elsevier, Amsterdam, 1998, pp. 147–179.
[34] Y. Sugita, Electrophysiological correlates of visual search asymmetry in humans, NeuroReport 6 (1995) 1693–1696.
[35] S.J. Thorpe, M. Fabre-Thorpe, Seeking categories in the brain, Science 291 (2001) 260–263.
[36] S.J. Thorpe, D. Fize, C. Marlot, Speed of processing in the human visual system, Nature 381 (1996) 520–522.
[37] S. Ullman, M. Vidal-Naquet, E. Sali, Visual features of intermediate complexity and their use in classification, Nat. Neurosci. 5 (2002) 682–687.

[38] M. Valdes-Sosa, M.A. Bobes, V. Rodriguez, T. Pinilla, Switching attention without shifting the spotlight: object-based attentional modulation of brain potentials, J. Cogn. Neurosci. 10 (1998) 137–151.
[39] C. Van Petten, A.J. Senkfor, Memory for words and novel visual patterns: repetition, recognition, and encoding effects in the event-related brain potential, Psychophysiology 33 (1996) 491–506.
[40] R. VanRullen, S.J. Thorpe, The time course of visual processing: from early perception to decision-making, J. Cogn. Neurosci. 13 (2001) 454–461.

Article n°3

Neuroreport, **16**, 349-354

Rapid categorization of natural scenes in monkeys: target predictability and processing speed

Marc J-M. Macé, Ghislaine Richard, Arnaud Delorme

& Michèle Fabre-Thorpe

Rapid categorization of natural scenes in monkeys: target predictability and processing speed

Marc J.-M. Macé,[CA] Ghislaine Richard, Arnaud Delorme[I] and Michèle Fabre-Thorpe

Centre de recherche Cerveau & Cognition, CNRS-UPS – UMR 5549, Faculté de Médecine de Rangueil, 133 route de Narbonne, 31062 Toulouse Cedex, France

[CA]Corresponding Author: marc.mace@cerco.ups-tlse.fr

[I]Present address: Computational Neurobiology Laboratory, Salk Institute, 10010 N. Torrey Pines Road, San Diego, CA 92037, USA

Received 9 December 2004; accepted 14 January 2005

Three monkeys performed a categorization task and a recognition task with briefly flashed natural images, using in alternation either a large variety of familiar target images (animal or food) or a single (totally predictable) target. The processing time was 20 ms shorter in the recognition task in which false alarms showed that monkeys relied on low-level cues (color, form, orientation, etc.). The 20-ms additional delay necessary in monkeys to perform the categorization task is compared with the 40-ms delay previously found for humans performing similar tasks. With such short additional processing time, it is argued that neither monkeys nor humans have time to develop a fully integrated object representation in the categorization task and must rely on coarse intermediate representations. *NeuroReport* 16:349–354 © 2005 Lippincott Williams & Wilkins.

Key words: Categorization; Early visual processing; Low-level cues; Macaques; Natural scenes; Recognition; Target predictability

INTRODUCTION

Although less documented than for pigeons or for avians in general, the ability of monkeys to categorize complex visual photographs has now been demonstrated for a variety of categories from subordinate to superordinate levels such as kingfishers, birds, fish, trees, primates, animals, food, objects, etc. [1–5]. Baboons have been shown to develop multimodal abstract concepts of human and baboon categories [6] and can make judgments of conceptual identity [7]. When performing categorization tasks with very severe temporal constraints, macaque monkeys are able to produce their motor response with very short reaction times (RTs). Their earliest correct responses are observed at a latency of 180 ms [5,8], a delay shown to challenge many models of object processing [9,10]. Performing very similar tasks, human participants, although very fast, are much slower than monkeys, with their earliest behavioral responses observed at about 280 ms after stimulus onset [11].

With such differences in minimal input–output processing time (180 ms in monkeys vs. 270 ms in humans), one should wonder how similar is the neural processing underlying visual categorization in humans and monkeys. These very short response latencies observed in monkeys might result from the processing of low-level cues rather than the use of abstract representations. Indeed, Torralba and Oliva [12] have shown that, in humans, the statistics of low-level features across natural images can be used to prime the presence or absence of objects in the scene and to predict their location before exploring the image. However, the initial use of low-level cues might be as important for monkeys as for humans. Alternative– and nonexclusive–explanations could also account for the rapidity of monkeys in these tasks. First, one cannot exclude a speed accuracy trade-off because monkeys are slightly less accurate than humans (about 90% vs. 94% correct). Second, it could simply result from shorter conduction delays because of macaques smaller brain dimensions.

This study had mainly two aims. First, we wanted to compare, in monkeys, the visual processing of natural images in two tasks: one during which the monkey might need to rely on the abstract representation of a high-level object category such as 'food' or 'animal', and another where the target was totally predictable, so that the monkey could respond using a limited number of specific low-level cues stored in short-term memory. Faced with similar tasks, human participants are faster when target predictability is total [11]. Because, in humans and monkeys, both tasks used natural photographs as stimuli and required the same motor response, any differences in the latencies of the motor responses should reflect central processing differences related to task demands. Thus, a second aim of the study was to compare how the different requirements of the two tasks would affect monkey and human performance.

METHODS

Participants: Three rhesus monkeys were trained to perform a rapid go/no-go visual superordinate categorization task with food objects (Rh1, male aged 7) or animals (Rh2 and Rh3, male and female aged 6 and 5) as targets. These monkeys have already been tested in different experiments, which showed their ability to categorize

familiar photographs and generalize to new photographs [5] and revealed that color did not play a crucial role in rapid categorization [8]. All procedures conformed to French and European standards concerning the use of experimental animals and the protocols were approved by the regional ethical committee.

Tasks and protocol: In the rapid categorization task, monkeys were presented with a random succession of different natural scenes, half of which were targets. The monkeys started stimulus presentation by placing one hand on a capacitive tactile key. When a target image was flashed, they had to quickly release the button and touch the screen (go response); otherwise they had to keep their hand on the button (no-go response). They were given a maximum of 1000 ms to respond, after which any response was considered as a no-go response.

In the second task, monkeys were still performing the same go/no-go task, except that only one single (food or animal) target was used and presented among varied nontarget images. As this single target was totally predictable, the task performed by the monkey became a 'recognition task'. Targets and nontargets were still equiprobable as in the categorization task.

To perform the tasks, monkeys were seated about 30–35 cm away from a tactile screen. A small fixation point appeared in the center of the screen and pictures were flashed around the fixation point on a black background for only 28 ms: a duration that prevented any exploratory eye movements. The tactile key used to start the sequence of images and to record the motor response was located below the screen at waist level. Two successive images were separated by a random 1.5–3 s intertrial period. Correct (go or no-go) responses were rewarded by a drop of fruit juice and a beep noise. Incorrect decisions were followed by a 3–4 s display of the incorrectly classified stimuli delaying the next trial and allowing time for ocular exploration. The monkeys worked daily for as long as they wanted (1–3 h), 5 days a week. At the end of each testing session and during weekends, *ad libitum* water was provided. They were restrained in a primate chair during testing (Crist Instruments, Georgia, USA).

The results presented here were recorded during 20 successive testing sessions. In a given testing session, monkeys performed the categorization task in alternation with the recognition task by blocks of 150 trials. Before the start of the testing session, monkeys performed the categorization task until they were calm and up to their usual level of performance. The testing session started first with a block of 150 trials of the categorization task using 150 different stimuli. Warning that the categorization task was going to become a recognition task was then given through a sequence of 10 trials presenting repetitively the single-target image that was going to be presented 75 times randomly among 75 different nontarget images in the subsequent recognition block. No warning was given in between the recognition block and the next categorization block.

For each of the 20 sessions, the 150 stimuli of the categorization task together with the 75 nontarget stimuli and the single-target stimulus of the recognition task were chosen at random from the pool of familiar images that the monkey had already categorized many times (the two monkeys working on the animal/nonanimal categorization were tested on the same single-target stimuli).

Thus, the animals alternated between a task in which training had optimized stimulus processing and a task in which target predictability was total. In each session, the monkey alternated categorization task blocks with recognition task blocks, until they stopped working on the task. A minimum of two blocks in each task was required for a session. Thus, a session was usually run on 1 day, exceptionally on 2 successive days.

Stimuli: All stimuli (examples in Fig. 1) used in the tasks were color photographs of natural scenes (Corel CD-ROM library). Targets and distractors included both closeups and general views. Food targets included photographs of fruit, vegetables, salads, cakes, biscuits, sweets, etc. presented against natural backgrounds. Animal targets included fish, birds, mammals and reptiles also presented in their natural environments. Distractors included landscapes, trees, flowers, objects, monuments, cars and some targets of the other categorization task.

Images (192×128 pixels, corresponding to an angular size of about $25^\circ \times 15^\circ$) were mostly horizontal photographs (73%). They were flashed for two frames at a refresh rate of 60 Hz (noninterlaced), using a programmable graphics board (VSG 2, Cambridge Research Systems, Rochester, Kent, UK) mounted in a PC-compatible computer.

RESULTS

We will first present the results obtained on the monkeys in the two tasks, and then briefly present the results obtained in a group of human participants to compare the effect of target predictability on processing speed in both species (the detailed human subject data have been published in Delorme *et al.* [11]).

Behavioral results: categorization versus recognition tasks: The analysis of the monkey behavioral performance included accuracy, speed of response and a study of the nontarget images that incorrectly induced a go response in the recognition task. Results concerning the food-target task will be given for three daily blocks of categorization and recognition tasks: a total of 9000 trials in each task for monkey Rh1. Results concerning the animal-target task will be given for two daily blocks of categorization and recognition tasks: a total of 6000 trials in each task and for each of the two animals.

Accuracy: Although very high in both tasks (92.4% correct in the categorization task; 96.3% correct in the recognition task), accuracy was significantly better in the recognition task (two-tailed χ^2, df=1, $p<0.0001$). This effect was present regardless of the target to find (food or animal), and was significant at $p<0.0001$ for each individual animal.

With animal targets, both monkeys were better at responding to target images than at ignoring distractors. This bias in favor of correct go responses was quite pronounced (about 4–5%) and was observed with the same strength in both tasks and for both monkeys (two-tailed χ^2, df=1, p always <0.0001).

The third monkey working with food targets did not show the same biased pattern. A small but significant bias

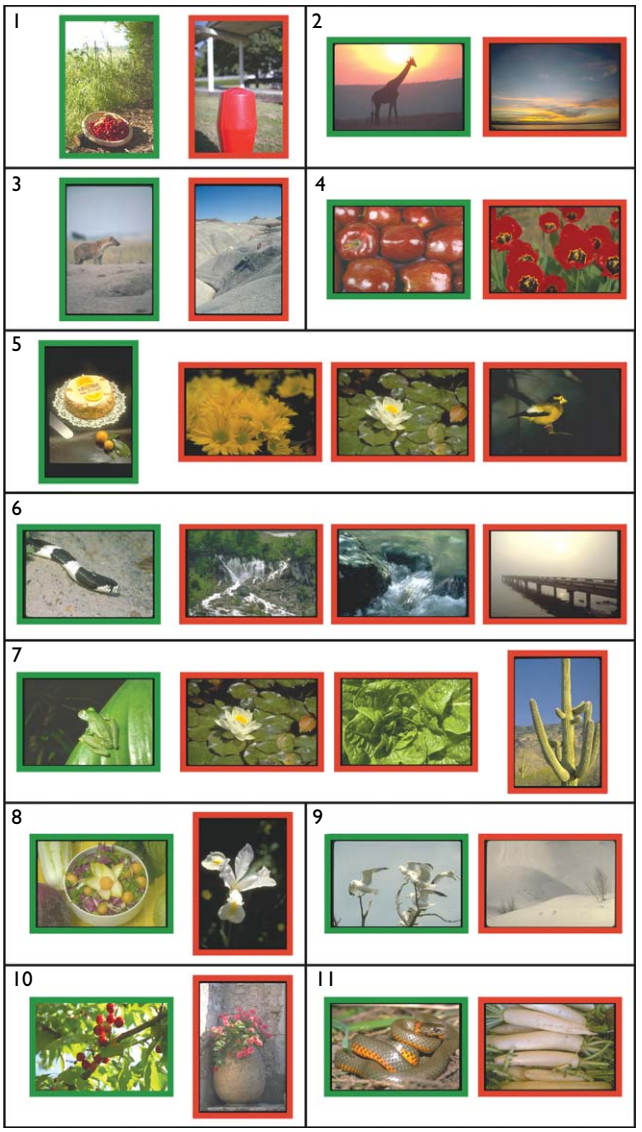


Fig. 1. Targets (with a green frame) and associated errors (with a red frame) in the recognition task. The figure shows the high variety of stimuli used in the 20 testing blocks in each of the tasks. On the right of each target image, 1–3 nontarget photograph(s) that induced a false alarm are shown. Some errors can clearly be related to prevailing color (1–7), global form (4, 6, 11), global orientation (3, 6, 11), color patches or specific form in specific locations (4–6, 8, 10), spatial layout of the scene (3, 6–11) or any combination. Similar natural images were used in the categorization task. Note that, with target no. 3, both monkeys made the same error.

(1.3%) was observed in the recognition task in favor of correct go responses whereas a stronger inverted bias (<3%) was seen toward correct no-go responses in the categorization task.

Reaction time: Monkeys were extremely fast in both tasks, but as illustrated in Fig. 2, RTs were always faster for the recognition task than for the categorization task. Overall, for the three monkeys, the processing time was shorter by 19 ms (mean RT: 244 vs. 263 ms). It was shorter by 17 ms for Rh1 working with food objects (mean RT 281 vs. 298 ms) and by 21 ms (14 ms for Rh2 and 28 ms for Rh3) with animal targets (mean RT 225 vs. 246 ms). For each individual animal, these differences were always significant (two-tailed *t*-test, $p<0.0001$).

Monkeys alternated between the categorization and the recognition tasks by blocks of 150 trials. Performance speed and accuracy were analyzed separately for each section of 50 trials to determine the stability of the performance throughout the progress of the 150 trials block. The performance was very stable, showing that the monkey modified its behavioral strategy as soon as the task changed.

d′ analysis: The average slower speed in the categorization task could result from the presence of some difficult photographs that need a longer time to process. Thus, although the average processing time could be shorter in the recognition task, the latencies of the earliest responses might be similar in both tasks. This is clearly not the case, and this shortening of RT latencies concerns the whole range of motor responses from the very first responses triggered. RT distribution in the recognition task can be seen (Fig. 2) as a shift of the entire RT distribution obtained on the categorization task toward shorter latencies. The dynamic d' calculated for both tasks and for each monkey (Fig. 2) illustrates clearly that the earliest responses triggered have shorter latencies in the recognition task than in the categorization task.

The minimal RT, evaluated as the first time bin from which correct hits significantly outnumbered false alarms, reflects the minimal processing time necessary to reach a decision in each task. The difference in minimal RT between the two tasks was 20 ms on average and was significant for each animal. It is clearly visible on the d' curves presented for each monkey in Fig. 2. Thus, the increase of latencies seen on the mean processing time is also present with the same strength on the very first behavioral responses observed.

Errors: A question that needs to be raised concerns the kind of errors that are produced in both tasks. So far, on all data collected when monkeys were performing the categorization task in this study and others, it has not been possible to determine the reasons for these false alarms. However, in the recognition task, the nontarget images that induced errors often shared some obvious low-level properties with the memorized target image. These features (see Fig. 1) appear to be often related to the prevailing color, the global form of objects or their coarse orientation, and sometimes to the spatial layout or complexity of the whole scene, etc. When performing the recognition task, monkeys could thus rely on low-level memorized visual cue(s). Because the distractors in the recognition task were chosen at random in the pool of familiar images already seen by the monkeys, it was rare that monkeys Rh2 and Rh3 had to deal with the same distractor image when looking for the same single-target image. Thus, it is worth noting that with the target image no. 3 (Fig. 1) both monkeys made the same error, which can be seen as induced by the prevailing color and/or the global layout of the scene.

Behavioral results: monkeys versus humans: The neural mechanisms underlying visual categorization are still poorly understood and their similarity in humans and animals is extremely controversial. Therefore, it is interesting to compare

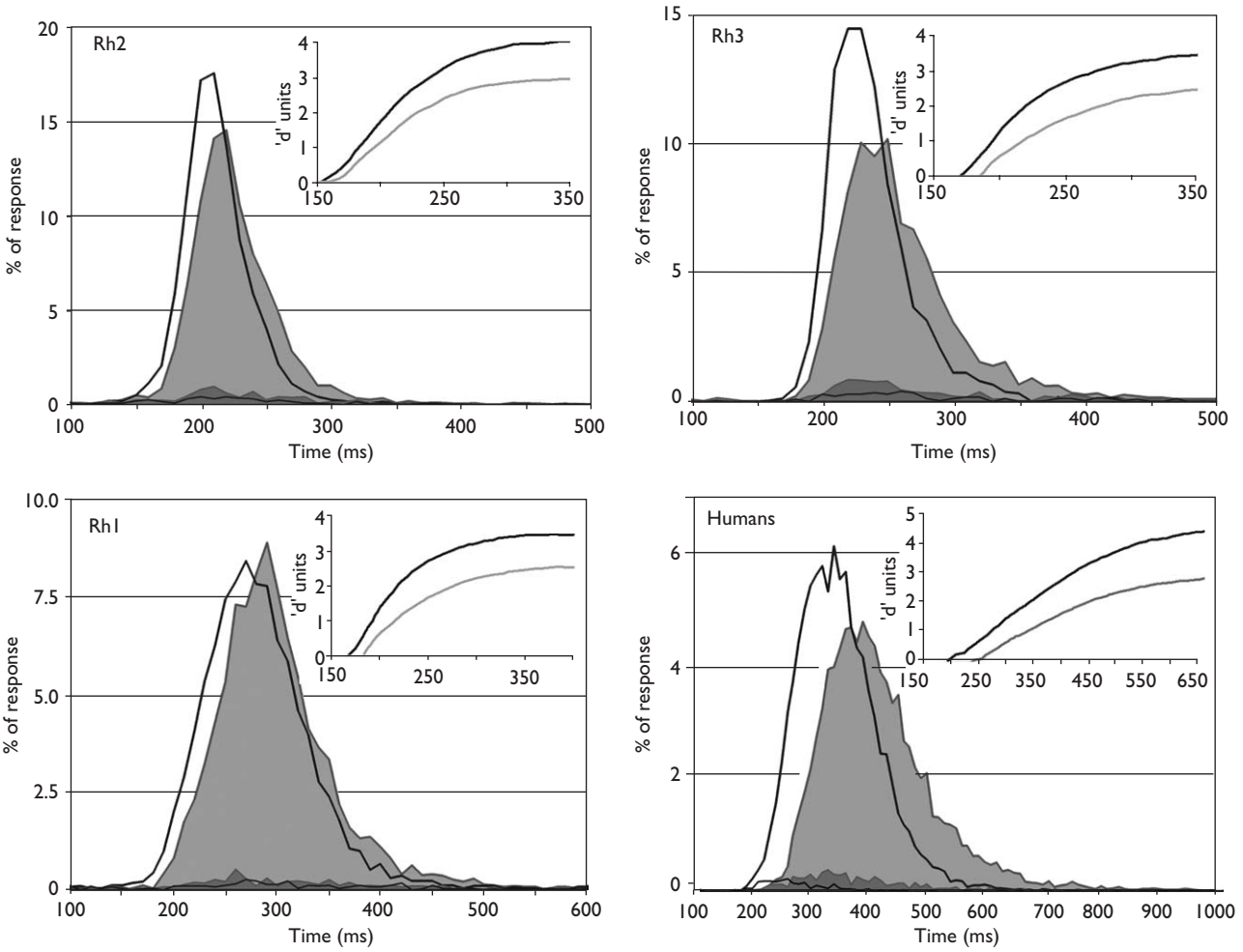


Fig. 2. Overall reaction time (RT) distributions of go responses in both the categorization task (gray traces and shaded distribution) and the recognition task (black lines) for each of the three monkeys (Rh1, Rh2 and Rh3) and for the group of 14 human participants. The two thick traces are for correct go responses toward targets, and the two thin traces are for false alarms induced by nontarget stimuli. Rh2, Rh3 and humans were tested with an animal task, and Rh1 with a food object task. In the top right-hand corner, a dynamic d' (see details in [22]) is calculated for each task from the cumulative number of hits and false alarms at each successive 10-ms time bin (gray trace: categorization task; black trace: recognition task). Note that although the scales of the axes are not identical in all illustrations, RT distributions and corresponding d' share the same time scale.

monkey and human performance in these two tasks to determine whether or not increasing target predictability has the same behavioral consequences in both species. A group of 14 humans had been studied on the same two tasks using animal targets as monkeys Rh2 and Rh3 of the present study (see Delorme *et al.* [11] for a detailed report of their performance and associated cerebral activity).

As with monkeys, human accuracy was higher in the recognition task (98.7%) than in the categorization task (93.1%). Humans were slightly better at ignoring distractors in the categorization task, whereas they were better at responding to targets in the recognition task. Thus, humans and monkeys showed the same bias in the recognition task, but the discrepancy in the categorization task merits further investigation in humans. Reanalysis of the human data in the categorization task showed that individual biases in humans were tightly correlated with response speed. Humans favoring speed over accuracy would display the same bias as Rh2 and Rh3, the two fastest monkeys. Conversely, human participants favoring accuracy over speed would display a bias similar to that of monkey Rh1, the slowest monkey. Thus, this performance bias is similar in humans and monkeys and is highly dependent on the individual speed–accuracy trade-off of the participant.

Concerning response speed, the mean RT was shortened by 63 ms in humans in the recognition task relatively to the categorization task, a larger difference than that found in monkeys. But as in monkeys, the entire RT distribution was also shifted toward shorter latencies. As illustrated in Fig. 2, the fastest responses were observed at about 220 ms in the recognition task, corresponding to a 40 ms speed advantage for this task (20 ms in monkeys). The analysis of the stimuli that induced errors in the recognition task also showed, as reported for monkeys in the present study, that humans were most likely relying on low-level visual characteristics of the memorized target.

DISCUSSION

One of the aims of the present study was to determine the time necessary for monkeys to process a natural image on

the basis of low-level cues and to evaluate how much more processing they would need to perform a categorization task in which they presumably rely on more abstract representations. Indeed, the results obtained show that the time needed by monkeys to process natural images depends on the task performed. First, in the recognition task in which monkeys had to recognize a given target image, the visual similarity between the target and the nontarget scenes that induced an erroneous go response strongly suggests that – as intended – monkeys relied on low-level cues that varied from target to target. Then, the comparison of monkey performance in the recognition task relative to the categorization task showed that monkeys were both more accurate (by about 4%) and faster when they responded to a single-target image. Testing humans with similar tasks had similar effects on performance, but with larger amplitude: their accuracy was increased by 5.6% and their processing time was decreased by 40 ms when considering the earliest responses produced and by about 60 ms for mean RT. The processing required for deciding whether an animal is present in a natural scene takes at least an additional delay of 20 ms in monkeys and 40–60 ms in humans.

Why is additional processing time longer in humans than in monkeys? A first interpretation lies in the stimuli used to run the categorization task. Whereas monkeys were tested with familiar images that they had already categorized many times, humans were tested with stimuli that they all saw for the first time [11]. However, this discrepancy can only explain the increased difference observed when comparing mean RTs (60 ms). When humans are tested with both new and familiar images, they produce their earliest responses at exactly the same latencies. The only effect of familiarity is to shorten the RT of long latency responses with a resulting effect of decreasing the mean RT by about 20 ms [13]. Thus, the additional processing delays that should be compared between the two species are the delays seen on the earliest responses: 20 ms in monkeys versus 40 ms in humans. A straightforward cognitive interpretation concerns the type of representations used by monkeys in the categorization task that might be less abstract and more figurative than in humans. But an alternative interpretation is to consider that the 20-ms delay in monkeys is simply the homolog of the 40-ms delay in humans. In fact, monkeys always produced their motor response faster than humans. This has mainly been reported for ocular movements. Express saccades, for example, are seen at 70 ms in monkeys and 100 ms in humans [14]; vergence reflex or tracking systems are observed at latencies of 55–60 ms in monkeys and 80 ms in humans [15–17]. Visuo-motor responses would thus be produced by monkeys at latencies that are about two-thirds of the human latencies. This is true also in the categorization task performed with familiar images that is used in the present study: monkeys have a mean RT of about 263 ms whereas human mean RT was observed at 424 ms [13]. Given that intracortical connections have been shown to be very slow [18,19], the differences between monkey and human latencies could be because of differences in brain sizes and reflect that, in monkeys, less time is lost in transferring information within a given cortical area or along the different cortical areas [5,9].

Thus, during the additional processing time required in between the recognition task (use of low-level cues) and the categorization task (abstract representation?), visual computations made by monkeys and humans might be very similar. This 20–40 ms temporal cost appears very limited when considering the discrepancy in task complexity. On the one hand, this additional delay argues strongly that, in the categorization task, monkeys and humans have to process visual information further than the simple low-level statistical differences in between target and distractor image sets shown by Torralba and Oliva [12]. On the other hand, this additional processing time is so short that neither monkeys nor humans would have time to rely on fully integrated high-level object representations. Thus, when responding in the categorization task, fast responses might rely on very coarse intermediate object representations. This is in agreement with experimental series showing that humans can categorize natural scenes at extreme eccentricities [20] and that natural scenes categorization in humans and monkeys is very robust even when using achromatic stimuli at very low contrast [21] [Macé MJ-M *et al.* in preparation. Rapid categorisation of achromatic natural scenes: how robust at very low contrasts? *Eur J Neurosci* (in preparation)]. It might be that categorization tasks in which a detailed object representation is necessary would induce a more spectacular increase in RT both in humans and in monkeys. Further experiments are needed to evaluate such an interpretation.

REFERENCES

1. Roberts WA, Mazmanian DS. Concept learning at different levels of abstraction by pigeons, monkeys, and people. *J Exp Psychol Anim Behav Process* 1988; **14**:247–260.
2. D'Amato MR, Van Sant P. The person concept in monkeys (Cebus apella). *J Exp Psychol Anim Behav Proc* 1988; **14**:43–55.
3. Schrier AM. Learning-set formation by three species of macaque monkeys. *J Comp Physiol Psychol* 1966; **61**:490–492.
4. Vogels R. Categorization of complex visual images by rhesus monkeys. Part 1: behavioural study. *Eur J Neurosci* 1999; **11**:1223–1238.
5. Fabre-Thorpe M, Richard G, Thorpe SJ. Rapid categorization of natural images by rhesus monkeys. *Neuroreport* 1998; **9**:303–308.
6. Martin-Malivel J, Fagot J. Cross-modal integration and conceptual categorization in baboons. *Behav Brain Res* 2001; **122**:209–213.
7. Bovet D, Vauclair J. Judgment of conceptual identity in monkeys. *Psychon Bull Rev* 2001; **8**:470–475.
8. Delorme A, Richard G, Fabre-Thorpe M. Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. *Vis Res* 2000; **40**:2187–2200.
9. Thorpe SJ, Fabre-Thorpe M. Neuroscience. Seeking categories in the brain. *Science* 2001; **291**:260–263.
10. VanRullen R, Thorpe SJ. Surfing a spike wave down the ventral stream. *Vis Res* 2002; **42**:2593–2615.
11. Delorme A, Rousselet GA, Macé MJ-M, Fabre-Thorpe M. Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Brain Res Cogn Brain Res* 2004; **19**:103–113.
12. Torralba A, Oliva A. Statistics of natural image categories. *Network* 2003; **14**:391–412.
13. Fabre-Thorpe M, Delorme A, Marlot C, Thorpe S. A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *J Cogn Neurosci* 2001; **13**:171–180.
14. Fischer B, Weber H. Express saccades and visual attention. *Behav Brain Sci* 1993; **16**:553–567.
15. Busettini C, Fitzgibbon EJ, Miles FA. Short-latency disparity vergence in humans. *J Neurophysiol* 2001; **85**:1129–1152.
16. Busettini C, Miles FA, Krauzlis RJ. Short-latency disparity vergence responses and their dependence on a prior saccadic eye movement. *J Neurophysiol* 1996; **75**:1392–1410.
17. Miles FA. The neural processing of 3-D visual information: evidence from eye movements. *Eur J Neurosci* 1998; **10**:811–822.

18. Nowak LG, Bullier J. In: Rockland KS, Kaas JH, Peters A (eds). *Extrastriate Visual Cortex in Primates*. New York: Plenum Press; 1997. pp. 205–241.
19. Frégnac Y, Bringuier V. In: Braitenberg AAV (ed.). *Brain Theory – Biological Basis and Computational Principles*. Amsterdam: Elsevier; 1996. pp. 143–199.
20. Thorpe SJ, Gegenfurtner KR, Fabre-Thorpe M, Bulthoff HH. Detection of animals in natural images using far peripheral vision. *Eur J Neurosci* 2001; **14**:869–876.
21. Macé MJ-M, Thorpe SJ, Fabre-Thorpe M. Rapid categorization of achromatic natural scenes: how robust at very low contrasts? *Eur J Neurosci*, in press.
22. Rousselet GA, Macé MJ-M, Fabre-Thorpe M. Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *J Vis* 2003; **3**:440–455.

Acknowledgments: This work was supported by the CNRS, the University Toulouse III, the Integrative and Computational Neuroscience ACI and the Cognitique program of the French government. M.M. and A.D. were supported by an MRT grant from the French government.

2.2.4 - En simplifiant la tâche à l'extrême

En simplifiant la tâche pour qu'elle ne soit qu'une reconnaissance d'une image précédemment encodée, nous avons vu que le TR minimal pouvait être diminué de 40 ms et la latence de l'activité différentielle de 35 ms. Mais pour déterminer la dynamique des traitements visuels, il nous faut recourir à une autre tâche. La tâche la plus simple que l'on puisse imaginer en conservant un protocole proche de celui que nous avons utilisé jusqu'à maintenant consiste à relever le doigt d'un bouton chaque fois qu'une scène naturelle est flashée à l'écran. Le sujet doit simplement détecter l'apparition de ce signal lumineux sans effectuer de traitement d'information.

La seule différence avec les expériences de psychophysique menées par le passé (Pieron, 1914 ; Galifret & Pieron, 1949) pour mesurer le temps de réaction minimal de l'homme à un stimulus visuel est que le sujet ne doit pas détecter l'apparition d'un stimulus simple (par exemple un point blanc sur fond noir) mais celle d'une image naturelle, afin de pouvoir comparer les résultats de cette expérience avec ceux des autres expériences utilisant des images naturelles. La difficulté la plus importante dans ce type d'expériences est que comme tous les essais sont des essais cibles et que la consigne est de répondre le plus rapidement possible, de nombreuses anticipations peuvent être commises. Pour diminuer ce biais, on peut augmenter de manière significative la durée aléatoire entre l'affichage de la croix de fixation et l'apparition de l'image. Cette expérience permet de mesurer le temps nécessaire à un sujet pour réagir à l'apparition d'une image naturelle sans avoir à effectuer le moindre traitement du signal. Il n'est pas possible de calculer précisément le TR minimal puisqu'il n'y a pas de fausses détections sur des distracteurs pour déterminer le moment à partir duquel les réponses correctes et incorrectes diffèrent. Cependant, la durée importante de l'intervalle de temps aléatoire avant l'image (2200 ± 800 ms) et la consigne d'éviter de commettre des anticipations font que les réponses avant 100 ms ne représentaient que 2% des essais (ils ont été retirés de l'analyse). 6 sujets ont participé à cette expérience et ont effectué 3 blocs de 100 essais. Même si la mesure du TR minimal ne peut pas être faite précisément, nous proposons de prendre comme TR minimal le temps à mi-hauteur de la courbe de distribution des TR, soit 160 ms (Figure 5). Le groupe de sujets était assez hétérogène et nous pouvons observer dans les résultats une courbe bi-modale avec une répartition en 2 groupes de 3 sujets rapides et 3 sujets lents. Cette différence de vitesse entre sujets induit dans cette expérience un TR moyen beaucoup plus élevée que le TR minimal : 233 ms (± 24,1 ms) contre 160 ms. Le TR moyen des 3 sujets les plus rapides était de 195,2 ms (± 7,3 ms) et celui des 3 sujets les plus lents de 266,7 ms (± 17,7 ms).

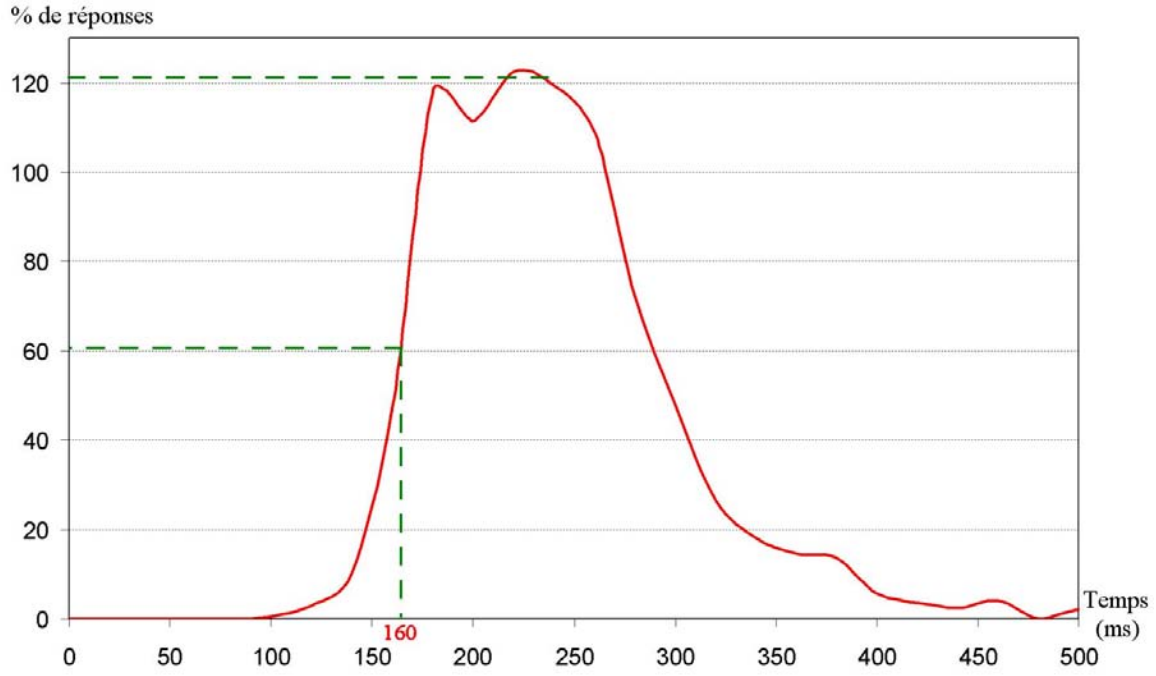


Figure 5 : Distribution des TR de 6 sujets humains dans la tâche de détection simple. Le TR minimal est déterminé par la mesure de la latence à mi-hauteur de la courbe, soit 160 ms.

Une autre expérience de détection d'images naturelles du même type que celle présentée ici a été menée par Rousselet (Rousselet *et al.*, 2002). Il n'est pas possible de comparer directement cette expérience avec la précédente puisque les images étaient présentées ici à 3,6° d'excentricité par rapport au point de fixation, entraînant ainsi des effets liés aux différences inter-hémisphériques qui existent dans les traitements visuels, en particulier pour les fréquences spatiales (Peyrin *et al.*, 2003). Cependant, une présentation centrale des images ne devrait avoir pour seul effet que d'améliorer légèrement la performance. Avec un nombre de sujets plus important et une plus grande homogénéité des résultats individuels, les données de cette expérience confirment celles que j'ai obtenues. Le TR moyen pour la simple détection d'une image située à 3,6° était de 211 ms (± 24,1) et le TR minimal, mesuré à mi-hauteur du pic de la distribution était de 160 ms (Figure 6).

Nous pouvons conclure de ces 2 expériences contrôle que le TR moyen pour détecter l'apparition d'une image naturelle sur un écran est d'environ 200 ms et que les premières réponses dans une telle tâche (hors anticipations) se situent très probablement autour de 160 ms (voir nos réserves plus haut). Ce délai est directement comparable à la "latence incompressible" de 150 à 160 ms trouvée en réponse à l'apparition d'un point lumineux (Galifret & Pieron, 1949 ; Vaughan *et al.*, 1966 ; Roufs, 1974). Ainsi le temps additionnel pour effectuer une tâche complexe de catégorisation visuelle rapide par rapport à une simple tâche de détection n'est que d'environ 90 ms. Ce qui contraint de façon importante l'ensemble des traitements additionnels nécessaires.

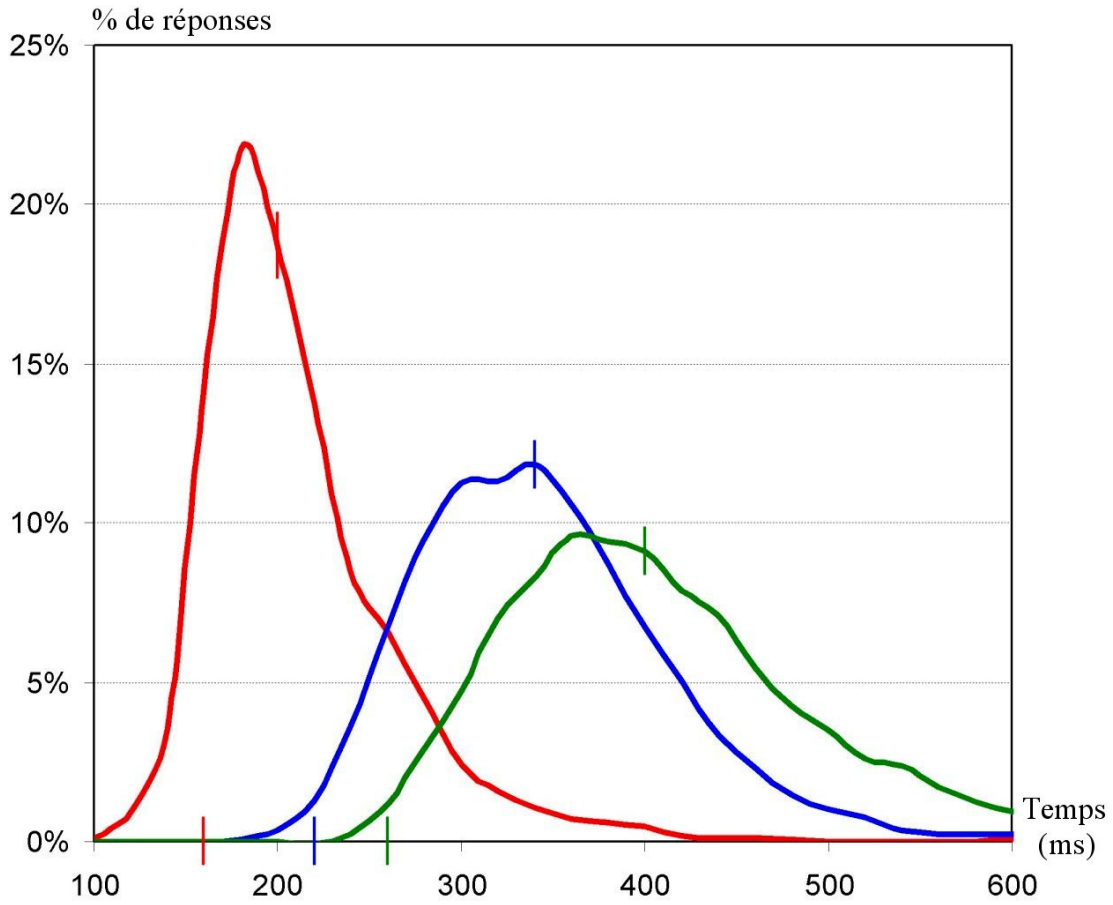


Figure 6 : Distribution des TR chez l'homme dans la tâche de détection simple (rouge) (d'après Rousselet et al., 2002), de reconnaissance (bleu) et de catégorisation (vert) (d'après Delorme et al., 2004). Les fausses détections pour les 2 dernières courbes ont été retranchées des réponses correctes. Le TR minimal dans ces 3 tâches est respectivement de 160, 220 et 260 ms (barres verticales sur l'axe des abscisses). Le TR médian de 200, 340 et 400 ms (barres verticales sur les courbes).

L'absence de distracteurs dans cette tâche ne permettait pas de calculer d'activités différentielles et nous n'avons donc pas enregistré l'activité cérébrale des sujets au cours de ces tâches de détection simples. Étant donné que la tâche du sujet est de détecter l'apparition d'un signal à l'écran, l'équivalent sur le plan électrophysiologique de cet événement est l'apparition d'un potentiel évoqué par le stimulus. Dans toutes les expériences que nous avons effectués avec les images naturelles, ce signal apparaît en général autour de 70 à 80 ms. Il correspond à l'activation des premières aires visuelles et ces données sont compatibles avec une réponse du sujet autour de 160 ms après que l'information visuelle ait parcouru le reste du système sensori-moteur. Les 60 ms qui séparent cette tâche de détection simple de la tâche de reconnaissance d'image (160 vs 220 ms) constituent un écart relativement important et ne nous permettent pas de conclure sur la similarité des mécanismes mis en jeu dans les deux tâches. Nous ne pouvons pas déterminer si le trajet des informations court-circuite certaines

aires pour permettre aux sujets de répondre plus rapidement dans le cas de la détection simple ou si une préactivation maximale de chaque aire visuelle est suffisante.

2.3 - A quoi correspondent ces 150 ms ? Quelle est l'origine de la différentielle ?

L'activité différentielle à 150 ms qui apparaît dans tant de tâches de catégorisation est intéressante parce qu'elle est le reflet de processus cognitifs à une latence relativement courte. Des études reposant sur l'IRMf (Fize *et al.*, 2000) ou la localisation de source (Delorme *et al.*, 2004) ont montré que cette activité différentielle prend principalement son origine dans le cortex inféro-temporal, une partie de la voie ventrale dans laquelle des neurones sélectifs à des objets ont été enregistrés. L'amplitude de cette activité différentielle est d'ailleurs fortement corrélée à la précision des sujets dans la tâche effectuée (expérience sur le contraste dans le 1er chapitre et expérience de masquage ci-dessous), ce qui peut laisser penser que la bonne perception d'un objet cible est directement liée au niveau d'activation des représentations des objets stockées dans le cortex inféro-temporal.

Plusieurs hypothèses ont été avancées pour expliquer à quoi correspond cette activité différentielle à 150 ms. La première idée est développée dans l'article de 1996 de Thorpe et al. qui se concentre principalement sur les activités différentielles frontales. Ne constatant pas de corrélations entre le temps et réaction et la latence de l'activité différentielle, les auteurs avaient proposé l'idée qu'elle résulte d'une inhibition massive de la réponse lorsque la scène présentée ne contient pas de cible. D'autres expériences avaient mis en évidence en frontal des effets très semblables d'inhibition de réponse (Gemba & Sasaki, 1989 ; Sasaki *et al.*, 1993).
Nous pouvons cependant douter de cette interprétation puisque nous savons désormais que : (1) l'activité différentielle frontale est majoritairement le reflet antérieur de dipôles occipito-temporaux (Rousselet *et al.*, 2004), et (2) les latences de décharge des neurones dans le cortex inféro-temporal sont bien plus fortement corrélées à l'apparition du stimulus qu'à la réponse motrice, rendant ainsi logique l'indépendance observée (DiCarlo & Maunsell, 2005). La catégorisation peut très bien intervenir autour de 150 ms de manière pratiquement indépendante du temps de réaction si la variabilité des TR prend son origine dans la variabilité motrice plutôt que dans celle du traitement des informations visuelles.

Il est alors naturel de ne pas trouver de corrélation entre la latence de l'activité différentielle et celle des temps de réaction si l'on admet que cette activité est générée par des neurones dans la voie ventrale. Le signal à 150 ms pourrait être constitué à la fois par la vague d'activation

feed-forward parcourrant le système visuel et par le flux d'informations en feedback provenant du cortex préfrontal (Barcelo *et al.*, 2000 ; Moore & Armstrong, 2003) ou d'aires impliquées dans la mémoire qui viendraient activer les représentations du cortex inféro-temporal utiles pour la tâche à accomplir. Ces deux flux d'informations pourraient interagir de manière complexe dans la voie ventrale et donner lieu à l'activité différentielle observée.
Rousselet et al. (Rousselet *et al.*, 2004) proposent une variante de cette hypothèse dans laquelle l'activité différentielle à 150 ms serait postérieure à la première vague d'activation du système visuel et correspondrait à la sélection spatiale par voie descendante de la zone du champ visuel contenant l'objet cible.

Il ne nous est pas possible à l'heure actuelle de trancher entre ces différentes hypothèses sur l'origine de l'activité différentielle, mais il est en revanche tout à fait possible d'étudier plus en détail la durée minimale pendant laquelle l'information visuelle doit être traitée pour permettre une catégorisation.

2.4 - Peut-on décomposer ces 150 premières millisecondes ? Article n°4

Dans les conclusions du chapitre précédent, nous avancions que les informations visuelles sont probablement transférées et traitées en parallèle dans les systèmes parvo- et magno-cellulaires. De plus, à l'encontre de ce qui est communément admis, nous proposions que les informations véhiculées par le système magnocellulaire pouvaient permettre d'atteindre une représentation de la scène visuelle suffisamment informative pour réaliser certaines tâches de catégorisation.
Une telle hypothèse a des conséquences importantes sur les propriétés de la catégorisation visuelle rapide. En effet, les limites du système visuel pourraient correspondre à celles du système magnocellulaire dans les domaines pour lesquels le système parvocellulaire est défaillant, comme c'est le cas lorsque le contraste est faible ou les fréquences temporelles élevées. Ainsi, selon les conditions de présentation, la représentation de la scène pourrait avoir un "grain" relativement grossier, caractéristique du système magnocellulaire.

Le système visuel est rapide pour extraire les informations utiles dans une image et y détecter des objets, mais il est également très rapide pour traiter plusieurs images successivement ou des images perçues pendant un temps très bref. La différence entre ces conditions de présentation est très importante. Quand le signal cérébral diverge autour de 150 ms entre les cibles et les distracteurs, c'est probablement que l'image a été suffisamment traitée pour

pouvoir déterminer si elle contient ou non une cible. Cependant ces 150 ms ne nous donnent d'indications que sur le temps global de traitement sans nous renseigner sur les différentes étapes de traitement qui ont lieu à l'intérieur du système visuel. La vision est faite pour interpréter et comprendre un flux d'information en continu et une telle capacité nécessite que plusieurs "images" soient simultanément en cours de traitement à un instant donné, à diverses étapes du système visuel (Keysers & Perrett, 2002). Étant donné les limites de l'EEG sur le plan spatial et de l'IRMf sur le plan temporel, il n'est pas possible de connaître en détail la durée de chaque étape de traitement, mais l'analogie avec une chaîne de montage nous montre que la vitesse de traitement ne dépend que de l'étape la plus lente. Ainsi, trouver le temps minimal qu'il faut laisser pour qu'il soit encore possible de catégoriser des images peut nous donner avec une assez grande précision la durée de l'étape de traitement la plus longue dans le système visuel. Grâce aux expériences qui ont permis de s'affranchir des différences physiques entre les images cibles et distracteurs, nous pouvons faire l'hypothèse que l'activité différentielle qui se développe à partir de 150 ms correspond à un traitement suffisamment avancé pour que le statut de l'image soit déterminé au dessus du niveau de la chance. C'est en quelque sorte une première évaluation du temps de traitement "total" nécessaire pour parcourir la chaîne de traitement visuel jusqu'à un niveau utile à la catégorisation. Si le traitement visuel ne s'effectuait qu'en une seule étape, le temps minimal à respecter entre 2 images présentées au système visuel devrait donc être de 150 ms pour que chaque image soit traitée correctement. Une présentation plus rapprochée des images entraînerait une forte baisse de la performance. A l'inverse, si le système visuel est découpé en plusieurs petites étapes de traitement, comme le montrent les données anatomiques, il doit être possible de présenter des images à un rythme plus rapide sans baisse de performance immédiate. En mesurant le temps minimal qu'il faut laisser au système visuel entre une image et un masque, nous pouvons déterminer le temps de traitement de l'étape la plus lente du système visuel.

Dans l'article suivant, nous décrivons une expérience dans laquelle les images sont présentées au sujet pendant un temps très bref, comme dans les expériences précédentes, mais où elles sont en plus suivies par un masque pour empêcher toute perception résiduelle sur l'écran (persistance dans l'activation des photophores) et sur la rétine (persistance rétinienne). Le masque est constitué de motifs fortement contrastés qui viennent également bloquer les traitements en cours sur l'image dans le système visuel.

Résumé de la publication : "The time course of visual processing: Backward masking and natural scene categorisation"

Les sujets devaient effectuer une tâche de catégorisation animal/non animal. Les images naturelles utilisées étaient suivies d'un masque pour bloquer toute perception et interrompre les traitements à différentes latences en manipulant le temps entre l'image et le masque (SOA : Stimulus onset asynchrony).

16 sujets ont participé à cette expérience dans laquelle le comportement et l'activité cérébrale étaient enregistrés. Les sujets effectuaient 1440 essais répartis en 16 séries de 90 essais. Les 9 conditions de SOA (de 6 à 106 ms + 1 condition contrôle avec le masque seul) étaient équiprobables dans une série. Le masque était composé d'une succession d'images représentant des motifs noirs et blancs fortement contrastés et à différentes échelles spatiales. L'avantage d'un masque dynamique composé de plusieurs images est qu'il permet de masquer à coup sûr n'importe quelle image de scène naturelle puisque son contenu fréquentiel et la position spatiale de ses traits recouvrent toute la surface de l'image dès que les premiers masques ont été affichés à l'écran. Une expérience contrôle a montré que 2 images de masques sont suffisants pour bloquer toute perception du stimulus.

Résultats résumés :

La précision des sujets en fonction du SOA montre que le système visuel est très rapide et qu'il peut catégoriser des images jusqu'à des fréquences élevées. Les sujets n'étaient au niveau chance que pour la condition la plus difficile dans laquelle l'apparition de l'image n'était séparée de celle du masque que de 6 ms. Dès la condition 12 ms, les réponses des sujets étaient significativement au dessus du niveau chance. La précision maximale était bien sûr obtenue sur la condition ayant le SOA le plus long (106 ms) ; pratiquement similaire à celle obtenue lors d'une expérience précédente sur des images en niveaux de gris mais non masquées (89 vs 91,4% correct) (Delorme *et al.*, 2000). L'augmentation de précision en fonction du SOA était forte entre 12 et 44 ms, mais la précision atteignait un plateau dès que le SOA atteignait 44 ms (85,6% correct).

La présence d'un masque dynamique au contenu variable et à différentes latences après l'apparition de l'image a fortement compliqué l'analyse des potentiels évoqués dans cette tâche. Ici l'avantage de travailler sur l'activité différentielle entre essais cibles et essais distracteurs s'est révélée particulièrement intéressante puisque les potentiels évoqués par le masque étaient soustraits à eux-mêmes lors du calcul de ce signal différentiel. L'amplitude de l'activité différentielle en fonction du SOA était très fortement corrélée à la précision obtenue

par les sujets (Figures 3 et 6 de l'article) dans les différentes conditions (coefficient de corrélation de pearson de 0,98).

Discussion :

Cette expérience montre à nouveau la remarquable efficacité du système visuel en terme de vitesse de traitement. Nous observons tout d'abord qu'un délai de seulement 12 ms entre l'image et le masque est suffisant pour qu'assez d'informations soient traitées et permettent d'effectuer une tâche de catégorisation complexe au-dessus du niveau de la chance. Nous observons ensuite qu'un délai de 40 ms permet aux sujets d'atteindre une précision proche du niveau maximal. En replaçant ces deux résultats dans notre analogie avec un traitement de l'information en pipeline, nous pouvons penser que les étapes de traitement les plus lentes dans le système visuel ont une durée comprise entre 10 et 40 ms. On retrouve ici des valeurs compatibles avec les conclusion tirées des résultats apportés par l'électrophysiologie. En effet, nous avons vus que les latences des activités différentielles enregistrées en EEG ne laissent qu'une centaine de millisecondes pour effectuer une dizaine d'étapes de traitement, soit environ 10 ms par étape (concept transposé chez le singe en figure 3 de l'introduction). Ces valeurs sont également en accord avec les résultats obtenus par Rolls et Tovee chez le macaque puisqu'ils montrent que la majeure partie de l'information encodée par un neurone se situe dans les 30 premières millisecondes de sa décharge (Tovee & Rolls, 1995 ; Rolls *et al.*, 1999). "Écouter" la sortie d'un neurone visuel pendant plus longtemps n'apporte que peu d'information supplémentaires sur le stimulus, ce qui explique pourquoi un masque présenté 40 ms après l'image n'a que peu d'impact sur la performance.

Article n°4

Vision Res, **45**, 1459-1469

The time course of visual processing: backward masking and natural scene categorisation

Nadège Bacon-Macé, **Marc J-M. Macé**, Michèle Fabre-Thorpe & Simon J. Thorpe

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Vision Research 45 (2005) 1459–1469
Vision Research
www.elsevier.com/locate/visres

The time course of visual processing: Backward masking and natural scene categorisation

Nadège Bacon-Macé *, Marc J.-M. Macé, Michèle Fabre-Thorpe, Simon J. Thorpe

Centre de Recherche Cerveau et Cognition (UMR 5549, CNRS-UPS), Faculté de Médecine de Rangueil, 133, Route de Narbonne, 31062 Toulouse, France

Received 23 April 2004; received in revised form 23 December 2004

Abstract

Human observers are very good at deciding whether briefly flashed novel images contain an animal and previous work has shown that the underlying visual processing can be performed in under 150 ms. Here we used a masking paradigm to determine how information accumulates over time during such high-level categorisation tasks. As the delay between test image and mask is increased, both behavioural accuracy and differential ERP amplitude rapidly increase to reach asymptotic levels around 40–60 ms. Such results imply that processing at each stage in the visual system is remarkably rapid, with information accumulating almost continuously following the onset of activation.

Keywords: Natural images; Backward masking; Early processing; Information integration; Event-related potentials (ERP)

1. Introduction

Human subjects are very quick and efficient at analysing briefly viewed natural scenes, an ability that has obvious survival value. We can determine whether a briefly flashed image contains an animal and make a behavioural response in as little as 250 ms, and this ability extends to other categories of visual stimulus such as faces or means of transport (Macé & Fabre-Thorpe, 2003; Rousselet, Macé, & Fabre-Thorpe, 2003; Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001a). Simultaneously recorded event-related potentials (ERP) diverge sharply between correct target and distractor trials just 150 ms after stimulus onset (Rousselet, Fabre-Thorpe, & Thorpe, 2002; Thorpe et al., 1996) which imposed even more severe temporal constraints. Extensive training failed to reduce this 150 ms latency, indicating that even with images never seen before, the system is operating virtually optimally and with a minimal number of processing stages (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001).

This sort of behavioural and electrophysiological evidence imposes an upper limit on the amount of time required for animal detection but provides relatively little direct information about the dynamics of the underlying processing. With only a 150 ms delay between the onset of activation in the retina and a cerebral differentiation between target and distractor pictures, it is a challenge to explain how visual information is processed and transmitted through the visual pathways. A distinction is often made between discrete or continuous models of information transmission (Eriksen & Schultz, 1979; Hasbroucq, Burle, Bonnet, Possamai, & Vidal, 2002;

* Corresponding author. Tel.: +33 5 62 17 37 75; fax: +33 5 62 17 28 09.
E-mail address: nadege.bacon-mace@cerco.ups-tlse.fr (N. Bacon-Macé).

0042-6989/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2005.01.004

McClelland, 1979), the first implying that there is a fixed minimum processing time at each stage before information can be sent to the next level, while the latter supposes that it can be transmitted continuously as soon as information becomes available. Both are consistent with a pipeline processing scheme in which every step can operate simultaneously and in parallel. Indeed, some form of pipeline processing seems necessary to account for the results of a recent study using RSVP (rapid serial visual presentation) showing that human subjects can detect images in sequences presented at rates of up to 75 images per second (Keysers & Perrett, 2002). Such data imply that less than 15 ms are enough to process a sufficient amount of information concerning each picture of the sequence.

RSVP experiments can be integrated into the broader approach of masking, which involves two or more temporally close stimuli to reduce the associated perception (Breitmeyer, 1984). Masking protocols are very useful to study the timing of information processing in the visual system since they allow processing to be interrupted at different times. Electrophysiological studies on monkeys have shown that the intensity and duration of neuronal responses are more and more affected as the mask gets closer to the stimulus, but that considerable information is available in monkeys from the first 30 ms of the neuronal responses (Rolls, Tovee, & Panzeri, 1999; Tovee & Rolls, 1995). In human subjects, there are many experiments that concern the influence of stimulus/mask interval on behavioural responses (Breitmeyer, 1984; Enns & Di Lollo, 2000), but few of them were used in the context of high-level tasks, such as categorisation. Moreover, few masking experiments have investigated the associated changes in cerebral activity, and most of those have involved fMRI methods. Nevertheless, there are reports of a correlation between the ability to detect or to name objects and the activation in occipital regions (Dehaene & Naccache, 2001; Grill-Spector, Kushnir, Hendler, & Malach, 2000; Vanni, Revonsuo, Saarinen, & Hari, 1996). As image and mask get temporally closer, both performance and cerebral activation decrease. This type of correlation can be particularly useful to understand the signals recorded from the brain during perceptual processing.

We present here the results of a backward masking experiment in a go/no-go categorisation task, in which natural scenes were followed by a very strong dynamic mask after a varying stimulus onset asynchrony (SOA). By interrupting processing after different delays, we could determine how information accumulates over time during the task. One of the novel features of the experiment was the use of a high screen refresh rate (160 Hz) that allowed us to present the test image for a single 6.25 ms frame and to vary the SOA by small 6.25 ms steps, a much higher resolution than is typically used in masking experiments.

2. Experimental procedure

2.1. Task

Sixteen subjects participated in this experiment (8 females, 8 males, with a mean age of 29 ranging from 21 to 50). They all volunteered for the study and gave their written informed consent. The go/no-go categorisation task was based on an experimental procedure introduced by Thorpe et al. (1996). Subjects were seated in a dimly lit room, at 1 m from a screen adjusted to an 800×600 pixel resolution and a 160 Hz refresh rate. Natural scene pictures (600×400 pixels in size) were flashed on the monitor for a single frame, which corresponds to 6.25 ms. Subjects were asked to release a button within 1 s if the picture contained an animal and maintain pressure otherwise.

Each subject was tested on 16 series of 90 trials, each of which contained the same number of target and distractor images. All subjects had previously completed at least 3 training blocks of 90 trials. They were asked to try to release the button on 50% of the trials, whatever the masking condition.

A trial began with the display of a white fixation cross in the middle of the black screen for 600–900 ms at random. Then the picture—target or distractor—was flashed, followed by the mask stimulus. Eight different values were used for the stimulus onset asynchrony (SOA) between the picture and the mask (6.25, 12.50, 18.75, 25.00, 31.25, 43.75, 81.25 and 106.25 ms) and display latencies were verified with a photodiode connected to an oscilloscope. Furthermore, we added a control condition in which only the mask was displayed after the fixation cross, without any picture presentation. These 9 conditions were counter-balanced in each series, with 10 trials per condition presented at random, producing a total of 90 trials per block. Any given subject only saw each picture once.

χ^2 tests were used to evaluate if behavioural accuracy was above chance level for each SOA condition. Masking effects between the conditions were assessed with analysis of variance (ANOVA) and post-hoc analyses were performed by using paired t-tests with a Bonferroni correction or Mann–Whitney tests.

2.2. Stimuli

A total of 1280 grey level natural images were used in this experiment. As demonstrated in previous work, ultra-rapid categorisation does not rely on colour cues, as performance is almost unaltered when stimuli are presented in grey level (Delorme, Richard, & Fabre-Thorpe, 2000). Moreover, masking effects were easier to obtain and control without colour information in the natural scenes. Half the images contained animals, and were as varied as possible (fish, insects, mammals,

birds or reptiles). The subjects had no knowledge about the size, position and number of the targets in a single picture. The other half of the images were distractors with a wide range of material including natural landscapes, indoor or outdoor scenes, man-made objects, etc... None of the pictures had been seen previously by the subjects and training pictures were not used in the test series.

2.3. Mask

To construct the mask, a white noise image was filtered at four different spatial scales, and the resulting images were thresholded to generate high contrast binary patterns. For each of the 4 spatial scales, 4 different versions were generated by mirroring and rotating the original image. A pool of 16 images was thus available for masking. The mask used in this experiment was a sequence of 8 images - a so-called "dynamic mask" (Fig. 1). The 8 images were chosen randomly from the pool, with each of the four spatial scales presented once during the first 4 images and again during the last four images. Thus, a pattern at each of the spatial scales appeared twice in the "dynamic" mask (see Fig. 1). All the images in the mask were presented for 2 refresh cycles, so that overall, the masking stimuli were displayed for 16 frames (around 100 ms).

2.4. ERP recordings

EEG data were recorded from a 32-electrode cap. Electrode locations were defined using the standard 10–20 Oxford system with 12 additional electrodes over occipital sites. Electrical activity was amplified by a NeuroScan Synamps amplifier linked to a PC computer, digitized at 1000 Hz, corresponding to a sample bin of 1 ms, and low-pass filtered at 100 Hz. Each recording epoch began 100 ms before the stimulus display on the screen and continued for 1000 ms after the stimulus onset. A baseline correction was carried out for each epoch using the 100 ms of pre-stimulus activity. Trials with artefacts related to ocular movements were rejected, by using a criterion of $[-80;+80\ \mu V]$ on two frontal electrodes (FP1 and FP2) between −100 and +400 ms. Within this time period, another artefact rejection was performed on trials with a strong alpha frequency activity, by using a $[-40;+40\ \mu V]$ criterion on parietal electrodes (Oz and Pz). Signals were then low-pass filtered at 40 Hz before the analysis. We were particularly interested in the occipito-temporal electrodes (standard O1, O2, OZ, IZ and non-standard PO7, PO8, PO9, PO10, O9, O10, P7, P8) and the frontal electrodes (standard FP1, FP2, F3, F4, Fz). Epochs corresponding to correct responses were averaged separately for targets and distractor trials on each masking condition.

Differential activities were determined by subtracting the average signal on correct distractor trials from the signal on correct target trials. Eight different curves were obtained, one for each SOA condition. The differential activity amplitude was calculated by a Matlab program that determined the most negative point between 150 and 250 ms after the onset of stimulus presentation (Rousselet, Thorpe, & Fabre-Thorpe, 2004; Thorpe et al., 1996). It was measured separately for each individual and also using the average signal across all

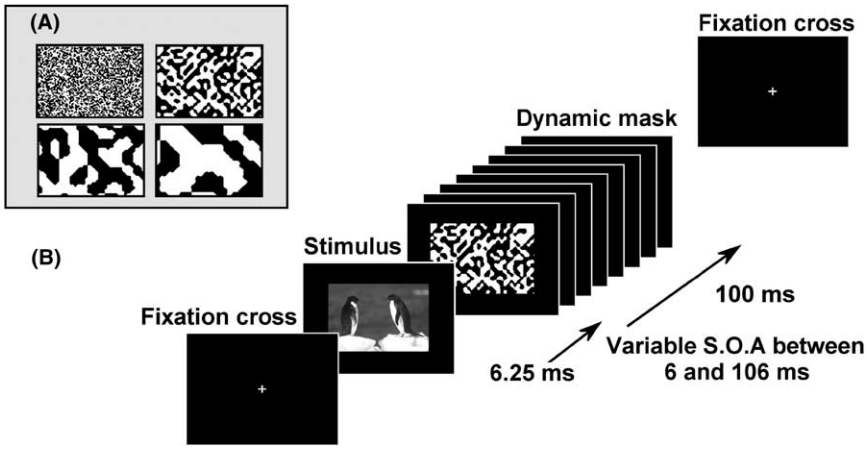


Fig. 1. Behavioural paradigm. (A) Four pictures with different spatial scales that constitute the dynamic mask. Each could be presented at 4 different orientations making a total of 16 different patterns. The images were intermixed to reduce the risk of generating retinal after-effects with the restriction that each spatial scale was used once during the first 4 pictures of the mask and once again during the 4 last ones. (B) In each series, subjects were tested on 90 trials organised as follows: first the fixation point is displayed on the centre of the screen for a random delay to avoid anticipated responses. Then the grayscale picture is flashed for only one frame using a monitor set at 160 Hz. After a variable 6.25–106.25 ms SOA, a dynamic mask is displayed, composed of eight 100% contrasted mask patterns at the four different spatial scales. The subjects then have 1000 ms to release the button if the picture contains an animal. Eight time steps were chosen for the SOA: 6.25, 12.50, 18.75, 25, 31.25, 43.75, 81.25 and 106.25 ms.

subjects. Differences in latencies and amplitude among SOA conditions were statistically evaluated by ANOVAs and post-hoc analyses were performed using *t*-tests with a Bonferroni correction. The correlations between electrophysiological measurements and behavioural data were performed with Pearson tests.

3. Results

3.1. Behavioural performance

3.1.1. Mask efficiency

We evaluated behavioural performance in terms of accuracy and reaction time as a function of the SOA. For each condition, a χ^2 test between correct and incorrect responses determined if accuracy was above chance level, set at 50% because targets and distractors were equally likely. Only the very shortest SOA interval (6 ms) resulted in performance at chance level, with a mean value of 51.9% (Fig. 2A). This result emphasizes the high efficiency of the mask, which effectively prevented visual processing when presented close to the stimulus. However for the next SOA (12 ms) condition, accuracy was already above chance level ($p < 0.01$) and rapidly increased to reach 85.6% with a 44 ms SOA. Accuracy then stabilized at a maximum value of 91.4% for the last condition (106 ms). However, increasing processing time above 44 ms had only a minor effect on performance since accuracy scores in the last three SOA conditions 44–81–106 ms were not significantly different ($p > 0.33$). Note that the maximum accuracy was very close to the accuracy obtained in a previous study (Delorme et al., 2000), using achromatic natural images flashed for 20 ms in the same go/no-go categorisation task without masking, and where subjects averaged 93% correct. Thus the mask has relatively little effect when it appears 40–60 ms after the image presentation onset, and visual processing remains extremely good despite the fact that the stimulus picture was flashed for only 6 ms.

3.1.2. Response inhibition with increasing difficulty

We noticed a strong reduction in response rate with the most difficult masking conditions. Before the experiment, subjects were asked to try to release the button on about half of the trials in each series. This instruction has been respected since the mean response rate, including correct and incorrect responses, was about 47% when grouped across all conditions. However, the response rate varied considerably with the SOA, as it exceeds 50% from 106 to 25 ms SOA, and drops strongly with SOAs below 25 ms (Fig. 2B).

Interestingly, these variations only affected the proportion of correct go responses. In contrast, the proportion of erroneous go responses to distractors was remarkably stable across all SOA conditions. It would appear that short SOAs did not lead subjects to make more false positives to distractor pictures but did prevent them extracting enough information to make a response on target trials. Another interesting result is given by comparing the response rate obtained with the shortest SOA (22.3%) with the control condition when only the mask was displayed without any picture (19.6%). These two conditions were not significantly different ($p > 0.24$) which suggests that with a 6 ms SOA, subjects behaved as if no image had been presented at all.

3.1.3. Reaction times

Mean reaction time decreased with longer SOAs, particularly for values over 44 ms ($F(7,127) = 2.591$, $p < 0.02$). In the conditions where the mask was close to the stimulus (SOA 6, 12, 19, 25 and 31 ms), reaction times were significantly longer ($p < 0.01$) than when the mask appeared later (44, 81 and 106 ms). A maximum difference of 54 ms was found between the 12 and 81 ms SOA conditions. This suggests that when the mask interrupts the visual processing earlier, the amount of

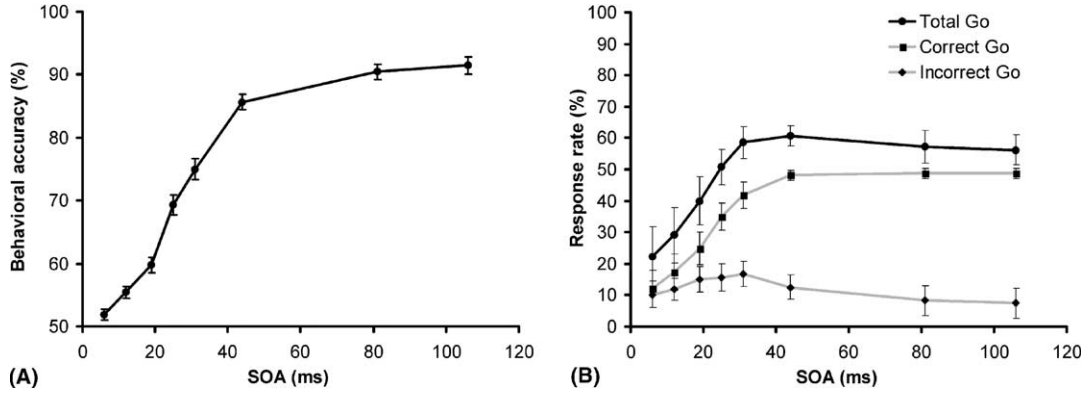


Fig. 2. Behavioural performance as a function of the SOA, averaged above 16 subjects. (A) Behavioural accuracy (±s.e.m.). (B) Mean percentage of go responses (±SD).

information is reduced and perceptual decisions require more time.

Masking effects can be observed throughout reaction time distributions by comparing the condition of optimal perception (106 ms) with each of the others (Fig. 3A). Above 44 ms, distributions are remarkably similar, but with 31 ms SOA, the median part of the distribution shows a pronounced plateau. Thus, there is a strong effect of the mask on the median reaction time, but the initial part of the distribution, corresponding to early responses, is not affected. With an SOA of 25 ms and less, early responses are also disrupted and the early part of the reaction time distributions no longer superimpose.

3.2. Electrophysiology

Disruptive effects on visual processing can also be observed on the ERP data. For each subject we averaged separately signals on distractor and target trials and subtracted one from the other to calculate a differential activity curve. As signals on targets and distractors contain information about the response to both the picture and the mask, subtracting these two signals is a good way to cancel out the activity associated with the physical encoding of the mask. The effects of the mask on image categorisation processing remain clearly observable on the residual signal (Fig. 4). Therefore, we will not present here a detailed analysis of the shape of the underlying ERP signals but rather focus on an analysis of the differential effects.

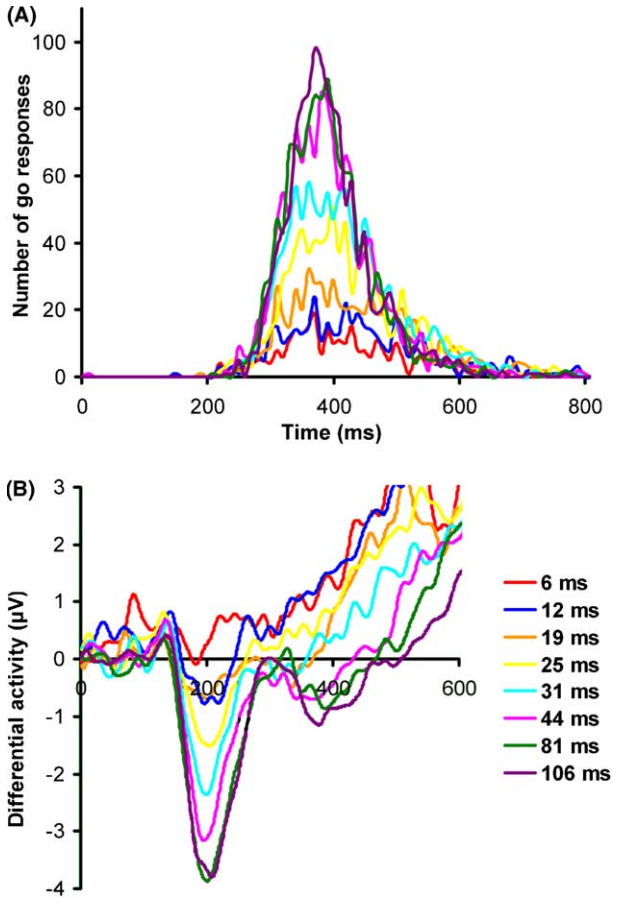


Fig. 3. Masking effects on behavioural reaction time and occipital cerebral activity. (A) Reaction time distribution of correct go responses as a function of the SOA (10 ms bin width), averaged on 16 subjects. (B) Differential activity averaged on 16 subjects for each SOA. The activity is calculated as the difference between signals on correct target and distractor trials, obtained from PO8 electrode.

We analyzed the differential activity with respect to the SOA. Fig. 3B shows averaged signals recorded on a representative occipital electrode (PO8). The onset of the differential activity appears to start at around the same latency (150 ms) but it is clear that the signal amplitude decreases with shorter SOAs ($F(7,1535) = 77.13$, $p < 0.001$). Moreover, with the exception of the two shortest SOAs (6 and 12 ms) for which the activity was rather weak, peak latencies are remarkably stable between 200 and 215 ms. In other words, when the mask is closer and closer to the picture, the reduction of perceptual differences between target and distractor stimuli strongly affects the amplitude of differential activity.

A lateralization effect can be observed in this task, as the mean amplitude of the differential activity was significantly larger for the electrodes over the right hemisphere compared to the left hemisphere (respectively 2.56 μV for the average of electrodes O2, PO10, PO8, O10, P8 versus 2.22 μV for the average of O1, PO9, PO7, O9, P7, all SOA conditions grouped; $F(1,1279) = 15.45$, $p < 0.0001$). This was the case for each of the SOA conditions except for the shortest one at 6 ms.

Although the masking effect is particularly visible on occipital electrodes, similar effects can be seen at most electrode sites (Fig. 5). At frontal sites, shortening the SOA induced a significant diminution of the maximal amplitude of the differential activity ($F(7,639) = 22.305$, $p < 0.0001$), appearing around 200 ms. But in contrast no lateralization effect could be observed at these sites ($p = 0.145$).

These electrophysiological results can be directly related to behavioural data, which also showed a clear diminution of performance with decreasing SOA. Differential activity amplitude and behavioural accuracy variations are in fact strongly correlated, as showed on Fig. 6. This observation reinforces the idea that the differential activity reflects the result of a perceptual decision and that differential ERP responses provide a powerful investigative tool.

3.3. Control experiment

The choice of the dynamic mask was made after a number of pilot experiments, and appeared to be very

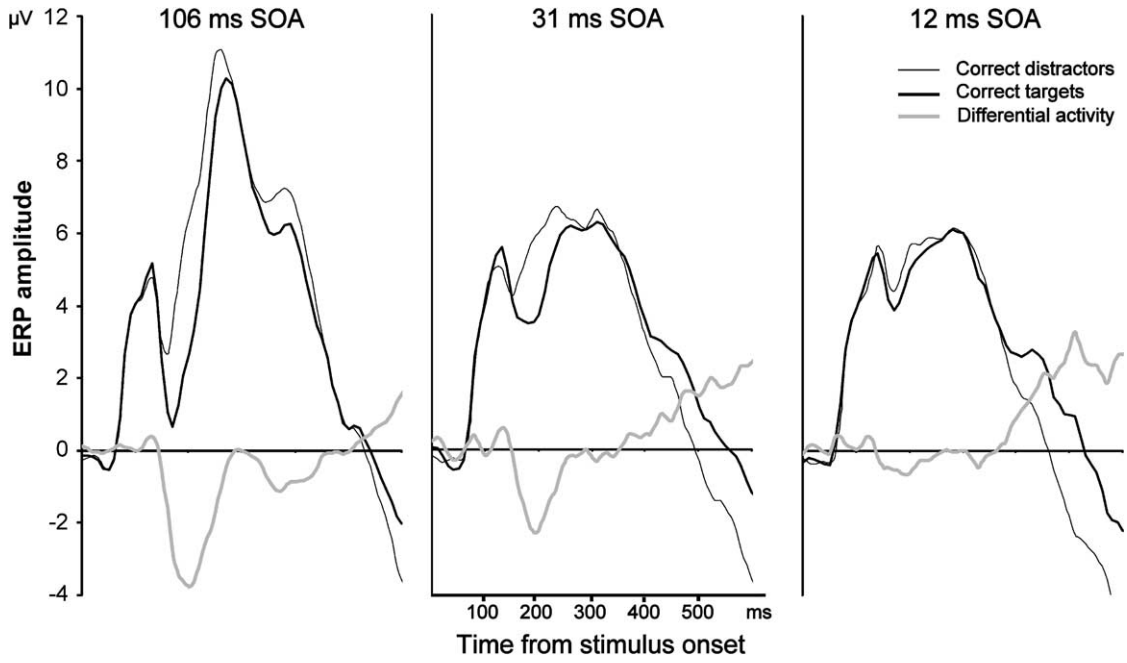


Fig. 4. Grand-average ERP on electrode PO8 for three SOA conditions. Differential activity is obtained by subtracting the signals on correct target and distractor responses.

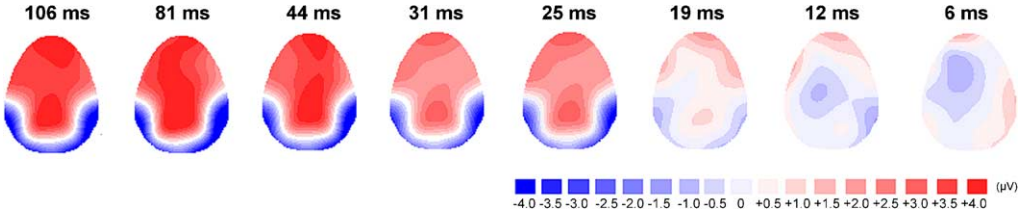


Fig. 5. Differential activity over the scalp at the maximal point of the amplitude, 200 ms after the image onset. Activity was averaged over 16 subjects.

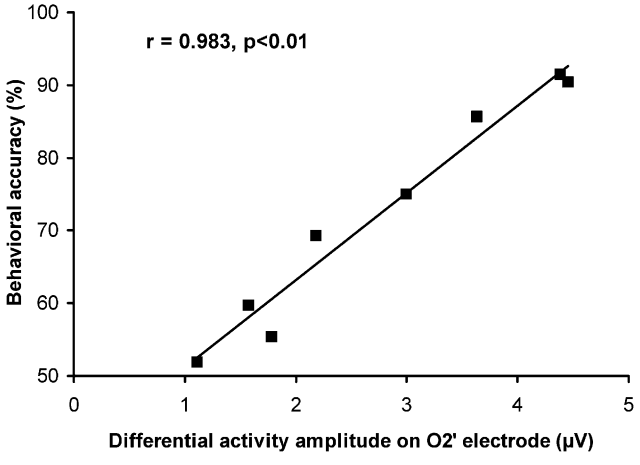


Fig. 6. There is a linear relation between behavioural accuracy and the amplitude of the occipital differential activity, particularly on the right hemisphere. Differential activity from PO8 electrode has been averaged over the 16 subjects, and correlated by a Pearson test ($p < 0.01$) to the mean behavioural accuracy among all SOA conditions.

efficient, as demonstrated by the fact that performance was at chance level with the shortest SOA. However, one possible problem of using a mask with multiple-frames is that one cannot be sure at what exact point the masking becomes totally effective. This could potentially add considerable uncertainty to the SOA considered as the disruptive latency.

We therefore have made a behavioural control experiment where 10 subjects performed the categorisation task at four different SOA values (4 conditions: 6, 12, 44 and 106 ms) and in which we varied the number of pictures in the mask from 1 to 8 (5 conditions: 1, 2, 3, 4 and 8 pictures). Across all the subjects, a total of 640 trials was performed for each of the 20 conditions. Furthermore, we encoded the spatial scale pattern used for each picture of the mask and particularly for the first one. These patterns were randomly chosen for each trial. Fig. 7 shows behavioural accuracy as a function of the first mask pattern. With only one image in the mask

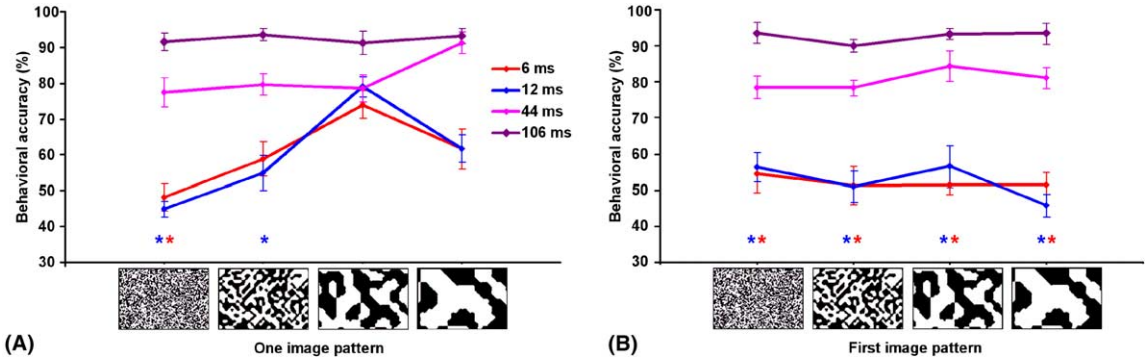


Fig. 7. Behavioural accuracy as a function of the spatial scale of the mask, for different SOAs (±s.e.m.) and as a function of the number of images in the mask. (A) Results when only one image is used in the mask. (B) Results when a sequence of two images is used in the mask. The pattern of the first image is illustrated here, all 3 other spatial patterns were used equally as the second image of the mask. The different curves represent behavioural accuracy for different SOAs. Ten subjects performed the experiment. Coloured asterisks indicate when the condition is at chance-level ($p < 0.01$).

(Fig. 7A), disruption effects depended strongly on the spatial scale of the mask at 6 and 12 ms SOA. When the finest spatial scale was used, masking was effectively complete since performance was at chance-level. Even with the second finest scale, performance was still very poor. This means that on virtually half the trials, just the first mask pattern was enough to completely disrupt processing, limiting the number of problematic trials. In contrast, at coarser spatial scales, the masking was less effective and subjects were able to perform more than 70% correct when just the mid-to-coarse scale mask image was used. However, when the mask contained two different images in succession (Fig. 7B), accuracy was no better than chance for both 6 and 12 ms SOA, irrespective of which spatial scale was flashed first. This result may appear to contradict the data from the previous experiment, in which performance was significantly above chance with an SOA of 12 ms, despite having used an even longer 8 image dynamic mask. However, it should be noted that there were much more trials per condition in the original experiment (2560 vs. 640 trials), which increases the statistical sensitivity of the test.

Together, the results of this control experiment demonstrate conclusively that the masking effects were indeed very strong and occurred very rapidly from the onset of the mask. Given that for the main experiment, the 8 image dynamic mask was continued for 100 ms, we can safely conclude that the disruption was complete throughout a critical period for target processing.

4. Discussion

4.1. Time course of information extraction

4.1.1. Visual information is extracted before masking

As might be expected, the behavioural data shows a strong masking effect on the visual processing involved in this high-level categorisation task, with both a drop in accuracy and an increase in reaction times. With a 6 ms SOA, processing appears to be completely blocked since the subjects were unable to perform significantly above chance. However, from 12 ms onwards, accuracy is already above chance level and performance increases until a ceiling effect is reached between 44 and 81 ms.

The high vertical refresh rate of the monitor (160 Hz), allowed us to observe a large range of masking levels, which leads to make several remarks about visual information extraction. First, it is noteworthy that the maximum accuracy reached by the subjects was close to the accuracy obtained in the same categorisation task used without masking (Delorme et al., 2000), indicating that the mask has no major effect after 81 ms. This also means that the presentation time of the picture, reduced to 6 ms in the present experiment from 20 ms in most of our previous experiments, had no appreciable effect on precision. It therefore appears that this 6 ms stimulation period is sufficient for the retina to extract enough information from the picture for animal detection to occur. Finally, the control experiment confirmed that the masking effects were strong even with just the first mask image, and that by the time the second mask pattern was presented, the disruption was complete. Given that the dynamic mask was maintained for 100 ms, we can be very confident that a long period of target processing is affected by the effective masking. This supposes that the delay available to extract relevant features from the stimulus is limited before masking takes place, and the visual system should base its analysis on a restricted amount of information to perform the task.

4.1.2. Visual information accumulates over time

The electrophysiological data also argue for a progressive accumulation of information. The differential activity, calculated by taking the difference between ERPs on correct target and distractor trials, is strongly affected by the SOA reduction. Its amplitude decreases

when the mask gets closer to the picture. We found a very high correlation between this reduction and behavioural accuracy. This can be related to another analysis of the differential activity in a previous categorisation task (Rousselet et al., 2004), where the status of the trials (Correct / False Alarm / Missed) could also be linked to the amplitude of the differential effects at occipital, frontal and parietal sites. The predictability of the behavioural outcome on the basis of the differential ERP signals constitutes a striking demonstration of a strong link between perception and cerebral activity, as previously shown with both fMRI (Dehaene & Naccache, 2001; Grill-Spector et al., 2000; Vanni et al., 1996) and unitary recordings in monkeys (Britten, Newsome, Shadlen, Celebrini, & Movshon, 1996; Leopold & Logothetis, 1996; Thompson & Schall, 1999).

The analysis of the differential signals between roughly 150 and 250 ms demonstrated that the more the activity averaged on targets differs from the distractor activity, the more subjects are able to detect the animals. This result could be related to the analysis of response rate since the mask has a higher effect on subjects' decisions for targets than for distractors (Fig. 2B). The overall data show that the difference between target and distractor responses is maximized when the mask is presented far from the stimulus, as if there were more and more cues accumulating to dissociate these two groups of images. The results are in accordance with the model of sensory information accumulation proposed by Schall (Schall, 2001) and derived from earlier work by Shadlen, Newsome and colleagues (Gold & Shadlen, 2000; Kim & Shadlen, 1999; Salzman & Newsome, 1994; Shadlen & Newsome, 1996). In their experiments, monkeys were trained to judge the main direction of motion in a collection of moving dots. The monkeys reported their responses by making an eye movement to one of two points, each indicating a given direction. The authors proposed that the decision depends on the accumulated signal corresponding to the increasing discrimination of the main direction of motion. In the same way in our experiment, we can suppose that the greater the separation between stimulus and mask, the greater the amount of processing that can be performed. Relevant information concerning the presence of an animal in the picture is accumulated until reaching a decisional threshold.

This model of cue accumulation over time fits with our data on reaction time. With SOAs between 25 and 44 ms, the early part of the reaction time distribution did not appear to be affected by the mask, but we observed a saturation effect on correct responses with median reaction times (Fig. 3A). This suggests that when pictures contain particularly salient cues, extensive information accumulation is not necessary and the subjects are capable of executing fast responses. However, when the target discrimination requires more analysis, information would not be available because processing is disrupted by the mask. Below 25 ms, the integration time was probably insufficient to process as much information because the mask affected even the earliest responses.

If the visual system bases the results of its analysis on the accumulated information, what does it imply for information encoding? How can the mechanisms of information extraction at different steps of the visual pathways be decomposed and what determines the impact of masking interference?

4.2. Information encoding in the visual pathways

4.2.1. Interference between stimulus and mask information: the where and how issues

Behavioural performance does not increase much with SOAs longer than 40 ms. We may relate this result with the latencies obtained from macaque neurophysiology showing that the first 30–40 ms includes the most selective part of the neuronal responses (Kovacs, Vogels, & Orban, 1995; Rolls et al., 1999; Tovee & Rolls, 1995). This data suggests that there is an upper limit on the time required at each processing stage to extract the relevant information that needs to be transmitted to the next step. Any processing that would take longer would be obliterated or smothered by the mask information.

Where would these masking effects take place? A first model would propose that the effects are more likely to occur relatively early in the visual pathways, for instance at the level of V1, depending directly on the structure where mask information could be encoded. Recordings in monkey infero-temporal cortex have demonstrated that the majority of neurons are maximally activated by stimuli more complex than bars or simple textures (Tanaka, Saito, Fukada, & Moriya, 1991), and showed a high degree of sensitivity to image scrambling, with activation decreasing together with performance (Vogels, 1999a, 1999b). Other studies, using functional imaging in humans, have compared the activation produced by objects and textures and found that a region of lateral occipital cortex was preferentially activated by objects even when the spatial frequencies and contrast of the object stimuli matched those of the texture stimuli (Grill-Spector, Kushnir, Edelman, Itzchak, & Malach, 1998; Malach et al., 1995). All these studies suggest that the mask, as a kind of texture stimulus, should mainly interact with picture information in earlier areas.

The next question concerns the *mechanism* by which the masking effect occurs. One simple explanation of masking assumes that the mask produces interference when neural responses to the mask and the test image overlap in time, and this effect is all the more important when it concerns spatially overlapping information related to critical features of the stimulus. There is good

evidence that the activation of sensory inputs to areas such as the striate cortex results in strong intracortical inhibition that could well interfere with the processing of subsequent inputs. The disrupting effects will thus depend on the spatio-temporal overlap between the neural responses to the test and mask stimuli.

Neurophysiological studies have shown that the onset latencies of neurons within a given visual structure vary from neuron to neuron, even when the visual stimulus is unchanged. For example, in primate visual cortex, onset latencies can vary from as little as 30 ms to 70 ms or more. The reasons for this variability are diverse, but one of the most important factors is undoubtedly stimulus contrast. It is notable that the shortest latencies ever seen have been obtained with very high contrast and high luminance stimuli. Given that the mask stimuli used in our experiments all have maximal contrast, we can assume that many neurons in V1 will respond to the mask with particularly short latencies (Albrecht, Geisler, Frazor, & Crane, 2002; Albrecht & Hamilton, 1982; Foxe & Simpson, 2002; Nowak & Bullier, 1997; Reich, Mechler, & Victor, 2001; Sestokas & Lehmkuhle, 1986). In contrast, the neural responses to the natural images used as test patterns are likely to be substantially more variable. Indeed, if we suppose that any given photograph of an animal will contain many different features that can be diagnostic for the presence of an animal, it is clear that the contrast associated with each feature will vary a lot. Thus, much of the critical information about the stimulus will be conveyed less rapidly than information about the mask, strengthening the effects of inhibitory mechanisms. Only information that can survive this spatio-temporal overlap would then be transmitted to the next step, and contribute to accumulate cues about the test stimulus. The first interpretation is thus based on the disruption of feed-forward processes, due to mask processing catching up with stimulus processing.

Another interpretation is based on the difference in transmission rates along the fast magnocellular (M) and the slower parvocellular (P) visual pathways. Detailed chromatic representation in the P stream reaches visual cortex roughly 20 ms after the M inputs that mainly transmit motion and coarse luminance-based information (Nowak & Bullier, 1997; Nowak, Munk, Girard, & Bullier, 1995). Taking into account this 20 ms delay between the two streams of information, the mask might have little effect on the magnocellular processing of the test image but would strongly interfere with its parvocellular processing. However, magnocellular information may be sufficient to allow good accuracy in the fast categorisation task used here (Delorme et al., 2000; Macé, Thorpe, & Fabre-Thorpe, in press), and the interference of the highly contrasted mask with the feed-forward processing of magnocellular information (Macé et al., in press) may still be significant.

Finally, mask processing could interrupt feedback processing of the stimulus, at least at two levels. Iterative loops are thought to be important for segmentation, and involve the convergence of feedback from higher areas to areas like V1 or V2 (Hupe et al., 1998; Lamme, Super, & Spekreijse, 1998). In such pattern masking experiments, the processing of feedback information is probably made difficult with a mask closely following the stimulus. Moreover, subjects often reported that they released the button without explicit understanding of the photograph, which is in accordance with the common idea that feedback processing may be crucial for conscious image perception (Bullier, 2001; Lamme & Roelfsema, 2000; Pascual-Leone & Walsh, 2001).

These interpretations are not mutually exclusive and could even explain the striking differences between the effect of short SOAs on the initial part of the RT distribution and the plateau effect obtained with the 31 ms SOA (Fig. 3A). In fact, two different kinds of perturbations may be reflected here. The plateau effect may result from disruption of feedback processing related to the detection of the target, or disruption of direct parvocellular inputs. In contrast, the shift observed in the initial part of the RT distribution with shorter SOAs could reflect the disruption of the initial wave of processing.

4.2.2. A pipeline architecture

If we assume that the visual system accumulates sensory information until a decision threshold is reached, the very progressive masking effect over time is another point of interest. Presumably, this sort of task requires information processing at several different levels of the visual system including the retina, LGN, V1, V2, V4 and inferotemporal cortex. We have argued in the past that this sort of fast processing may leave only a short time at each processing stage before the next level has to respond, maybe as little as 10 ms or so (Bullier & Nowak, 1995; Fabre-Thorpe et al., 2001; Nowak et al., 1995; Thorpe & Fabre-Thorpe, 2002). These results confirm that visual processing can rely on such short latencies and challenges traditional views that use firing rate codes to convey information (Thorpe, Delorme, & Van-Rullen, 2001; VanRullen & Thorpe, 2001b; VanRullen & Thorpe, 2002). Further investigations will be necessary to understand how visual processing can be performed in such temporally constrained conditions. In the case of a serial model of information transmission, we might have expected sharply contrasted responses in which performance and ERPs are strongly affected below the decision threshold and less disrupted above this threshold, although the averages of performance across trials and subjects may obscure some types of more discrete transition (Miller, 1988). In contrast, in the case of a continuous model, masking effects may be progressively lessened with increased SOA. Our data showed that as SOA is increased above 12 ms, subjects

gradually increased their ability to detect an animal in the picture, which suggests that information can be conveyed to extrastriate areas in a continuous and asynchronous way. The lack of a clear threshold may also indicate that the information does not need to be fully processed at each stage (that is for all points of the space at the same time), but could be forwarded to the next stage and processed more progressively. Although the results do not argue conclusively in favour of a continuous model, they nevertheless strongly suggest that information transfer occurs progressively at each stage using a form of pipeline architecture.

A final point concerns the nature of the information used to perform the animal/non-animal task. While, in principle, we think that this should be considered as a true high-level visual task, there have been suggestions that even relatively high-level categorisations such as "natural vs man-made scenes" can be made on the basis of relatively low-level information. For example, Torralba and Oliva have reported that a linear combination of the outputs of a series of orientation and spatial frequency tuned channels can allow performance at over 80% correct (Torralba & Oliva, 2003). We certainly cannot exclude the possibility that our subjects are using this sort of information. However, so far at least, none of these purely low-level strategies has succeeded in achieving performance levels of above 90%, nor have they been used to differentiate between classes of objects. We therefore feel that other more complex visual processing strategies are probably at work. The current set of experiments does not allow us to distinguish between these possibilities. Nevertheless, they do demonstrate that, whatever the nature of the information used to perform the task, it is information that the visual system can extract extremely rapidly.

Acknowledgements

This work was supported by the CNRS, the ACI Integrative and Computational Neuroscience. Financial support was provided to both N. Bacon-Macé and M.J-M. Macé by a Ph.D. grant from the French government. We thank G.A. Rousselet for technical assistance and helpful discussions on the results of the experiment.

Supplementary data

Correlation between behavioural accuracy and the amplitude of the occipital differential activity, on the right hemisphere occipital electrodes. The values shows individual correlation calculated with a Pearson test ($p < 0.01$) between behavioural accuracy and the differential activity amplitude on five occipital electrodes. Differential activity amplitude was determined by the most negative point between 150 ms and 250 ms on averaged signals by condition, for each subject. The bottom line indicates an even stronger r-value by averaging the parameters for all 16 subjects before correlating them. Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.visres.2005.01.004.

References

Albrecht, D. G., Geisler, W. S., Frazor, R. A., & Crane, A. M. (2002). Visual cortex neurons of monkeys and cats: temporal dynamics of the contrast response function. *Journal of Neurophysiology, 88*(2), 888–913.

Albrecht, D. G., & Hamilton, D. B. (1982). Striate cortex of monkey and cat: contrast response function. *Journal of Neurophysiology, 48*(1), 217–237.

Breitmeyer, B. G. (1984). *Visual masking: an integrative approach* (p. 454). Oxford, New York: Oxford University Press.

Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S., & Movshon, J. A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Visual Neuroscience, 13*(1), 87–100.

Bullier, J. (2001). Feedback connections and conscious vision. *Trends in Cognitive Sciences, 5*(9), 369–370.

Bullier, J., & Nowak, L. G. (1995). Parallel versus serial processing: new vistas on the distributed organization of the visual system. *Current Opinion in Neurobiology, 5*(4), 497–503.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition, 79*(1–2), 1–37.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. *Vision Research, 40*(16), 2187–2200.

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in Cognitive Sciences, 4*(9), 345–352.

Eriksen, C. W., & Schultz, D. W. (1979). Information processing in visual search: a continuous flow conception and experimental results. *Perception and Psychophysics, 25*(4), 249–263.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*(2), 171–180.

Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans. A framework for defining "early" visual processing. *Experimental Brain Research, 142*(1), 139–150.

Gold, J. I., & Shadlen, M. N. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature, 404*(6776), 390–394.

Grill-Spector, K., Kushnir, T., Edelman, S., Itzchak, Y., & Malach, R. (1998). Cue-invariant activation in object-related areas of the human occipital lobe. *Neuron, 21*(1), 191–202.

Grill-Spector, K., Kushnir, T., Hendler, T., & Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nature Neuroscience, 3*(8), 837–843.

Hasbroucq, T., Burle, B., Bonnet, M., Possamai, C. A., & Vidal, F. (2002). Dynamique du traitement de l'information sensorimotrice: apport de l'electrophysiologie. *Canadian Journal of Experimental Psychology, 56*(2), 75–97.

Hupe, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P., & Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature, 394*(6695), 784–787.

Keysers, C., & Perrett, D. I. (2002). Visual masking and RSVP reveal neural competition. *Trends in Cognitive Sciences, 6*(3), 120–125.

Kim, J. N., & Shadlen, M. N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience, 2*(2), 176–185.

Kovacs, G., Vogels, R., & Orban, G. A. (1995). Cortical correlate of pattern backward masking. *Proceedings of National Academic Science USA, 92*(12), 5587–5591.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neuroscience, 23*(11), 571–579.

Lamme, V. A., Super, H., & Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology, 8*(4), 529–535.

Leopold, D. A., & Logothetis, N. K. (1996). Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature, 379*(6565), 549–553.

Macé, M. J.-M., Thorpe, S. J., & Fabre-Thorpe, M. (in press). Rapid categorisation of achromatic natural scenes: how robust at very low contrasts? *European Journal of Neuroscience*.

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., et al. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of National Academic Science USA, 92*(18), 8135–8139.

McClelland, J. L. (1979). On the time relations of mental processes: an examination of systems of processes in cascade. *Psychological Review, 86*(4), 287–330.

Miller, J. (1988). Discrete and continuous models of human information processing: theoretical distinctions and empirical results. *Acta Psychologica, 67*(3), 191–257.

Nowak, L. G., & Bullier, J. (1997). The timing of information transfer in the visual system. In K. S. Rockland, J. H. Kaas, & A. Peters (Eds.). *Extrastriate visual cortex in primates* (Vol. 12, pp. 205–241). New York: Plenum Press.

Nowak, L. G., Munk, M. H., Girard, P., & Bullier, J. (1995). Visual latencies in areas V1 and V2 of the macaque monkey. *Visual Neurosciences, 12*(2), 371–384.

Pascual-Leone, A., & Walsh, V. (2001). Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science, 292*(5516), 510–512.

Reich, D. S., Mechler, F., & Victor, J. D. (2001). Temporal coding of contrast in primary visual cortex: when, what, and why. *Journal of Neurophysiology, 85*(3), 1039–1050.

Rolls, E. T., Tovee, M. J., & Panzeri, S. (1999). The neurophysiology of backward visual masking: information analysis. *Journal of Cognitive Neuroscience, 11*(3), 300–311.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5*(7), 629–630.

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision, 3*(6), 440–455.

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2004). Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vision Research, 44*(9), 877–894.

Salzman, C. D., & Newsome, W. T. (1994). Neural mechanisms for forming a perceptual decision. *Science, 264*(5156), 231–237.

Schall, J. D. (2001). Neural basis of deciding, choosing and acting. *National Review of Neuroscience, 2*(1), 33–42.

Sestokas, A. K., & Lehmkuhle, S. (1986). Visual response latency of X- and Y-cells in the dorsal lateral geniculate nucleus of the cat. *Vision Research, 26*(7), 1041–1054.

Shadlen, M. N., & Newsome, W. T. (1996). Motion perception: seeing and deciding. *Proceedings of National Academic Science USA, 93*(2), 628–633.

Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology, 66*(1), 170–189.

Thompson, K. G., & Schall, J. D. (1999). The detection of visual signals by macaque frontal eye field during masking. *Nature Neuroscience, 2*(3), 283–288.

Thorpe, S., Delorme, A., & Van-Rullen, R. (2001). Spike-based strategies for rapid processing. *Neural Network, 14*(6–7), 715–725.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520–522.

Thorpe, S. J., & Fabre-Thorpe, M. (2002). Fast visual processing and its implications. In M. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed.). Cambridge, MA: MIT Press.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network, 14*(3), 391–412.

Tovee, M. J., & Rolls, E. T. (1995). Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. *Visual Cognition, 2*(1), 35–58.

Vanni, S., Revonsuo, A., Saarinen, J., & Hari, R. (1996). Visual awareness of objects correlates with activity of right occipital cortex. *Neuroreport, 8*(1), 183–186.

VanRullen, R., & Thorpe, S. J. (2001a). Is it a bird. Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception, 30*(6), 655–668.

VanRullen, R., & Thorpe, S. J. (2001b). Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. *Neural Computation, 13*(6), 1255–1283.

VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research, 42*(23), 2593–2615.

Vogels, R. (1999a). Categorization of complex visual images by rhesus monkeys. Part1: behavioural study. *European Journal of Neuroscience, 11*(4), 1223–1238.

Vogels, R. (1999b). Effect of image scrambling on inferior temporal cortical responses. *Neuroreport, 10*(9), 1811–1816.

2.5 - 150 ms de traitement ... une surévaluation ?

Les différentes expériences que nous avons menées nous ont finalement permis de trouver des tâches ayant des temps de réaction plus courts que ceux de la catégorisation animal/non animal ; malheureusement au prix d'une grande réduction de leur complexité puisqu'il s'agit de tâches de détection simple et de reconnaissance. Il existe cependant un moyen d'obtenir des temps de réaction bien plus courts en conservant une tâche de difficulté comparable à celle de notre tâche de catégorisation classique. Il faut pour cela changer le mode de réponse du sujet. La main n'est pas l'effecteur le plus rapide qui soit et il est possible d'adapter notre protocole de catégorisation d'images naturelles pour que la réponse motrice soit donnée grâce à un simple mouvement des yeux.

La distance allant du colliculus supérieur (structure qui dirige le mouvement des yeux) jusqu'à l'œil est beaucoup plus réduite que celle allant du cortex moteur à la main et les latences des structures oculomotrices sont toujours très courtes(Busettini *et al.*, 1997 ; Masson *et al.*, 2000). Kirchner et Thorpe ont réalisé cette expérience (Kirchner & Thorpe, 2006) en présentant simultanément à l'écran une cible et un distracteur ; le sujet donnant sa réponse en déclenchant une saccade oculaire du côté de la cible. En utilisant un tel protocole, il est possible de réduire de plus de 100 ms le TR minimal par rapport à une réponse manuelle de type go/no-go (TR min de 120 à 130 ms : Figure 7).

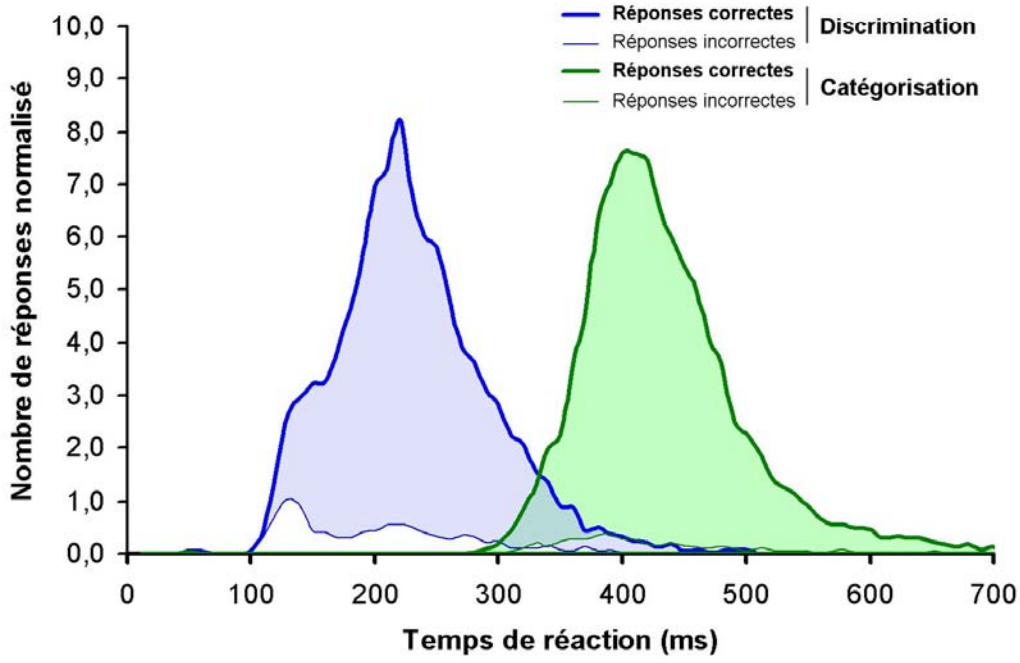


Figure 7 : Distribution des TR manuels dans une tâche de catégorisation (vert) et des latences de saccades oculaires dans une tâche de discrimination (bleu). Les erreurs dans ces deux tâches sont en traits fins respectivement vert et bleu. Avec l'aimable autorisation de Nadège Bacon-Macé, d'après des données de Kirchner et al., Vis Res (2006) et Delorme et al., Cog Brain Res (2004).

Il est étonnant de trouver pour les mouvements oculaires des latences si courtes qu'elles sont pour certaines inférieures à celles observées en électrophysiologie, alors que les EEG permettent déjà de s'affranchir de la composante motrice du temps de réaction. On peut supposer que ces latences de réponse très précoces pour les mouvement oculaires sont explicables par une extraction rapide et une comparaison des caractéristiques bas-niveau des images présentées. Cependant, il semble que les indices utilisés par le système visuel dans cette tâche ne sont pas triviaux puisque les auteurs n'ont pas pu trouver de différences statistiques simples entre les groupes d'images qui permettent de les classer avec une fiabilité élevée.

La précocité des latences observées, outre la grande vitesse du système oculomoteur, s'explique peut-être également par les différences de difficulté entre la tâche "go/no-go" manuelle et la tâche "go" oculaire dans laquelle le sujet sait qu'il y a toujours une cible parmi les deux images présentées. Le sujet n'a pas à comparer une image présentée à l'écran avec des représentations stockées en mémoire, il peut se "contenter" de discriminer parmi deux images celle qui a le plus de chance d'appartenir au groupe des images cibles. Même s'il ne fait pas de doute que les structures cérébrales impliquées dans la catégorisation manuelle et saccadique diffèrent relativement tôt (après V4 ? (Kirchner & Thorpe, 2006)), une étude de Bacon-Macé et al. (Bacon-Macé *et al.*, en révision) utilisant un protocole de masquage montre que les premières étapes de traitement sont probablement communes entre ces deux modes de réponse. Ceci laisse également penser que les tâches sont d'une complexité équivalente en ce qui concerne les premières étapes d'extraction de l'information visuelle.

Il reste néanmoins que cette tâche de discrimination saccadique met en évidence des processus extrêmement rapides aboutissant avant que l'activité différentielle à 150 ms ne soit encore apparue. Le fait que l'activité cérébrale différentielle enregistrée dans les autres protocoles de catégorisation visuelle ne semble pas nécessaire dans cette tâche de discrimination saccadique appelle à de nombreuses interrogations sur l'origine et le rôle de ce signal.

2.6 - Conclusion générale sur la vitesse de traitement

Nous avons vu dans ce chapitre que la latence de l'activité différentielle à 150 ms ne constitue qu'une première étape permettant de cerner tout au plus la durée totale des traitements nécessaires à la tâche et qu'il est possible d'étudier plus finement le traitement visuel rapide en ayant recours à différents protocoles (détection, reconnaissance et masquage) ou à d'autres effecteurs (réponse oculaire). Ces expériences fournissent de précieuses indications sur la vitesse de traitement des informations visuelles et permettent de décomposer plus avant les mécanismes utilisés dans le système visuel pour extraire le sens des images.
Les expériences de détection et de reconnaissance confirment à quel point les processus de catégorisation visuelle rapide sont optimisés en terme de vitesse. Alors que la tâche est grandement simplifiée dans le cas de la reconnaissance d'une cible unique, par rapport à une catégorisation complexe, le TR n'est diminué que de 40 ms chez l'homme et de 20 ms chez le singe. De plus, la totalité des processus aboutissant à la catégorisation n'ajoute que 90 ms de temps de traitement par rapport au simple fait de signaler qu'une image est apparue à l'écran.
L'ensemble des processus de catégorisation nécessite peu de temps, et grâce au masquage, il est possible d'analyser finement le décours temporel des traitements visuels.

Ces très grandes vitesses, à la fois pour le temps total de traitement et pour le temps d'intégration des informations pour les images masquées sont encore des arguments en faveur d'un rôle du système magnocellulaire dans la catégorisation visuelle. Il est en effet plus rapide que le système parvocellulaire pour amener les information visuelles au cortex, mais également plus à même de fournir des informations sur les images masquées de par ses performances supérieures aux fréquences temporelles élevées.

3 - Représentations accessibles avec les informations précoces

Les premières informations visuelles

L'architecture générale du système visuel apparaît optimisée pour que les informations en provenance de la rétine parcourent très rapidement les diverses étapes des voies visuelles. Certains auteurs avancent sur la base de localisations précises de sources EEG (Foxe & Simpson, 2002 ; Di Russo *et al.,* 2001) que l'ensemble de la voie ventrale pourrait être activée en seulement 100 à 120 ms chez l'homme. Les premières informations disponibles dès cette période sont cependant en quantité limitée et la description du monde qu'elles reflètent est rudimentaire et incomplète. Nous avons vu dans les chapitres précédents que ces informations grossières peuvent néanmoins suffire pour effectuer des tâches complexes, comme la catégorisation d'images naturelles, avec un pourcentage de réussite tout à fait satisfaisant.

Les expériences rapportées plus haut sur le rôle négligeable de la couleur dans une tâche de catégorisation ainsi que l'influence modérée de l'excentricité et de la réduction de contraste permettent d'avancer que le système magnocellulaire, dont les informations circulent plus rapidement dans le système visuel, joue un rôle important dans les premiers traitements visuels en fournissant une ébauche de la scène visuelle dans laquelle viennent s'ancrer ultérieurement des informations plus détaillées (Sherman, 1985). A travers ce protocole de catégorisation, nous cherchons à comprendre comment le système visuel extrait les informations de la scène visuelle pour construire une représentation de plus en plus aboutie et quels sont les différents éléments qui permettent une compréhension globale de la scène. Jusqu'à quel niveau de représentation faut-il détailler la scène visuelle pour effectuer une tâche de catégorisation ? Ce niveau de détail est-il le même quel que soit l'objet cible considéré ?

Nous avons vu dans le 1er chapitre que les singes réalisent la tâche avec des performances tout à fait comparables à celles des humains (et même bien meilleures si l'on considère les temps de réaction), comment faut-il interpréter le fait qu'une espèce dépourvue de langage puisse effectuer des tâches de catégorisation supposées impliquer une capacité d'abstraction relativement avancée ? Quelles sont les étapes de traitement qui permettent à leur système visuel de reconnaître et catégoriser des objets ?

Reconnaissance d'objets et catégorisation

Reconnaître qu'un objet est une chaise ou placer un objet dans la catégorie des chaises semble assez similaire. Il s'agit dans les deux cas de répondre à la question : "quel est cet objet ?" en faisant correspondre les informations perçues avec des informations stockées en mémoire. Pourtant ces deux manières de répondre à la même question ont donné lieu dans l'histoire des neurosciences à deux champs d'études largement séparés : la reconnaissance d'objets et la catégorisation. Les relations entre ces deux domaines ont été très limitées au cours de leur développement respectif, probablement à cause de différences de point de vue : la reconnaissance d'objets s'appuie exclusivement sur la perception alors que la construction des catégories requiert d'autres informations comme la fonction de l'objet ou le contexte dans lequel il se situe. Cette dissociation entre deux domaines si proche est regrettable et le rapprochement qui s'est effectué ces dernières années ne saurait être que bénéfique pour chacun d'eux (Schyns, 1998 ; Palmeri & Gauthier, 2004). L'exemple le plus évident est peut être sur le plan théorique : les modèles de reconnaissance d'objets sont bien plus nombreux et plus aboutis que les modèles de catégorisation. Étant donné que dans nombre de ces modèles, la catégorisation peut découler d'une "simple" généralisation de l'objet à reconnaître, il est possible de bénéficier des avancées théoriques de la reconnaissance d'objets pour étudier la catégorisation. Lorsque nous présenterons différents modèles de reconnaissance d'objets, nous préciserons quels sont ceux qui possèdent de bonnes aptitudes à la catégorisation.

Humphreys (Humphreys *et al.*, 1999) propose comme architecture générale de la reconnaissance d'objets un système à plusieurs étapes passant par une description structurale, puis une représentation sémantique, en interaction avec les représentations mnésiques pour enfin aboutir à une représentation lexicale (Figure 1). L'encodage structural de l'image est bien sûr effectué par les premières aires de la vision, alors que l'extraction sémantique serait davantage du ressort du cortex inféro-temporal, du cortex périrhinal et de la région hippocampique.

Catégoriser ou identifier un objet consiste alors à faire coïncider sa description structurale (ie la représentation perceptuelle construite à partir des informations externes) avec les représentations stockées en mémoire lui correspondant. C'est cette mise en relation des deux représentations qui permet d'associer l'objet perçu à une classe précise d'objets connus. Le long de la voie ventrale, les différentes aires visuelles contiennent des cellules dont les réponses sont sélectives à des formes de plus en plus complexes et la représentation qu'elles encodent est donc elle aussi de plus en plus élaborée. Les représentations construites

précocement sur la base des informations magnocellulaires sont probablement utilisées dans le système visuel pour commencer le traitement des images sans attendre le transfert de la totalité de l'information en provenance de la rétine.

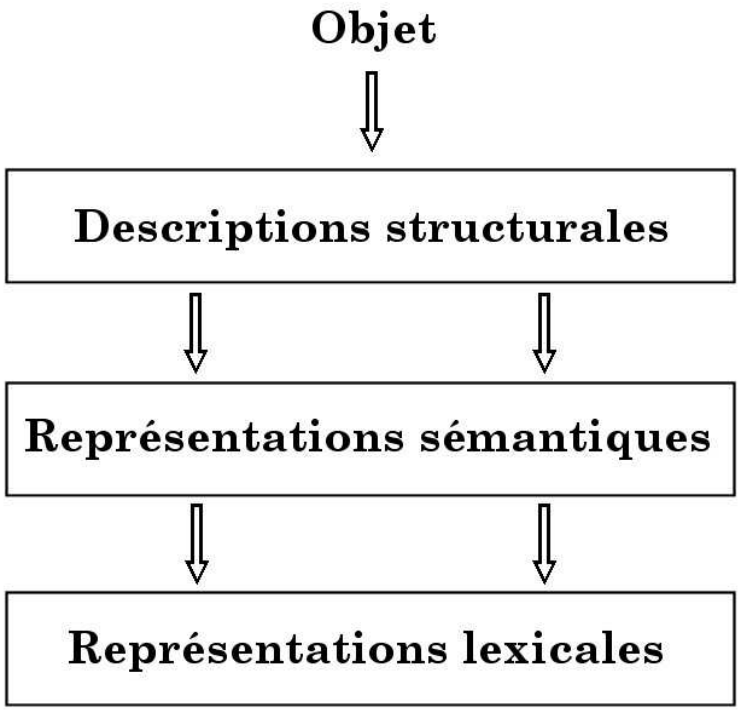


Figure 1 : Représentation schématique des étapes de traitement nécessaires pour nommer un objet. Reproduit d'après Humphreys et al., Psychol Res (1999).

Nous avons vu que la catégorisation visuelle est très robuste à des dégradations d'images. Cette caractéristique du système visuel permet de préserver des performances élevées dans des conditions de vision dégradées que l'on rencontre dans la vie courante : contraste/luminance extrêmes (brouillard et pénombre), parties des objets occultées, vue atypiques, etc. Les représentations visuelles sont de plus en plus abstraites le long des voies visuelles et l'indépendance de plus en plus grande entre les réponses des neurones et les caractéristiques physiques de l'image permettent cette souplesse d'analyse nécessaire dans bien des situations.

3.1 - Modèles de la reconnaissance d'objets

De nombreux modèles ont été proposés pour expliquer comment le système visuel reconnaît les objets. La puissance d'une théorie est validée à la fois par l'ensemble des faits qu'elle explique, mais aussi par les prédictions dont elle est capable. Les théories expliquant la manière dont le cerveau élabore les représentations internes des objets et la forme de ces

représentations permettent de faire des prédictions sur le fonctionnement du système visuel qu'il est ensuite possible de vérifier grâce à divers protocoles expérimentaux.

Le paysage des théories computationnelles de la reconnaissance d'objets est particulièrement vaste, mais nous allons nous limiter ici à quelques grandes catégories : la décomposition structurale (Marr, 1982 ; Biederman, 1987), l'évaluation des contraintes géométriques (Ullman, 1998) et l'analyse par vues ou par éléments de l'image (Edelman (Edelman & Duvdevani-Bar, 1997), Wallis et Rolls (Wallis & Rolls, 1997), Gautrais et Thorpe (Gautrais & Thorpe, 1998), Riesenhuber et Poggio (Riesenhuber & Poggio, 1999)).

3.1.1 - Théorie de Marr

Les premiers essais de programmes informatiques de reconnaissance d'objets et d'analyse de scènes ont eu lieu dans les années 70 (une revue de ces premiers essais dans Mackworth (Mackworth, 1972)). Ces programmes s'appuyaient sur des algorithmes de détection de contours rudimentaires qui étaient combinés en structures de plus en plus complexes selon des règles formelles. Ces algorithmes étaient si imprécis qu'ils ne pouvaient finalement que traiter des images pré-segmentées dans lesquelles les traits et les angles étaient rendus explicites. Les faibles résultats de ces premiers essais de vision artificielle expliquent l'effervescence de la communauté des neurosciences au début des années 1980 devant les avancées théoriques effectuées par David Marr, l'un des pères des neurosciences computationnelles. Ses articles sur le cervelet et le néocortex avaient déjà marqué les esprits, mais son ouvrage posthume sur la vision (Marr, 1982) a initié une véritable révolution en étant le premier à proposer une théorie complète de la reconnaissance d'objets. Marr part du principe que le but des traitement visuels est de reconstruire des représentations 3D centrées sur l'objet, donc indépendantes du point de vue, à partir des objets perçus en 2D par chaque rétine. Ces représentations 3D peuvent alors être manipulées pour les faire correspondre, par translations, rotations ou inversions à des modèles internes des objets stockés en mémoire. Au début de la reconstruction, seules les caractéristiques les plus élémentaires de l'image rétinienne sont traitées, comme les variations de luminance qui déterminent les bords des objets et leur forme. Après deux niveaux successifs de cette représentation schématique, c'est une représentation dite en "2½D" qui est calculée en utilisant les autres indices présents dans la scène visuelle comme la texture, les ombrages, les mouvements et la disparité binoculaire. Ce n'est qu'après ces différentes étapes que la reconstruction finale en 3D de la scène visuelle peut avoir lieu, en calculant la profondeur et l'orientation de chaque surface présente dans la scène.

3.1.2 - Théorie des géons

C'est sur ce cadre théorique de reconstruction géométrique très stricte de l'espace que s'appuiera Biederman dans les années 80 pour élaborer sa théorie sur la reconnaissance par composant (Biederman, 1987). Selon cette hypothèse, les objets sont tous constitués par l'assemblage plus ou moins complexe d'un nombre réduit de primitives volumiques (cylindre, cube, cône, pyramide…) baptisées géons. Pour reconnaître un objet, le système visuel doit auparavant segmenter la scène, puis identifier tous les géons qui constituent un objet sur la base de leurs propriétés non-accidentelles (propriétés visuelles invariantes en 3D) afin d'accéder à une représentation sémantique par regroupement des différents géons constitutifs des objets. On perçoit très nettement les analogies qui existent entre cette théorie et le langage, avec les géons qui jouent le rôle d'un alphabet des formes visuelles pour constituer tous les objets par combinaison, tels les lettres qui constituent des mots. Hummel et Biederman (Hummel & Biederman, 1992) ont proposé une implémentation de leur modèle dans laquelle les objets sont analysés au travers de sept couches successives (Figure 2). L'implémentation de ce modèle ne s'inspire que très peu des traitements effectivement réalisés par les neurones ou de l'architecture globale du système visuel. Les capacités de reconnaissance de ce réseau restent très limitées, mais ce modèle de reconnaissance d'objets garde un grand intérêt historique en étant le premier implémenté formellement.

Difficultés pratiques

Les deux théories de la reconnaissance d'objets présentées ci-dessus sont de type bottom-up et nécessitent un nombre important d'étapes, dont un passage par une segmentation de l'image, avant que la reconstruction des objets de la scène visuelle et leur interprétation ne puisse avoir lieu. Cette multiplicité des étapes de traitement, même si elle reste compatible avec la vitesse de reconnaissance des objets dans certaines implémentations, constitue un véritable désavantage par rapport aux autres théories.

Les nombreux modèles de la reconnaissance d'objets apparus à l'époque, tous plus ou moins inspirés des idées de Marr, ont donné naissance à bon nombre de systèmes de vision artificielle. Malgré les importants efforts consentis pour améliorer l'algorithmique et l'implémentation de ces modèles, les performances des systèmes de vision artificielle sont restées très basses, sans qu'il n'y ait beaucoup d'espoir sur leurs évolutions futures. Les modèles géométriques ne parvenaient à de bons résultats que pour des tâches de détection ou d'identification dans des environnements contrôlés. L'ouverture à des images naturelles ou des tâches de catégorisation conduisait immanquablement à des performances catastrophiques.

Ces résultats décevants s'expliquent par la difficulté inhérente à l'extraction des différentes parties d'un objet dans une photographie. Et même lorsque l'image d'entrée est extrêmement simple, la décomposition en composantes fondamentales est sujette à une grande instabilité. En effet, quel que soit le mode de décomposition du signal choisi, on se trouve toujours face à un large choix entre plusieurs possibilités équiprobables (par exemple, un simple A peut être découpé en 3 ou 5 segments). Pour des objets complexes, cette instabilité dans la résolution des parties composantes se traduit par une véritable explosion combinatoire dans le nombre de solutions des décompositions possibles.

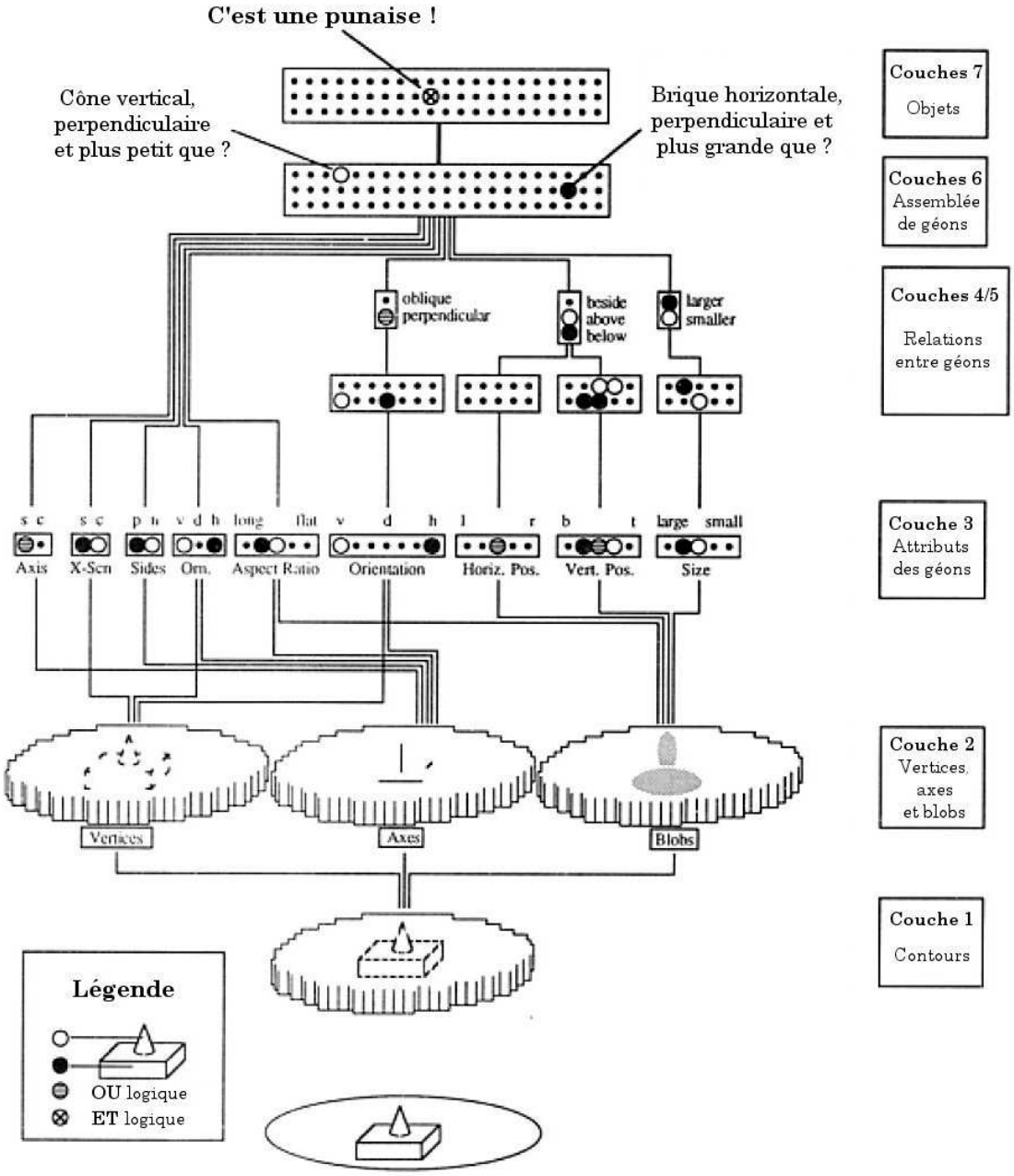


Figure 2 : Organisation hiérarchique du modèle de reconnaissance d'objets d'Hummel et Biederman. La première couche détecte les contours et la deuxième les coins, les axes et les surfaces. Les couches suivantes extraient les éléments géométriques de bases (géons), déterminent leurs relations spatiales et les combinent pour détecter les objets. Reproduit d'après Hummel et al., Psychol Rev (1992).

3.1.3 - Modèle d'Ullman

Ullmann (Ullman, 1998), lui aussi dans le cadre théorique d'une reconstruction géométrique des objets, s'appuie sur d'autres propriétés invariantes de l'image (3D) pour construire une représentation du monde environnant. L'idée la plus intéressante dans le modèle d'Ullman concerne le flux de données à l'intérieur du système visuel. L'appariement entre les objets reconstruits depuis l'entrée du système et les modèles stockés en mémoire ne se ferait pas obligatoirement en haut de l'architecture visuelle, mais à des niveaux intermédiaires en fonction de la complexité des objets à traiter et de leur degré de préactivation. Les représentations des objets dont la présence est la plus probable dans la scène seraient sélectionnées par une première vague de traitement rapide à travers le système grâce à des projections directes vers des aires intermédiaires (V1-V4 / V2-IT). L'activation de chacun de ces modèles pourrait générer un grand nombre de variations à partir de la vue stockée. Ces variations seraient propagées dans le système visuel par des voies descendantes en direction des représentations en cours de construction. Des algorithmes d'alignements de vues et de comparaisons seraient alors utilisés pour trouver la meilleure correspondance entre les représentations construites à partir de la perception et les représentations construites à partir des modèles. On peut souligner tout de suite un important problème qui apparaît avec ce modèle : il est théoriquement impossible de reconnaître un objet du monde si son modèle n'a pas été pré-activé puis propagé à travers le réseau. Une autre difficulté survient si l'on fait l'hypothèse que certaines situations induisent une préactivation massive d'un grand nombre de modèles (foule, environnement complexe ou changements rapides) qui pourraient engendrer une confusion totale dans le système.

La théorie d'Ullman souffre des mêmes problèmes que celle de Biederman en ce qui concerne l'extraction de l'information dans l'image de départ. En effet, l'extraction des éléments caractéristiques dans une image naturelle est très difficile et limite cette méthode de reconnaissance d'objets à des environnements très contrôlés. Le modèle d'Ullman est également mis en défaut sur le plan de la catégorisation. En effet, l'alignement géométrique précis de l'objet perçu et du modèle ne favorise pas l'abstraction des détails de l'image qui permet de généraliser la reconnaissance à toute une catégorie d'objets ou à des objets nouveaux.

3.1.4 - Remise en cause de l'invariance à la vue et de la reconstruction géométrique

Les premières expériences de psychologie visant à prouver la validité des modèles de reconstruction géométrique des objets ont semblé encourageantes. La principale prédiction imposée par l'architecture des modèles de reconnaissance d'objets présentés ci-dessus est que la vitesse de reconnaissance doit être indépendante de l'orientation de l'objet. Cette propriété découle de la nature même de ces modèles dans lesquels l'objet est reconstruit selon une vue centrée sur lui. Les premiers résultats obtenus par Biederman et Gerhardstein (Biederman & Gerhardstein, 1993) allaient dans ce sens, mais ils ont été rapidement critiqués pour avoir utilisé des objets familiers. Le choix des objets présentés s'avère en effet crucial et l'utilisation d'objets nouveaux a régulièrement montré que l'invariance aux changements de point de vue est plutôt l'exception que la règle aussi bien pour des rotations dans le plan qu'en profondeur (Bulthoff & Edelman, 1992 ; Edelman & Bulthoff, 1992 ; Tarr *et al.*, 1998). Des résultats électrophysiologiques chez le singe sont venus renforcer cette idée en montrant que les neurones de IT ne sont pratiquement jamais invariants aux rotations et aux changements d'illumination dans la scène (Logothetis & Pauls, 1995 ; Perrett *et al.*, 1991). Les seules invariances observées sont limitées à des changements de faible amplitude ou à des objets bien connus du sujet. Des objets familiers sont effectivement reconnus indépendamment du point de vue (Booth & Rolls, 1998), probablement parce qu'il est possible de combiner les réponses de plusieurs neurones sensibles au point de vue obtenus dans un premier temps pour construire la spécificité des neurones indépendants du point de vue.
Ces résultats ont été répliqués en psychophysique (Bulthoff *et al.*, 1995) et en IRMf chez l'homme (Tanaka, 1997 ; Grill-Spector *et al.*, 2001).

En plus de ces remises en cause expérimentales sont apparues des critiques plus théoriques sur le postulat principal des théories de Marr de la reconnaissance d'objets ; à savoir que la seule représentation interne véritablement exploitable pour le système visuel doit être une reconstruction géométrique fidèle du monde extérieur. Cette idée peut sembler intellectuellement séduisante, mais la meilleure représentation est celle qui est adaptée à la tâche à effectuer et elle n'est pas forcément similaire à ce qui se trouve dans le monde où à la représentation à laquelle il est possible d'accéder consciemment. De plus, ce n'est pas parce qu'une forme a été reconstruite en un modèle interne qu'elle est *reconnue* au sens propre du terme puisque cette reconstruction ne résout pas en elle-même la question de *l'interprétation* de l'image.

3.1.5 - Modèles de reconnaissance par indices ou par vues

Les fissures apparues dans le dogme d'une représentation interne qui serait une sorte de copie idéalisée de la réalité ont permis aux modèles fondés sur les indices visuels ou sur les vues des objets de se faire une place à partir des années 90. Les modèles proposés par Thorpe et Gautrais (propagation asynchrone de l'information, Gautrais & Thorpe, 1998), Vetter, Hurlbert et Poggio (réseau de régularisation, Vetter *et al.*, 1995), Edelman (reconnaissance par vues, Edelman & Duvdevani-Bar, 1997), Perrett (Perrett *et al.*, 1998) ou Riesenhuber et Poggio (Riesenhuber & Poggio, 2002) ont pour principal point commun de fonctionner en utilisant une reconnaissance par indices visuels ou par vues (souvent partielle) des objets.

Les implémentations de ces modèles sont très diverses, mais leurs grands principes de fonctionnement sont les mêmes du point de vue des traitements. La première étape consiste à reconnaître des parties ou des vues d'objets en utilisant des filtres de Gabor ou des détecteurs ayant des propriétés similaires à ceux des champs récepteurs des neurones rétiniens (circulaires ON/OFF) ou corticaux (détections de bords et d'angles). Des détecteurs plus complexes peuvent être construits en combinant les sorties de plusieurs cellules simples, à la manière de ce qu'Hubel et Wiesel proposaient en leur temps (Hubel & Wiesel, 1962) pour V1.

3.1.6 - Modèle de Thorpe et Gautrais : codage par rang

Si nous prenons pour exemple le modèle avancé par Thorpe et Gautrais (Gautrais & Thorpe, 1998), l'image est traitée par une rétine et une couche corticale correspondant à V1 (en très simplifiée) avant que l'information ne soit propagée vers des couches encodant une représentation de la scène plus abstraite (Figure 3). Ces auteurs avaient constaté expérimentalement que le système visuel est trop rapide pour fonctionner avec un codage de type fréquentiel. Dans ce modèle, le code utilise la latence relative de décharge des neurones pour transférer de l'information. Ainsi le premier potentiel d'action qui arrive dans le système visuel donne comme information que cette zone de l'image contient un fort contraste de luminance, probablement parce qu'elle contient un bord. Les couches suivantes peuvent se comporter comme de simples détecteurs de parties d'objets ou d'objets entiers. Les avantages principaux de ce modèle sont sa très grande vitesse, son efficacité computationnelle et enfin sa robustesse aux changements de contraste, de luminance, à l'addition de bruit, etc. Ce modèle est en revanche relativement sensible aux variations de taille, de rotation et de position puisqu'il ne contient pas de couches intermédiaires pour s'abstraire de ces changements. Il est cependant possible de générer automatiquement un grand nombre de vues d'un objet afin qu'il soit reconnu par le réseau dans n'importe quelle condition.

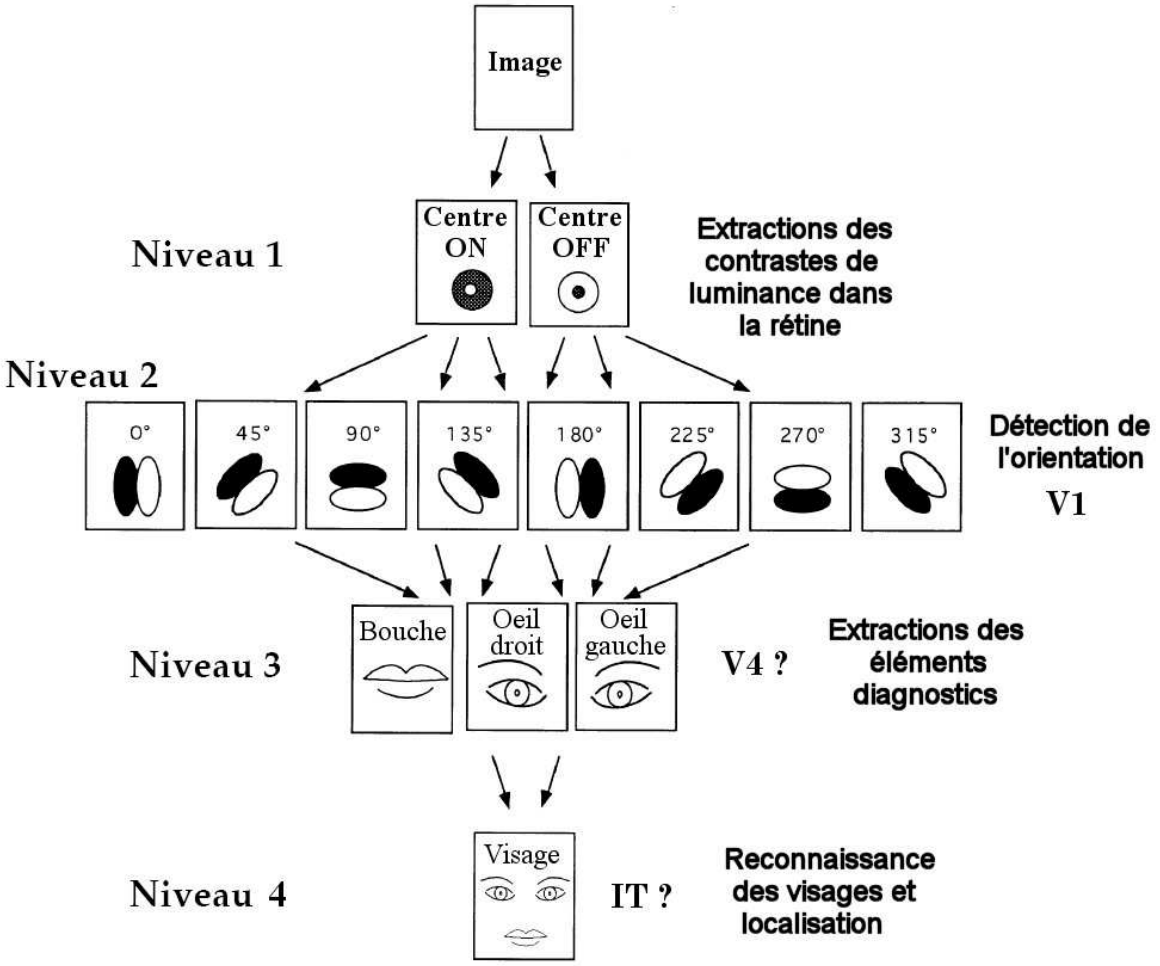


Figure 3 : Organisation hiérarchique du modèle de Thorpe et Gautrais (baptisé Spikenet). Les couches 1 et 2 correspondent assez bien aux traitements effectués par la rétine et par V1 dans le système visuel. Les couches 3 et 4 peuvent être vues comme des analogues des aires V4 et IT, bien qu'éloignées de leur contrepartie biologique. Reproduit d'après les thèses d'Arnaud Delorme et Rufin VanRullen (2000).

3.1.7 - Modèle de Riesenhuber et Poggio

Nous prendrons comme second exemple des modèles de reconnaissance par vues celui de Riesenhuber et Poggio (Riesenhuber & Poggio, 2002) qui fait une synthèse de différents modèles comme celui de Wallis et Rolls et celui d'Edelman (Figure 4). Contrairement au modèle de Thorpe et Gautrais, qui est plutôt un modèle de reconnaissance par éléments de l'image, le modèle de Riesenhuber et Poggio est un modèle par vue, comme celui d'Edelman. Les détecteurs restent très simples en entrée du système (fonction de Gabor, cellules simples et complexes). Les éléments détectés dans une image sont confrontés grâce à des réseaux de régularisation à des vues stockées en mémoire sous la forme de RBF (Radial Basis Function).

Cette transformation mathématique permet de mémoriser des vues des objets en ayant des possibilités de généralisation et d'interpolation pour les faire correspondre aux entrées visuelles (Poggio & Girosi, 1990 ; Poggio & Edelman, 1990). Le recours à ces fonctions de transformation pour l'encodage en mémoire permet d'obtenir aussi bien une identification qu'une catégorisation en fonction de la largeur donnée à la fonction de base choisie. Ces fonctions assurent également une certaine invariance de la réponse pour des translations, des variations de taille et des rotations de l'image.

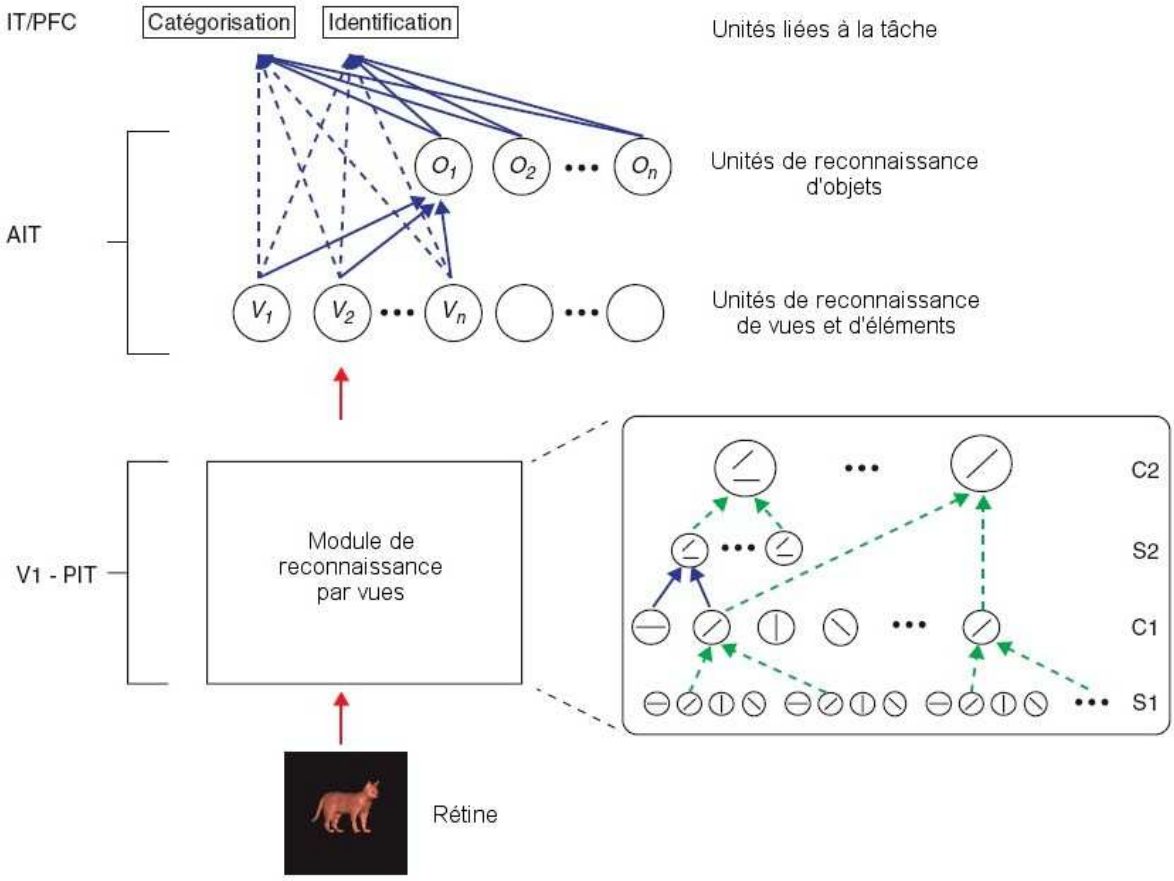


Figure 4 : Organisation hiérarchique du modèle de Riesenhuber et Poggio. Les traitements visuels dans le premier bloc (V1-PIT) permettent de maintenir une bonne sélectivité (flèches bleues) tout en permettant une invariance à la position et à l'échelle (ce qui reproduit l'augmentation de taille des champs récepteurs dans le système visuel). Les unités activées en sortie de ce module de reconnaissance sont combinées pour répondre de manière adéquate à la tâche demandée. Reproduit d'après Riesenhuber et al., Curr Opin Neurobiol (2002).

Nous pouvons terminer cette présentation succincte des modèles de reconnaissance des objets en citant des méthodes d'analyse statistiques des images qui obtiennent de bonnes performances en catégorisation pour une complexité calculatoire minimale. C'est le cas par exemple de la méthode des histogrammes (Schiele & Crowley, 1996 ; Mel, 1997 ; Carson *et al.*, 1997) qui s'appuie sur la distribution des contrastes de luminances ou de couleur dans les

images (sans tenir compte de leur organisation spatiale). De la même manière, Oliva et Torralba (Oliva & Torralba, 2001) utilisent une autre méthode statistique appliquée à toute l'image pour catégoriser des scènes naturelles. Grâce à l'analyse de la distribution des fréquences spatiales (l'enveloppe spatiale), leur modèle de catégorisation est capable d'obtenir une précision de 80% lors d'une tâche de classification du contexte des images. Ces modèles sont intéressants puisqu'ils permettent d'obtenir des performances élevées grâce à des calculs statistiques simples et sans faire appel à une représentation de l'objet. Ils ne résolvent cependant pas le problème de l'interprétation de la scène visuelle et montrent leurs limites lorsque les images deviennent ambiguës ou très complexes. Ils ne permettent pas non plus de savoir où se trouve l'objet dans l'image puisque les informations spatiales ne sont pas conservées lors des calculs.

3.1.8 - Avantages et inconvénients des modèles par vues

Un avantage important des modèles qui ne nécessitent pas de reconstruction explicite de la scène visuelle est qu'elles sont en général très rapides, parfois même implémentables en une seule passe à travers le système visuel. Ces modèles de reconnaissance d'objets sont aussi très robustes au bruit, à la réduction de contraste, etc. Les occlusions sont également bien gérées dans ces modèles puisque de très nombreux éléments diagnostiques ou de très nombreuses vues de chaque objet sont stockés, ce qui permet d'assurer une reconnaissance de l'objet même quand une partie de celui-ci est masquée (par le biais également de la réactivation complète d'un réseau nerveux partiellement stimulé).

Les théories de reconnaissance par vues/indices visuels souffrent comme toutes les autres théories de quelques problèmes computationnels. Le nombre d'indices visuels traités et de vues stockées en mémoire déterminent le nombre de dimensions dans l'espace de recherche que l'on va utiliser. Lorsque ces nombres sont trop importants, on assiste très vite à une explosion combinatoire.
Les implémentations qui ont été faites à partir de ces modèles étaient à l'origine limitées, mais des avancées importantes dans la manière de choisir les indices diagnostiques dans une tâche donnée (et donc de réduite le nombre de dimensions de l'espace de recherche) et l'augmentation des capacités de calcul ont rapidement permis à ces modèles d'obtenir des résultats très impressionnants en reconnaissance d'objets, *même avec des images naturelles.*

3.1.9 - Identification et catégorisation dans les modèles par vues

Gauthier, Tarr et al., (Gauthier *et al.*, 1997) ont montré que l'identification et la catégorisation sont aux deux extrémités d'un continuum de reconnaissance qui peut être effectué par un seul système de traitement. Les modèles à bases d'indices visuels sont en général plus à même de rendre compte à la fois des capacités de catégorisation et d'identification (discrimination fine) en faisant varier les paramètres dépendants que sont la sélectivité et l'invariance (compromis sensibilité/stabilité pour Marr). Il est délicat d'obtenir à la fois une représentation qui permet d'identifier précisément un objet et de reconnaître les objets qui lui sont proches comme appartenant à la même catégorie. C'est pourtant un enjeu fondamental à prendre en compte pour notre compréhension de la reconnaissance d'objets par le système visuel. Les exigences contradictoires de l'identification et de la catégorisation sont peut-être résolues en faisant appel à d'autres structures que le cortex visuel. Freedman (Freedman *et al.*, 2001) a par exemple enregistré des neurones dans le cortex préfrontal répondant à la présentation de catégories précises (chien/chat). Le cortex préfrontal, très impliqué dans la compréhension des règles et la planification peut également jouer un rôle dans la pré-activation des représentations visuelles intermédiaires grâce à des connexions descendantes vers les aires visuelles supérieures (AIT et cortex périrhinal). Des connexions symétriques existent également puisqu'il est possible de déplacer de manière flexible les frontières des catégories en contrôlant le contenu des images à catégoriser (Freedman *et al.*, 2002).

Parallèlement aux expériences sur les rotations 2D et 3D des objets ainsi que sur l'occlusion de leurs différentes parties afin de valider les prédictions des modèles de la reconnaissance d'objets, une autre approche consiste à manipuler les images afin de déterminer plus finement la nature et le contenu des représentations visuelles utilisées par le système visuel. En manipulant les indices présents dans un stimulus, il est possible de cibler différents niveaux de l'architecture visuelle. Par exemple, dans des expériences de "visual search" dans lesquelles des sujets doivent retrouver une cible précise parmi un ensemble de distracteurs, la recherche d'une barre orientée, pour laquelle des cellules de V1 sont sélectives est probablement effectuée à un niveau plus élémentaire que la recherche d'un visage donné pour lequel les premières réponses sélectives ne sont retrouvées qu'en IT.

Nous avons choisi une autre approche dans les expériences présentées ci-dessous. Les images présentées en entrée sont toujours les mêmes, mais la tâche dans laquelle est engagée le sujet varie. En modifiant le niveau de catégorisation de l'image (par exemple en passant du niveau

superordonné au niveau de base), il est possible de changer la difficulté de la catégorisation et peut être de faire varier les indices diagnostiques utilisés pour répondre, ou encore les niveaux de l'architecture visuelle impliqués.

3.2 - Comparaison entre niveaux de catégorisation : article n°5

Les travaux effectués dans les années 60-70 sur les catégories ont montré que les catégories s'organisent à la fois sur un plan horizontal et un plan vertical. Le plan horizontal correspond aux différentes catégories entre elles (chats, voitures, chaises…) et le plan vertical aux différents niveaux de catégorisation (Ferrari, voitures, moyens de transport) (Figure 5). De nombreuses expériences ont montré que les différents niveaux de catégorisation ne sont pas tous équivalents et qu'il existerait un niveau de catégorisation privilégié appelé "niveau de base". Les sujets sont plus rapides pour catégoriser des objets à ce niveau particulier alors que l'accès à un niveau de catégorie plus large (superordonné) ou plus restreint (subordonné) prend plus de temps. Le niveau de base serait ainsi le premier niveau auquel accède le sujet. L'accès au niveau superordonné résulterait d'une généralisation (abstraction…) à partir du niveau de base et l'accès au niveau subordonné résulterait d'une discrimination plus fine (plus figurative) utilisant des informations visuelles supplémentaires. Cette architecture sur le plan vertical est toutefois remise en cause dans le cas d'objets atypiques dans leur catégorie (les manchots parmi les oiseaux par exemple) qui sont catégorisés à la même vitesse au niveau de base et au niveau subordonné (Jolicoeur *et al.*, 1984b), le niveau d'entrée pouvant donc être parfois trouvé au niveau subordonné...

Nous avons vu dans les expériences présentées ci-dessus que les sujets sont à la fois précis et très rapides pour effectuer une tâche de catégorisation de type animal/non animal et que cette rapidité pose déjà d'importants problèmes pour la plupart des modèles de reconnaissance d'objets. Or, la catégorie "animal" est par essence une catégorie superordonnée et les résultats communément admis sur l'architecture des catégories permettent de faire l'hypothèse que les sujets pourraient donc être encore plus rapides pour une tâche de catégorisation effectuée au niveau de base (par exemple, oiseau/non oiseau ou chien/non chien). Cependant, lorsque l'on considère les contraintes déjà imposées par la vitesse du système visuel dans la tâche superordonnée un nouveau gain temporel pourrait s'avérer particulièrement difficile à expliquer. Dans l'ensemble des expériences présentées ci-dessous, nous comparons directement la vitesse de catégorisation au niveau superordonné et au niveau de base.

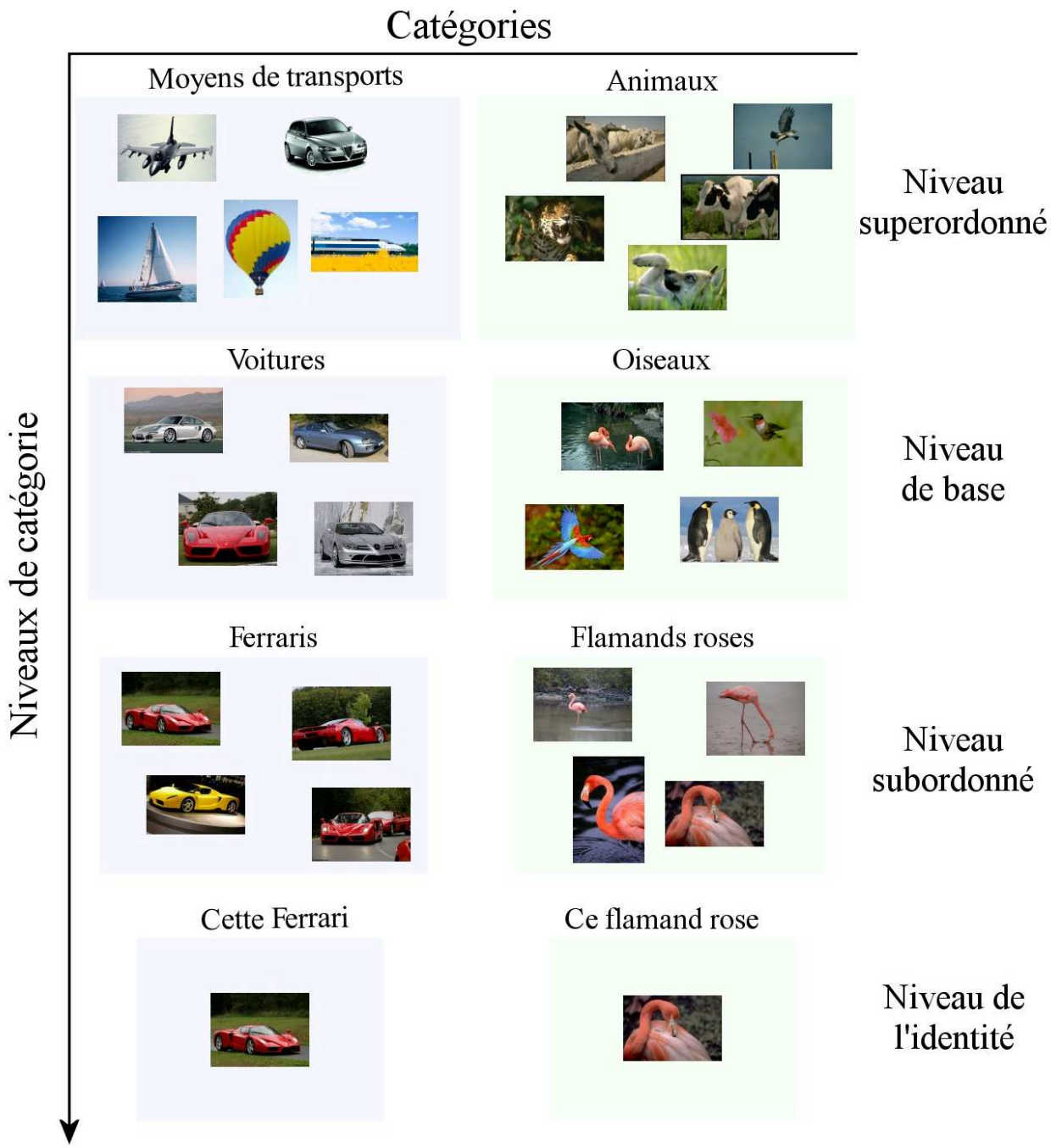


Figure 5 : Représentation "imagée" de l'architecture des catégories. Les différentes catégories sont situées sur le plan horizontal et les différents niveaux de catégorisation sur le plan vertical. Le niveau d'entrée est classiquement défini comme étant situé au niveau de base sauf dans des cas particuliers dans lequel il se situe au niveau subordonné. Nos données montrent que le niveau d'entrée pour catégoriser des images naturelles se situe plutôt au niveau superordonné avec dans ce cas là une véritable organisation verticale des catégories jusqu'au niveau de l'identité de l'objet.

Résumé de la publication : "What's seen first: The animal or the bird ?"

18 sujets ont participé à chacune des deux expériences. Dans chaque expérience, ils devaient effectuer une tâche de catégorisation au niveau superordonné (animal/non animal) et une tâche de catégorisation au niveau de base. Dans l'expérience 1, la tâche au niveau de base portait sur une catégorisation oiseau/non oiseau et dans l'expérience 2, chien/non chien. Pour s'affranchir des biais liés à l'utilisation d'images naturelles et comparer la performance des sujets sur les images d'oiseaux et de chiens par rapport aux autres images d'animaux, les cibles "animaux" dans la tâche de catégorisation superordonnée étaient pour moitié des oiseaux (Exp1) ou des chiens (Exp2), l'autre moitié étant composée d'images variées d'autres animaux. Dans la tâche au niveau de base, la moitié des distracteurs étaient des animaux (non-oiseaux, Exp1 ou non-chiens, Exp2) et l'autre moitié des distracteurs "neutres" (paysages, outils, fleurs, divers objets, etc…). Chaque image n'était vue qu'une fois par un sujet donné pour éviter tout effet d'apprentissage. De plus, pour éviter les biais liés à la sélection des images présentées à différents niveaux de catégorisation, toutes les images présentées dans la tâche de catégorisation superordonnée étaient également présentées, à d'autres sujets, dans la tâche de catégorisation au niveau de base.

Nous présentons dans cet article les données comportementales et électrophysiologiques recueillies dans ces expériences.

Résultats résumés :

La tâche contrôle de catégorisation au niveau superordonné permettait de retrouver les résultats habituels avec des sujets à la fois rapides (TR moyen de 391 ms dans l'expérience 1 et de 386 ms dans l'expérience 2) et précis (respectivement 95,8 et 95,5 % correct). Le TR minimal était de 250 ms dans les 2 expériences. Le fait que la moitié des photographies d'animaux utilisées dans ces tâches représentaient des oiseaux (Exp1) ou des chiens (Exp2) permet de contrôler la précision et la vitesse des sujets sur ces images par rapport aux images d'autres animaux. Les images d'oiseaux étaient catégorisées avec une précision et des TR meilleurs que les autres images d'animaux alors que les images de chiens étaient catégorisées avec la même performance en précision et en vitesse (Figure 2 de l'article). Ceci signifie que les images d'oiseaux et d'animaux choisies dans les expériences étaient bien représentatives de l'ensemble des animaux avec une performance très similaire à celle obtenue sur les autres animaux. Si l'on regarde maintenant la performance obtenue pour ces mêmes images d'oiseaux et de chiens, mais lorsque la tâche requiert une catégorisation au niveau de base, nous observons non pas à une réduction du TR moyen et minimal mais au contraire une très

forte augmentation de celui-ci, bien que le niveau de base soit censé être le niveau d'entrée de la catégorisation visuelle.

En ce qui concerne les potentiels évoqués, les premiers effets significatifs sont présents sur la P2, autour de 150 ms. Sur ce pic, l'amplitude du signal enregistré sur les distracteurs neutres aussi bien en occipital qu'en frontal est toujours supérieure à celle enregistrée sur des images d'animaux. Dans les tâches de catégorisation superordonnée, le signal enregistré sur les images d'oiseaux avait une amplitude légèrement supérieure à celui enregistré sur les images d'animaux. Le signal correspondant aux images de chiens présentait une amplitude comparable à celui enregistré sur les animaux. L'analyse des activités différentielles entre les cibles et les distracteurs permet d'obtenir des informations plus précises sur la manière dont le traitement des informations visuelles est effectué. Dans les deux expériences, les activités différentielles enregistrées au cours de la tâche au niveau de base avaient la même latence que celles enregistrées lors de la tâche superordonnée (environ 160 ms). En revanche l'amplitude de ces activités différentielles était toujours moindre au niveau de base, reflétant peut être la plus grande difficulté de la tâche. En effet, des corrélations entre la précision et l'amplitude de l'activité différentielle ont déjà été mises en évidence dans l'expérience de catégorisation à bas contraste (Macé *et al.*, 2005b) et dans celle utilisant un masquage des images (Bacon-Macé *et al.*, 2005b).

Discussion :

En grande majorité, les études montrant une rapidité d'accès plus importante au niveau de base qu'aux autres niveaux hiérarchiques des catégories impliquaient à un degré plus ou moins important un traitement lexical de l'information. Les sujets devaient souvent rapporter leur réponse verbalement ou vérifier qu'un objet correspondait ou non à un mot présenté juste avant l'essai. L'originalité du travail présenté ici réside dans l'utilisation d'une tâche très simple pouvant être réalisée par le macaque et dans laquelle le traitement lexical de l'information intervient le moins possible. En utilisant de tels paramètres dans le protocole expérimental, nous montrons que les sujets catégorisent des images au niveau superordonné avec en moyenne un gain de temps de 40 à 60 ms par rapport à une catégorisation au niveau de base. Le TR minimal est lui aussi plus court de 30 à 40 ms pour les catégories superordonnées comparées aux catégories de base. Ces différences comportementales ne se retrouvent pas sur la latence de l'activité différentielle enregistrée entre les potentiels évoqués par les cibles et par les distracteurs dont la valeur est pratiquement la même dans toutes les conditions. Les seuls effets observables affectent l'amplitude de cette activité différentielle.

Ces résultats sont en accord avec l'idée d'une représentation de la scène visuelle qui s'appuierait sur les premières informations disponibles dans le système visuel et dont nos données laissent à penser qu'il pourrait s'agir des seules informations magnocellulaires. Les premières représentations des scènes visuelles reconstruites dans le système visuel sont donc probablement en noir et blanc et dépourvues de détails car basées sur des basses fréquences spatiales. Elles sont également relativement abstraites (reconnaissance d'un animal mais pas du type d'animal !). Cette idée d'une représentation grossière reconstruite très rapidement s'accorde bien avec l'idée d'un traitement "coarse to fine" de la scène visuelle tel que proposé par Schyns (Schyns & Oliva, 1994) et mis en évidence expérimentalement par Sugase et al. (Sugase *et al.*, 1999). Dans l'article, nous discutons aussi de l'interprétation de ces résultats au regard de la hiérarchie des catégories et de leur accès par le système visuel lors de l'analyse des objets. Il est probable que si le niveau de base des catégories est le plus souvent le niveau d'entrée lorsque la catégorisation de l'objet doit être verbalisée ("c'est un chien", "c'est un animal"), dans le cas de catégorisations visuelles, le niveau d'entrée se situerait d'abord au niveau le plus général (le plus abstrait) avant que des traitements visuels plus détaillés mais plus coûteux temporellement ne permette d'affiner la représentation de l'objet avec une précision de plus en plus grande. Bien que nous n'ayons pas encore de réelle réponse expérimentale, une prédiction simple voudrait que la catégorisation d'un chien ou d'un oiseau au niveau subordonné (lévrier, martin pêcheur) ajoute encore un coût temporel dans notre tâche de catégorisation rapide. La hiérarchie de catégories telle que proposée par Rosch pourrait refléter la hiérarchie lexicale des catégories dans laquelle l'accès rapide observé pour les catégories de base s'expliquerait par la fréquence des mots les plus couramment utilisés pour dénommer les objets. D'autres différences importantes résident dans l'utilisation d'images naturelles par rapport à des objets isolés, puisqu'il a été montré que l'ajout d'un contexte à un objet permettait de réduire l'avantage du niveau de base sur le niveau superordonné dans une tâche de catégorisation (Murphy & Wisniewski, 1989).

Article n°5 : *en préparation*

What's seen first: The animal or the bird?

Marc J.-M. Macé [CA], Nadège M. Bacon-Macé,

Jean-Luc Nespoulous & Michèle Fabre-Thorpe

Centre de Recherche Cerveau et COgnition (UMR 5549, CNRS-UPS), Faculté de Médecine de Rangueil. 133, route de Narbonne. 31062 Toulouse, France.

[CA]Corresponding author : marc.mace@cerco.ups-tlse.fr

ABSTRACT

Since the princept work of Rosch and colleagues in 1976 it is commonly accepted that in perceptual tasks, basic level categories (dog, chair…) are accessed before superordinate level categories (animal, furniture…) and a large number of studies involving various experimental paradigms have since confirmed this results. But the basic level advantage has been recently challenged by the speed at which objects can be categorized at the superordinate level in a go/no-go paradigm using natural images (Thorpe *et al.*, 1996; VanRullen & Thorpe, 2001b). Using the same protocol in this study, we specifically compared the speed of processing when subjects are performing a go/no-go visual categorization task at the basic or at the superordinate level. We found that categorizing at the basic level is 40-65 ms slower than at the superordinate level. The discrepancies of our results with previous experiments are explained by important paradigms differences such as lower language processing and higher visual system pre-activation. The faster access to a more general/abstract category than basic level could reflect a 2-stage process where the visual system can first analyze the image at a coarse level, with enough information to decide whether the object belongs to the superordinate category and then, with additional detailed processing, gain access to basic-level related information.

Keywords: Categorization, human performance, natural image, category levels.

INTRODUCTION

In 1973, Rosch and colleagues proposed that the different levels of categorization, organized as a taxonomic system, present a favored access stage whatever the perceptual system used (Rosch *et al.*, 1976). This so-called *basic level* is defined as the most abstract level where objects still share a common shape. It has been proposed that this level corresponds to an optimum in term of cognitive efficiency of categorization. This advantage of the basic level (ie, dog or chair) upon the *superordinate* (animal or furniture) and *subordinate levels* (shepherd or rocking chair) was confirmed in object naming and category membership verification experiments (Rosch *et al.*, 1976). Developmental studies also observed that basic level categories are the first learned by children (Mervis & A., 1982) and include the most spontaneously used terms by adults to name objects (Rosch *et al.*, 1976). From these observations, it was inferred that the basic level should correspond to the stored mnemonic representation that is activated first when an object is perceived. Superordinate levels were then considered as abstract generalization of the basic level and subordinate levels as perceptually more inclusive categories. This idea has later been refined by Jolicoeur (Jolicoeur *et al.*, 1984a) and Murphy (Murphy & Brownell, 1985) who introduced the concept of *entry level category* to explain the reaction times advantages found for members belonging to atypical subordinate categories (a penguin is categorized faster as a penguin than as an bird, contrary to more common birds for which basic level is accessed first). Besides, other authors observed that expertise was also a factor that could modulate the entry level towards subordinate level and abolish in some specific cases the basic level advantage (Tanaka & Taylor, 1991; Johnson & Mervis, 1997). However, these observations concerned only the basic and subordinate levels, and this is only recently that the prevalence of basic level over superordinate level has also been challenged. The first element came from Murphy & Wisniewski (Murphy & Wisniewski, 1989) who reported that the RT advantage of basic over superordinate levels was reduced -but did not disappear- by presenting objects in full scenes instead of the usually isolated stimuli on a neutral background. More recently, experiments on rapid visual categorization at the superordinate level definitely challenged the traditional view. The surprising speed at which subjects can detect animals or vehicles in natural scenes in a simple go/no go task, raised many questions about the basic level dominance, at least in the visual modality (Thorpe *et al.*, 1996; VanRullen & Thorpe, 2001a). Cerebral activity associated with this task performance revealed that a differential signal between targets and non-targets appears as soon as 150 ms over the occipital electrodes, already challenging most models of object recognition. VanRullen et al., first pointed the fact that it would be very difficult to expect the basic-level categorization to be even faster than this. Contrary to the original experiments that used a reduced set of drawing objects in isolation, these last studies involve a large variety of targets presented in natural scenes photographs. Another difference which could be fundamental is that subjects gave manual responses, as opposed to the frequently used verbal responses from the previous experiments. In fact, all studies that reported a basic level advantage relied, at different degrees, on some lexical processing. Neurons that respond both to the picture of a celebrity and to its written name have been recorded in human medial temporal lobe, but their latency of response is usually around 300 ms at the earliest (Quian Quiroga *et al.*, 2005). At which point does the semantic processing of a stimulus require an access to the lexical structure of categories? Experiments with macaque monkeys have raised many questions since it has been shown that they are nearly as accurate as human subjects to categorize objects like animals or food items, despite being largely faster (Fabre-Thorpe *et al.*, 1998). The lack of language in this

specie does not prevent a high performance in the categorization task and suggests that a superordinate experiment strictly based on visual -and not lexical- processing can potentially lead to different results than previously obtained. Very recently, Large et al. (Large *et al.*, 2004) followed this idea and found an advantage for superordinate level in visual categorization of isolated objects in a yes/no task. However, the effect was relatively weak (less than 15 ms) and could have been due to a speed/accuracy trade-off, as subjects where 2% better in the slower basic level categorization task. As in the previous studies, they used a small number of isolated objects drawings and did not benefit from the contextual effect reported by Murphy et al. in 1989.

In the two present experiments, we wanted to optimize the visual processing of the stimulus and minimize the influence of language by using (1) categorization of full-scene natural images in separate blocks, either at the superordinate (animal/non animal) or the basic level (bird/non bird or dog/non dog) and (2) go/no-go motor responses to obtain fast responses and reduce higher level cognitive interactions during the task. It has been shown, using peripheral stimuli (Thorpe, 2001 ; Boucart *et al.*, Submitted) or dual-task paradigms (Li *et al.*, 2002), that this task involve a large portion of non-conscious processing. Here, the brief presentation of the images (<30 ms) and the instruction to respond as fast as possible (and within 1s) constitute temporal constraints that were likely to increase the amount of implicit processing in the task.

In the category-level studies mentioned above, subjects were generally tested on several categories and the non-targets were chosen among the other categories, some of them belonging to the same higher level categories. This precaution is necessary to ensure that subjects categorize the images only at the requested level. We used the same constraint here and half of the non-targets in the basic level tasks were images from the same superordinate category (non-bird or non-dog animals).

We also recorded the EEG signal during task performance to obtain precise temporal information on the categorization process. To our knowledge, only two studies that investigated categorization levels hierarchy have used EEG. Tanaka's main result (Tanaka *et al.*, 1999) was that superordinate level categorization leaded to a greater negativity than basic level between 306 and 356 ms, probably reflecting more important semantic/cognitive processing when access to the superordinate level was required. But unfortunately, no comparison could be made with behavior. The second study (Large *et al.*, 2004 is a complete behavioral/EEG study with 3 different levels of categorization. They found that, in accordance with the observed behavioral RT decrease, the first EEG components associated with targets at the superordinate level had earlier latencies when compared to the other levels of categorisation. In the present paper, we will mainly look at the differences between target and non-target ERPs to calculate the precise moment when images are sufficiently processed by the visual system to allow their categorization.

MATERIALS AND METHODS

Participants

18 subjects (9 women and 9 men) were tested in each experiment (mean age: 32 in experiment 1 and 30 in experiment 2). Five of the subjects were tested in both experiments. They all volunteered and gave their written informed consent. They all had normal or corrected–to-normal vision.

Procedure

Subjects were seated at 1 meter from a computer screen in a dimly lighted room. They started the experiment by placing a finger over a response pad for at least one second. A fixation cross appeared for 300-900ms, immediately followed by a photograph of a natural scene, flashed at the centre of the screen for 26 ms (view angle: 20° x 13.5°). With non-target

photographs, subjects had to keep their finger on the button (no-go response) and with a target images, they had to release the button as quickly and accurately as possible (go-response). They had 1 second to trigger their go response after which delay their response was considered as a no-go. The inter-stimulus interval time (ISI) was random between 1.6 and 2.2 seconds (mean: 1.9s).

Two different experiments have been performed in distinct sessions. Within each of the experiment, the aim was to compare the performance in superordinate and basic level tasks. In both experiments, the superordinate level task was an animal/non-animal categorization. At the basic level, subjects were required to perform a bird/non-bird (Bird experiment, Fig 1A) or a dog/non-dog (Dog experiment, Fig 1B) categorization. Each subject completed 16 blocks of 96 trials: 10 at the basic level and 6 at the superordinate level. Each categorization task was preceded by a training block of 48 trials.

Half of the subjects began with the superordinate task, the other half started with the basic level task. In the superordinate animal/non-animal task, half of the animal targets were either birds (Bird experiment) or dogs (Dog experiment). Conversely, in the basic level (bird or dog) categorization task, half of the non-target were non-bird or non-dog animals, while the other half were non-animal (neutral non-targets) pictures. As each image was only seen once by a given subject, the main concern was to avoid any bias induced by the selection of natural photographs for the tasks. To prevent this bias, we counterbalanced images across conditions and subjects: all bird (or dog) photographs were seen by some subjects as targets in the animal task and by the others as targets in the bird (or dog) task. Similarly, all images of non-bird (or non-dog) animals were seen by some subjects as targets in the superordinate task and by the other subjects as non-targets in the basic level task. Non-animal (neutral) images were also seen by different subjects in the 2 tasks. With such protocol, the effect observed on performance can be confidently attributed to task requirements and not to an image selection bias.

Stimuli

We used a set of 1536 images in each task, chosen as varied as possible from a Corel Database (Fig 1). Images of birds and dogs contained a large panel of the species (birds of prey, parrots, sparrows, wading birds, gulls… or shepherds, poodles, mastiffs, spaniels, dachshunds…). Other animal images were also varied and could contain mammals, insects, fish, reptiles etc. Subjects had no *a priori* knowledge about the size, position or number of target(s) in the pictures. "Neutral" non-targets did not contain animals but were as varied, including plants, buildings, people, man-made objects, various landscapes…

EEG recording and analysis

Brain electrical activity was recorded from a 32 electrodes cap in accordance with the 10-20 system and completed by additional occipital electrodes connected to a Synamps amplifier system (Neuroscan Inc.). The ground electrode was placed along the midline, ahead of Fz. Impedances were kept below 5 kΩ. The signal was sampled at 1000 Hz and low-pass filtered at 100 Hz with a notch at 50Hz. Potentials were on-line referenced relative to electrode Cz and average re-referenced off-line. Baseline correction was performed using the 100 ms pre-stimulus interval. Two artifact rejections were applied over the [-100 ms; +400 ms] time period, first on frontal electrodes FP1 and FP2 with a criterion of [-50; +50 μV] to reject trials with eye movements, and second on parietal electrodes Oz and Pz with a criterion of [-30; +30 μV] to remove trials with excessive alpha rhythms. Statistical tests were performed on raw data but a 30 Hz low-pass filter was applied for illustrations (Fig 4&5). Epochs were computed separately for correct target and non-target trials, for each categorization task. The differential activity was

calculated by subtracting the averaged signal on correct non-target trials from the signal on correct target signal. Significant differences between the two conditions were assessed by paired t-tests at the $p<0.01$ level, at each scalp location for each 1-ms time bin. The time bin for which a significant t-test value was reached and followed by at least 15 consecutive significant bins was taken as the onset latency of a differential activity. All values reported in the text met this criterion.

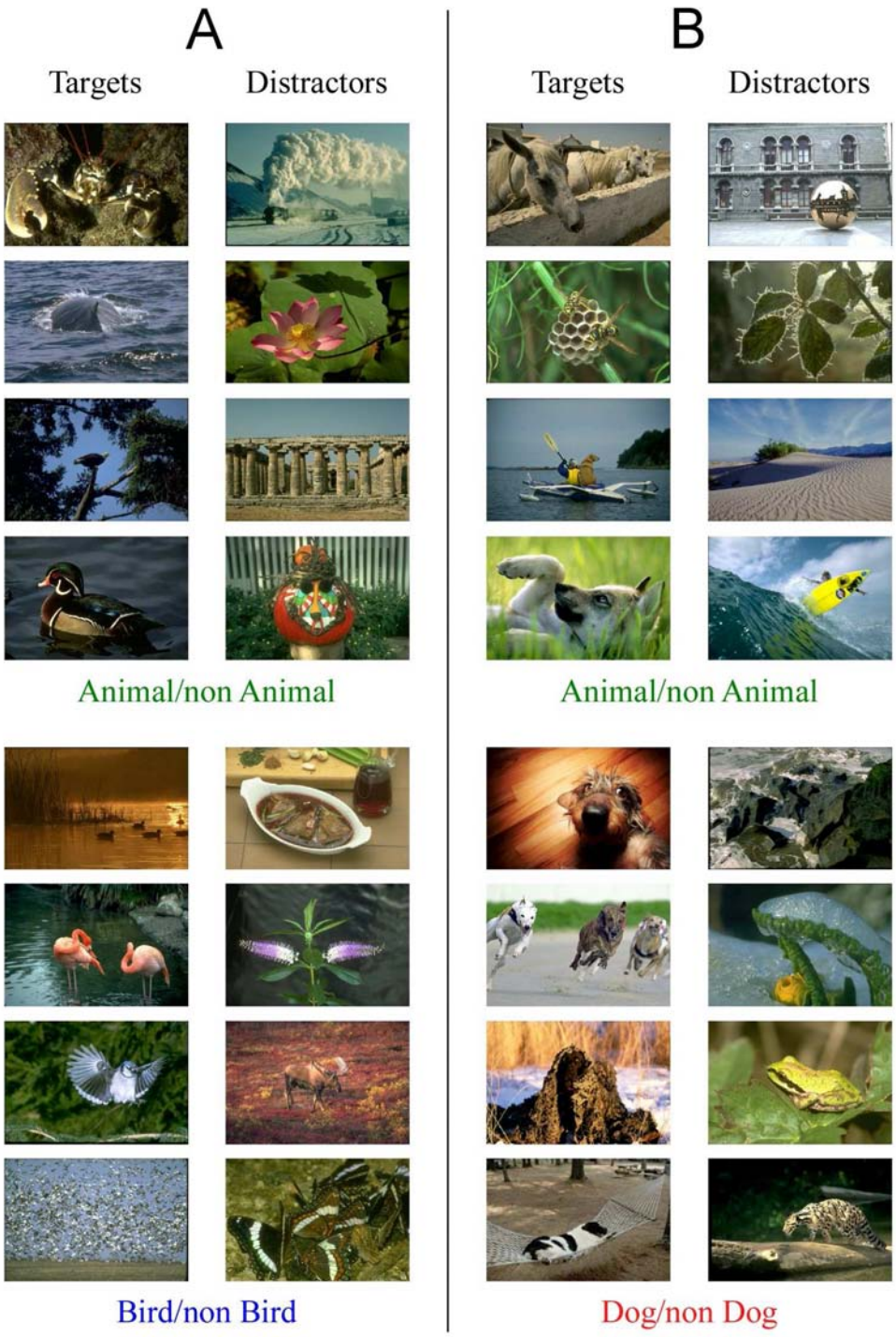


Figure 1: Examples of target and non-target images in the Bird experiment (A) and the Dog experiment (B) used either in the superordinate or in the basic level task. Note that in each experiment, half of the targets in the animal/non animal task are images of the corresponding basic level category (birds or dogs). In basic level tasks, half of the non-targets are images from the same superordinate category (non-bird or non-dog animals).

RESULTS

Performance in the different tasks was evaluated using both accuracy and go-response reaction times measurements. Targets and non-targets were equiprobable in each series, which set the chance level at 50%.

Behavior: superordinate level tasks (A/nA)

Accuracy: The control condition in both Bird and Dog experiments was an Animal/non-Animal task (A/nA) that has often been used in other studies (Thorpe *et al.*, 1996; Fabre-Thorpe *et al.*, 2001; VanRullen & Thorpe, 2001a). A particularity of the present experiment was that half of the targets were very varied photographs of different types of animals whereas the other half included only birds in the Bird experiment and dogs in the Dog experiment. These bird and dog photographs were also used as targets in the basic level categorization tasks (for different subjects to avoid any repetition bias), to allow analysis of the very same set of images when processed either at the superordinate or at the basic level. The global accuracy in the control A/nA task was similar between the two experimental series (95.8% and 95.5%). In control task of the Bird experiment, bird pictures where categorized as "animal" with an accuracy that was slightly higher than for the non-bird animals (accuracy on targets : 98.5 vs 94.8%; Khi², $p<0.05$). In the control task of the Dog experiment, subjects categorized the dog photographs as animal with an accuracy that was virtually the same than for all non-dog animal pictures (accuracy on targets : 96.4 and 96.8%, ns). In both control tasks, subjects were slightly better at responding on animal targets (hit responses) than at ignoring non-targets (correctly withhold responses): Bird experiment: 96.6 vs 95.0; Dog experiment: 96.6 vs 94.4; Khi², $p<0.05$ for both).

The important conclusion of these results is that the varied set of birds and dogs images chosen for the categorization at the basic level were categorized with the same accuracy (dogs) or even a better accuracy (birds) than the set including all other animal photographs.

Speed: Concerning the speed of responses, in both the Bird and Dog experiments, the animal control task was performed with comparable mean reaction time (mean RT : 394 ms in the Bird and 386 ms in the Dog experiment). Interestingly in both control tasks, the pictures of birds and dogs were categorized faster than the pictures of other animals. Birds where categorized as animal with a mean RT of 385 ms (402 ms for the other animals; t-test, $p<0.01$) and dogs with a mean RT of 377 ms (394 ms for the other animals, t-test, $p<0.01$). These differences are accounted for by some very long latency responses recorded on some non-bird or non-dog animal targets. Such long latency responses are due to specific photographs that need long processing times to be analyzed (Fabre-Thorpe *et al.*, 2001). In contrast, the very first responses appeared at the same latencies (fig 2). To evaluate the reaction time of the earliest responses that cannot be attributed to anticipation we use as index the minimal processing time (MinRT). This value corresponds to the first time bin in the RT distribution from which correct responses significantly outnumber false alarms. It reflects the shortest processing time necessary for the visual system to discriminate between targets and non-targets in a given task. In both control A/nA tasks the MinRT was the same for the bird and dog photographs compare to all other types of animals (Bird experiment: 270 ms for birds and other animals (250 ms when taken altogether); Dog experiment: 260 ms for dogs and other animals (also 250 ms altogether)).

This control task allows to conclude that the bird and dog photographs were analyzed as any other animal image in the control A/nA task. If anything, they might be slightly easier to process as they were categorized in average 10 ms faster, with fewer long

latency responses and even with a higher percentage of correct responses (+3%) for the bird photographs.

Behavior: basic level tasks

Accuracy: In the basic level categorization task , subjects scored in average 95.6% correct in the Bird task and 92.6% in the Dog task. They reached accuracy scores that were slightly lower when categorizing the bird photographs at the basic level (97.2%) than at the superordinate level (98.5%; Khi^2, $p< 0.05$). The same effect was observed with dogs categorized at the basic level compared to the superordinate level (94.7% vs. 96.8%; Khi^2, $p<0.05$). As in the superordinate tasks, subjects were better at responding on targets than at ignoring non-targets in both basic level tasks (97.2% vs 93.9% in the bird task, 94.7% vs 90.5% in the dog task; Khi^2, $p<0.05$).

As the non-targets pictures contained at the same time neutral images and non-target animals, it is important to specifically look at the performance on photograph of non-target animals. Indeed, the false alarms were in a very large majority elicited by non-target animals. In the Bird task, 90% of the false alarms where committed on non-bird animals; a proportion that was even higher in the Dog experiment with 95% of the false alarms on non-dog animals. This result also means that when subjects performed the categorization task at the basic level, they were able to correctly ignore neutral non-animal photographs with a very high degree of accuracy (99% correct or over in the bird and the dog tasks). Interestingly, it seems that some animal categories were more likely to induce false alarms but strongly depends on the basic level categorization task considered. Figure 3 shows the difference of false alarms percentage between the Bird and the Dog task as a function of the animal non-target categories (% of FA in the Dog task minus % of FA in the Bird task). We see that some categories of animals had completely opposed patterns of errors in the 2 experiments. For example insects and sea animals elicited far more error in the Bird task than in the Dog task and this was exactly the opposite for bears and felines.

Speed: The mean reaction time at the basic level was 434 ms in the Bird task and 452 ms in the Dog task. These mean RT values are considerably longer than those obtained for the same images of birds and dogs categorized at the superordinate level (control A/nA tasks). Globally, subjects were respectively 40 ms and 65 ms slower in the Bird and in the Dog experiments to categorize images at the basic level. This effect was not limited to the mean RT. As illustrated in figure 2, the whole RT distributions in the Bird and the Dog experiments were shifted towards longer latencies at the basic level. The shape of the RT distribution in the Dog task was also asymmetrical with a particularly large number of long latency responses.

This shift in latencies was also present on the earliest responses : this is reflected by the increased minimal processing time (MinRT) in the basic level tasks compare to the control tasks: 300 ms in the Bird task and 290 ms in the Dog task. Thus, MinRTs were respectively increased by 50 ms and 40 ms when compared to the control superordinate tasks (Fig 2).

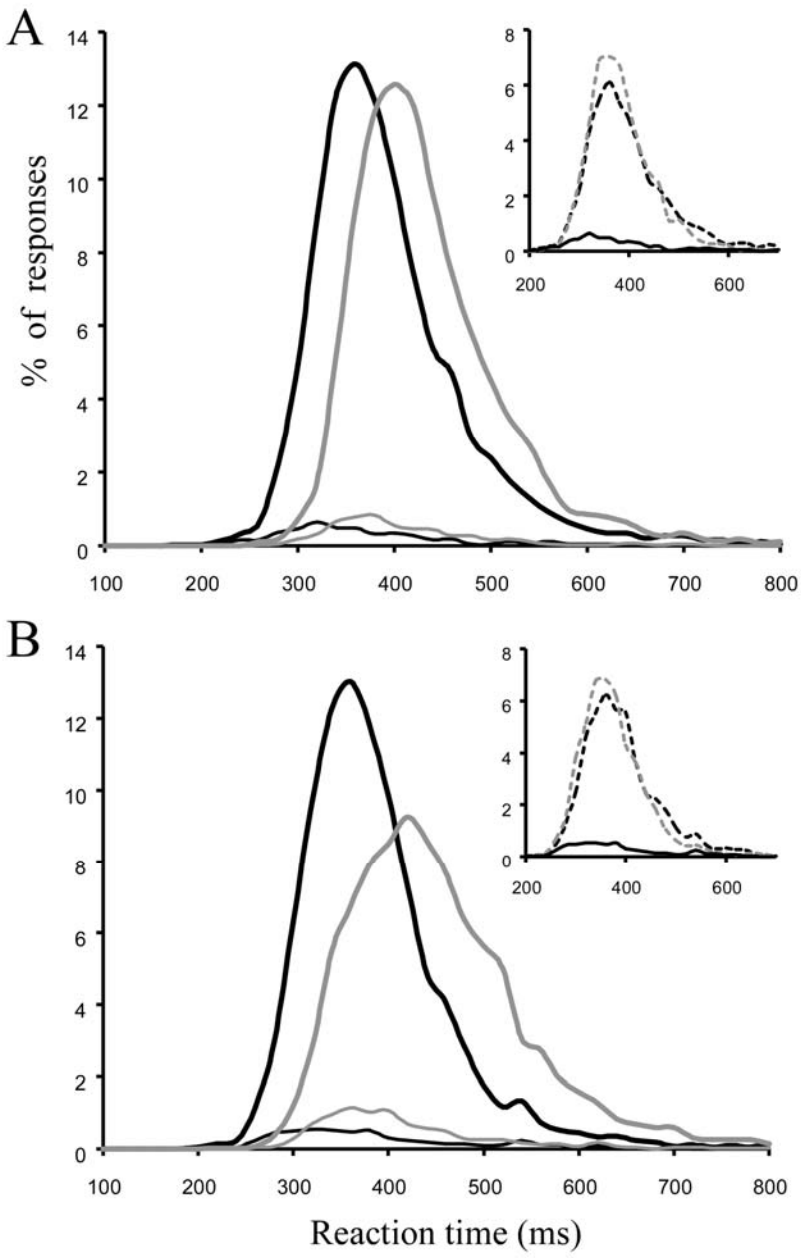


Figure 2: Reaction time distributions in the Bird (A) and in the Dog (B) experiments on correct (thick lines) and incorrect (thin lines) trials, calculated with 10 ms bin width. Reaction time distributions were computed separately in the superordinate level task (black curves) and the basic level task (gray curves). The inserts correspond to RT distributions within the animal/non animal task for birds (A) and dogs (B) in dotted gray lines and the other animals in dotted black lines.

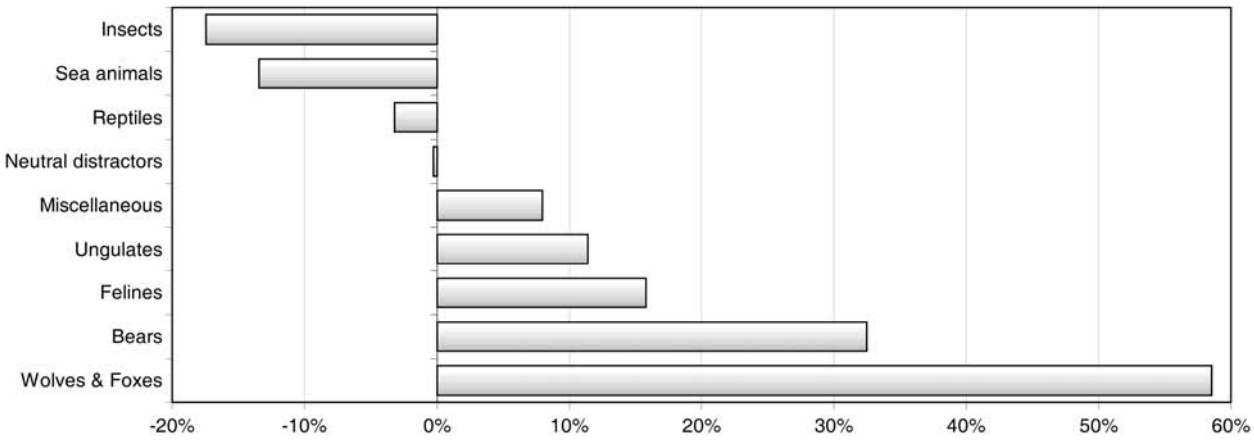


Figure 3: Bars represent the difference of false alarm rates between the Bird and the Dog tasks for several subgroups among animal non-targets. Negative values (insects, sea animals…) mean that subject had a higher false alarm rate for this category of animals when they where looking for birds compare to dogs. For example, the positive value of 32.5% for bears is the result of the subtraction between 38.5 (false alarm rate for bears in the Dog task) and 6.0 (false alarm rate for bears in the Bird task). These discrepancies in the nature of the committed errors could take it's origin in the specific pre-activation state of the visual system in order to recognize rapidly its targets.

Electrophysiology: basic and superordinate level tasks

We averaged separately the signal on target and non-target trials. In the control tasks, we also separated dogs and birds signal from the other animals and in the basic level tasks, animal non-targets signal from the other non-targets. Latencies and amplitudes of the early components P1 and N1, supposed to reflect the physical encoding of the stimulus, were similar between the different conditions within all the tasks (Fig 4 A-D). The first significant effects were found on the P2 wave (150-300 ms) with a larger amplitude on neutral non-targets compare to the other images in all tasks, both at occipital and frontal sites. In the control tasks, there was a slight difference in the amplitude of the P2 between non-bird animals and birds, but not between non-dog animals and dogs (respectively Fig 4A and 4C). In the basic level tasks, P2 amplitude difference between bird and non-bird animals was slightly increased and a small difference appears on P2 amplitude between non-dog animals and dogs (respectively Fig 4B and 4D).

To look closely at these differences, we computed the differential activity (DA) by subtracting the average signal on correct non-target trials from the average signal on correct target trials (Fig 5). A small early DA was found around 70 ms both in the Bird and in the Dog tasks (half of posterior electrodes were significantly different from the baseline) whereas no such early differences where seen in both control tasks. These very early differences have been shown to reflect physical differences between the target and non-target image sets (VanRullen & Thorpe, 2001b). This component could have been expected, as images of birds or dogs share a large amount of similar features within them, whereas the set of images containing all the other animals is more heterogeneous and thereby less distant from the non-target pictures. We won't discuss further these early ERP differences, as they may be related to low level physical encoding of the images rather than to an early processing mechanism that would be used to analyze the same stimulus but only in the basic-level categorization task.

In both experiments, we found the first large differential activity that probably reflects the advanced cognitive processing which allows the detection of the animal targets between 150 and 250 ms (Thorpe *et al.*, 1996; VanRullen & Thorpe, 2001b; Rousselet *et al.*, 2004). The average onset latency of this task-related DA was measured on 7 posterior electrodes (P7-P8-PO7-PO8-O9-O10-Oz). In the Bird experiment, it was found on average at 164 ms in the control task and 160 ms in the basic level task. In the Dog experiment, these latencies were respectively 159 ms and 161 ms. Thus the DA onset latencies in the basic level tasks are very similar to those observed in the superordinate tasks. If we look more specifically at the trials with bird and dog photographs within the control tasks, the onset latencies are nearly unchanged (166 and 161 ms respectively) but the amplitude of the DA is considerably higher when compared to the target trials containing the other animal species (Fig 5 A&B: dark grey vs black).

The fact that we could not find any difference in the onset latency of the different conditions might be surprising in regard to the behavioral results. With MinRTs shifted by 40-50 ms towards longer latencies in the basic level categorization tasks compare to superordinate level tasks, one would expect a delayed onset of the differential activity. This is clearly not the case and an explanation may come from recent experiments (Bacon-Macé *et al.*, 2005a; Macé *et al.*, 2005a) which give us evidence that the difficulty of a task is more related to the amplitude of the DA than to its latency. This was the case in both basic level tasks which had lower DA amplitude when compared to superordinate tasks, together with lower behavioral results (Fig 5 A&B: light grey vs dark grey/black). However, this correlation does not hold here, as in the control task of the Dog experiment, DA was clearly larger on dogs than on the other animals while behavioral accuracy was at a similar level (Fig 5 A&B: dark grey vs black).

Although the differential activity elicited in go/no-go A/nA control tasks is clearly visible on EEG recordings, it seems from the present results that no clear differential activity could be recorded between non-target animals and target animals in the basic level tasks.

On figure 5, a difference between the two experiments in the profile of their basic level DA curves is visible. The DA for the dogs reached a lower amplitude than for the birds. The latency and the time course are quite similar up to a point were a marked plateau appears in the dog task. This effect could be due to the higher feature similarity between dogs and the other animals than between birds and the other animals, a characteristic that could explain both this lower DA signal and the lower behavioral performance.

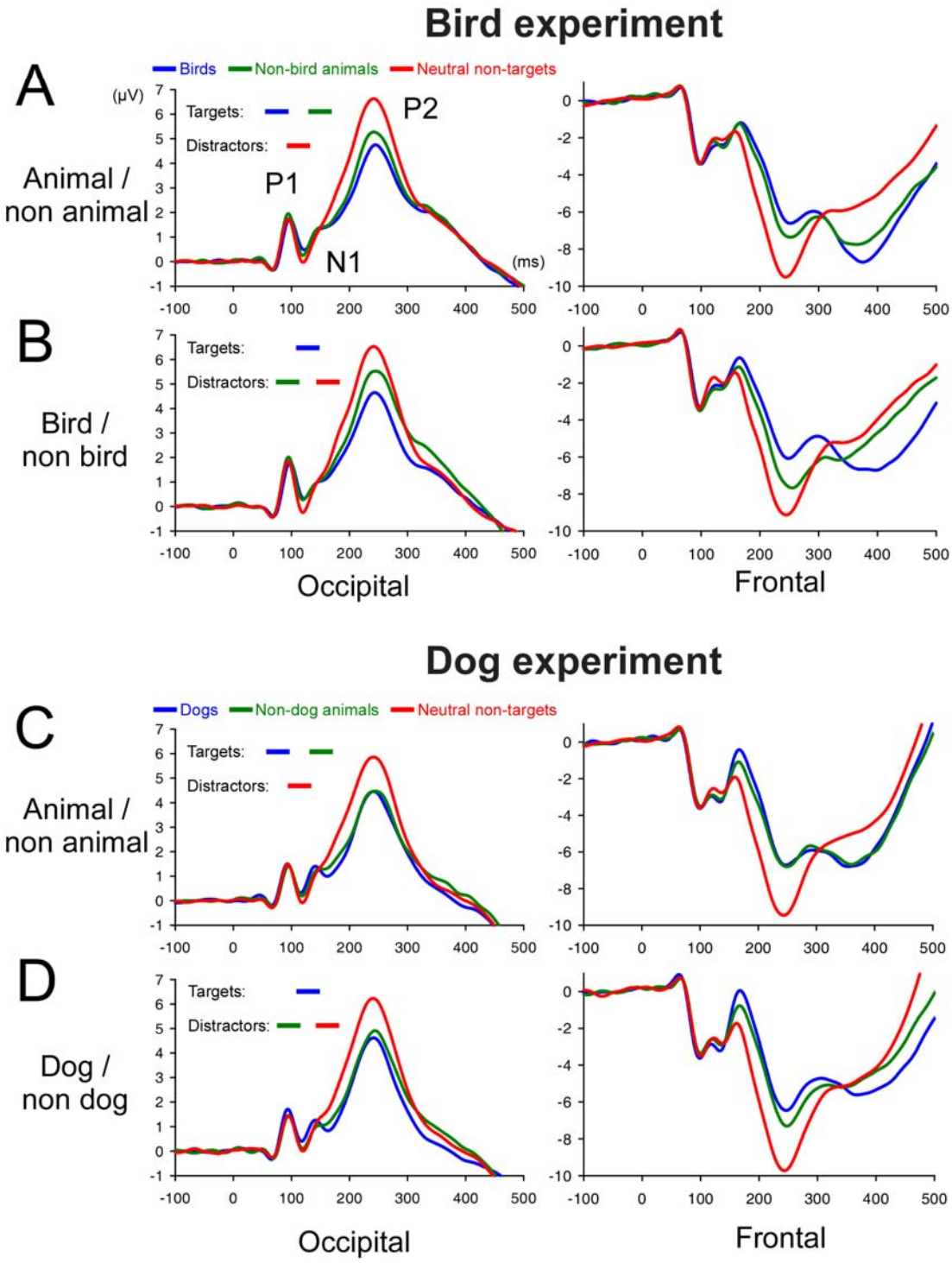


Figure 4: Event Related Potentials (ERP) recorded in Bird (A&B) and Dog experiment (C&D) on over 6 occipital (O1-O2-O9-O10-Iz-Oz) and frontal (FP1-FP2-F7-F8-F3-F4) electrodes. ERPs on birds (A&B) or dogs (C&D) are in blue, non-bird (A&B) or non-dog (C&D) animals in green and neutral non-targets in red. The images of birds/dogs and other animals were either categorized at the superordinate level (A&C) or at the basic level (B&D).

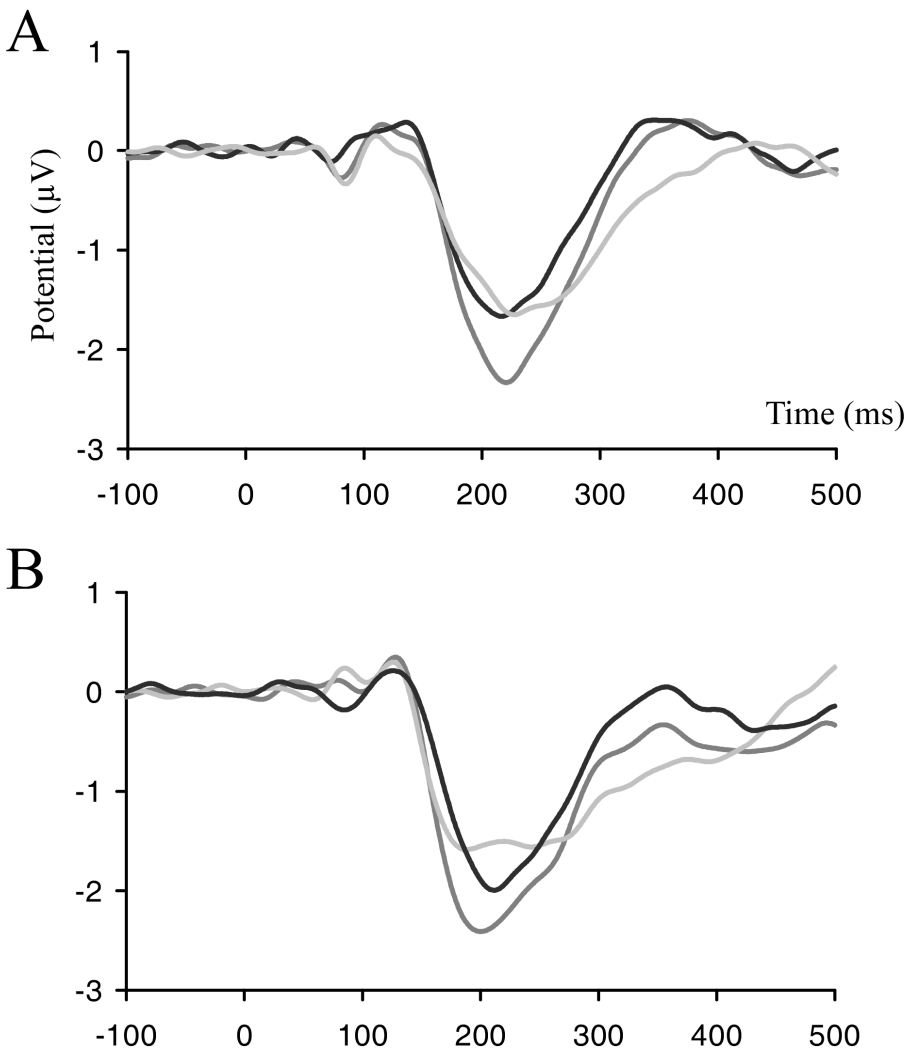


Figure 5: Differential activity (DA) signal calculated by subtracting the signal recorded on correct non-targets from correct targets on 7 posterior electrodes (P7-P8-PO7-PO8-O9-O10-Oz) in Bird (A) and Dog (B) experiments. The DA latencies were similar for all conditions in most of the electrodes and only the amplitude of the DA differs across the conditions. In the control tasks, the amplitude of Birds or Dogs minus non targets (dark grey) was larger than non-bird or non-dog animals minus non-targets (black). In the basic level tasks, we calculated the DA between the targets and the non-targets (light grey).

DISCUSSION

The two experiments presented here strongly suggest that the accepted idea supporting that basic level is perceptually accessed before superordinate level must be challenged. In a go/no-go categorization task using complex natural scenes photographs, subjects are largely faster (and also better in Dog experiment) at finding animals than birds or dogs. This result is surprising in regard to the large set of previous experiments that have demonstrated a basic level advantage both in terms of speed and accuracy. Part of our 40-65 ms reversed advantage could be explained by the stimulus we used here. Murphy et al., (Murphy & Wisniewski, 1989) showed that the

display of detailed scenes reduced the basic level advantage measured on isolated objects. However, according to this study, the "full scene effect" should not be important enough on its own to completely revert the traditional speed advantage of the basic level. Murphy et al. proposed that this effect is due to coherent interactions between the background, potential objects in the scene and the target. However, we propose a complementary hypothesis to explain this radical inversion in processing speed, by raising the differences in terms of lexical requirements between the experiments. The large majority of previous works on levels of categorization require a lexical access additionally to the visual processing of objects (category verification tasks, lexical priming, category naming, lexical input to switch target category at each trial, etc). The consequence could have been that the revealed architecture of category structure was more derived from the lexical category structure than actual visual processing. Although no doubt subsists regarding the superiority of basic level in language, the results found here could reflect a category structure more strictly linked to visual perception. Our experiment was designed to minimize the interactions with language, and monkey's high performance in this task constitutes in itself a good demonstration of this point. In the visual system, the superordinate level may not constitute an abstraction from the basic level as previously proposed (Rosch *et al.*, 1976; Jolicoeur *et al.*, 1984a), but rather represents the less detailed level at which some visual properties can be grouped to primarily construct distinct categories. This idea is very close to the hypothesis of an architecture that would process information from coarser to finer representations, as theoretically proposed by Schyns et al. (Schyns & Oliva, 1994). A number of experimental findings support this model: electrophysiological studies in monkeys (Sugase *et al.*, 1999), functional imaging in humans (Liu *et al.*, 2002) or psychophysical data (Sergent, 1985, Parker *et al.*, 1992). During the reconstruction of the scene along the visual pathway, objects contours are the highest saliency locations (strongest contrast), and should be processed first, conveyed by the low spatial frequency information. This information is available at the very beginning of the integration process and may be sufficient to infer the coarse representation of the scene necessary to perform a superordinate categorization task. In contrast, the basic level categorization may need to wait for more details on the stimulus that require additional processing.

Recent RSVP (Rapid Serial Visual Presentation) experiments (Bacon-Macé, in preparation) showed that only coarse information is included in the very first visual representation, and that more time is needed to access the details of the image. For example, in a task where subjects had to categorize animals, they were able to respond for the presence of an animal with great confidence while being unable to report whether this animal was a rabbit or a horse. So, in conditions where the information capture is severely limited by temporal constraints, the categorization task that shows best performance is at the superordinate level. At these extreme conditions, only the information that is extracted, conveyed and processed at first could be used to respond. A comparable result had been found in a masking experiment on natural images (Grill-Spector & Kanwisher, 2005). Grill-Spector et al. showed that a categorization at a level intermediate between basic and superordinate level can be done with the same accuracy as a simple detection task. These experiments provide more evidence that the superordinate level is accessible as soon as the very first information are available, contrary to what happen for more inclusive levels.

We calculated differential activities based on the EEG signal of correct trials. Contrary to the behavioral results, no clear difference was found

between the latencies of differential activities onsets in the superordinate and the basic level tasks for the two experiments. If we suppose that an image of a bird or a dog is categorized first at the superordinate level and afterwards at the basic level, there is no reason why the onset latency between these two tasks should differ. What is certainly harder to discriminate are the basic level images from the images that belong to the same superordinate category. This case corresponds to the light gray curve in the Figure 5, and no differences are clearly visible in it's DA onset latency compare to the black curve (superordinate task). This could be because some features that birds or dogs exclusively share among them allow some images to be distinguished early from the other animals and neutral non-targets. However, this DA has a reduced amplitude compare to the bird or dog DA (dark gray) in the superordinate task. The visual system seems uncomfortable at discriminating the images of birds and dogs from the other animals and incidentally, the number of false alarms on animals non-targets was 10 to 25 times greater than on neutral non-targets.

It is interesting to note that the behavioral results on the two experiments were not totally similar, the dog/non dog categorization seams to be harder to perform compared to the bird/non bird task. This effect may originate from the high prototypicallity of dogs relatively to the other animals. Dogs are very typical mammals, not only they occupy a special place in the animal kingdom for humans, but they also share more features with the other animals than the birds may do (like 4 legs and some fur). In contrast, the feathers, wings, beaks and aerodynamic shapes of birds can lead to certainly more peculiar and thus distinguishable views. The expected consequence is that more visual features are diagnostic for birds and they could be more easily distinguished from the other animals comparatively to dogs. As an illustration, typical errors in the bird/non bird task were made on insects and fishes (an aero/hydro-dynamic global shape effect ?) whereas errors in the dog task were committed on canines, bears, felines and ungulates (Fig 5). Besides, we also observed that the amplitude of the differential activity was more reduced on the dog/non-dog task compared to the animal/non-animal task, reflecting the greater difficulty for the visual system to separate the dog representations from the other animals. This degree of difficulty in the tasks raised the question of the choice of pictures in this experiment. We tried to use images as varied as possible and the set we used could be considered as relatively representative of the animal kingdom, a point that could be criticized in previous studies working on the categorization levels with a very small number of stimuli. Even if some differences exist between our two basic level tasks, performance on both bird and dog categorization were strongly affected compared to the animal categorization. Following Schyns and colleagues, (Schyns, 1998) we argue that finding the most diagnostic features that characterize the targets relatively to the non-targets images is the essential point to perform such categorization tasks. A logical sequel of this study would consist in a basic level categorization task where the proportion of “tricky” non-targets (same superordinate category as the targets) would by changed. By this way, we could measure the effects of varying the number of shared features between targets and non-targets. Using a target category at a more inclusive level than the basic level (subordinate or exemplar level) could lead to even shorter reaction time if no images from the same higher order level are used. In such tasks, the visual system could be more and more confident at doing prediction regarding the target features and better pre-activation of the visual representations could shorten the RT. There could even be a shift towards simple detection of low-level feature if predictability becomes complete, as in a previous

experiment where a single natural image was used as target (Delorme *et al.*, 2004a). Subject were around 40 ms shorter to detect this single image than to categorize at the superordinate level. This result also means that unconstrained categorization tasks at levels inferior to the superordinate level could at most be 20-40 ms faster.

Manipulating category structures and boundaries is not a new idea and has previously been performed to explain typicality and expertise effects (Tanaka & Taylor, 1991; Murphy & Brownell, 1985). The underlying hypothesis is that training on a particular set of stimuli can possibly modify the representations in the inner visual processing and facilitate recognition by increasing encoding speed. Regarding the atypical members of a category, the effect should be due to some better combination of diagnostic features that could help to boost their recognition. However, the results we present here are not limited modifications in the category structure with repeatedly presented stimuli or a peculiar set of features. They imply that the process of categorization inside the visual system is different from what had been thought for several decades. The representations involved in our task are linked to low level cortical structure organization, with a limited plasticity (no great differences in purely visual capabilities after millions of years of separate evolution between human beings and macaques). We have now to investigate how these different category structures are linked and cooperate to rapidly attach a semantic and a lexical content to any objects.

ACKNOWLEDGMENT

This work was supported by the Integrative and Computational Neuroscience ACI program of the CNRS. Financial support was provided to M.J-M. Macé by a Ph.D. grant from the French government. The authors declare that they have no competing financial interests.

REFERENCES

Bacon-Macé, N., Mace, M.J., Fabre-Thorpe, M. & Thorpe, S.J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Res*, **45**, 1459-1469.

Boucart, M., Fatima, N., Despretz, P., Defoort-Dhelemmes, S., Macé, M.J.-M. & Fabre-Thorpe, M. (Submitted). Implicit but not explicit recognition at very large visual eccentricities. *J Exp Psychol Hum Percept Perform*.

Delorme, A., Rousselet, G.A., Mace, M.J. & Fabre-Thorpe, M. (2004). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Brain Res Cogn Brain Res*, **19**, 103-113.

Fabre-Thorpe, M., Delorme, A., Marlot, C. & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *J Cogn Neurosci*, **13**, 171-180.

Fabre-Thorpe, M., Richard, G. & Thorpe, S.J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport*, **9**, 303-308.

Grill-Spector, K. & Kanwisher, N. (2005). Visual recognition. *Psychol Sci*, **16**, 152-160.

Johnson, K.E. & Mervis, C.B. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal-of-experimental-psychology-General*, **126**, 248-277.

Jolicoeur, P., Gluck, M.A. & Kosslyn, S.M. (1984). Pictures and names: making the connection. *Cognit Psychol*, **16**, 243-275.

Large, M.E., Kiss, I. & McMullen, P.A. (2004). Electrophysiological correlates of object categorization: back to basics. *Brain Res Cogn Brain Res*, **20**, 415-426.

Li, F.F., VanRullen, R., Koch, C. & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proc Natl Acad Sci U S A*, **99**, 9596-9601.

Liu, J., Harris, A. & Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. *Nat Neurosci*, **5**, 910-916.

Macé, M.J., Thorpe, S.J. & Fabre-Thorpe, M. (2005). Rapid categorization of achromatic natural scenes: how robust at very low contrasts? *Eur J Neurosci*, **21**, 2007-2018.

Mervis, C.B. & A., C.M. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, **53**.

Murphy, G.L. & Brownell, H.H. (1985). Category differentiation in object recognition: typicality

constraints on the basic category advantage. *J Exp Psychol Learn Mem Cogn*, **11**, 70-84.
Murphy, G.L. & Wisniewski, E.J. (1989). Categorizing objects in isolation and in scenes: what a superordinate is good for. *J Exp Psychol Learn Mem Cogn*, **15**, 572-586.
Parker, D.M., Lishman, J.R. & Hughes, J. (1992). Temporal integration of spatially filtered visual images. *Perception*, **21**, 147-160.
Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C. & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, **435**, 1102-1107.
Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**, 382-439.
Rousselet, G.A., Thorpe, S.J. & Fabre-Thorpe, M. (2004). Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vision Res*, **44**, 877-894.
Schyns, P.G. (1998). Diagnostic recognition: task constraints, object information, and their interactions. *Cognition*, **67**, 147-179.
Schyns, P.G. & Oliva, A. (1994). From blobs to boundary edges: Evidence for time and spatial scale dependent scene recognition. *Psychol Sci*.
Sergent, J. (1985). Influence of task and input factors on hemispheric involvement in face processing. *J Exp Psychol Hum Percept Perform*, **11**, 846-861.
Sugase, Y., Yamane, S., Ueno, S. & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, **400**, 869-873.
Tanaka, J., Luu, P., Weisbrod, M. & Kiefer, M. (1999). Tracking the time course of object categorization using event-related potentials. *Neuroreport*, **10**, 829-835.
Tanaka, J.W. & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognit Psychol*, **23**, 457-482.
Thorpe, S., Fize, D. & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, **381**, 520-522.
Thorpe, S.J., Gegenfurtner, K. R., Fabre-Thorpe, M., Bulthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *Eur J Neurosci*, **14**, 869-876.
VanRullen, R. & Thorpe, S.J. (2001a). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, **30**, 655-668.
VanRullen, R. & Thorpe, S.J. (2001b). The time course of visual processing: from early perception to decision-making. *J Cogn Neurosci*, **13**, 454-461.

3.3 - Diagnosticité

Dans les deux expériences précédentes, nous avons vu que la catégorisation au niveau de base s'avère dans le cas de catégorisations visuelles, moins rapide que la catégorisation au niveau superordonné. Dans ces deux expériences, nous nous étions assurés que le sujet effectuait bien la tâche de catégorisation au niveau de base en utilisant des animaux (non-oiseaux ou non-chiens) comme distracteurs. Alors qu'ils ne représentaient que 50% des distracteurs, plus de 95% des fausses détections étaient commises sur ces images appartenant à la même catégorie superordonnée (animaux) que les oiseaux et les chiens, probablement parce que le délai de traitement supplémentaire par rapport à la tâche animal/non animal permettait de limiter les erreurs aux images ressemblant le plus aux cibles. En analysant les erreurs des sujets, nous nous sommes aperçus que les distracteurs animaux qui induisaient le plus de fausses détections étaient différents entre les deux expériences. Lorsque les sujets recherchaient des oiseaux, la majorité de leurs erreurs se produisaient sur des images d'insectes ou de poissons, alors que quand les sujets recherchaient des chiens, ce sont les renards, les félins et les ours qui induisaient le plus d'erreurs. Ces différences très importantes dans la répartition des erreurs sont sans aucun doute liées à une préactivation différente du système visuel dans chacune des deux tâches. Tous ces éléments renforcent l'idée que la précision observée dans une tâche de catégorisation dépend de la distance physique entre les objets-cibles et des objets-distracteurs représentés tous deux par des ensembles d'éléments (idée proposée par Duncan et Humphreys pour le visual search (Duncan & Humphreys, 1989)). Ainsi une tâche de catégorisation au niveau de base dans laquelle les distracteurs font tous partie de la même catégorie superordonnée devrait être plus difficile à réaliser que la même tâche avec des distracteurs faisant tous partie d'autres catégories superordonnées. C'est ce que nous avons testé dans une seconde série expérimentale. Nous avons de plus enregistré l'activité cérébrale pour étayer l'hypothèse d'une variation de l'amplitude de l'activité différentielle et non des latences en fonction de la difficulté de la tâche de catégorisation à effectuer.

La moitié des sujets étaient des experts dans la tâche (ils avaient participé à plusieurs expériences animal/non animal) et l'autre moitié étaient des sujets naïfs. Quelques experts faisant partie de l'équipe avaient connaissance du mode de construction des séries, mais tous les autres sujets ignoraient l'existence de variations dans la proportion des distracteurs animaux, leur seule tâche étant de catégoriser des chiens le plus vite et le plus précisément

possible. Certains de ces sujets rapportèrent des variations de difficulté dans la tâche mais aucun n'a pu déterminer précisément l'origine de ce ressenti. Les sujets commençaient tous par catégoriser des animaux au niveau superordonné pendant 3 séries avant de commencer la tâche de catégorisation chien/non chien. Ils effectuaient 3 blocs contenant des proportions de 0, 50 ou 100% de distracteurs animaux. L'ordre de ces blocs était contrebalancé entre les sujets et les biais d'images étaient évités en contrebalançant, sur l'ensemble des sujets, le statut des images et leur utilisation dans les différents blocs.

Résultats résumés :

Dans la tâche contrôle au niveau superordonné, les sujets avaient un TR moyen de 396 ms, une précision de 94,6% correct et un TR minimal de 260 ms. Dans la tâche de catégorisation au niveau de base les performances variaient fortement en fonction de la composition du groupe des images distracteurs. Lorsque la moitié des distracteurs étaient des animaux (non-chiens), les sujets avaient une précision moyenne de 92,5 % pour un temps de réaction moyen et minimal de respectivement 458 ms et 280 ms, reproduisant ainsi le résultat obtenu dans l'expérience précédente avec une augmentation de plus de 60 ms du temps de réaction moyen pour catégoriser des images de chiens au niveau de base par rapport au niveau superordonné. En augmentant encore la proportion de distracteurs animaux, la précision diminuait encore à 90,5 % et le TR moyen augmentait encore à 477 ms, le TR minimal étant autour de 320 ms. A l'inverse, lorsque les distracteurs ne contenaient aucune image d'animaux, la précision des sujets augmentait de 2 % par rapport à la catégorisation au niveau superordonnée (96,5 %) pour un TR moyen inférieur de 5 ms (391 ms) et un TR minimal inférieur de 10 ms (250 ms).
La latence à partir de laquelle apparaît l'activité différentielle dans les différentes conditions ne varie que très peu. C'est sur l'amplitude du signal cérébral différentiel entre essais-cibles et essais-distracteurs qu'est enregistré l'effet le plus important avec une réduction de l'amplitude corrélée à l'augmentation de la proportion de distracteurs animaux jusqu'à devenir tout juste significative.

Discussion :

En analysant le signal EEG enregistré lors de ces diverses expériences de catégorisation au niveau de base ainsi que les erreurs commises par les sujets, nous comprenons que l'organisation hiérarchique des catégories est loin d'être le seul facteur qui intervient pour déterminer le temps nécessaire pour accéder à un niveau de représentation donné. Un autre facteur primordial dans la catégorisation est celui de la diagnosticité qui stipule qu'une cible

sera identifiée d'autant plus rapidement qu'il existe d'indices permettant de la distinguer des distracteurs. Dans une tâche de catégorisation de chiens (niveau de base) pour laquelle les distracteurs n'incluaient aucune image d'animal, les sujets pouvaient effectuer leur catégorisation au niveau de base, mais aussi au niveau superordonné en utilisant des processus ultra rapides de traitement comme dans la tâche animal/non animal. Ils étaient même en moyenne légèrement plus rapides et plus précis, probablement parce que l'ensemble des chiens est moins varié que l'ensemble des animaux, ce qui réduit la complexité de la tâche. La réduction de l'espace des attributs des cibles pourrait être poursuivie ; on pourrait par exemple imaginer des traitements encore plus rapides en demandant au sujet de ne plus catégoriser que les labradors (sans distracteurs animaux ou autres races de chiens) ou encore un labrador particulier. Le cas extrême de cette réduction de la variabilité de la cible conduit à rechercher une cible unique, c'est à dire à effectuer une tâche de reconnaissance, comme celle rapportée par Delorme et al. (Delorme *et al.*, 2004b). Dans ce cas, la prédictibilité des attributs de la cible devient totale et il est possible de préactiver le système visuel afin de déclencher la réponse motrice dès qu'un indice bas niveau bien précis est perçu.

Comme pour les cibles, la variabilité des distracteurs influence la performance. Lorsque les distracteurs sont très hétérogènes, ils possèdent de nombreux attributs qui peuvent réduire la liste des indices diagnostiques de la catégorie cible. A l'inverse, des distracteurs très homogènes entre eux devraient faciliter la tâche du sujet qui pourrait les ignorer plus facilement (à confirmer par l'expérience). Mais le facteur le plus important dans une tâche de catégorisation concerne en fait la distance entre l'espace des attributs des cibles et l'espace des attributs des distracteurs. Plus les éléments de ces deux espaces se recouvrent et plus la tâche devient difficile pour les sujets. C'est ce que l'on observe dans cette expérience avec une tâche qui devient de plus en plus difficile (TR en hausse et baisse de la précision) lorsque l'on introduit une proportion croissante de distracteurs animaux dans la tâche chien/non chien. Ces observations sur l'hétérogénéité des cibles et des distracteurs ont déjà été proposées par Duncan et Humphreys (Duncan & Humphreys, 1989) pour expliquer les performances des sujets dans diverses conditions de "visual search". La figure 6 récapitule l'influence de ces différents éléments.

Cette expérience va également dans le sens du modèle d'accumulation de l'information évoqué dans la discussion des précédentes expériences. Les sujets sont obligés d'adapter leur temps de réaction en ralentissant fortement dans les conditions où des distracteurs partagent de nombreuses caractéristiques visuelles avec les cibles. Dans une telle situation, le système visuel doit attendre des informations diagnostiques supplémentaires qui ne lui parviennent que

plus tardivement. Il est ainsi probable que la tâche de catégorisation visuelle au niveau de base, en présence de distracteurs "animaux", ne puisse être réalisée qu'en utilisant des informations plus détaillées que celles qui sont fournies par le système magnocellulaire.

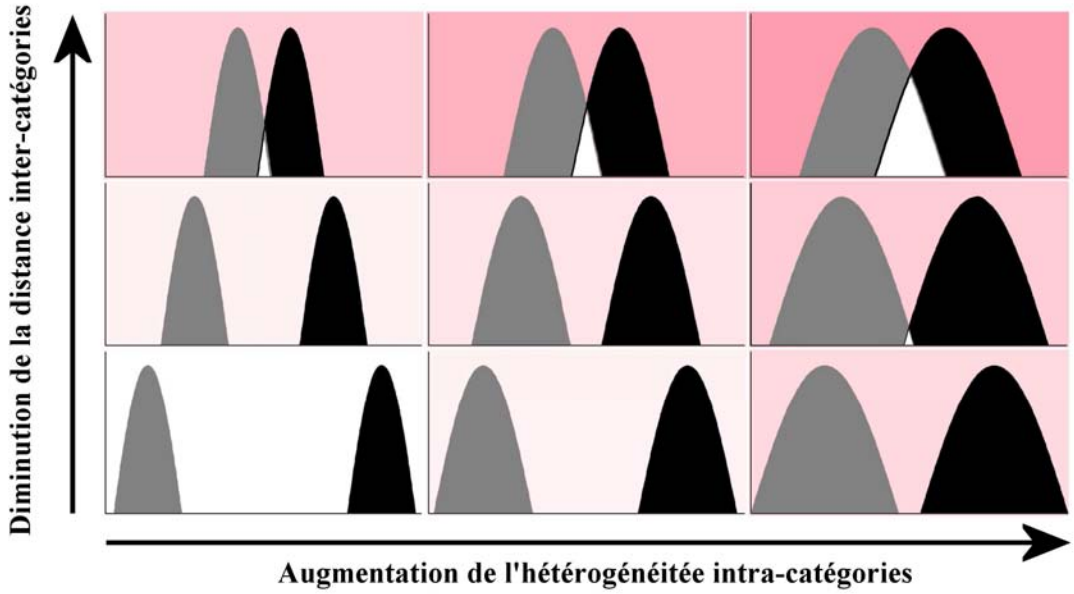


Figure 6 : Les distributions grises et noires représentent deux catégories. La couleur de fond des cases indique la difficulté à discriminer ces deux catégories (blanc : facile et rouge : difficile). La difficulté d'une tâche de catégorisation dépend de plusieurs facteurs. Le plus important (en ordonné) est déterminé par la distance minimale entre les deux ensembles à catégoriser. On cherchera dans une tâche donnée les indices les plus diagnostiques afin d'augmenter au maximum cette distance inter-catégories. Les $2^{ème}$ et $3^{ème}$ critères co-varient dans notre représentation, mais sont indépendants dans la réalité. Il s'agit de la variabilité intrinsèque des cibles entre elles et des distracteurs entre eux (hétérogénéité intra-catégories). Plus la variabilité des cibles diminue et plus la tâche devient facile (les prédictions sur les caractéristiques de la cible sont meilleures : la préactivation du système visuel est donc plus efficace). De la même manière, plus la variabilité des distracteurs diminue et plus la tâche devient facile, parce que la probabilité de trouver des éléments visuels communs entre les cibles et les distracteurs diminue (le nombre d'indices diagnostiques de la présence d'une cible augmente).

Cette expérience vient tout d'abord confirmer les résultats des expériences précédentes en montrant que pour le système visuel, la catégorie d'entrée n'est pas la catégorie de base, contrairement au système lexical. Dans le système visuel, l'accès le plus rapide se trouve au niveau le plus abstrait, celui des catégories superordonnées, la catégorisation s'effectuant probablement sur un modèle "coarse to fine". Elle progresserait vers des catégories de plus en plus précises (de bases puis subordonnées) en même temps que les informations visuelles traitées plus lentement mais de façon plus détaillée s'accumulent dans le système visuel pour faire émerger une représentation de plus en plus fine de la scène visuelle.

- (Ces résultats feront l'objet d'un article actuellement en préparation) -

3.4 - Un cas particulier : la catégorisation des visages. Articles n°6 & 7

Nous avons montré que la catégorisation devient difficile lorsque les cibles et les distracteurs partagent de nombreuses caractéristiques physiques, comme par exemple les chiens et l'ensemble des autres animaux. Mais la diagnosticité des indices visuels n'est pas le seul facteur qui influence la vitesse de catégorisation et l'expertise que possède le sujet avec les objets à catégoriser joue également un rôle important. Il n'est généralement pas nécessaire de préciser aux sujets naïfs qui participent à nos expériences que l'on ne doit pas considérer les humains comme des cibles dans une tâche animal/non animal. Les humains bénéficient en effet d'un statut particulier qui fait qu'un distracteur qui contient un humain ne provoque que rarement une fausse détection, alors que les attributs visuels des humains ne sont objectivement pas très éloignés de ceux des singes. Je me suis intéressé à ce statut particulier des humains et Guillaume Rousselet souhaitait de son côté étudier les effets généralement rapportés pour des visages en gros plans (N170, effet d'inversion, etc…) en utilisant comme cibles des visages humains à toutes les échelles spatiales, à l'endroit ou à l'envers, en évitant les effets de répétition spatiale et en veillant à ce que les visages ne soient pas tous présentés au centre de l'image. Nous avons donc élaboré une expérience dans laquelle les visages étaient "contextualisés" dans des scènes naturelles pour satisfaire nos critères communs (images naturelles, visages et corps humains à toutes les échelles, etc…). La consigne des sujets était de ne répondre que lorsqu'un visage était présent dans la scène, mais la faible proportion de distracteurs dans lesquels un humain était visible sans que son visage ne le soit et la plus grande facilité à catégoriser des corps humains que des visages font qu'on peut considérer que les sujets effectuaient dans cette première expérience une tâche humain/non humain plus qu'une tâche visage/non visage. Les distracteurs dans cette tâche étaient constitués pour moitié de distracteurs neutres et pour moitié d'images d'animaux. Les sujets alternaient entre la tâche de catégorisation des visages humains contextualisés et la tâche de catégorisation des animaux. Mis à part la plus grande expertise des sujets avec les stimuli représentant des humains, cette tâche de catégorisation humain/non humain était tout à fait comparable à la tâche de catégorisation oiseau/non oiseau ou chien/non chien en ce qui concerne le niveau de catégorisation.

Résumé des deux publications : " Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes " et "Temporal course of ERP in fast object categorization in natural scenes: a story more complicated than expected?"

Résultats comportementaux résumés : expérience n°1, visages contextualisés

Les performances des sujets étaient très élevées dans les deux tâches de catégorisation avec une précision de 96,4 % sur les visages d'hommes et de 96,3 % sur les images d'animaux. Les TR moyens étaient similaires pour les deux tâches de catégorisation : 382 ms. Ce résultat est étonnant au regard des expériences précédentes dans lesquelles les oiseaux et les chiens étaient catégorisés à des latences bien plus longues lorsqu'ils étaient traités au niveau de base par rapport au niveau superordonné. Cette différence entre les expériences de catégorisation d'oiseaux et de chiens et celle-ci s'explique peut être par les effets d'expertise qui se manifestent lors de la catégorisation des formes humaines. Cette expertise pourrait compenser l'augmentation de difficulté de la tâche due à la grande ressemblance entre les cibles et les distracteurs pour maintenir un niveau de performance équivalent dans les deux tâches de catégorisation.

Digression sur les visages

Dans l'expérience ci-dessus, nous avons mis en évidence un effet d'expertise lorsque les sujets doivent effectuer une tâche de catégorisation de visages humains contextualisés qui s'apparente plutôt à une tâche de catégorisation humain/non humain. On peut encore augmenter simultanément la difficulté de la tâche et le niveau d'expertise en proposant au sujet une tâche de catégorisation de visages d'humains en gros plan parmi des têtes d'animaux également en gros plans. Parmi les stimuli que nous utilisons, ceux pour lesquels l'ensemble des sujets possède le plus d'expertise sont sans conteste les visages. Ce sont des objets qui possèdent à la fois une structure complexe et une grande importance écologique de par les informations à caractère social qu'ils véhiculent. Malgré les similitudes dans la structure globale des visages et dans leur organisation interne, nous n'avons aucun mal à en différencier plusieurs milliers au cours de notre existence avec une grande précision et une grande rapidité. Déterminer le genre, l'âge et l'expression d'un visage est également une tâche dont les humains s'acquittent avec une très bonne précision. Cette grande expertise de l'homme dans le traitement des visages pourrait se traduire par des performances singulièrement élevées dans une tâche de catégorisation impliquant les visages. Les visages d'animaux constituent une catégorie de stimuli possédant une structure globale et une organisation interne proche de

celle des visages d'humains. La faible quantité d'indices diagnostiques fiables pour catégoriser des visages humains parmi des visages animaux (ou inversement) pourrait être à nouveau compensée par une meilleure exploitation des attributs visuels de ces objets grâce à l'expertise du système visuel des sujets avec ces stimuli.
Dans cette deuxième expérience qui vient compléter la première collaboration avec Guillaume Rousselet, les visages humains et les têtes d'animaux étaient présentés à l'endroit ou à l'envers pour qu'il puisse étudier la spécificité du traitement des visages humains à travers l'effet d'inversion. Mon objectif était de rechercher les possibles interactions entre diagnosticité et expertise dans une tâche de catégorisation visuelle rapide. Comme nous le verrons dans l'article qui suit, malgré la difficulté de la tâche en terme de diagnosticité, les performances comportementales et les activités cérébrales différentielles restent relativement bien préservées, ce qui suggère que l'expertise permet aux visages de bénéficier de traitements avancés.

Il existe une abondante littérature sur les visages et de nombreux débats sont encore en cours comme celui concernant l'existence d'un module spécifique dédié à la reconnaissance des visages. Ces observations empiriques sur le statut particulier des visages semblent confirmées par des résultats d'expériences montrant que les visages sont traités différemment des autres objets et par des structures cérébrales distinctes (Sergent *et al.*, 1992 ; Allison *et al.*, 1994 ; Bentin *et al.*, 1996). Il n'est pas question de faire ici un exposé exhaustif de la littérature portant sur le traitement des visages Mais comme nous allons le voir, les choses ne sont pas forcément aussi simples qu'il n'y parait...

Les premiers arguments en faveur d'une zone spécialisée dans l'analyse des visages sont apparus avec des études de cas de patients présentant des troubles "spécifiques" de la reconnaissance des visages (prosopagnosie). Ces personnes ont une vision normale, mais il leur est impossible de reconnaître une personne, même très familière, à partir de son visage (Farah *et al.*, 1995). Le cas parallèle existe également avec une incapacité à reconnaître une catégorie particulière d'objets (Moscovitch *et al.*, 1997). Ces observations vont bien sûr dans le sens d'une ségrégation des structures pour traiter les différentes catégories d'objets. Mais c'est une simplification dangereuse de dire que seule la reconnaissance de telle ou telle catégorie précise est affectée dans le cas d'une lésion car une preuve formelle nécessiterait de tester tous les objets possibles, ce qui est bien sûr techniquement irréaliste. De plus, même s'il ne fait pas de doute que l'atteinte principale d'un patient prosopagnosique concerne la

reconnaissance des visages, elle est souvent associée à d'autres troubles moins marqués (Damasio *et al.*, 1982 ; de Gelder *et al.*, 1998). Certains patients prosopagnosiques ont ainsi des difficultés pour effectuer des discriminations fines entre stimuli autres que des visages, sous-tendant l'idée que les atteintes des patients prosopagnosiques touchent des aspects fonctionnels de la reconnaissance en termes génériques plutôt que des structures spécialisées dans la reconnaissance d'objets particuliers.

Les visages ont été les premiers objets complexes pour lesquels des réponses neuronales sélectives ont été enregistrées chez le singe (Gross *et al.*, 1972 ; Perrett *et al.*, 1982). Chez l'homme, diverses zones du cortex visuel occipito-temporal ont montré en IRMf une activation plus importante pour des visages que pour d'autres objets (Puce *et al.*, 1995 ; Kanwisher *et al.*, 1997 ; McCarthy *et al.*, 1997). Bien que de légères différences existent entre les régions activées dans ces différentes études, un consensus émerge facilement sur l'implication d'une portion du gyrus fusiforme dans la reconnaissance des visages (Farah & Aguirre, 1999) ainsi que du STS (superior temporal sulcus) et du gyrus occipital inférieur (Puce *et al.*, 1996). La région médiane du gyrus fusiforme latéral a même été baptisée "Fusiform Face Area" (FFA) par Kanwischer (Kanwisher *et al.*, 1997) qui y voit une zone exclusivement dédiée au traitement des visages (Figure 7). Une controverse est apparue sur le rôle précis de cette structure, entre les partisans de la simple détection de visages (Kanwisher *et al.*, 1998) et ceux de l'identification précise (George *et al.*, 1999 ; Haxby *et al.*, 2000). Une étude montre cependant qu'une lésion du gyrus fusiforme altère la reconnaissance des visages mais pas leur détection.

Il est délicat d'établir si ces aires sont spécifiquement dédiées à l'analyse des visages ou si elles peuvent participer à la reconnaissance d'autres objets. Est-ce que ce sont les visages et uniquement les visages qui activent spécifiquement ces aires ou bien est-ce que ce sont les traitements qu'il faut opérer sur les visages pour les reconnaître ? Si la deuxième solution est la bonne, on peut imaginer que des expériences dans lesquelles des discriminations visuelles fines doivent être effectuées sur des objets pour lesquels les sujets possèdent une expertise vont activer les mêmes structures. C'est bien ce que trouvent Gauthier, Tarr et al. (Gauthier *et al.*, 2000a) lorsque leurs sujets doivent identifier des greebles, sortent de marionnettes virtuelles complexes qu'ils ont appris à distinguer (Figure 8). La question n'est cependant pas encore tranchée et de nombreux articles contradictoires ont été publiés par différentes équipes (Farah, 1996 ; Kanwisher, 2000 ; Gauthier *et al.*, 2000b ; Grill-Spector *et al.*, 2004). Que les

visages soient traités par une région particulière du cerveau ou non, que cette zone du cerveau reflète uniquement le traitement des visages ou celui de discriminations fines, il reste néanmoins que la catégorie des visages est extrêmement intéressante grâce à l'expertise dont elle bénéficie comparativement à d'autres catégories.

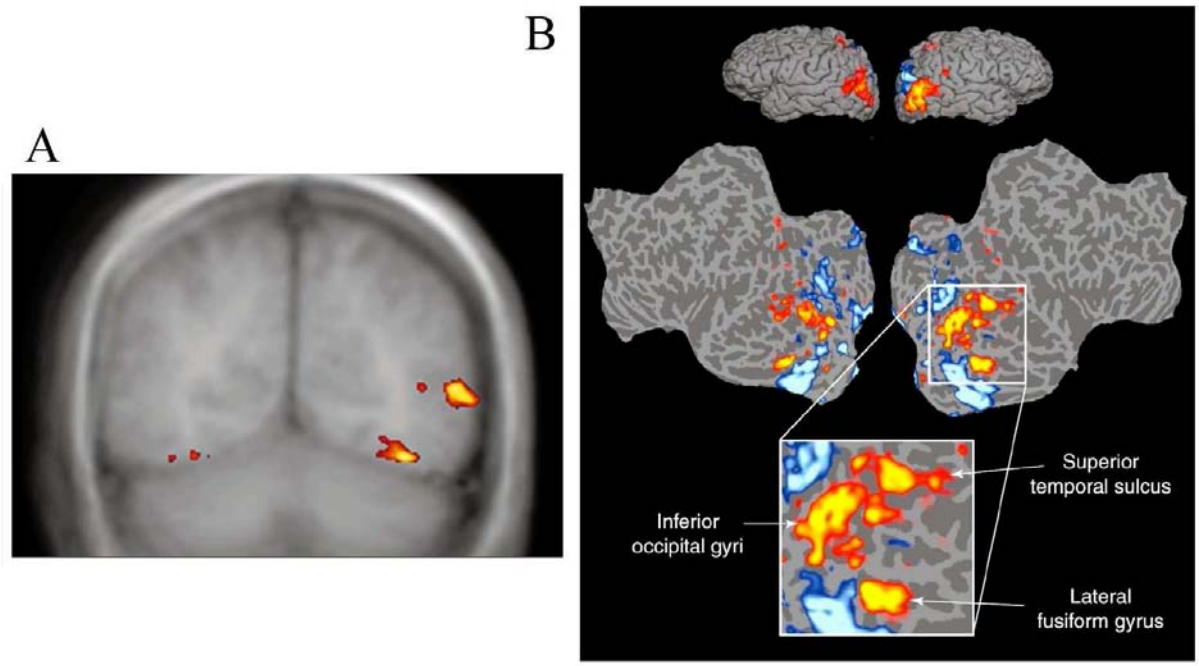


Figure 7 : A- Coupe coronale montrant l'activation du gyrus fusiforme (en position ventrale) et du sulcus temporal supérieur (STS, en position latérale) en IRMf lors de la perception de visages statiques. B- Vue générale du cerveau et vue du cortex expansé puis aplati montrant une activation plus importante du STS, du gyrus fusiforme latéral et du gyrus occipital inférieur lors de la perception d'un visage par rapport à une maison. Les activations chez ce sujet sont fortement bilatérales, contrairement au sujet en A. Reproduit d'après (A) Allison et al., Vis Neurosci (2000) et (B) Haxby et al., Trends Cogn Sci (2000).

De manière plus générale, l'ordre d'accès aux différents niveaux de catégorisation peut varier selon le niveau d'expertise du système visuel dans une tâche donnée. Dans les expériences précédentes sur les niveaux de catégorisation, les sujets étaient plus rapides pour accéder au niveau superordonné. On peut se demander si les visages, qui bénéficient d'un important effet d'expertise, ne peuvent pas obtenir un avantage de traitement par rapport aux autres animaux au niveau superordonné. C'est la raison pour laquelle nous avons réalisé une deuxième expérience dans laquelle les performances des sujets étaient comparées dans une tâche de catégorisation animal/non animal et dans une tâche de catégorisation de visages humains. Dans cette expérience, les têtes d'animaux et les visages d'hommes étaient tous présentés en gros plan, pour réduire la variabilité dans les images et augmenter la difficulté de la tâche en réduisant les différences entre les cibles et les distracteurs.

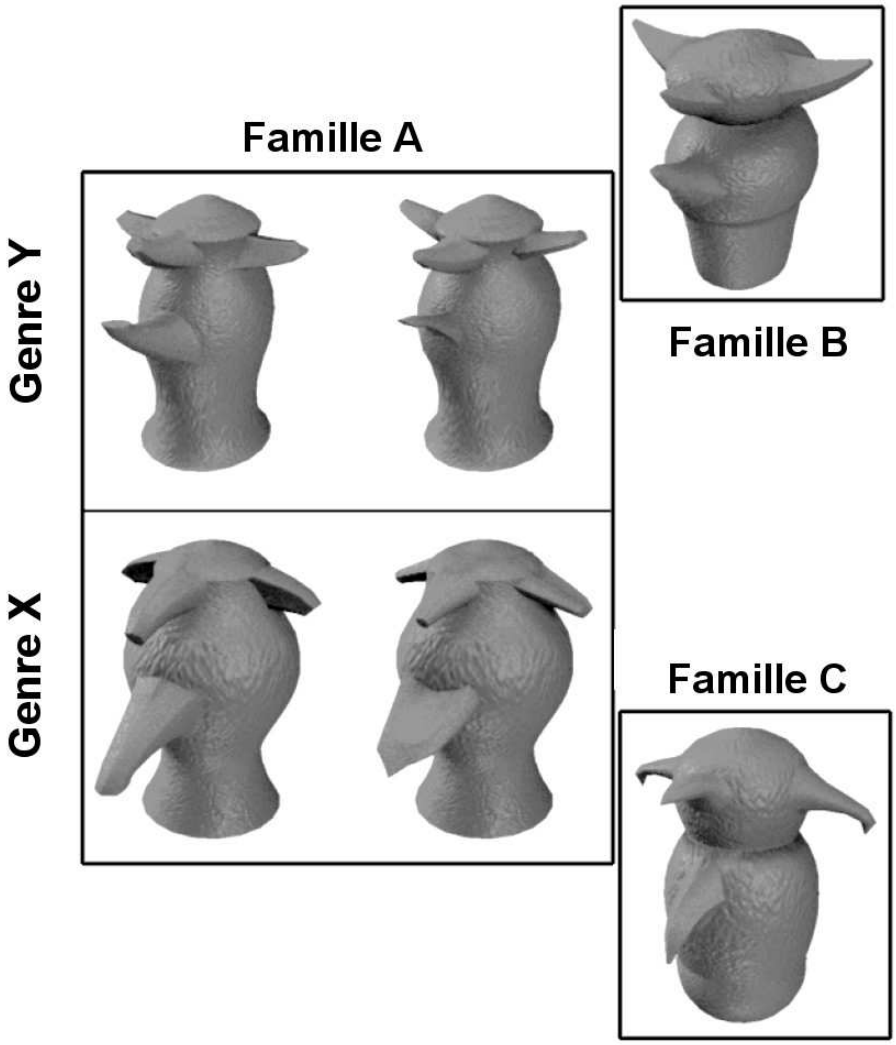


Figure 8 : Quelques exemples de greebles appartenant à 3 familles. Deux exemplaires de chaque genre sont présentés pour la famille A. Reproduit d'après un projet de recherche de M. Tarr.

Résultats comportementaux résumés : expérience n°2, visages en gros plans

Dans cette 2ème expérience, la précision était de 97.7 % sur les visages d'hommes et de 97,9 % sur les visages d'animaux. Les sujets étaient en moyenne légèrement plus rapides pour catégoriser les hommes que les animaux (382 ms contre 392 ms).

Les résultats des enregistrements électrophysiologiques associés à ces deux expériences sont présentés dans le deuxième article ci-dessous. Nous discuterons plus en détail (dans une partie située après les articles et dédiée à l'analyse des activités différentielles précoces) les résultats électrophysiologiques qui exploitent le principal avantage d'une double tâche de catégorisation pour s'affranchir des différences physiques entre les images-cibles et distracteurs.

Résultats électrophysiologiques pour les deux expériences :

Contrairement aux résultats comportementaux qui montrent une grande similitude entre la catégorisation animal/non animal et la catégorisation humain/non humain, le signal EEG associé à l'exécution de ces deux tâches est très différent. Dans l'expérience 1, les différentielles de type 1 (cibles - distracteurs) enregistrées dans la tâche animal/non animal reproduisent les résultats obtenus précédemment avec une activité différentielle apparaissant autour de 150 ms. Dans les différentielles de type 2 (images cibles tâche 1 - mêmes images utilisées comme distracteurs tâche 2), la soustraction des signaux enregistrés sur les mêmes images présentées soit comme des cibles, soit comme des distracteurs permet de s'affranchir des différences physiques entre les groupes d'images tout en conservant le signal directement liée au processus de catégorisation. Pour les images d'animaux, les différentielles de type 2 sont observées à une latence similaire à celle des différentielles de type 1. Le signal enregistré sur les visages est en revanche très différent. Pour la différentielle de type 1, l'activité différentielle apparaît vers 130 ms pour disparaître ensuite autour de 160 ms. De manière surprenante, on ne trouve aucune activité différentielle de type 2 pour les visages humains avant 200 ms, ce qui semble indiquer que le système visuel traite de la même manière -jusqu'au même niveau de détail ?- un visage humain qu'il soit cible ou distracteur dans la tâche.

Dans l'expérience 2, dans laquelle les visages d'hommes et d'animaux étaient présentés en gros plan, les différentielles de type 1 sont très similaires à celles de l'expérience 1, mais les différentielles de type 2 sont inexistantes pour les 2 catégories de visages.

Rappelons que dans ces deux expériences, les images étaient présentées à l'endroit et à l'envers (rotation de 180°) pour l'étude des effets d'inversion sur les performances de catégorisation menée par G. Rousselet. La très faible incidence de l'inversion des images sur les performances comportementales constitue un argument important en faveur d'un modèle de reconnaissance des objets qui ne fait pas intervenir de rotation mentale. En effet, les sujets ne sont qu'environ 10 ms plus lents pour catégoriser les images à l'envers, ce qui impliquerait des vitesses de rotation mentale de l'ordre de 18000 degrés par secondes, ce qui est supérieur de plusieurs ordres de grandeur aux valeurs trouvées dans des tâches où une rotation mentale est clairement nécessaire pour répondre.

Discussion :

Les visages d'humains ou d'animaux peuvent être analysés sur la base de traitements visuels ultra-rapides, ils donnent lieu aux mêmes performances comportementales qu'une catégorie superordonnée telle que les animaux alors qu'ils se situent probablement à un niveau d'analyse plus fin. Il est possible que l'expertise que possède le système visuel pour cette catégorie particulière d'objets lui permette d'atteindre des performances similaires. Au regard du comportement des sujets, les visages humains étaient traités avec la même vitesse et la même précision que les animaux. Cependant, le signal EEG enregistré sur les humains et sur les animaux différait très largement. Les visages humains présentaient des différentielles de type 1 entre 130 et 160 ms mais aucune de ces différences ne subsistait quand les différences physiques entre les images cibles et distracteurs étaient supprimées. Ce résultat surprenant nous amène à penser que le système visuel effectue peut être un traitement particulier sur les visages d'hommes et les humains, quel que soit leur statut dans la tâche. Ainsi, que les visages soient cibles ou distracteurs, le système visuel les traiterait -par défaut- jusqu'au même niveau de détail, sans doute à cause de l'importance biologique et sociale que revêt ce type d'analyse.

Dans l'expérience 2, les cibles et les distracteurs partagent plus de caractéristiques physiques. Étant donné qu'une partie de l'activité différentielle de type 1 est due à des différences physiques entre les images, il est normal de constater une diminution de l'amplitude de ce signal différentiel entre les deux expériences. La disparition de l'activité différentielle de type 2 sur les animaux, à l'image de celle obtenue sur les visages d'hommes, s'explique peut être par un traitement par défaut plus approfondi de tous les objets qui ressemblent potentiellement à un visage humain, tels que les visages d'animaux.

Ainsi, il n'est pas possible de trancher entre les deux hypothèses : soit le cerveau traite différemment les visages des autres stimuli, soit ce sont les visages eux-mêmes qui nécessitent des traitements particuliers. Dans tous les cas, les visages ne sont pas catégorisés mieux ou plus rapidement, mais le niveau de détail qui semble nécessaire pour différentier un visage humain par rapport à un visage animal ne permet pas d'exclure une supériorité dans leur traitement.

Article n°6

J Vis, **3**, 440-455

Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes

Guillaume A.Rousselet, **Marc J-M. Macé**

& Michèle Fabre-Thorpe

Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes

Guillaume A. Rousselet Centre de Recherche Cerveau et Cognition, CNRS-UPS UMR 5549, Toulouse, France

Marc J.-M. Macé Centre de Recherche Cerveau et Cognition, CNRS-UPS UMR 5549, Toulouse, France

Michèle Fabre-Thorpe Centre de Recherche Cerveau et Cognition, CNRS-UPS UMR 5549, Toulouse, France

Object categorization can be extremely fast. But among all objects, human faces might hold a special status that could depend on a specialized module. Visual processing could thus be faster for faces than for any other kind of object. Moreover, because face processing might rely on facial configuration, it could be more disrupted by stimulus inversion. Here we report two experiments that compared the rapid categorization of human faces and animals or animal faces in the context of upright and inverted natural scenes. In Experiment 1, the natural scenes contained human faces and animals in a full range of scales from close-up to far views. In Experiment 2, targets were restricted to close-ups of human faces and animal faces. Both experiments revealed the remarkable object processing efficiency of our visual system and further showed (1) virtually no advantage for faces over animals; (2) very little performance impairment with inversion; and (3) greater sensitivity of faces to inversion. These results are interpreted within the framework of a unique system for object processing in the ventral pathway. In this system, evidence would accumulate very quickly and efficiently to categorize visual objects, without involving a face module or a mental rotation mechanism. It is further suggested that rapid object categorization in natural scenes might not rely on high-level features but rather on features of intermediate complexity.

Keywords: rapid visual categorization, human performance, natural scenes, human faces, animals and animal faces, inversion effect, mental rotation, configural processing

Introduction

Recent biologically plausible models of object visual processing have emphasized that much of the computation underlying scene categorization might rely on essentially parallel feed-forward mechanisms (Riesenhuber & Poggio, 2000; Thorpe & Imbert, 1989; VanRullen, Gautrais, Delorme, & Thorpe, 1998; Wallis & Rolls, 1997). These suggestions are supported by the finding that in humans, a differential brain activity develops between target and distractor trials from 150 ms in various categorization tasks using natural images (Thorpe, Fize, & Marlot, 1996; Rousselet, Fabre-Thorpe, & Thorpe, 2002). This processing time seems to correspond to an optimum, because it cannot be speeded up even with highly familiar natural images (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001). Moreover, when considering the number of processing steps between the retina and the high-level visual cortical areas of the ventral pathway, this 150-ms delay challenges most models of visual processing because it appears compatible only with a first feed-forward wave of information processing (Thorpe & Fabre-Thorpe, 2001). Thus, this delay appears as the minimal processing time from which discriminability between two categories of stimuli can develop. However, even if the human visual system is able to extract a great deal of information in under 150 ms, visual perception does not end up after a first pass through the visual system that might not even allow access to a conscious representation (Dehaene & Naccache, 2001; Thorpe, Gegenfurtner, Fabre-Thorpe, & Bulthoff, 2001); in many cases, reaching a decision will require more time consuming detailed analysis.

In parallel, growing evidence suggests that faces may have a special computational status (Farah, Wilson, Drain, & Tanaka, 1998; Kanwisher, 2000; but see Tarr & Gauthier, 2000) that would allow them to be processed more efficiently and even faster than any other class of objects. However, the precise speed of face processing remains a controversial question. Indeed, very rapid categorization of isolated and relatively homogenous face stimuli has been reported in the literature, with brain activity onsets appearing as early as 50-80 ms poststimulus (George, Jemel, Fiori, & Renault, 1997; Mouchetant-Rostaing, Giard, Bentin, Aguera, & Pernier, 2000a, 2000b; Seeck et al., 1997). These findings have been disputed as other groups have reported early face processing in the 100-130-ms latency range (Debruille, Guillem, & Renault, 1998; Halgren, Raij, Marinkovic, Jousmaki, & Hari, 2000; Halit, de Haan, & Johnson,

DOI 10:1167/3.6.5 Received March 11, 2003; published July 31, 2003 ISSN 1534-7362

2000; Itier & Taylor, 2002; Linkenkaer-Hansen et al., 1998; Pizzagalli, Regard, & Lehmann, 1999; Schendan, Ganis, & Kutas, 1998; Yamamoto & Kashikura, 1999; Liu, Harris, & Kanwisher, 2002) or even later in the 150-200-ms latency range (Bentin, Allison, Puce, Perez, & McCarthy, 1996; Carmel & Bentin, 2002; Eimer, 2000; Jeffreys, 1996; Rossion et al., 2000; Taylor, Edmonds, McCarthy, & Allison, 2001).

However, the vast majority of experiments with faces used isolated, homogeneous, and well-centered stimuli. Such a bias in stimulus sets could explain early face selective brain activity that could be due either to a higher predictability of the expected stimuli that would speed up processing (Delorme, Rousselet, Macé, & Fabre-Thorpe, 2003) or to the bottom-up extraction of low-level physical properties from a set of homogenous stimuli (VanRullen & Thorpe, 2001b). Thus, the data obtained with isolated face stimuli may not necessarily apply to real-world situations. For instance, it is known from single-unit recordings in monkeys that the responses of neurons tuned to faces and other object categories are affected by the presence of other competing objects, and by the presence of a background (Chelazzi, Duncan, Miller, & Desimone 1998; Trappenberg, Rolls, & Stringer, 2002). Thus, it is interesting to investigate the functioning of the biological visual system in more realistic situations when faces are presented in the context of natural scenes. In order to obtained such a "realistic" estimate of face processing speed, we used a rapid go/no-go categorization task with briefly presented (20 ms) photographs of real-world scenes in which subjects had to react when the photograph contained a human face. Such a go/no-go design involves the simplest motor output possible, allowing subjects to respond as fast as they could with the minimal motor constraints. For comparison with another class of targets, subjects alternated between this face categorization task and an animal categorization task used in a series of earlier studies from our group.

The second issue we wanted to address concerned the characteristics of the object representations activated during rapid categorization tasks. These early representations could be specific to canonical presentations of the stimuli used in the tasks. Alternatively, they might rely on relatively view invariant representations. One way to address this issue is to analyze how processing is affected with inverted pictures. Indeed, face processing has been shown to be more sensitive to inversion than other object categories (Bentin et al., 1996; Rossion et al., 2000; Yin, 1969). This pattern of results has been taken as evidence that face perception relies on specific mechanisms dedicated to the processing of the configural information present in upright faces (Maurer, Le Grand, & Mondloch, 2002). To explain the additional time necessary to process inverted pictures, some models of object recognition postulate the existence of a normalization stage at which an object orientation must be aligned with a memory template before matching can take place (see review in Tarr & Bülthoff, 1998; Ullman, 1996). Such normalization stage might be associated with a time consuming mental rotation of misaligned objects (Jolicoeur, 1988; Tarr & Pinker, 1989; Vannucci & Viggiano, 2000). Here we wanted to assess whether this inversion effect would affect the rapid categorization of human faces or animals presented in the context of natural scenes. To address this last issue, half of the pictures (faces, animals, and other natural scenes), whether targets or distractors, were presented upside-down.

Behavioral performance was analyzed in subjects alternating between rapid categorization of human faces and of animals presented randomly, upright or inverted, in the context of natural scenes. The processing speed and the magnitude of the inversion effect were compared for human faces and animals in two experiments, in which the main difference was in the presentation scale of the targets.

Experiment 1

The first experiment was designed to compare directly the animal task used by our group in several previous experiments to a homologue human face task. In both tasks, target images were photographs of real-world scenes in which human faces or animals were shown at different scales, orientations, and positions (Figure 1). Because "face" stimuli did not contain isolated items, but faces in the context of human bodies embedded in natural scenes, we will refer in the remaining of the text to "human" pictures and "contextual face task."

Methods

Participants

The 24 adult volunteers in this study (12 women and 12 men; mean age 31 years, ranging from 19 to 53 years; 5 left-handed) gave their informed written consent. All participants had normal or corrected-to-normal vision.

Experimental procedure

Subjects were seated in a dimly lit room at 100 cm from a computer screen (resolution, 800 x 600; vertical refresh rate, 75 Hz) piloted from a PC computer. To start a block of trials, they had to place their finger on a response pad for 1 s. A trial was organized as follows: a fixation cross (0.1° of visual angle) appeared for 300-900 ms and was immediately followed by the stimulus presented during two frames (i.e., about 23 ms in the center of the screen). Participants had to lift their finger as quickly and as accurately as possible (go response) each time a target was presented and to withhold their response (no-go response) when the photographs did not contain a target. Responses were detected using infrared

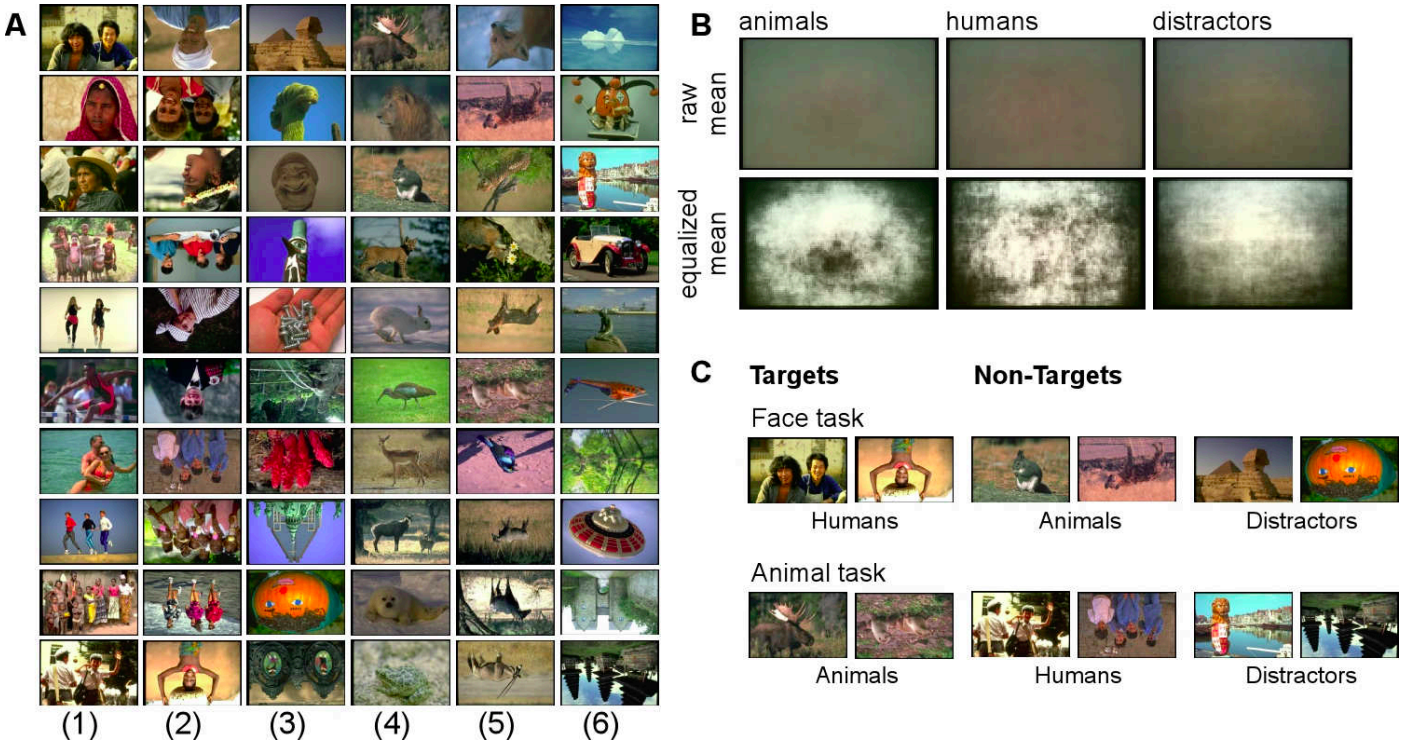


Figure 1. Tasks and stimuli. A. Examples of pictures used in Experiment 1. The 10 upright and inverted target pictures never missed by the subjects and associated with the fastest reaction time are presented for the face categorization task (columns 1 and 2, respectively) and for the animal categorization task (columns 4 and 5). Some examples of upright and inverted distractors that did not contain humans nor animals ("neutral" distractors) and on which subjects made no error are also illustrated in the upper and lower parts of column 3 for the face task and of column 6 for the animal task. B. Pixel-by-pixel average picture (raw mean) for each stimulus category (distractors refer to the neutral distractors) with equalized version computed using a commercial graphic software. The raw mean images were virtually uniform gray fields. The equalized images were obtained using the equalize function in a commercial graphic software. For each color channel and the luminance channel, the function attributes a “black" value to the darkest pixel and a "white" value to the brightest one. It then redistributes regularly the intermediate pixel values of the distribution between these two extremes. C. Tasks. While performing one of the two tasks, half of the non-targets were targets of the other task, and the other half were neutral distractors. Note the variety of stimuli used in this experiment.

diodes. Subjects were given 1000 ms to respond; longer reaction times were considered no-go responses. This maximum response time delay was followed by a 300-ms black screen, before the fixation point of the next trial was presented again for a variable duration, resulting in a random 1600-2200-ms intertrial interval.

An experimental session included 16 blocks of 96 trials. In 8 blocks, the target was an animal and in the remaining 8 blocks, the target was a human face. In each block, target and non-target trials were equally likely. Among the 48 non-targets, 24 contained targets of the other categorization task. Thus, when performing the face categorization task on a 96-trial block, 48 pictures contained at least one face, 24 non-target scenes contained animals, the last 24 non-targets “neutral distractors” being other types of natural scenes (see stimuli). Moreover, half of the targets and half of each of the non-target subsets were presented upright while the other half was presented inverted (180° rotation). Each image was seen only once by a given subject, with one orientation (upright or inverted) and one status (target or non-target), but the design was counterbalanced so that across all 24 subjects (1) each image (“neutral” distractor, animal or face image) was seen 12 times both in upright and inverted positions, and (2) each animal or face image was seen 16 times as a target and 8 times as a non-target. Half of the subjects started with the animal categorization task, the other half with the human face categorization task and conditions alternated by blocks of two. Subjects had two training blocks of 48 images before starting the test session. Training pictures were not repeated during testing.

Performance was evaluated by determining the percentage of correct trials and the latency at which subjects triggered their finger movement response, computed between stimulus onset and finger lift. An ANOVA was run on reaction times (RT) and rates of correct responses with category (animals vs. humans) and orientation (upright vs. inverted) as within-subject factors. A Greenhouse-Geisser correction for nonsphericity was applied.

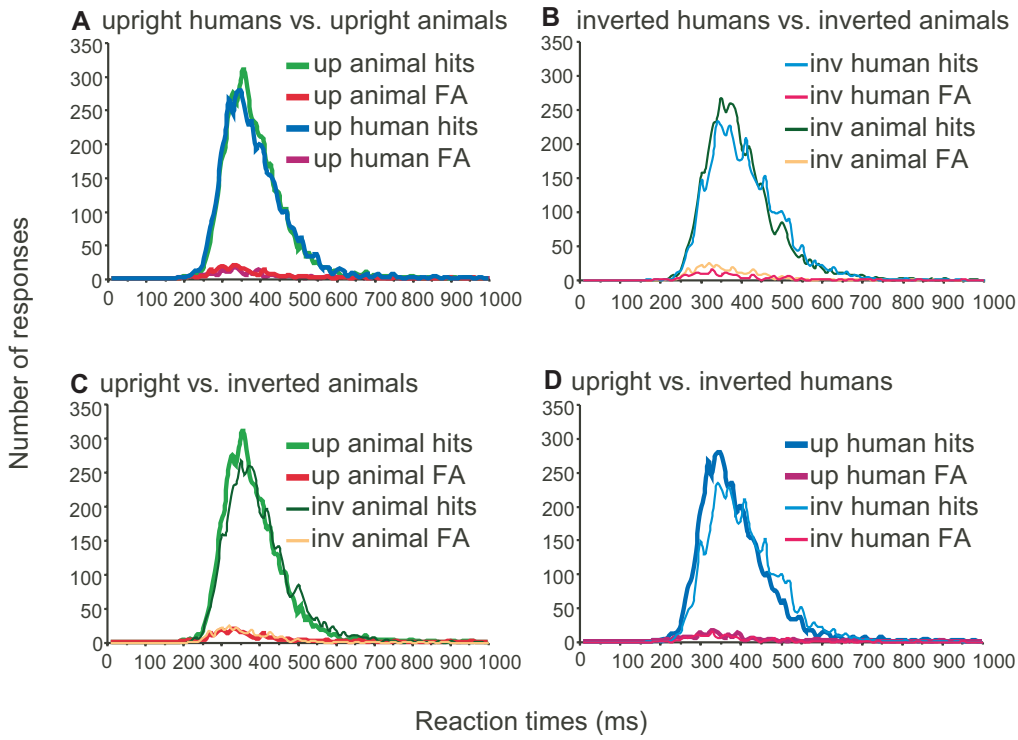


Figure 2. Reaction time (RT) distributions on correct and incorrect go-responses. RT distributions are presented with the number of responses expressed over time, with 10-ms time bins. Overall, no effect of the categorization task is seen on the early part of the RT distributions. Whether upright or inverted, responses to faces followed virtually the same time course as responses to animals (A and B). Inversion slightly disrupted the processing time course of both target-categories (C and D), an effect that was slightly more pronounced for faces.

Stimuli

We used photographs of natural scenes taken from a large commercial CD-ROM library (Corel Stock Photo Library, see Figure 1). From this database, we selected 576 images that contained human faces, 576 images that contained animals, and 384 images that contained neither human faces nor animals. They were all horizontal photographs (768 by 512 pixels, sustaining a visual angle of about 19.9° x 13.5°) and chosen to be as varied as possible. Animals included mammals, birds, fish, and reptiles. Human faces were presented in real-world situations with views ranging from whole bodies at different scales to face close-ups and including Caucasian and non-Caucasian people. There was also a wide range of non-target images that included outdoor and indoor scenes, natural landscapes (mountains, fields, forests, beaches, etc.), street scenes, pictures of food, fruits, vegetables, plants, buildings, tools, and other man-made objects, as well as some trickier distractors (e.g., dolls, sculptures, and statues, and a few non-target images containing humans for which the faces were not visible).

Subjects had no a priori information about the presence, the size, the position, or the number of targets in an image. Unique presentation of images prevented learning, and brief presentations prevented exploratory eye movements.

Results

In this section we will address three different aspects of processing: (1) processing of upright stimuli, comparing task performance for upright humans and upright animals; (2) processing of inverted stimuli, comparing inverted humans and inverted animals; and (3) effects of inversion on processing, comparing upright and inverted stimuli.

Overall, subjects were very accurate on both tasks, scoring 95.6% in the human task and 95.5% in the animal task (n.s.d.) and very fast (mean RT of 393 ms vs. 388 ms, respectively, n.s.d.). ANOVA tests performed on the overall results revealed that subjects categorized human targets with a lower accuracy than animal targets (95.7% vs. 98.3%, respectively; $F = 16$, $p = .001$), whereas they correctly ignored a higher proportion of distractors in the contextual face task than in the animal task (95.3% vs. 92.8%, respectively; $F = 20.8$, $p < .0001$). There was no main effect of category on mean and median RT. However, both measures presented a significant interaction between the category and orientation factors (both: $F = 18.0$, $p < .0001$). These main effects are

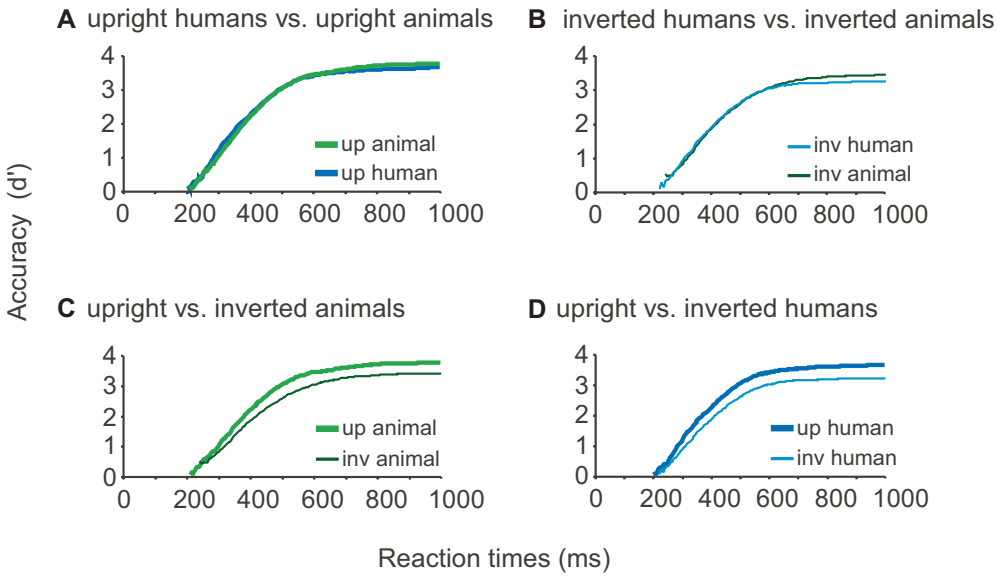


Figure 3. Time course of performance. Average performance accuracy (in d' units) is plotted as a function of processing time with 10-ms time bins. Cumulative numbers of responses were used. The d' was calculated from the formula $d' = z_n - z_s$, where z_n is chosen such that the area of the normal distribution above that value is equal to the false-alarm rate, and where z_s is chosen to match the hit rate. Note that the d' calculated here is not presumed to represent the actual distributions of signal and noise that underlie performance in the response time task. By taking into account the hit and false alarm rates in a single value at each time point, this time course of performance gives an estimation of the processing dynamics for the entire subject population. The plateau values correspond to the d' calculated from the overall accuracy results. Confirming results from Figure 2, performance time course functions were virtually identical for contextual human face and animal categories, independent of the orientation (i.e., upright or inverted). The inversion effect was very similar in both cases with a slightly earlier onset for human pictures.

explored in details in the two next sections using post hoc ANOVA, paired *t* tests, and Wilcoxon tests.

Contextual faces versus animals: upright stimuli

Here only the trials (over 9,200) performed in each task with upright scenes are considered. Mean accuracy was virtually identical in the two tasks (96.4% and 96.3% for faces and animals) (Figure 2A and Figure 3A).

Accuracy, however, was biased differently in each of them. Subjects categorized upright human targets with a lower accuracy than upright animal targets (humans = 97.5%, animals = 98.7%, Wilcoxon test, $z = -2.3$, $p = .02$), whereas no significant effect was present at the level of upright distractors (humans = 95.3%, animals = 93.9%, n.s.d.).

Regarding processing speed, upright contextual faces were not categorized faster than upright animals. First, this was shown by the RT distributions of correct go-responses in both tasks (Figure 2A). Second, there was no task effect on either mean (382 ms in both conditions) or median RT (368 ms for faces and 371 for animals) (Figure 2A and Figure 3A). Thus, on average, animals and faces were processed at the same speed according to mean and median RT. Given the problems associated with using only mean RT values to evaluate processing speed (Perrett, Oram, & Ashbridge, 1998; McElree & Carrasco, 1999), we used two more appropriate values: the time course of performance (Figure 3) and the minimal RT. The analysis of these two factors confirmed that contextual faces and animals were categorized at the same speed within natural images. Comparing the time course performances of each task (Figure 3A) clearly shows that early responses were produced at similar latencies regardless of the task and that performances follow time courses that are virtually undistinguishable. The minimal behavioral processing time was evaluated by determining the latency at which correct go-responses started to significantly outnumber incorrect go-responses ($\chi 2$, $p < .001$) using a noncumulated RT histogram with 10-ms time bins (Figure 2). These early responses cannot be considered as anticipations because if behavior was random on target and distractor trials (which are equally likely), hits and false alarms should have the same probability. The latency at which go-responses are statistically biased toward hits gives an indication of the minimal processing time required to trigger a motor response in the task while eliminating any bias due to anticipations. The analyses were performed either on the overall data (set by pulling together all trials from all subjects) or for each subject separately. No significant differences between the contextual face and the animal categorization tasks were found. The minimal processing

Table 1. Average Results From Experiment 1

	Contextual human face task		Animal Task	
	Upright scenes	Inverted scenes	Upright scenes	Inverted scenes
Accuracy (%)				
Mean	96.4 (1.7) [92.2-99.2]	94.7 (2.3) [88.3-98.2]	96.3 (2.0) [91.1-99.2]	94.8 (2.3) [89.8-98.4]
Correct go	97.5 (2.6) [90.1-100]	93.9 (4.9) [78.7-99.5]	98.7 (1.3) [95.3-100]	97.9 (1.4) [95.3-100]
Correct nogo (tD)	94.5 (5.9)	94.9 (5.0)	94.7 (4.1)	92.8 (4.2)
Correct nogo (nD)	96.1 (2.3)	95.8 (2.0)	93.1 (4.2)	90.5 (4.6)
RT (ms)				
Mean	382 (43) [317-468]	405 (49) [338-500]	382 (41) [312-465]	395 (43) [324-486]
Median	368 (43) [309-457]	391 (50) [317-484]	371 (42) [305-460]	380 (44) [298-470]
Minimal RT (ms)				
Overall data	260	260	260	260
Individual data	329 (43) [250-370]	353 (50) [270-430]	333 (35) [260-380]	348 (41) [270-460]

(tD) and (nD) refers respectively to the distractors that were used as targets in the other task or to the neutral distractors used in both tasks. SD is indicated in brackets. Range of individual responses (min and max) is indicated in square brackets.

time was 260 ms with the overall data set (for both faces and animals) and 329 ms (contextual faces) versus 333 ms (animals) for individual data. These results do not support any processing speed advantage for human faces.

Contextual faces versus animals: inverted stimuli

The comparison of performance did not show any difference between the processing of contextual human faces and animals when presented in an upright orientation. In our protocol, half of the stimuli were also presented upside down and the present section compares the processing of inverted contextual faces and inverted animals to investigate whether the similarity found with upright stimuli extends to inverted ones. As in the preceding section, the comparison is carried out on over 9,200 trials for each condition.

Mean accuracy was virtually identical for inverted faces (94.7%) and inverted animals (94.8%) (Figure 2B and Figure 3B). Accuracy showed the same biases than with upright stimuli, with a higher accuracy (97.9% vs. 93.9%; Wilcoxon test, $z = -4.1$, $p < .0001$) on inverted animal targets than on inverted contextual faces. Moreover, the higher accuracy on inverted distractors observed in the contextual face task (95.4%) when compared to the animal task (91.7%) was highly significant (Wilcoxon test, $z = -3.9$, $p < .0001$).

Figure 4 illustrates the higher number of errors performed on inverted distractors in the animal task both when compared to the set of upright stimuli in the animal task and when compared to the set of inverted distractors processed in the contextual face task. The figure also illustrates that, regardless of their orientation, neutral distractors induce a higher number of false alarms in the animal categorization task. Again this is true when compared to the other set of distractors in the animal task, or when compared to the performance on neutral distractors in the contextual face task.

When considering the average categorization speed, inverted faces were categorized about 10 ms slower than inverted animals. This was true (both paired t test $p < .006$) for both mean RT (405 ms and 395 ms, respectively, for contextual faces and animals) and median RT (391 ms and 380 ms, respectively) (Figure 2B and Figure 3B). However, this processing speed difference failed to reach statistical significance for the minimal processing time (as defined in the preceding section). Minimal RT was 260 ms, regardless of the kind of targets to categorize, when calculated on the overall data set. Mean minimal RT calculated on all individual subject data was 348 ms for animals and 353 ms for faces. The RT distributions and the performance time course functions for each task also show a good overlap of early responses regardless of the task. Differences are observed later (around mean RT or for late responses).

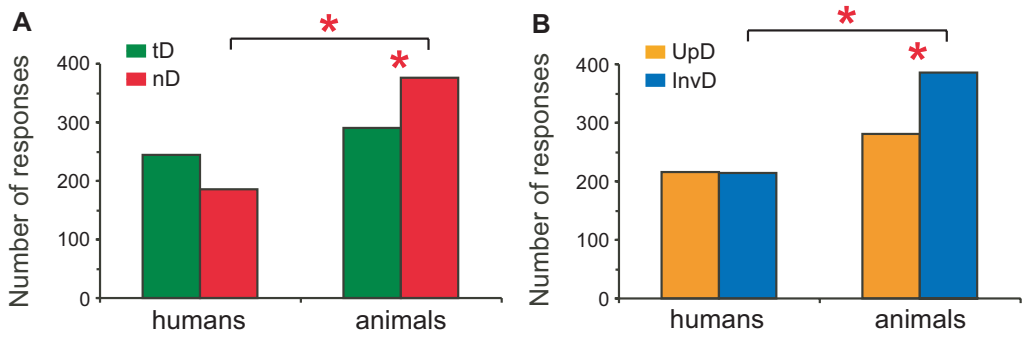


Figure 4. Analysis of incorrect go-responses made toward distractors in the "contextual human face" task and in the "animal" task. The data indicate a different processing of the distractors depending on the task performed by the subject. Statistically significant differences between two conditions are illustrated with an asterisk. A. Comparison of incorrect go-responses triggered by neutral distractors (nD in red) and by distractors that were targets in the other categorization task (tD in green). Independent of picture orientation, the responses on distractors showed a significant bias (interaction between task and type of distractor factors, F = .0, p = .002). More errors were made on neutral distractors in the animal task compared to the human face task (F = 36.9, p = .0001). Within the animal task, neutral distractors induced more errors than human faces (tD) (F = 6.8, p = .016). B. Comparison of incorrect go-responses triggered by upright (UpD in orange) and inverted (InvD in blue) distractors. An interaction between task and orientation factors (F = 7.0, p = .014) showed that more errors were made on inverted distractors in the animal task (F = 18.7, p = .0001), whereas no difference was seen in the contextual human face task (n.s.d.). Inverted distractors were also better categorized in the human face task than in the animal task (F = 37.5, p = .0001).

Contextual faces versus animals: the inversion effect

In this section, we focus more specifically on the presence and the strength of the inversion effect as a function of the target category.

Inversion produced a very weak decrease of global accuracy (<2%) that was very similar for both animals and human faces (orientation effect: F = 37.1, p < .0001; no interaction between task and orientation factors) (Figure 2C and 2D and Figure 3C and 3D). The percentage of correct go-responses decreased significantly with inversion for both animals (98.7% vs. 97.9%, Wilcoxon test, z = -2.7, p = .006) and contextual faces (97.5% vs. 93.9%, z = -4.1, p < .0001). Statistically, this was shown by a main orientation effect (F = 27.6, p < .0001) that was stronger for faces (interaction between orientation and task factors: F = 19.7, p < .0001). In parallel with the slight decrease of global accuracy, inverted pictures were also categorized on average with significantly longer RT (Figure 2C and 2D and Figure 3C and 3D) than upright pictures (mean RT: F = 140.7, p < .0001; median RT: F = 72.9, p < .0001). This held true for both categories but with an inversion effect on speed that was reliably more pronounced for faces (+23 ms on both mean and median RT, both paired t test: p < .0001) than for animals (+13 ms on mean RT, p < .0001; +9 ms on median RT, p = .001). Although the global reaction time increase appears robust with both kinds of inverted targets at the level of mean and median RT, it is far from being as obvious when considering the minimal processing time. When determined on the overall data, no effect was seen regardless of the categorization task. At the individual level, however, there was a small inversion effect for both categories with a nonsignificant tendency to be more pronounced for faces (+24 ms, p < .0001) than for animals (+15 ms, p = .004). The time course of performance showed that the stimulus inversion did not simply shift the curve toward longer latencies but rather decreased the slope of the functions that originate at similar early latencies.

Discussion

Overall, subjects were able to respond both very accurately and rapidly in the two tasks. This level of performance is impressive given the extreme variability of the photographs used in this experiment. It can be taken as the hallmark of the sophistication of the fast mechanisms implemented in the ventral pathway of the human brain (Riesenhuber & Poggio, 2000; Thorpe & Imbert, 1989; VanRullen et al., 1998; Wallis & Rolls, 1997). If this conclusion had already been reached from results of earlier studies, here we extend these findings by showing that (1) the fast coarse categorization of objects in natural scenes is very weakly affected by inversion; (2) contextual human faces cannot be processed faster or more efficiently than another relevant visual category such as animals; and (3) the inversion effect, although very weak in both tasks, is slightly more pronounced for faces.

The fact that animals are processed with the same speed and accuracy as contextual human faces when both types of targets are presented at different scales, in varied

number and position in the image, argues against a hardwired face mechanism that would be more efficient than other non-face object mechanisms (Tarr & Cheng, 2003). Because it has been shown previously that animals could not be processed faster than another relevant, nonbiological category, such as means of transport (VanRullen & Thorpe, 2001a,2001b), contextual faces cannot be said to benefit from specific temporal advantages, at least in our task. We do not want to argue that this kind of rapid categorization process would apply to any object category; instead, it might depend on a certain level of expertise (that needs to be determined) beyond which the categorization of any behaviorally relevant object could rely on such fast processes.

Although we found evidence that inversion of natural scenes did produce reliable effects on performance, with responses delayed (13 ms vs. 23 ms for animal and faces) and accuracy impaired for inverted pictures (1% vs. 3.5% for animal and faces), it is important to note that these effects were both very weak (although slightly more pronounced for faces). With such temporal constraints, very little time would be available to implement a mental rotation mechanism during the time course of the categorization process. On the other hand, the speed of recognition of an object might depend on the rate of accumulation of activity from object selective neurons (Perrett et al., 1998; Ashbridge, Perrett, Oram, & Jellema, 2000). Neurons in higher-level occipito-temporal visual areas respond to complex stimuli such as animals and faces. At the level of neuronal populations, the strength of the population response is correlated to the number of activated neurons. Now, we can hold the very plausible assumption that the population response must reach a given constant threshold activation level (Hanes & Schall, 1996) in order for a behavioral response to be triggered. Through experience, more neurons, each one more selectively tuned, respond to animals, human faces, and body parts in the upright position compared to inverted positions. Groups of neurons responding to upright and inverted objects would start to respond at about the same latency but responses would accumulate more slowly in the case of inverted stimuli, leading to an increase in response latency. This hypothesis is supported by the time course of performance (Figure 3) that originated at similar latencies but increased with different slopes depending on whether the stimuli were presented upright or upside down. It follows that, on average, it takes slightly more time to reach the threshold for inverted stimuli, and therefore to categorize them.

If the processing of upright faces and animals followed the same behavioral temporal course, what is special in faces that led to differences in the processing of inverted stimuli? The inversion effect is usually taken as evidence that face processing relies preferentially on configural mechanisms distinct from part-based mechanisms thought to be more important in the processing of other objects (e.g., see review in Itier & Taylor, 2002; Rossion & Gauthier, 2002). When faces appear in their typical upright orientation, configural information is extracted. This extraction is disrupted by inversion, except for objects whose discrimination relies on characteristic features that are not affected by inversion. However, following Perrett's hypothesis, the fact that faces were more sensitive to inversion than animals can be explained by a face population selectivity more strictly linked to the canonical upright view through experience (see support for such a view in Rossion & Gauthier, 2002; Tarr & Gauthier, 2000). Accordingly, neurons would fire less efficiently in response to inverted than upright faces, leading to a smaller accumulation of activity for inverted faces compared to inverted animals (because the latter might be represented by a cell population less strictly tuned to the upright orientation). As a consequence, the stronger inversion effect for faces often explained by the specificity of face processing (Farah et al., 1995, 1998) can be alternatively explained by the rate of accumulation of selective neural activity.

However, it remains possible that different strategies or brain mechanisms were used in the two tasks. Inversion had different effects on each category: when looking for animals, subjects made a high number of incorrect responses on inverted distractors, whereas when looking for human faces, they tended to miss more inverted targets. This could be the consequence of a greater similarity between animals and distractors than between faces and distractors, and the use of more specific representations to perform the face task than the animal task. This hypothesis is supported by the fact that more errors on neutral distractors and on inverted distractors were performed during the animal task than during the face task.

Finally, animals were slightly more easily detected in natural scenes than faces, which might indicate that the two sets of images were not equated in difficulty and might potentially have masked a processing speed advantage in favor of faces. Furthermore, this discrepancy might also potentially explain the very weak inversion effect found for faces. To test these alternative explanations and further characterize the processing of faces in natural scenes, we designed a second experiment.

Experiment 2

Experiment 2 was designed to compare the rapid categorization of faces and animals with more homogenous sets of images. In Experiment 2, subjects were only presented with close-up views of human and animal heads and were required to categorize human faces and animal faces. Human and animal faces were chosen to be as varied as possible but always in the context of natural scenes; furthermore, neutral distractor pictures (that did not contain animal or human faces)

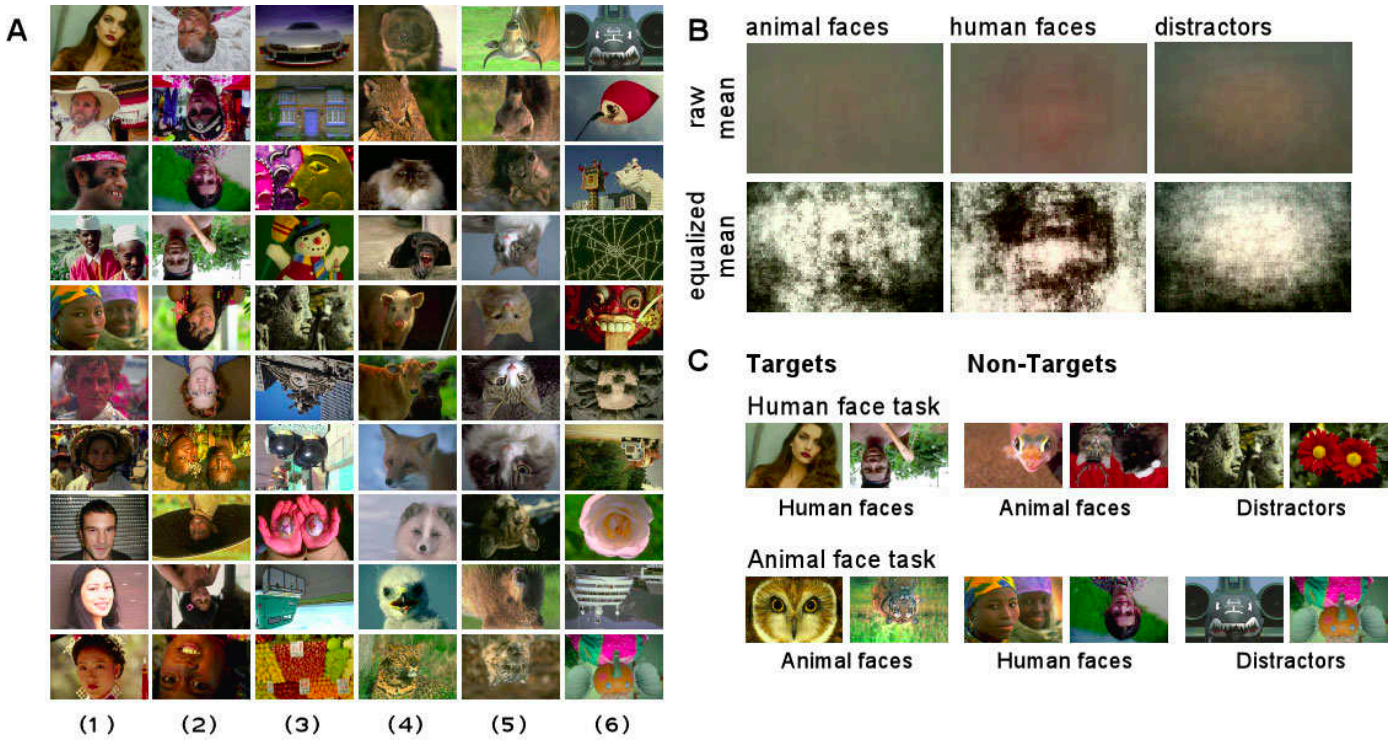


Figure 5. Picture examples and experimental design. Nomenclature as in Figure 1.

were chosen to include "tricks," such as dolls, statues, flowers, and other headlike "blobs."

Methods

Except where otherwise mentioned, methods were identical to those used in Experiment 1.

Participants

The 24 human participants (12 women and 12 men, mean age 30 years, ranging from 19 to 51 years, 3 left handed) who volunteered in this study gave their informed written consent. Nine of them had participated in the first experiment. All participants had normal or corrected-to-normal vision.

Experimental procedure

An experimental session included 8 blocks of 96 trials. Subjects performed two categorization tasks: in 4 blocks the target was an animal face and in the 4 other blocks the target was a human face. In each block, target and non-target trials were equally likely. Among the 48 non-targets, 24 were targets in the other categorization task. Thus, when performing a human face categorization task on a 96 trial block, 48 pictures contained at least one human face, 24 non-target scenes contained animal faces, the last 24 non-targets being neutral distractors (i.e., other types of natural scenes and "trick" stimuli) (see Stimuli and Figure 5). Half of the targets and half of each non-target subset were presented upright while the other half was presented inverted. The design was counterbalanced so that in the overall group of subject, each image was seen in upright and inverted positions and processed as a target and as a non-target. Half of the subjects started with the animal face categorization, the other half with the human face categorization. Subjects had one training block before starting each of the two test sessions. Training pictures were not used during testing.

Stimuli

A total of 768 photographs were selected from the Corel Stock Photo Library; 288 contained human faces, 288 additional images contained animal faces, and the last 192 photographs contained neither human nor animal faces (Figure 5). They were all horizontal photographs (768 by 512 pixels, sustaining about 19.9° by 13.5° of visual angle) and chosen to be as varied as possible. Faces were always highly visible with views ranging from close-up to views showing the most upper part of the body. Animals included mammals, birds, fish, and reptiles. They did not include arthropods and were chosen so that a face configuration could always be seen (eyes, mouth, and nose). Human faces were presented in real-world situations and included humans from all over the world. There was also a very wide range of non-target images that included outdoor and indoor scenes, natural landscapes, street scenes, pictures of food, fruits, vegetables, plants, flowers, buildings, tools, and other man-made objects, as well as many "tricky" distractors, such as dolls, sculptures, and statues. A particular attempt

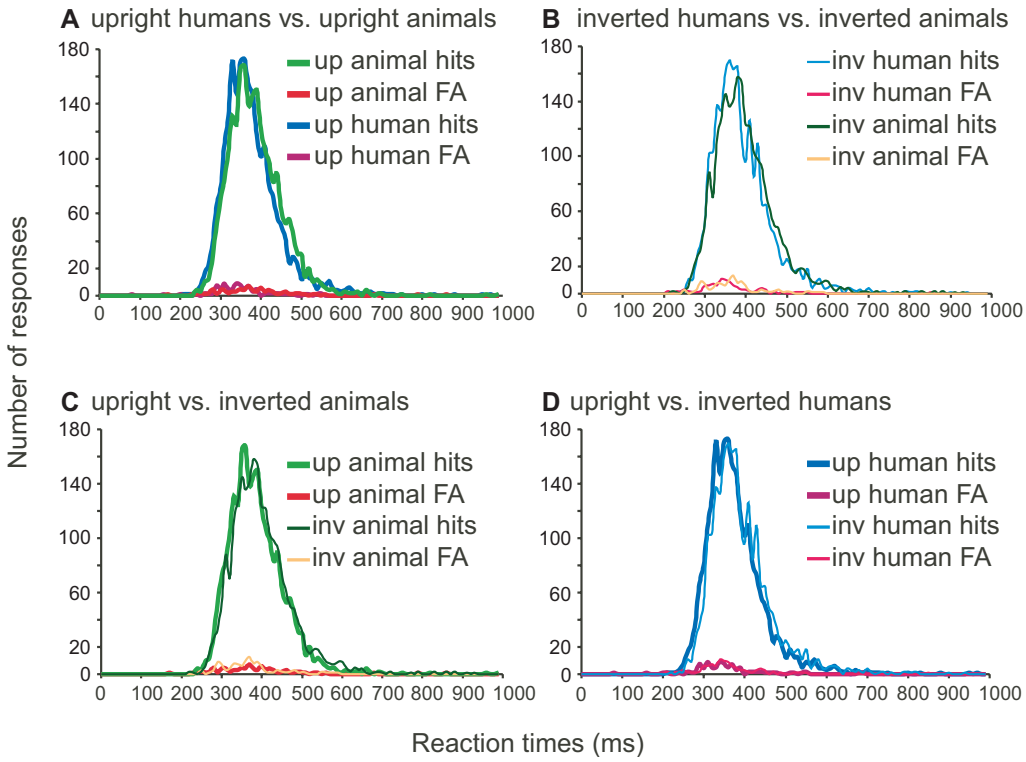


Figure 6. Reaction times (RT) distributions on correct and incorrect go-responses. (See caption Figure 2.) Overall, no effect on processing speed is seen on the early part of the RT distributions except in D, where the hits on upright human faces start to diverge early from the hits on inverted faces. Whether upright or inverted, responses to human faces followed virtually the same time course as responses to animal faces (A and B). Inversion slightly disrupted the processing time course of both target-categories (C and D), an effect that was slightly more pronounced for faces.

was made for most distractors to have one or more headlike "blobs" positioned centrally or laterally in the picture, as were human and animal faces.

Subjects had no a priori information about the presence, the size, the position, or number of targets in an image, and to prevent learning, each image was seen only once in one orientation (upright or inverted), either as a target or as a non-target, by each subject.

Results

In Experiment 2, despite the greater target/distractor similarity compared to Experiment 1, the use of close-up views led to excellent performances both in terms of accuracy and speed. ANOVA tests performed on the overall results showed no category effect on global accuracy (97.4% for both human and animal faces), target accuracy (99.3% for both) or distractor accuracy (95.5% for both). However, median RT were shorter in response to human faces (377 ms) than to animal faces (387 ms) ($F = 4.6$, $p = .043$), a main effect that was not significant for mean RT (humans: 389 ms; animals: 397 ms). The next two sections will present a detailed analysis of these global results using post hoc ANOVA, paired t tests, and Wilcoxon tests. The first section will compare the processing of upright human faces to the processing of upright animal faces. The second section will concentrate on inverted stimuli. The third section will present specifically the differences between upright and inverted stimuli on the processing of human faces and animal faces.

Human faces versus animal faces: upright stimuli

Mean accuracy was virtually identical for both kinds of upright pictures with 97.7% in the human face task versus 97.9% in the animal face task (Figure 6A and Figure 7A). Targets were better categorized than non-targets (99.5% vs. 96%, respectively, $F = 37.7$, $p < .0001$), with similar proportions of go-responses for upright humans (99.6%) and upright animals (99.5%). Contrary to Experiment 1, subjects tended, on average, to respond about 10-ms faster for human than for animal faces (Figure 6A and Figure 7A). This slight advantage reached significance for median RT (371 ms vs. 384 ms, paired t test: $p = .031$) but not for mean RT (382 ms vs. 392 ms, n.s.d.). This effect is relatively clear on the RT distribution for intermediate and long latency responses. On the other hand, although it is barely visible on the initial part of the RT distribution of Figure 6A or at the

onset of the performance time course functions of Figure 7A, the 10-ms global advantage in favor of human compared to animal pictures was also observed with the minimal processing time computed on cumulated population data (260 ms vs. 270 ms, respectively). The same tendency in favor of human pictures was seen for individual minimal processing time in both tasks, but it did not reach significance (327 ms vs. 338 ms, n.s.d.).

Human faces versus animal faces: inverted stimuli

No statistical difference could be seen between the accuracy scores computed for each task. Indeed, subjects again reached very similar performances (Figure 6B and Figure 7B) scoring 97.2% with inverted human faces and 96.9% with animal faces. Correct go responses were triggered in similar proportion in both tasks (99.0% vs. 99.2%).

The overall mean RT showed a 6-ms lag between human face (396 ms) and animal face processing (402 ms) that did not reach significance. This lag reached 8 ms when calculated on the overall median RT between human faces (median RT: 382 ms) and animal faces (median RT: 391 ms), an effect that did not reach significance either.

When it was calculated on the overall population data, the earliest responses were found earlier for animal faces (270 ms) than for human faces (280 ms). A pattern that was not consistent when individual data were considered as mean individual data showed a nonsignificant advantage for inverted animal faces (345 ms) versus human faces (335 ms) (Figure 6B and Figure 7B).

As in the first experiment, the incorrect go responses produced on distractors were analyzed (Figure 8) and outlined different biases depending on the task performed by the subject. As in Experiment 1, subjects made fewer errors on neutral distractors in the human face task than in the animal face task, regardless of their orientation (Figure 8A), but a bias was found within the human face task for the two different subsets of distractors: subjects made more errors on pictures that contained animals than on neutral distractors. Finally, Figure 8B shows the same bias as that already seen in Experiment 1, with more errors on inverted stimuli in the animal task.

Human faces versus animal faces: the inversion effect

As in Experiment 1, inversion had a reliable but weak effect on performance. Inversion decreased global

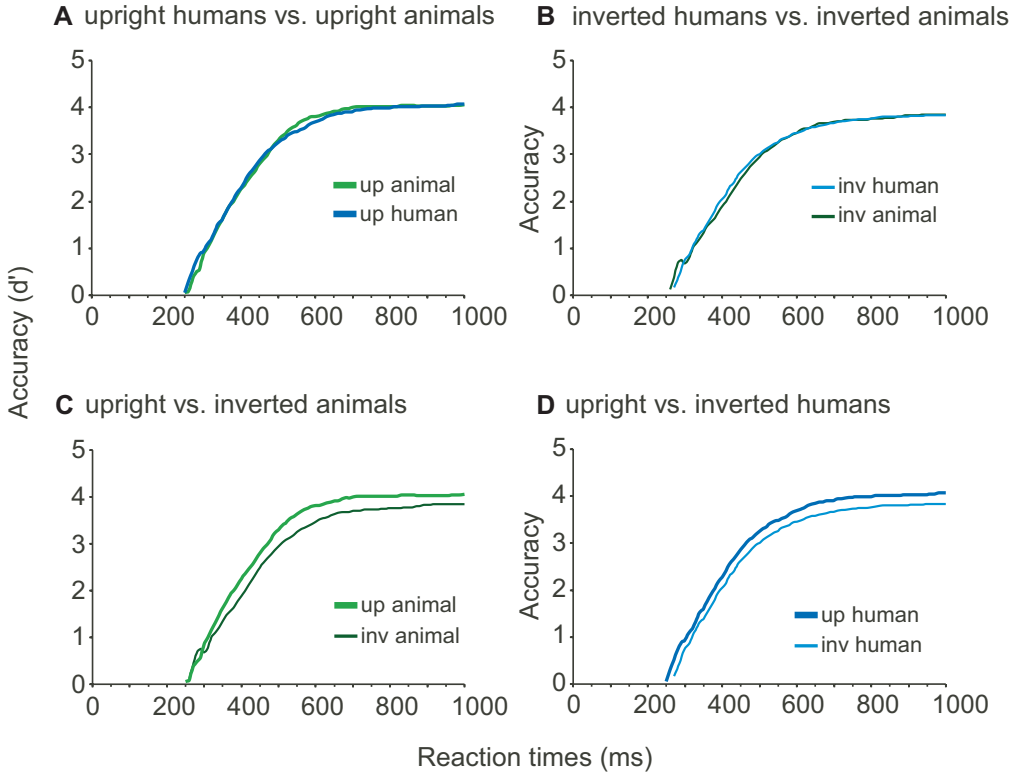


Figure 7. Performance time course. (See caption Figure 3.) A and B show that human and animal faces follow the same type of processing course. C and D show the slight decrease of accuracy in both tasks and the temporal cost associated with inverted stimuli. The temporal cost is seen from the very beginning with human faces whereas the d' curves for upright and inverted animal faces, initially superimposed, diverge later on.

Table 2. Summary of Results From Experiment 2

	Human face task		Animal face task	
	Upright stimuli	Inverted stimuli	Upright stimuli	Inverted stimuli
Accuracy (%)				
Mean	97.7 (1.8) [92.1-100]	97.2 (1.7) [91.0-99.5]	97.9 (1.3) [95.7-100]	96.9 (1.6) [93.2-99.5]
Correct go	99.6 (1.3) [93.6-100]	99.0 (1.2) [95.8-100]	99.5 (0.9) [95.8-100]	99.2 (0.8) [97.9-100]
Correct nogo (tD)	94.6 (6.1)	93.9 (6.4)	96.6 (3.7)	94.5 (4.1)
Correct nogo (nD)	97.0 (3.1)	96.8 (2.5)	95.8 (3.0)	94.9 (4.2)
RT (ms)				
Mean	382 (33) [338-445]	396 (28) [352-444]	392 (35) [328-479]	402 (36) [337-493]
Median	371 (31) [330-428]	382 (26) [338-431]	384 (37) [312-464]	391 (34) [328-468]
Minimal RT (ms)				
Overall data	260	280	270	270
Individual data	327 (27) [290-380]	335 (22) [290-400]	338 (26) [290-410]	345 (31) [270-420]

(tD) and (nD) refers respectively to the distractors that were used as targets in the other task or to the neutral distractors used in both tasks. SD is indicated in brackets. Range of individual responses (min and max) is indicated in square brackets.

accuracy in both tasks (-0.5% in the human face task, -1% in the animal face task, $F(1,23) = 8.3$, $p = .008$) (see Figure 6C and 6D and Figure 7C and 7D). This effect was only significantly reliable for animal faces (Wilcoxon test, $z = -2.5$, $p = .013$; human faces: n.s.d.). When considering accuracy on targets and distractors separately, the inversion effect, albeit very small, reached significance only for go-responses on human faces ($z = -2.1$, $p = .039$) and for no-go responses on animal faces ($z = -2.0$, $p = .042$).

Inversion also slightly delayed RT (mean: +14 ms and +10 ms, $F(1,23) = 58.3$, $p < .0001$; median: +11 ms and +7 ms, $F(1,23) = 34.7$, $p < .0001$, for human and animal faces, respectively), an effect that was not significantly stronger for human than for animal pictures, as shown by an absence of interaction between task and orientation factors. However, the result concerning minimal RT calculated from the overall population data showed a difference between early processing of human and animal faces. There was no effect of orientation for animal faces (270 ms for upright and inverted stimuli), but the minimal RT was 20 ms shorter with upright faces (260 ms) than inverted faces (280 ms). This small differential effect between the two tasks can be seen in Figure 7 by comparing the initial part of the d' curves in Figure 7C and 7D. The performance curve with inverted human faces is shifted toward longer latency with the same slope than for upright faces, whereas with animal faces, the earliest responses appear at the same latency, and only the slope of the performance curve is affected when inverting the animal faces. However, this result on the overall data set was not confirmed by the analysis of individual minimal reaction time showing the same inversion effect for human faces (+9 ms) and animal faces (+7 ms) ($F = 16.5$, $p < .0001$, no interaction with the category factor).

Discussion

Experiment 2 tried to provide a more direct comparison of human face versus animal face processing in natural scenes by using more homogenous sets of images. Levels of difficulty in the two tasks were similar regarding target detection accuracy. Despite high feature similarities between targets, and despite our considerable effort to use confusing distractors sharing global features with close-ups of faces, subjects performed remarkably well in these two tasks, in which processing efficiency was virtually identical. The high accuracy level reached in this experiment might be explained by the fact that humans (and faces in particular) constitute a very special object class, automatically categorized and segregated by our visual system, hence producing no interference with other object categories. Indeed, as in Experiment 1, we found evidence that neutral distractors were associated with more errors in the animal task than in the human task, which might imply that there was a higher similarity between neutral distractors and animals than between neutral distractors and humans. However, does this mean that human faces would benefit from computational advantages that would make them easier or faster to detect? We found no clear evidence in favor of this hypothesis. In the present experiment, contrary to the first one, there was a tendency for human faces to be processed on average about 10-ms faster than animal faces, an advantage that was present for both upright and inverted orientations, but appeared only for upright stimuli when considering the earliest behavioral responses. Such a small but reliable effect might be explained at the neuronal population level by a larger number of neurons coding for human faces than for different animal faces, thus slightly reducing the time to threshold decision as previously postulated in the

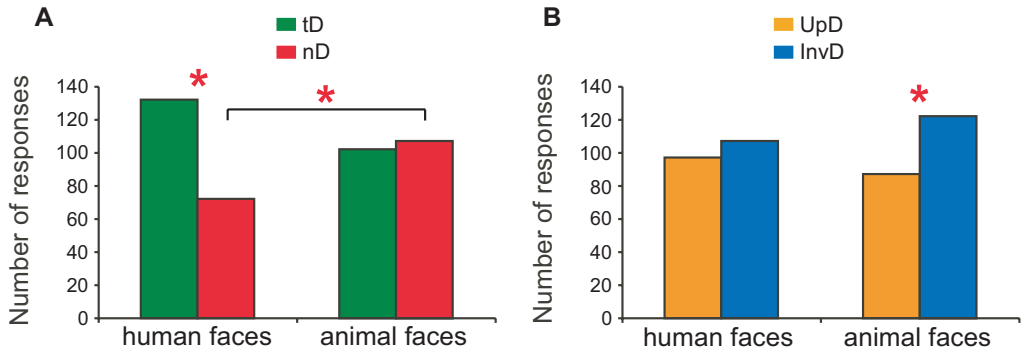


Figure 8. Analysis of incorrect go-responses made on distractors in the human and in the animal face tasks. (See Figure 4 caption for details.) A. Independently of picture orientation, the responses on distractors showed a significant bias (interaction between the task and type of distractor factors, F = 4.8, p = .04). Neutral distractors were slightly better categorized in the face task than in the animal task (96.9% vs. 95.3%, respectively, F = 7.5, p = .012). Within the human face task, animal faces (tD) induced more errors than neutral distractors (F = 4.5, p = .045). B. Furthermore, the orientation of the distractors induced a bias only in the animal task in which more errors were induced by inverted than by upright distractors (F = 7.0, p = .014).

discussion of Experiment 1. Indeed, a 10-ms difference in processing speed does not fit with the involvement of a different mechanism for the processing of human faces compared to animal faces, but rather point to a quantitative difference in the processing of the two categories rather than to a qualitative difference. The time course of performance in the two tasks strengthen this interpretation. Thus, these results are best explained in a framework in which the ventral pathway is conceived as implementing a unitary mechanism processing all object categories (Tarr & Cheng, 2003). Under such a framework, the speed to categorical decision threshold would depend on the number of neurons tuned to a specific category. According to this working hypothesis, time to threshold would be not surprisingly shorter for an extensively represented category such as human faces compared to another object category such as animal faces. Following this idea, a delay as short as 10 ms can find an explanation at the level of a neuronal population more sensitive to a category than the other, rather than in the involvement of a totally different mechanism. This delay was rather small probably due to the task used in the two experiments used here. Indeed, a superordinate categorization task might rely on coarsely defined diagnostic features (Schyns, 1998; Ullman, Vidal-Naquet, & Sali, 2002). A processing time course similar to the one found here for animals and humans has also been reported for another category like means of transport (VanRullen & Thorpe, 2001a), which suggests that the same level of complexity might be reached in a large range of natural scene categorization tasks. The use of more demanding categorization tasks relying on more specific features might reveal more dramatically an existing bias at the neuronal population level between two categories. If subjects had been asked to realize a gender discrimination task with human and animal faces, the difference between the two categories would certainly have been much larger. However, even in this condition, the same simple mechanism of accumulation of evidence working at the level of a large neuronal population might be sufficient to explain the results. This kind of experiment will be important in the future to distinguish between different models of organization of the ventral pathway.

A complementary interpretation on the small difference in processing speed between human and animal faces lies in the smaller range of variability between different human faces compared to the large differences between faces of vertebrate animals (birds, monkeys, antelopes, reptiles, etc.). This seemed to be partly the case, given that more structure appeared in the "mean image" for humans than for animals (Figure 5B). It might be that reducing the number of different animal species would have allowed a more specific pre-setting of the neuronal population responding to animals, thus eliminating any differences at all between animal and human faces.

As in Experiment 1, another weak but consistent effect was seen with inversion in both tasks. Whereas the accuracy impairment appeared to be of similar magnitude for animal and human faces, the earliest response to inverted human faces could appear with a 20-ms delay when compared with upright human faces. This might be the hallmark of face configural processing, more disrupted by inversion than other object processing routines (Yin, 1969). However, as already developed in the discussion of the first experiment, a more simple explanation, emphasizing experience-induced bias at the neuronal population level, could constitute a viable alternative. According to this model, there is no need to call for the involvement of a mental rotation mechanism or a mechanism specifically dedicated to the processing of upright human faces. One might argue that models of object recognition relying on a time consuming normalization stage between sensory inputs and memory

templates might explain the inversion effects in our two experiments (Tarr & Bülthoff, 1998; Ullman, 1996). However, although we found reliable inversion effects, the maximal increase in processing time was about 20 ms. Thus, if a normalization mechanism (e.g., mental rotation) had to be done at the neuronal level it would have to fit in this demanding 20-ms time window. Instead, it has been suggested that whatever the orientation, neuronal responses start to accumulate at the same latency at the population level (Perrett et al., 1998). Life experience, in which stimuli appear more often in the upright orientation, would bias the population selectivity so that more cells respond to upright than inverted stimuli (Ashbridge et al., 2000). As a consequence, neuronal responses would accumulate faster to reach the categorization threshold in the former rather than in the later case. By integrating both category and orientation biases in this simple mechanism, it is possible to explain the larger orientation effect on processing speed in the human than in animal face task. Our results support this view because we did find a robust inversion effect for animals. Again, this explanation directly supports models of object processing in which there are quantitative rather than qualitative differences between human faces and other object categories. From the point of view emphasized in the first section of this discussion, larger inversion effects for human faces might be found as task requirements become more demanding. Indeed, if the strength of the inversion effect was stronger for human faces than for animal faces in the superordinate categorization task used here, this difference was not extremely important, and might be related to task instructions. A more important disruption of human face processing compared to other objects is found when subjects are asked to perform a recognition task (Diamond & Carey, 1986; Yin, 1969). This effect might be explained by the use of more specific representations that are themselves more specifically tuned to the orientation in which they have been learned. In keeping with this hypothesis, it has been shown that non-face object categories can present the same inversion effect as faces in a recognition task if subjects are experts at distinguishing between individuals of these categories (Diamond & Carey, 1986; Gauthier & Tarr, 1997). It follows that an apparent dichotomy between face and non-face object processing, such as the strength of the inversion effect, is not necessarily the hallmark of an independent face system; alternatively, it could reflect one point along a continuum of dynamically changing computational strategies (Riesenhuber & Poggio, 2002; Tarr & Cheng, 2003; Tarr & Gauthier, 2000).

These two experiments showed that in the context of natural scenes, faces are categorized following a time course very similar to another biological object category such as animals. Because it has been demonstrated that a nonbiological object category such as vehicles could be processed as efficiently as the animal category (VanRullen & Thorpe, 2001a, 2001b), it might well be that every well known object category could be selected in a "glimpse" by a wave of processing in the ventral pathway (Riesenhuber & Poggio, 2000, 2002; VanRullen et al., 1998). Given the strong temporal constraints in these tasks, with selective responses appearing as early as 260 ms, such a fast coarse categorization process might rely on the activation of neurons selective to visual diagnostic properties by an essentially feed-forward flow of activation. Furthermore, the relatively weak inversion effects found in these experiments indicate that the representations activated to categorize a natural scene are relatively coarse, at least coarser than several high-level properties that have been found to be strongly affected by inversion (Tarr & Bülthoff, 1998). It might thus suggest that this kind of fast visual categorization of complex stimuli do not necessarily rely on similarly complex high-level representations, but might rather be achieved through the detection of diagnostic features of intermediate complexity (Ullman et al., 2002). Further experiments will be necessary to precisely determine the nature of these representations. This pattern of results is overall compatible with models that suggest the existence of a single object processing system whose performance is modulated by expertise, level of recognition, and information availability (Perrett et al., 1998; Schyns, 1998; Tarr & Cheng, 2003). The interplay between these different factors would determine the efficiency of the system, without requiring any face-specific module, or any mental rotation mechanism.

Acknowledgments

We kindly acknowledge Nadège M. Bacon for her help in programming image presentation in Experiment 2 and Caitlin R. Sternberg and Anne-Sophie Paroissien for their help in testing subjects. We thank Roxane J. Itier and Rufin VanRullen for their valuable comments on an earlier version of the manuscript. This work was supported by the CNRS and the Cognitique grant n°IC2. Financial support was provided to both G.A.R. Rousselet and M.J.-M. Macé by a Ph.D. grant from the French government. Commercial relationships: none.

References

Ashbridge, E., Perrett, D. I., Oram, M. W., & Jellema, T. (2000). Effect of image orientation and size on object recognition: Responses of single units in the macaque monkey temporal cortex. *Cognitive Neuropsychology, 17*, 13-34.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience, 8*, 551-565.

Carmel, D., & Bentin, S. (2002). Domain specificity versus expertise: Factors influencing distinct processing of faces. *Cognition, 83*, 1-29.[PubMed]

Chelazzi, L., Duncan, J., Miller, E. K., & Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of Neurophysiology, 80*, 2918-2940. [PubMed]

Debruille, J. B., Guillem, F., & Renault, B. (1998). ERPs and chronometry of face recognition: Following-up Seeck et al. and George et al. *Neuroreport, 9*, 3349-3353. [PubMed]

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*, 1-37. [PubMed]

Delorme, A., Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2003). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. Manuscript submitted for publication.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115*, 107-117. [PubMed]

Eimer, M. (2000). The face-specific N170 component reflects late stages in the structural encoding of faces. *Neuroreport, 11*, 2319-2324. [PubMed]

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*, 171-180. [PubMed]

Farah, M. J., Wilson, K. D., Drain, H. M., & Tanaka, J. R. (1995). The inverted face inversion effect in prosopagnosia: Evidence for mandatory, face-specific perceptual mechanisms. *Vision Research, 35*, 2089-2093. [PubMed]

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychology Review, 105*, 482-498. [PubMed]

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research, 37*, 1673-1682. [Pubmed]

George, N., Jemel, B., Fiori, N., & Renault, B. (1997). Face and shape repetition effects in humans: A spatio-temporal ERP study. *Neuroreport, 8*, 1417-1423. [PubMed]

Halgren, E., Raij, T., Marinkovic, K., Jousmaki, V., & Hari, R. (2000). Cognitive response profile of the human fusiform face area as determined by MEG. *Cerebral Cortex, 10*, 69-81. [PubMed]

Halit, H., de Haan, M., & Johnson, M. H. (2000). Modulation of event-related potentials by prototypical and atypical faces. *Neuroreport, 11*, 1871-1875. [PubMed]

Hanes, D. P., & Schall, J. D. (1996). Neural control of voluntary movement initiation. *Science, 274*, 427-430. [PubMed]

Itier, R. J., & Taylor, M. J. (2002). Inversion and contrast polarity reversal affect both encoding and recognition processes of unfamiliar faces: A repetition study using ERPs. *Neuroimage, 15*, 353-372. [PubMed]

Jeffreys, D. (1996). Evoked potential studies of face and object processing. *Visual Cognition, 3*, 1-38.

Jolicoeur, P. (1988). Mental rotation and the identification of disoriented objects. *Canadian Journal of Psychology, 42*, 461-478. [PubMed]

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience, 3*, 759-763. [PubMed]

Linkenkaer-Hansen, K., Palva, J. M., Sams, M., Hietanen, J. K., Aronen, H. J., & Ilmoniemi, R. J. (1998). Face-selective processing in human extrastriate cortex around 120 ms after stimulus onset revealed by magneto- and electroencephalography. *Neuroscience Letters, 253*, 147-150. [PubMed]

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: An MEG study. *Nature Neuroscience, 5*, 910-916. [PubMed]

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Science, 6*, 255-260. [PubMed]

McElree, B., & Carrasco, M. (1999). The temporal dynamics of visual search: Evidence for parallel processing in feature and conjunction searches. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1517-1539. [PubMed]

Mouchetant-Rostaing, Y., Giard, M. H., Bentin, S., Aguera, P. E., & Pernier, J. (2000a). Neurophysiological correlates of face gender processing in humans. *European Journal of Neuroscience, 12*, 303-310. [PubMed]

Mouchetant-Rostaing, Y., Giard, M. H., Delpuech, C., Echallier, J. F., & Pernier, J. (2000b). Early signs of visual categorization for biological and non-biological stimuli in humans. *Neuroreport, 11*, 2521-2525. [PubMed]

Perrett, D. I., Oram, M. W., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: An account of generalisation of recognition without mental transformations. *Cognition, 67*, 111-145. [PubMed]

Pizzagalli, D., Regard, M., & Lehmann, D. (1999). Rapid emotional face processing in the human right and left brain hemispheres: An ERP study. *Neuroreport, 10*, 2691-2698. [PubMed]

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3* (Suppl.), 1199-1204. [PubMed]

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology, 12*, 162-168. [PubMed]

Rossion, B., Gauthier, I., Tarr, M. J., Despland, P., Bruyer, R., Linotte, S., & Crommelinck, M. (2000). The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: An electrophysiological account of face-specific processes in the human brain. *Neuroreport, 11*, 69-74. [PubMed]

Rossion, B., & Gauthier, I. (2002). How does the brain process upright and inverted faces? *Behavioral and Cognitive Neuroscience Reviews, 1*, 62-74.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5*, 629-630. [PubMed]

Schendan, H. E., Ganis, G., & Kutas, M. (1998). Neurophysiological evidence for visual perceptual categorization of words and faces within 150 ms. *Psychophysiology, 35*, 240-251. [PubMed]

Schyns, P. G. (1998). Diagnostic recognition: Task constraints, object information, and their interactions. *Cognition, 67*, 147-179. [PubMed]

Seeck, M., Michel, C. M., Mainwaring, N., Cosgrove, R., Blume, H., Ives, J., Landis, T., & Schomer, D. L. (1997). Evidence for rapid face recognition from human scalp and intracranial electrodes. *Neuroreport, 8*, 2749-2754. [PubMed]

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. *Cognition, 67*, 1-20. [PubMed]

Tarr, M. J., & Cheng, Y. D. (2003). Learning to see faces and objects. *Trends in Cognitive Sciences, 7*, 23-30. [PubMed]

Tarr, M. J., & Gauthier, I. (2000). FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience, 3*, 764-769. [PubMed]

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21*, 233-282. [PubMed]

Taylor, M. J., Edmonds, G. E., McCarthy, G., & Allison, T. (2001). Eyes first! Eye processing develops before face processing in children. *Neuroreport, 12*, 1671-1676. [PubMed]

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520-522. [PubMed]

Thorpe, S., & Imbert, M. (1989). Biological constraints on connectionist models. In R. Pfeifer, Z. Schreter, F. Fogelman-Soulié, & L. Steels (Eds.), *Connectionism in perspective* (pp. 63-92). Amsterdam: Elsevier.

Thorpe, S. J., & Fabre-Thorpe, M. (2001). Seeking categories in the brain. *Science, 291*, 260-263. [PubMed]

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bulthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience, 14*, 869-876. [PubMed]

Trappenberg, T. P., Rolls, E. T., & Stringer, S. M. (2002). Effective Size of Receptive Fields of Inferior Temporal Visual Cortex in Natural Scenes. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Ullman, S. (1996). High-level vision. Cambridge, MA: MIT Press.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*, 682-687. [PubMed]

VanRullen, R., Gautrais, J., Delorme, A., & Thorpe, S. (1998). Face processing using one spike per neurone. *Biosystems, 48*, 229-239. [PubMed]

VanRullen, R., & Thorpe, S. J. (2001a). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception, 30*, 655-668. [PubMed]

VanRullen, R., & Thorpe, S. J. (2001b). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience, 13*, 454-461. [PubMed]

Vannucci, M., & Viggiano, M. P. (2000). Category effects on the processing of plane-rotated objects. *Perception, 29*, 287-302. [PubMed]

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51*, 167-194. [PubMed]

Yamamoto, S., & Kashikura, K. (1999). Speed of face recognition in humans: An event-related potentials study. *Neuroreport, 10*, 3531-3534. [PubMed]

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, 81, 141-145.

Temporal course of ERP in fast object categorization in natural scenes: a story more complicated than expected?

Guillaume A. Rousselet
Centre de Recherche Cerveau et Cognition (UMR 5549) CNRS - Université Paul Sabatier Toulouse - France
McMaster University, Department of Psychology, Neuroscience & Behaviour, Hamilton, Canada

Marc J.-M. Macé
Centre de Recherche Cerveau et Cognition (UMR 5549) CNRS - Université Paul Sabatier Toulouse - France

Simon J. Thorpe
Centre de Recherche Cerveau et Cognition (UMR 5549) CNRS - Université Paul Sabatier Toulouse - France

Michèle Fabre-Thorpe
Centre de Recherche Cerveau et Cognition (UMR 5549) CNRS - Université Paul Sabatier Toulouse - France

We report results from two experiments in which subjects had to categorize briefly presented upright or inverted natural scenes. In the first experiment, subjects decided whether images contained animals or human faces presented at different scales. In the second experiment, subjects responded to close-up views of animal faces or human faces. We compared the ERP to the same images when seen as targets and non-targets in different tasks. First, ERP differential activities were only weakly affected by inversion. Second, and more importantly, all task-dependent differences in ERP were surprisingly weak and of relatively long latencies. This contrasts strongly with the remarkably accurate behavioral responses of the subjects and their very short behavioral reaction times, implying that strong and early task-dependent ERP differences are not required for performing such high level visual tasks. Instead, we argue that some of the strong differential effects occurring from 135 ms between physically different sets of stimuli almost certainly reflect processing that is nevertheless intimately related to the identification and recognition processes.

Keywords: Rapid visual categorization, ERP, differential activity, natural scenes, inversion effect

Introduction

Both behavioral and electrophysiological evidence can be used to provide information about the speed of visual processing. Behavioral responses to visual stimuli have the distinct advantage of being directly relevant to survival. Thus, the fact that humans can initiate go/no-go responses to the presence of an animal in a briefly flashed natural scene in as little as 250 ms puts a clear upper limit on the time required for visual processing in some conditions (VanRullen & Thorpe, 2001a). And the fact that monkeys can perform the same sort of task with behavioral reactions that are even shorter (starting from 180 ms), imposes even more severe temporal constraints (Fabre-Thorpe, Richard & Thorpe, 1998). However, any behavioral reaction time measurement includes not only the time required for sensory processing, but also the time to initiate and execute the motor response. In such cases, electrophysiological measurements can be used to help determine the time course of the intervening processes. In animals, single unit recording can be used to determine precisely when individual neurons respond during a particular task

and much can be learned from the time course of neuronal responses in regions such as inferotemporal cortex (e.g. Sheinberg & Logothetis, 2001; Tanaka, 1996; DiCarlo, 2005). There is also a limited amount of evidence from single unit recordings made in human patients during surgical procedures for the treatment of epilepsy, but the fact that such subjects are often under heavily pharmacological treatment means that the latencies obtained may well be abnormally long (e.g. Allison, Puce, Spencer & McCarthy, 1999; Kreiman, Koch & Fried, 2000). One approach that has been used successfully in normal human subjects involves Event Related Potential (ERP) recording. By analyzing the averaged waveforms produced in response to images containing targets, and subtracting the average waveform produced in response to non-target images, one obtains a difference waveform that can, in appropriate conditions, be used to determine the moment when responses to targets and non-targets start to differ. The time at which the difference waveform starts to diverge significantly from 0 can provide an estimate of the time necessary to discriminate targets from non-targets.

In an early study, Thorpe, Fize & Marlot (1996) found that the difference between targets and non-targets ERPs at frontal sites becomes statistically significant from 150 ms following the onset of the images.

However, interpreting these differential response functions is not without difficulties. In some conditions one can obtain significant differences in the ERPs to two classes of images that could simply be due to low-level differences in the physical properties of the pictures, and not to recognition per se. For example, a set of images physically darker than an other one could easily produce differences in the neural responses in areas such as V1 and lead to ERP differences with remarkably short latencies. One way to avoid this potential confound is to change the target status of the images so that one can compare the ERP responses to the same images treated either as targets, or as non-targets. In such a case, the same physical images are compared, thus any differences in the differential signal cannot be due to low-level features. This approach was first developed to study the effects of attention in the auditory system (e.g. Hillyard, Hink, Schwent & Picton, 1973) and then in the visual system using relatively simple stimuli (e.g., Hillyard & Münte, 1984). VanRullen & Thorpe (2001b) extended this approach to natural scenes and showed that while very early differential effects were abolished by such a manipulation, differential signal that started from 150 ms was preserved.

In VanRullen and Thorpe's experiment, there were two basic target categories – animals and means of transport. Subjects alternated between blocks in which animals were targets, and blocks in which 'means of transport' were targets. In each condition, half of the non-target images were targets from the other category and by carefully counterbalancing the experimental design, each individual image was treated either as a target or as a non-target by different subjects.

In the present paper, we apply the same sort of analysis to another set of data, for which the behavioral results have been published previously in this journal (Rousselet, Macé, & Fabre-Thorpe, 2003). In the first experiment, subjects had to decide whether the image contained either an animal or a human face. The animals and faces could be at almost any size and position within a natural scene. In the second experiment, subjects had to either respond to animal faces or human faces, but in this case the images were all relatively close up views of just the head region. As reported previously, performance was exceptionally good, despite the wide range of stimuli used (Rousselet, et al., 2003). Furthermore, it was found that inverting the images had remarkably little effect on performance, a point that is of major importance

for understanding the nature of the neuronal mechanisms underlying face and object processing.

Regarding the issue of ERP differential effects, the main conclusion from this study is that, particularly in the case of the face stimuli, the task-dependent differences in ERP were surprisingly weak and of relatively long latencies. This result, which contrasts strongly with the remarkably accurate behavioral responses of the subjects and their very short behavioral reaction times implies that *strong task-dependent ERP differences are not required for performing such high level visual tasks*. Instead, we argue that some of the very strong differential effects occurring from 135 ms between physically different sets of stimuli almost certainly reflect processing that is nevertheless intimately related to the recognition and identification processes.

Methods

Forty-eight subjects volunteered in these two studies and gave their written informed consent. All had normal or corrected to normal vision. Nine subjects participated in both experiments.

Task setup

Subjects were sat in a dimly lit room at 100 cm from a computer screen (resolution: 800 x 600, vertical refresh rate: 75 Hz) piloted from a PC computer. To start a block of trials, they had to place a finger on a response pad for one second, then a fixation cross (0.1° of visual angle) appeared for 300-900 ms and was followed by the stimulus presented in the middle of the screen for two frames, i.e. about 26 ms. Participants had to lift their finger as quickly and as accurately as possible (go response) each time a target was presented. Responses were detected using infrared LEDs. Subjects had 1000 ms to respond after which their response was considered as a no-go response. A 300 ms black screen followed this maximum response time delay, before the fixation point was presented again for a variable duration, resulting in a random 1600-2200 ms inter-trial interval. When the photographs contained no target, subjects had to keep their finger on the pad for at least 1000 ms (no-go response).

In experiment 1, a session included 16 blocks of 96 trials and subjects alternated between two categorization tasks. In 8 of the blocks, the targets were animals and in the 8 other blocks, the targets were human faces. Half of the subjects started with the animal categorization, the other half with the human face categorization and conditions alternated every two blocks. In experiment 2, there were 8 blocks. In the first 4 blocks, the targets were animal faces and in the other 4 blocks the targets were human faces (starting blocks counterbalanced across subjects). For both experiment, in each block, target and non-target trials were equally likely. Among the 48 non-targets, 24 were targets of the other categorization task. Thus, when performing a human face categorization task, on a 96 trial block, 48 pictures contained human faces, 24 non-target scenes contained animals, the last 24 non-targets being other types of natural scenes. Moreover, half of the images in each subset were presented upright while the other half were presented inverted (rotation 180°). A given subject saw each image only once, with one orientation (upright or inverted) and one status (target or non-target). Subjects had two training blocks of 48 images before starting the test session. Training pictures were not used during the test.

Stimuli

We used photographs of natural scenes taken from a large commercial picture library (Corel Stock

Photo Library). They were all horizontal photographs (768 by 512 pixels, sustaining about 20° by 13.5° of visual angle) and chosen to be as varied as possible (see sample images in Rousselet, et al., 2003). Animals included essentially mammals, but also birds, fish, and reptiles. Human faces were presented in real-world situations with views ranging from whole bodies at different scales to face close-ups and including Caucasian and non-Caucasian people. There was also a very wide range of non-target images that included outdoor and indoor scenes, natural landscapes (mountains, fields, forests, beaches...), street scenes, pictures of food, fruits, vegetables, plants, buildings, tools and other man-made objects, as well as some tricky distracters (e.g. dolls, sculptures, statues... and non-target images containing humans for which the faces were not visible). In experiment 2, only close-up views of target objects were used and a special attempt was made to use many tricky non-targets and "blob" objects appearing in positions similar to human and animal faces. Subjects had no a priori information about the presence, the size, the position or the number of targets in an image, brief presentation prevented ocular exploration and trial unique presentation prevented learning.

EEG recording and analysis

A SynAmps amplifier system (Neuroscan Inc.) was used to record brain electrical activity with 32 electrodes mounted in an elastic cap (Oxford Instruments) in accordance with the 10-20 system with the addition of extra occipital electrodes from the 10-10 system (FP1/2, F3/4, F7/8, Fz, C3/4, Cz, T7/8, Pz, P3/4, PO3/4, POz, TP7/8, T5/T6, PO7/8, O1/2, Oz, Iz, PO9/10, O9/10). The ground electrode was placed along the midline, ahead of Fz and impedance was systematically kept below 5 kΩ. Signals were digitized at a sampling rate of 1000 Hz and low-pass filtered at 40 Hz before analysis. Potentials were on-line referenced on electrode Cz and re-referenced off-line by subtracting the average of all electrodes from each individual electrode signal. Baseline correction was performed using the 100 ms of pre-stimulus activity. Two artifact rejections were applied over the [-100 ms; +400 ms] time period, first on frontal electrodes with a criterion of [-80; +80 μV] to reject trials with eye movements, second on parietal electrodes with a criterion of [-40; +40 μV] to remove trials with excessive activity in the alpha range. Only correct trials were averaged.

Differences between pairs of conditions were assessed using a percentile method with 999 permutations. This procedure provided a confidence interval around the mean difference under the null hypothesis that the 2 conditions were actually sampled from the same population. Permutation procedures do not assume a specific data distribution and are much more robust than parametric statistics such as paired t-tests, providing more power and less sensitivity to false alarms. Differences reported in this paper are significant at $p<0.01$, corrected for multiple comparisons (Bonferroni: 0.01/32).

As stated in the introduction, the design of the experiments was similar to that of VanRullen & Thorpe (2001b) and allowed us to compute 2 types of differential signal, with or without physical differences between targets and non-targets. The "classical" differential activity, called "type 1" throughout this article, corresponds to the subtraction of the signal recorded on non-targets from the signal recorded on targets *in a given categorization task*. These images are different and hence type 1 differential activity can result from both stimulus-related and task-related differences. The second differential activity, called "type 2", corresponds also to a target minus non-target subtraction, but with the same images seen as target or non-targets *in two different categorization tasks*. Thus type 2 differential activity only reflects task-related differences.

Results

In the first experiment, subjects (n = 24, 12 women, 12 men, mean age 31) performed remarkably well. A detailed analysis of the behavioral results has been published separately (Rousselet, et al., 2003). Upright faces and animals were processed on average as efficiently (96.4% and 96.3%, respectively) and at the same speed (median reaction time: 368 ms and 371 ms, respectively).

The time at which enough information was available to discriminate a given category from the others was assessed from event-related potentials (ERP) on correct trials. Target ERPs were compared to non-target ERPs using a permutation method in which differences were tested every millisecond on the whole set of scalp electrodes. Early and large differences were found over posterior electrodes in both hemispheres and at frontal electrodes for the two categorization tasks, with differential effects that were strongest at lateral occipito-temporal sites (Figure 1). The peak latency of the differential activity was very different across the four conditions with up to 100 ms between upright human and inverted animals DA peaks.

Regarding the differential activity signal, responses to upright target animals differed significantly from non-targets as early as 150 ms, a result that constitutes a direct replication of previous studies performed in our laboratory (Fabre-Thorpe, Delorme, Marlot & Thorpe, 2001; Thorpe, Fize & Marlot, 1996; VanRullen & Thorpe, 2001b) (Figure 2A, Figure 3A). However, differential effects when faces were targets started even earlier, with significant effects as early as 100-130 ms (Figure 2C, Figure 3B).

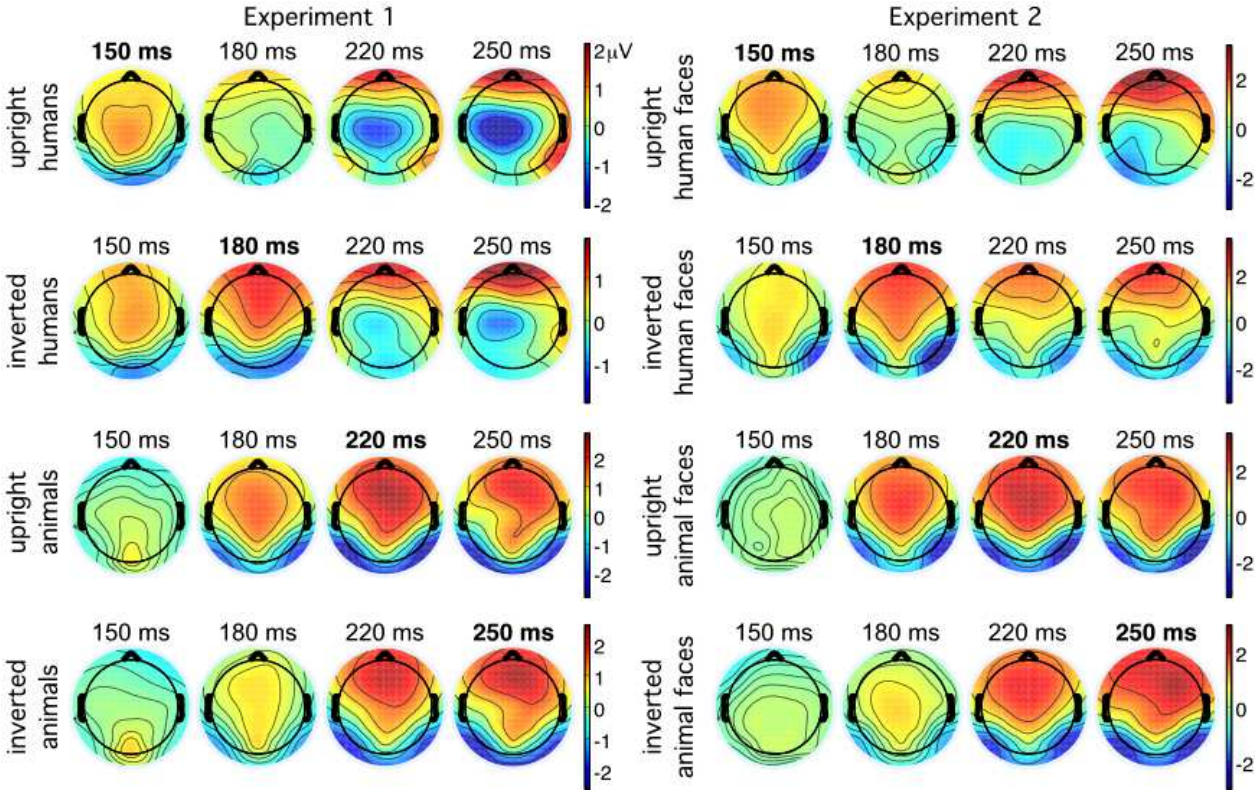


Figure 1. Two dimensional linear interpolation maps of the differential activities in each experiment and for each condition. The maps represent the signal recorded at the peak latencies of the occipital differential activity in the four conditions (latencies in bold).

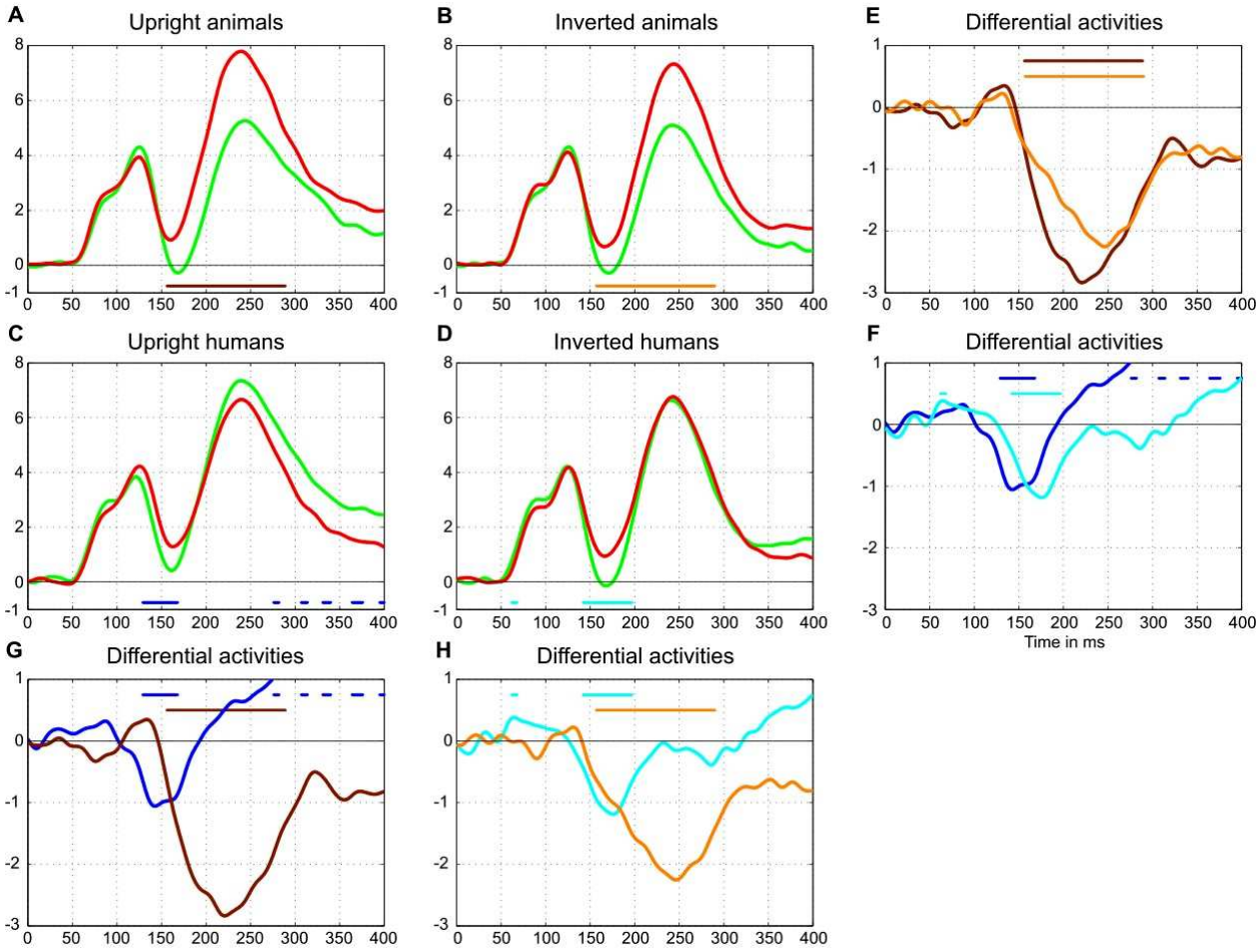


Figure 2. Comparison of the ERP associated with the processing of targets and non-targets in experiment 1. Each graph represents the grand-average signal (in micro-volts) recorded from the right posterior electrode PO10. This electrode was chosen because it showed up the largest differential effects in amplitude. For each target category, the ERPs are presented for upright (A&C) and inverted (B&D) stimuli. Target ERPs (in green) were computed from trials in which the indicated category was seen as target. Non-target ERPs (in red) were computed from trials in which pictures with the same orientation as targets were seen as non-targets. They include neutral non-targets and pictures from the target category of the other task. The type 1 differential activities were computed by subtracting non-target trial ERPs from target trial ERPs separately for each category and each orientation. The two graphs on the right (E&F) show the effect of inversion on type 1 differential activities separately for both categories. The two graphs at the bottom (G&H) allow the comparison of the type 1 differential activities associated with humans and animals separately for both orientations. Colored horizontal lines indicate time points of significant differential activities (permutation, $p<0.01$).

Furthermore, we investigated the effects of inversion on processing speed in such a task, a manipulation that is known to slow down particularly the identification of faces (Rossion & Gauthier, 2002). It appeared that both behavior and the onset of ERP differences were only very weakly affected by inversion. Inversion produced a global decrease of accuracy that was very similar for both faces and animals (<2%). Inverted pictures led to significantly longer RT than upright pictures and the inversion effect was reliably more pronounced for faces (+23 ms on median RT) than for animals (+9 ms on median RT). These weak effects at the behavioral level are confirmed by ERP results. The differential activity for inverted animals starts virtually at the same latency (≈150 ms) but develops with a shallower slope and reaches lower amplitude than for upright animals

(Figure 2E). The differential activity onset for faces is slightly delayed by inversion (Figure 2F).

So far, these results seem to imply that face specific processing can start very shortly after stimulus presentation, as early as 100-130 ms, hence faster than the categorization of animals which seems to require at least an additional 20-30 ms. Furthermore, this capacity relies on relatively view invariant representations as shown by the very weak inversion effects on processing efficiency. We also found small but reliable ERP differences before 100 ms (Figure 3E & F). They appeared as early as 70-90 ms for upright and inverted animals and 50-60 ms for upright and inverted faces. The onset latencies of all the significant differences for all the conditions in the two tasks and the two experiments are reported in Figure 3. We suspected these very early differences might be due to uncontrolled low-level differences between the sets of target and non-target images, as previously suggested for the categorization of natural images (Johnson & Olshausen, 2003, 2005; VanRullen & Thorpe, 2001b).

In a first attempt to reduce physical differences between target images, a new experiment was designed in which subjects (n = 24, 12 women, 12 men, mean age 30, 9 of which participated in the first study) were required to categorize human faces or animal faces in pictures depicting close-up views of these targets. Pictures of both human and animal faces were chosen to be as varied as possible in an attempt to decrease the physical differences between the two sets of target images. Pictures that did not contain faces were specifically chosen to contain many tricky objects like dolls, statues, and flowers... At the behavioral level, the use of close-up views unexpectedly led to excellent performance levels, with slightly higher accuracy and slightly longer reaction times for both categories compared to the first experiment (Rousselet, et al., 2003).

The stimulus selection in experiment 2 had several consequences at the ERP level. The key finding was that the very early differences recorded in experiment 1 (Figure 3B) for faces where not found anymore (Figure 3F). Upright animals were still associated with very early differential activities but the effects were restricted mainly to occipital midline electrode Oz (Figure 3E). However, the large lateral occipito-temporal differential activities reported in experiment 1 were still present and even reached higher amplitude in this second experiment (Figure 4). These differences appeared approximately at the same time as in the previous experiment reaching statistical significance respectively at about 150 ms and 130 ms for upright animal and human faces. Inversion had a very limited effect on these onset latencies (Figure 3E & F, Figure 4E & F). In addition, as reported in experiment 1, the slope of the differential activity was steeper for upright animals compared to inverted ones (Figure 4E).

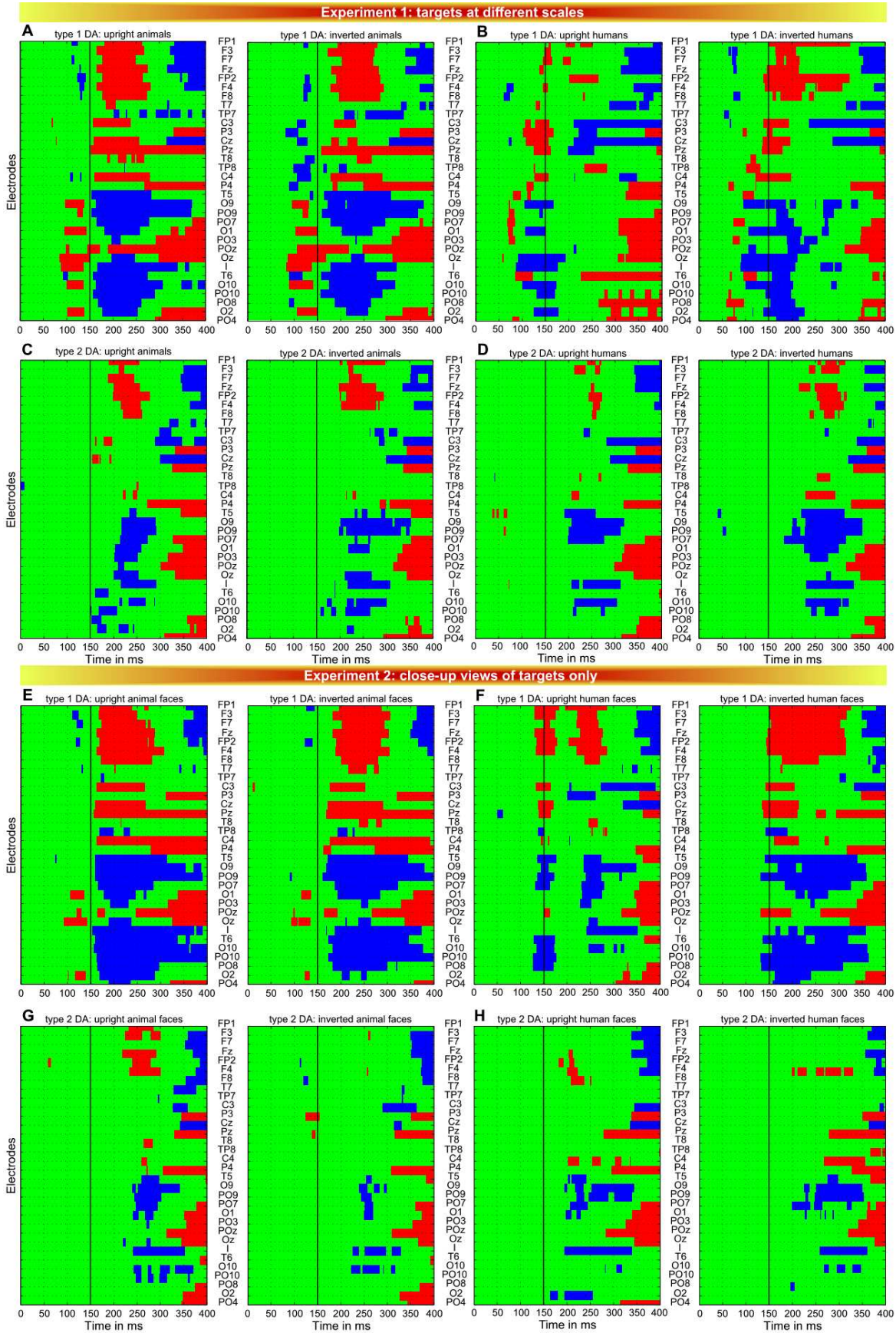

Experiment 1: targets at different scales
A
type 1 DA: upright animals
type 1 DA: inverted animals
B
type 1 DA: upright humans
type 1 DA: inverted humans
C
type 2 DA: upright animals
type 2 DA: inverted animals
D
type 2 DA: upright humans
type 2 DA: inverted humans
Experiment 2: close-up views of targets only
E
type 1 DA: upright animal faces
type 1 DA: inverted animal faces
F
type 1 DA: upright human faces
type 1 DA: inverted human faces
G
type 2 DA: upright animal faces
type 2 DA: inverted animal faces
H
type 2 DA: upright human faces
type 2 DA: inverted human faces
Electrodes
Time in ms

Figure 3. Latencies of the differential activities at all electrodes and in all conditions. Along the vertical axes, electrodes have been sorted from anterior (top) to posterior (bottom) sites. A-D: Experiment 1. Panels A & B report the latencies of the type 1 differential activities computed by subtracting the ERP associated with all non-targets from the targets for animals (A) and humans (B), either upright or inverted. Panels C & D report the latencies of the type 2 differential activities, when physical differences were removed. E-H panels show the same results for experiment 2. Non-significant differences are represented in green. Significant differences ($p<0.01$) for which the differential activity was negative (positive) are represented in blue (red).

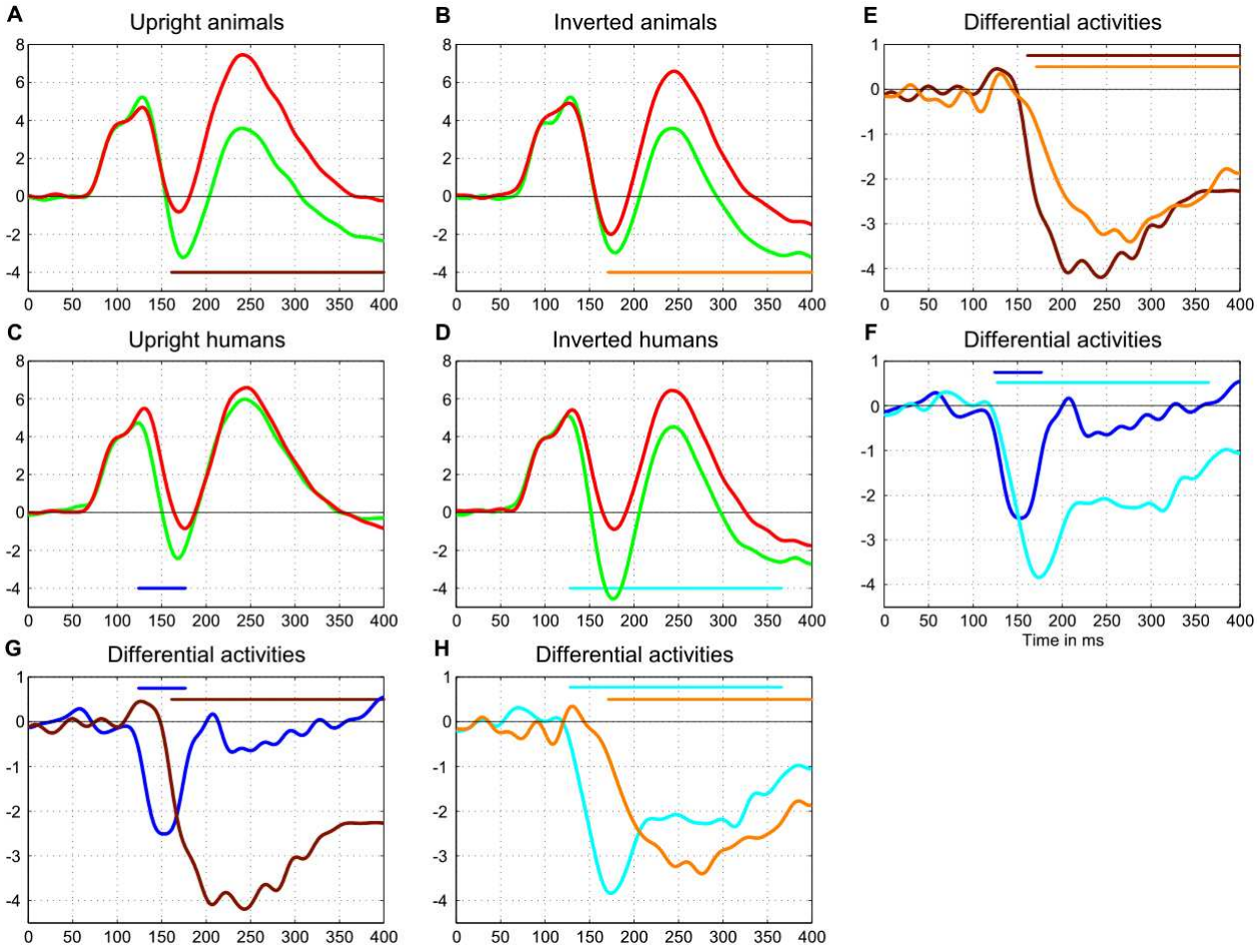


Figure 4. Comparison of the ERP associated with the processing of targets and non-targets in experiment 2 (close-ups of animal and human faces). Caption as in Figure 2.

The second experiment directly demonstrated that very early differences recorded with human faces as targets were due to uncontrolled physical differences between stimuli. However, some of these very early differences remained when animal faces were targets. To address this issue, the processing speed of the different categories was estimated independently of their visual attributes. A new set of differential activities (type 2 differential activity) was computed in which target ERP for a given category and a given orientation was compared to ERP associated with the same category and the same orientation when it was seen as a non-target. This manipulation controlled for physical differences since across subjects the same pictures were seen as targets and non-targets. The only differences that remain in the electrophysiological signal are due to task status and should thus give us an estimate of the time required to access task related categorical information independently of physical differences. The results

across all electrodes are depicted in Figure 3C & D for experiment 1 and Figure 3G & H for experiment 2. Type 2 differential activities had very small amplitude compared to type 1 differential activities (compare Figure 2 & 5 and Figure 4 & 6). In experiment 1, the animal task was found again to affect ERP at around 150 ms (Figure 3C, Figure 5E) confirming a previous report that used this technique (Van Rullen & Thorpe, 2001b). This latency was almost unaffected by inversion. Surprisingly, the human face task did not affect type 2 DA before about 180 ms for upright pictures (Figure 3D, Figure 5F). Task status had a slightly earlier effect on inverted human ERP at electrode PO7. The results from experiment 1 suggest that the early differential activities recorded for faces were unrelated to subject performance since they disappeared when physical properties between target and non-target ERP were equated. Only the large differential activities at 150 ms in the animal task seem to be related to the extraction of task related categorical information. However, results from experiment 2 cast doubt on this interpretation. Indeed, in experiment 2, the effects of task status on ERPs to both human faces and animal faces were all surprisingly late (Figure 3G & H, Figure 6), especially in regards to the high behavioral performance. Appart from an isolated "early" difference at about 160 ms for upright human faces on electrode O2, all other electrodes across all categories presented their earliest effects after 200 ms.

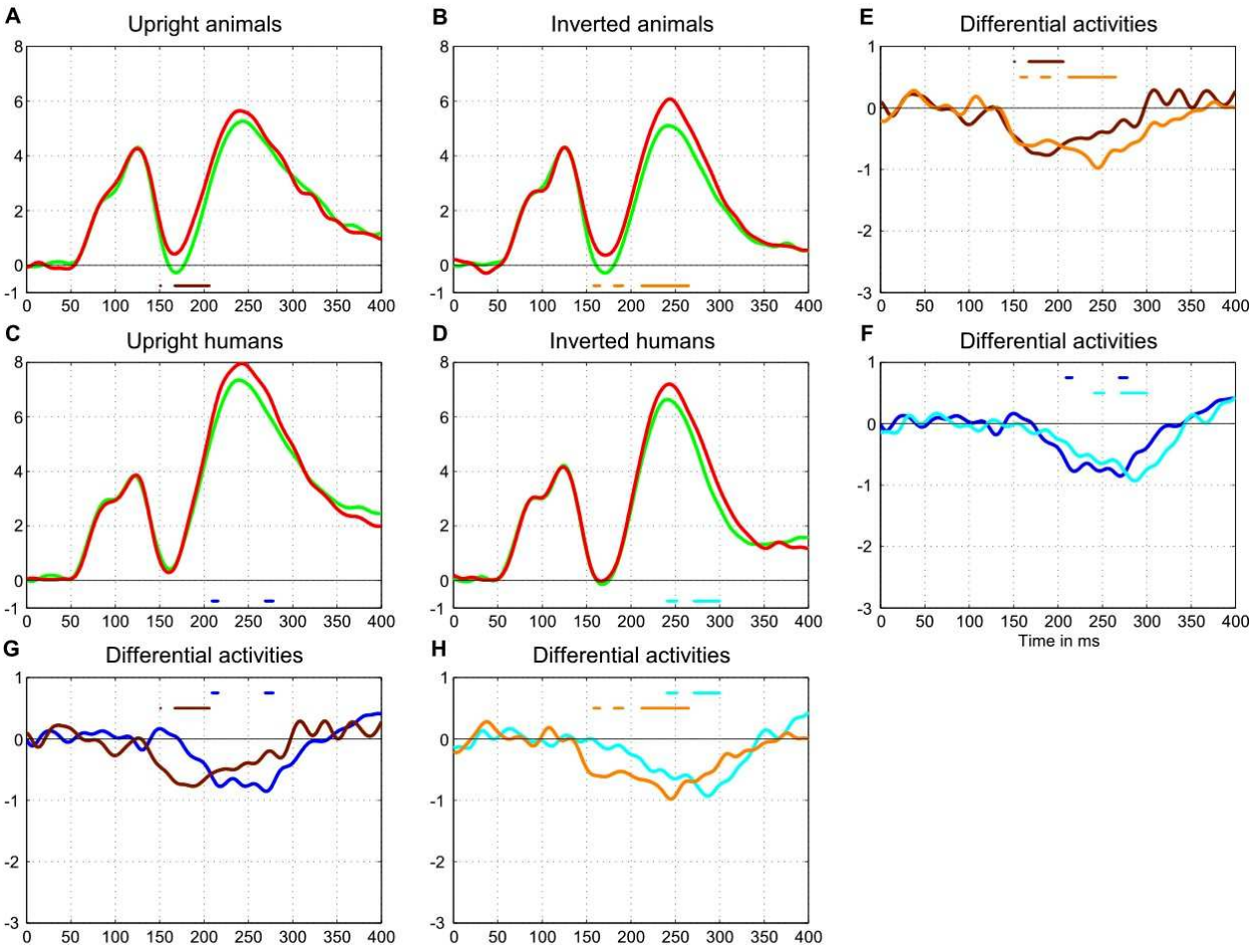


Figure 5. Type 2 differential activities showing the effects of task status independently of physical differences in experiment 1 at electrode PO10. These type 2 differential activities (in micro-volts) were computed by subtracting the ERP associated with images of a given category when processed as a non-target (in red) from the ERP associated with the same image-category when processed as target (in green).

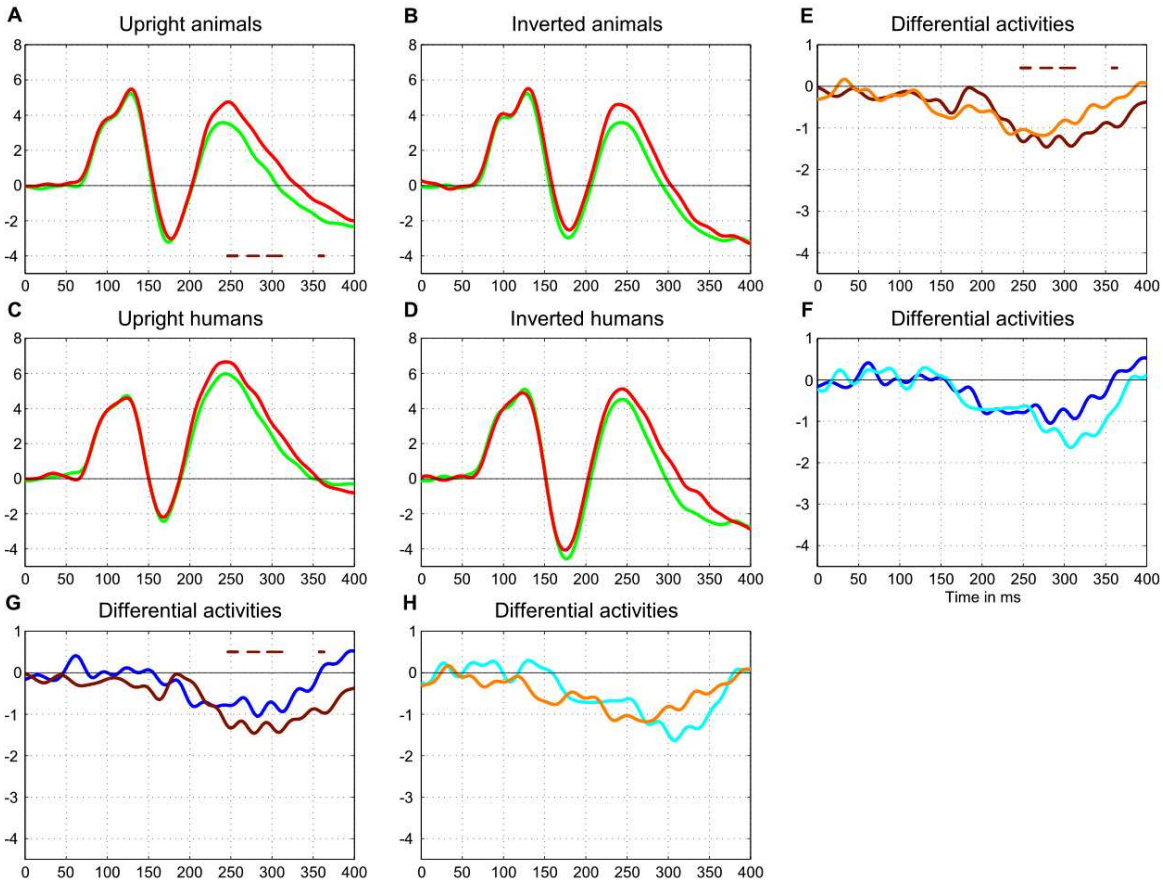


Figure 6. Type 2 differential activities showing the effects of task status independently of physical differences in experiment 2 at electrode PO10. Caption as in figure 5

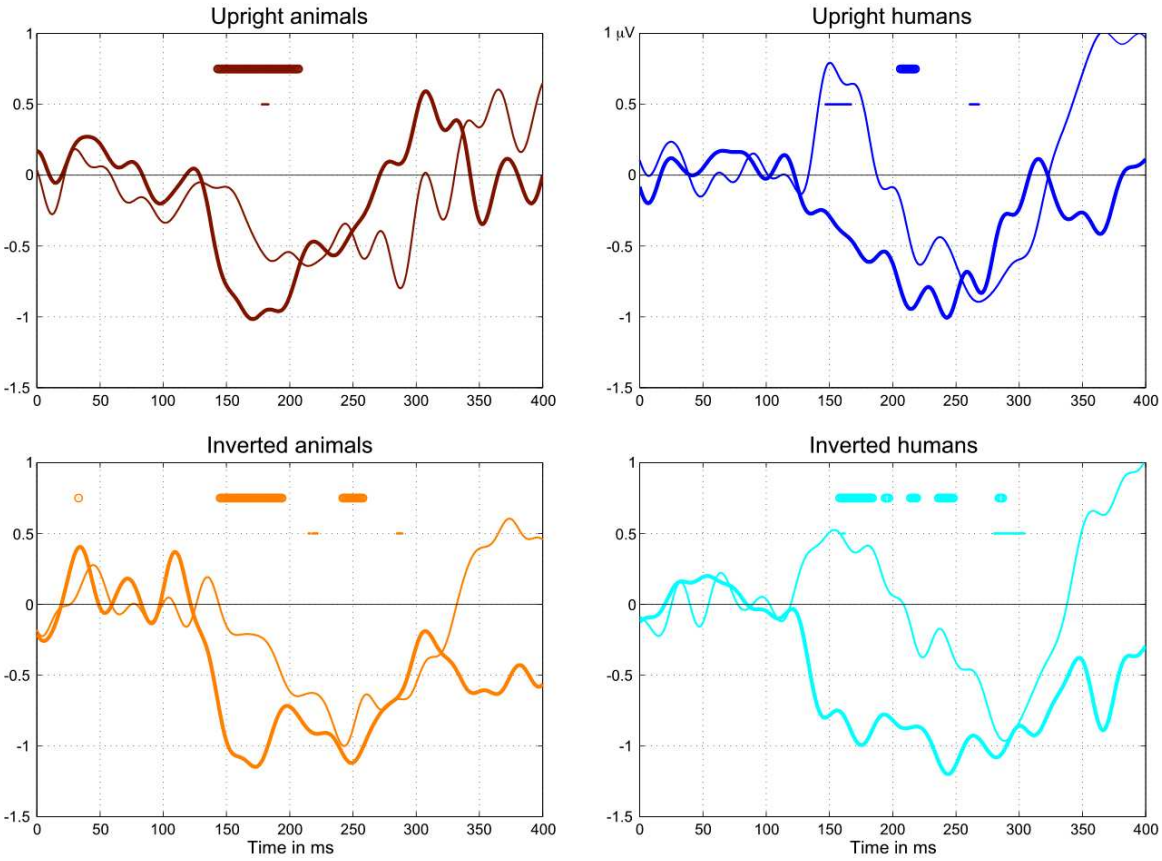


Figure 7. Type 2 differential activities as a function of RT in experiment 1 at electrode PO10. Thick lines = shorter RT; thin lines = longer RT. Horizontal marks indicate time points of significant differences ($p<0.01$).

In a final analysis, differential activities were computed as a function of subjects' RT. For each subject, the RT histogram was divided into 2 equal parts (median split). For each part, the corresponding target ERP was averaged separately. Then, non-target ERPs were subtracted from target ERPs to generate 2 types of differential activities corresponding to faster and slower RT according to the RT distributions. This analysis confirms that there is a relationship between behavioral RT and type 2 differential activity onset (Figure 7, Figure 8) in keeping with Johnson & Olshausen (2003, 2005). The only exception seems to be the differential activity evoked by upright human faces in experiment 2, which seems to hold a special status in this regard (Rousselet, Macé, & Fabre-Thorpe, 2004). Contrary to what was found by Thorpe et al. (1996) and Johnson & Olshausen (2003), this analysis also demonstrates that such a relationship also exists in the case of type 1 differential activities (Figure 9, Figure 10). Although the relationship between type 1 differential activity and RT exists, it must be noted that there does not seem to be a direct mapping between short and long RTs across categories and differential activity onsets.

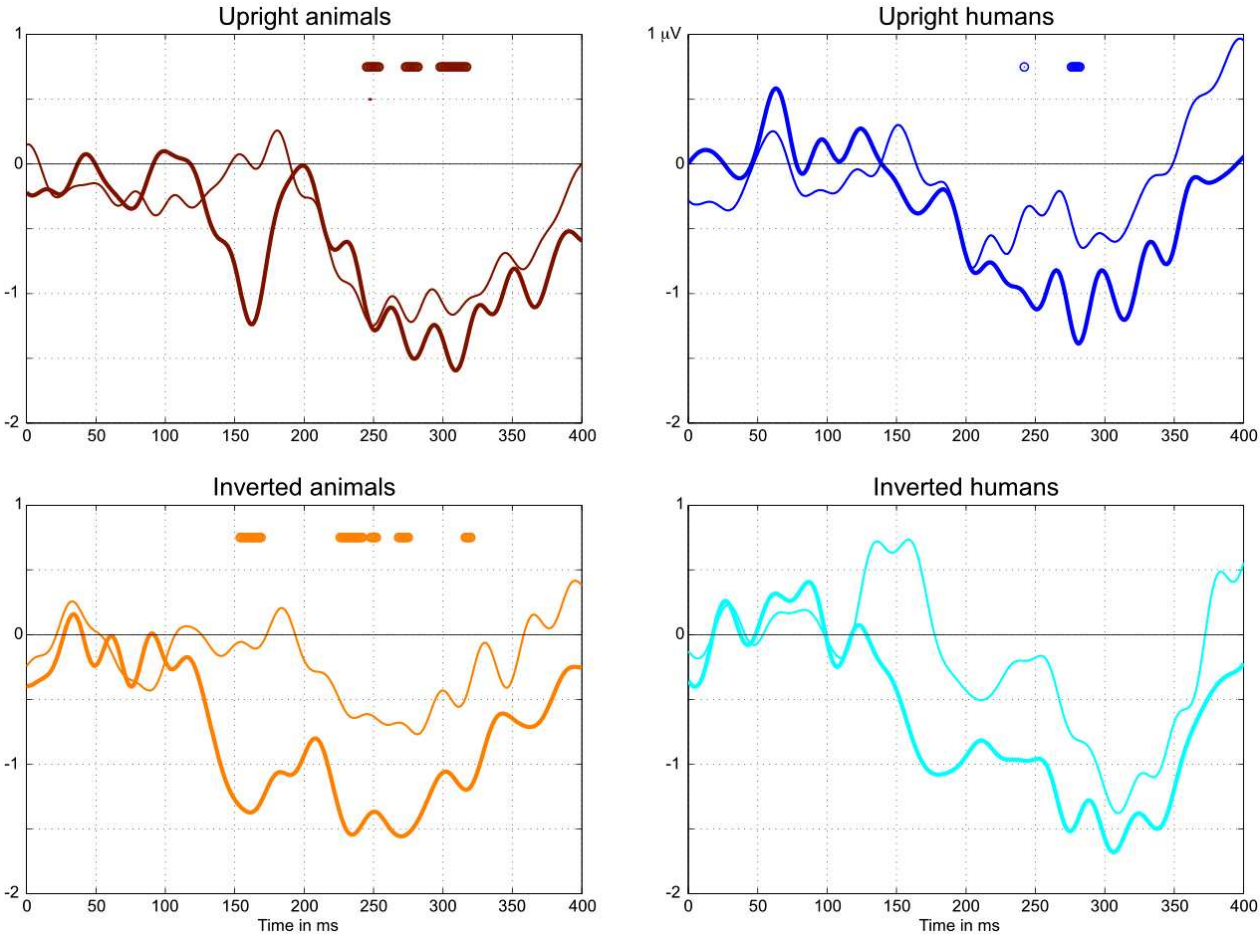


Figure 8. Type 2 differential activities as a function of RT in experiment 2 at electrode PO10. Thick lines = shorter RT; thin lines = longer RT. Horizontal marks indicate time points of significant differences ($p<0.01$).

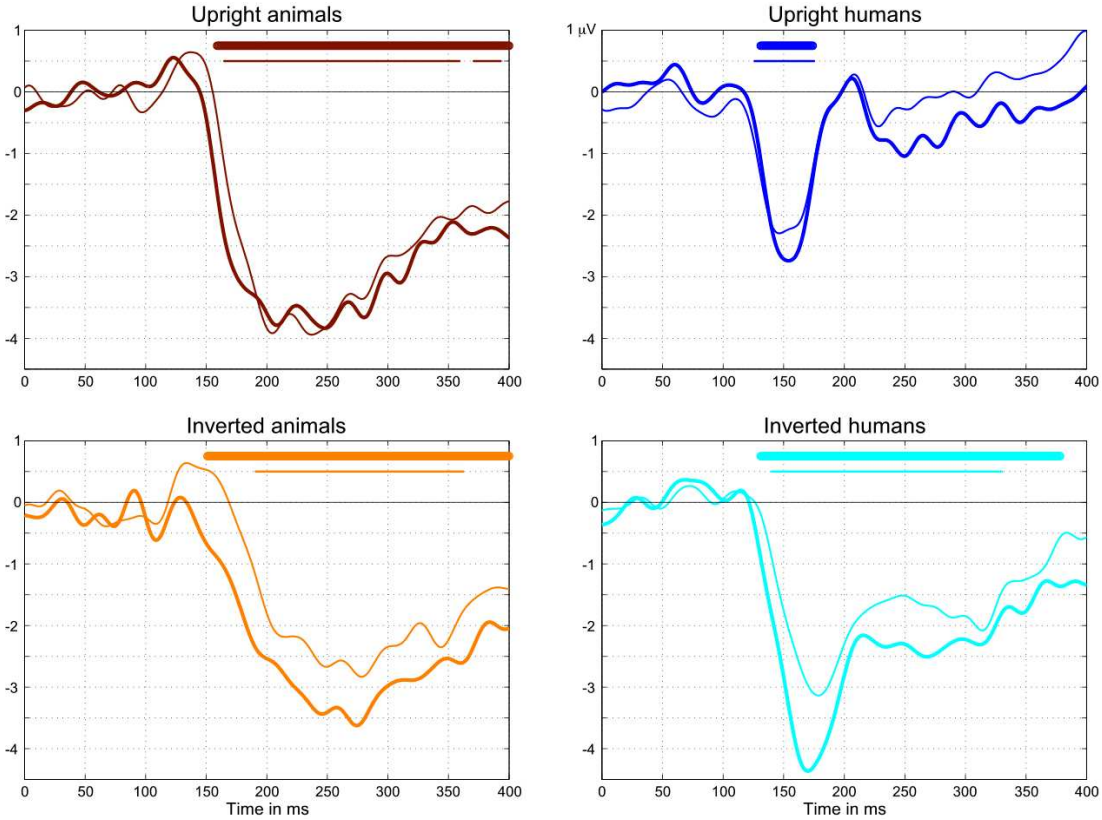


Figure 9. Type 1 differential activities as a function of RT in experiment 1 at electrode PO10. Thick lines = shorter RT; thin lines = longer RT. Horizontal marks indicate time points of significant differences (p<0.01).

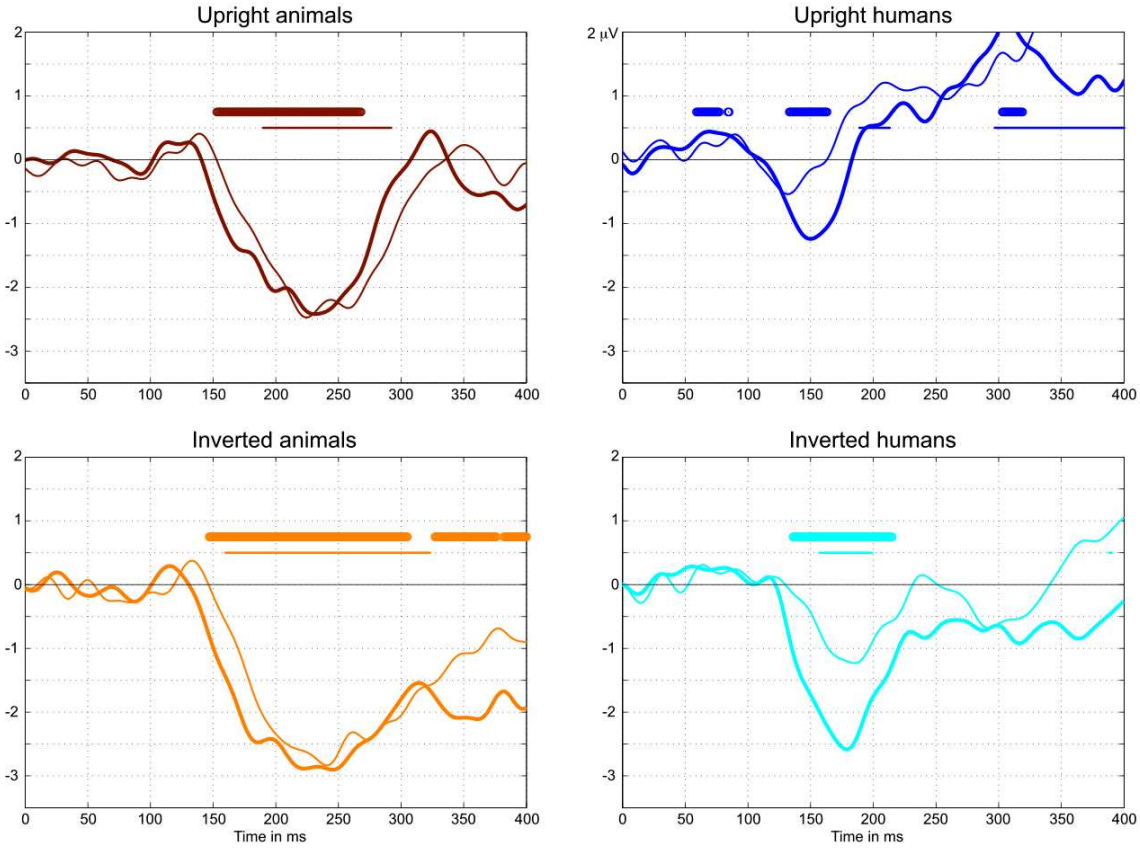


Figure 10. Type 1 differential activities as a function of RT in experiment 2 at electrode PO10. Thick lines = shorter RT; thin lines = longer RT. Horizontal marks indicate time points of significant differences (p<0.01).

Discussion

By examining the averaged ERP responses in the various task conditions we were able to find clear and statistically significant differences between the brain responses to different visual stimulus classes at numerous electrode sites. Some of these differences had a distribution and a time course that was very similar to those seen in previous studies on rapid scene processing (Fabre-Thorpe et al., 2001; Johnson & Olshausen, 2003; Thorpe et al., 1996; VanRullen & Thorpe, 2001b). In this section we will discuss the various hypotheses that can be evoked to account for these differences.

A first point concerns the anatomical distribution of the differential responses. The original 1996 paper by Thorpe et al. concentrated on the differential signals observed at frontal recording sites, which showed a clear enhanced negativity on no-go trials. This finding fitted with a number of other studies that had shown cortical negativity associated with no-go trials. Furthermore, in that original study, the fact that the temporal profile of the differential activity was essentially identical when calculated for trials with short reaction times and trials with long reaction times led the authors to speculate that the activity might be specifically related to response inhibition on no-go trials. However, more recent studies that have examined differential activity in forced choice tasks in which the subject has to make a response on every trial suggest that this explanation may be inadequate. For example, Johnson and Olshausen (2003) recently found a very similar pattern of differential effects at frontal sites when they compared a go/no-go and a forced choice response paradigm. Similar differential effects at frontal and parietal sites were also reported in a force-choice task by Antal et al. (2000). Such results are clearly inconsistent with the simple notion that the differences are caused by response inhibition per se.

Other results also argue against a response related explanation of the effect. In the original 1996 study, the restricted number of electrodes meant that very little data was available for more posterior sites close to the occipital cortex. Another explanation comes from the use of a linked ears reference in Thorpe et al. (1996, as well as in Fabre-Thorpe et al., 2001), a method that tends to mask occipital activities in favor of frontal activities. In fact, later studies using an averaged reference showed that in parallel with the frontal differential activity, there was a clear differential activity with the opposite polarity at lateral occipito-temporal sites (Rousselet et al., 2002; VanRullen & Thorpe, 2001b). This bipolar arrangement can be seen clearly in Figure 1 of the present study. The close similarity between the onset times of the occipital and frontal differential responses as well as the results from source analysis using BESA is consistent with the idea that a considerable proportion of the differential responses at both frontal and occipito-temporal sites is produced by the same set of sources in occipitotemporal cortex (Delorme, Rousselet, Macé, & Fabre-Thorpe, 2004; Fize, Fabre-Thorpe, Richard et al., in press). However, at least some of the later differential effects could depend specifically on activity in prefrontal areas (Rousselet, Thorpe, & Fabre-Thorpe, 2004).

What underlying processes could give rise to this differential activation in occipitotemporal areas? It is useful to distinguish at least three different potential causes, each characterizing activity at a particular level in the visual system. First, consider neurons at the earliest levels of the visual processing hierarchy, selective for relatively low-level stimulus features such as contour orientation and the presence or absence of terminations. Suppose that we take a set of images from a given class (for example, photographs of human faces) and determine the average response of neurons in V1, and then do the same for another set

of images from another class (for example, photographs of landscapes). If the images of landscapes contained a higher proportion of horizontal edges (for instance, because of the presence of a horizon), then a statistically significant difference between the average response to the two image classes might be present even though none of the neurons involved coded anything specific about either faces or landscapes. Attributing the categorization label to such differential activity would in this case be an error.

Consider now what might happen if we were looking at neurons at a later stage of visual processing that are selective to facial features. There is abundant evidence for such neurons from single unit recording studies in awake behaving monkeys, and it is known that at least some neurons can respond selectively as a function of stimulus gaze direction (Perrett et al., 1992). If one was to measure the average response of this sort of population of neurons in response to the two different image categories (faces vs. landscapes), there could also be a strong difference in response. But in this case, the difference would have considerable significance for the task, because it would reflect the activity of populations of face selective cells that could be involved in recognition and categorization.

Is there a way to distinguish between "interesting" and "uninteresting" differential activities? The methodology used by VanRullen and Thorpe and used again here provides one way of tackling this issue. By switching between two different target categories, the same images can be presented either as targets or non-targets. If a difference still exists under these conditions, it is clear that no simple low-level difference between the images could explain the effects because the two image sets are physically identical. The differential responses obtained in the present experiments show that all the experimental conditions produced roughly the same effect, but the point at which this effect became significant differed markedly between conditions. In the standard "animal/non-animal" task (experiment 1), clear differential effects emerged at about 150 ms for both upright and inverted stimuli (Figure 3C). This result thus reinforces the study by VanRullen and Thorpe (2001b), who also found significant effects at similar latencies with this type of analysis. Together, such findings demonstrate clearly that information related to the category must have started to be encoded by around 150 ms, as proposed by Thorpe et al. (1996).

However, the results for the other conditions are more difficult to interpret. The comparison of responses to humans as targets with humans as non-targets in experiment 1 did not start to become significant until about 180 ms. And in experiment 2, all the differential responses started later, with the earliest significant effects for animals and humans not appearing until over 200 ms. This result is surprising because it breaks any obvious relationship between onset latency for this differential effect and the ability of the subjects to trigger their responses. The conclusion would seem to be that while this form of differential activity can (if successful) put an upper limit on the time required to extract a certain type of visual information, it does not necessarily provide a good predictor of when the subject will respond (it is an upper limit because there is always the possibility that earlier effects might not be captured by the ERP waveforms). If the differential activity was directly related to the decision process, one would expect subjects to be as much as 50 ms slower at detecting the presence of an animal in experiment 2 than they were in experiment 1. This was very clearly not the case: accuracy, mean reaction times and minimal reaction times were very similar for both experiments (Rousselet et al., 2003).

How could it be that subjects can perform the challenging visual task in experiment 2 without showing clear signs of task-related activity in ERP

recordings? To understand this, consider again a population of "face-selective" cells in inferotemporal cortex. Let us suppose that these neurons have responses that are relatively "hard-wired" in that they will respond to the presence of a face essentially irrespective of the task being performed by the subject. In such a case, one can imagine that changing the target category for the subject might have little or no effect on the magnitude of the cumulative response of these neurons (no “type 2” differential activity would be observed). And yet, despite this, the neurons could still perfectly signal whether or not the scene contains a face. If the output of the neurons was being used to drive a decision mechanism (located perhaps in a brain area outside the visual processing pathways per se, such as in prefrontal cortex; Freedman, Riesenhuber, Poggio, & Miller, 2003), one could imagine that the subject would perform the task well without there being any clear sign of task status (type 2) differential effects in the visual system itself. On the other hand, a comparison of the responses to a wide set of non-target images without faces and target images with faces could well reveal clear type 1 differences because of the large number of face-selective cells that are activated.

Our suggestion is that with special target categories such as faces that are processed very efficiently, there is no need for modulation of responsiveness within the visual system itself, with the result that no "type 2" differential effect would be visible. This view is in keeping with the argument of an equivalent ‘depth of processing’ for faces whether they are targets or non-targets, which might be developmentally driven by their intrinsic social interest (Rousselet, Macé, & Fabre-Thorpe, 2004; see also e.g. Puce, Allison, & McCarthy, 1999). In experiment 2, close-up views of animal faces could also benefit from the same "face effect" as humans, and thus explain the long latencies and small amplitude of the type 2 differential activity. Top down influences might also play a crucial role and an alternative processing model can thus be proposed (Figure 11). Suppose that in order to reliably detect any one of a large number of animal forms, some form of top-down "priming" of neurons selective for particular animal features was required. The top-down priming would have the effect that the neurons would respond more strongly when the corresponding features were present, and this enhanced activation could be detected by a later decision stage. The increased response when a target was present in the scene might be visible at the level of the global ERP response (DA of type I) because the amount of neural activity would be increased (e.g. Chelazzi, Duncan, Miller, & Desimone, 1998). However, in this case, changing the target category from "animal" to something else ("means of transport" as in the study by VanRullen & Thorpe (2001b), or "human" as in experiment 1 of the present study) would have the effect of modifying the priming effect and revealing a “type 2” differential effect.

We could also tweak this processing model by introducing an intermediate population of neurons somewhere between high-level feature detectors (IT?) and the decision level (Figure 11). This population could consist of a relatively small number of neurons, possibly located in the prefrontal cortex (Freedman et al., 2003) and receiving the majority of their inputs from IT cells. Following this hypothesis, switching between categorization tasks would not be done by massively shifting top-down activation from one set of high-level feature detectors to another but by switching the activation from of a small subset of cells in this intermediate population to another one. The reduced number of cells involved in this operation could well explain why no early differences could be found in the type 2 differential activity before 180-200 ms for faces. The discrepancy between animals and faces regarding type 2 differential activities in experiment 1 could be linked to the extensive experience we have with faces that could modify the importance of this intermediate

stage of processing. Hence, the amount of EEG-detectable top-down pre-activation could be linked to our level of expertise for a given category of objects. Specifically, for very important categories such as human vs. animal, the setting of relatively small populations of category specific cells could allow task switching with considerably less "effort" than for categories such as means of transport.

Note that all these processing strategies would allow the subjects to perform the task reliably, but only when a top-down priming strategy is used would one expect to see changes in the responsiveness of neurons within the visual system as a function of the target class. However, it is interesting to note that with faces, the early type 1 differential effects tend to be considerably less long lasting than for animals, a result that might fit with the idea of a more hard-wired "automatic" processing in this case.

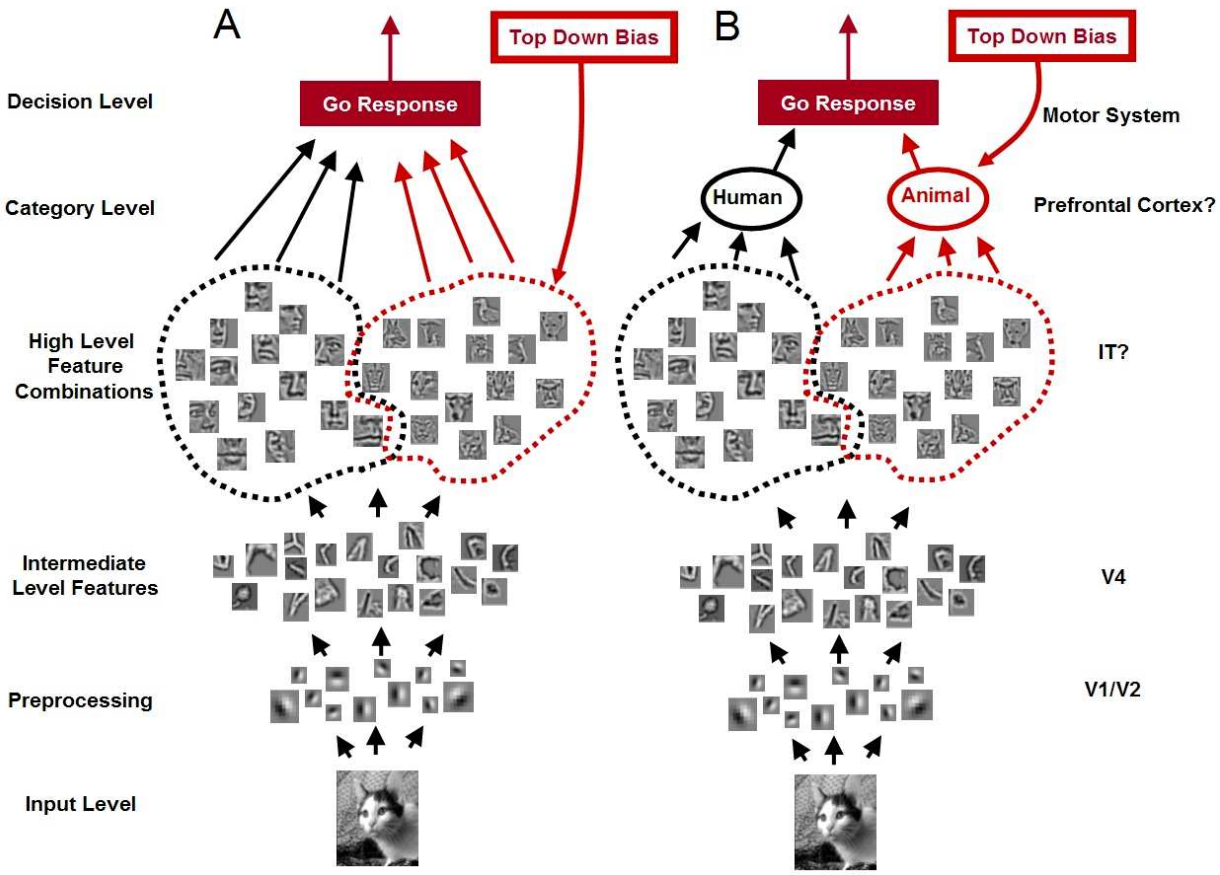


Figure 11. Two different ways in which top-down presetting could be used to switch between tasks with either human or animal faces as targets. In both cases, the basic processing architecture is the same, with preprocessing in V1/V2, the coding of intermediate features in areas such as V4, and the coding of category specific high-level feature combinations in areas such as IT. In A, the top-down biasing directly affects the high-level feature level, causing substantial changes in the population response. In contrast, in B there is a specific category level (possibly in prefrontal cortex) that would allow top-down biasing to switch between tasks without affecting the activity in high-level visual areas.

One criticism that has been raised concerning the relevance of the 150 ms type 1 differential effects is that there is no relation between the onset latency of the effect and behavioral reaction times. In the original Thorpe et al. (1996) study, it was shown that the differential activity at frontal sites has the same time course when the curves are plotted using average waveforms calculated from trials with fast and slow reaction times. At that time, it seemed highly unlikely that the processing time required for analyzing a given image did not depend strongly on the nature of the image. But more recently, evidence has accumulated in favor of the view that processing time might in fact be relatively constant for many natural images. One argument comes from the study by Fabre-Thorpe et al. (2001) who found that the distribution of reaction times on images with short reaction times was essentially random across subjects, as if the main contribution to variability doesn't originate from the images themselves and their neuronal processing, but from other trial-dependant factors such as attentional state or readiness. Therefore, *it is not impossible that high-level, categorization related, neuronal activity is actually reflected in ERP differential activity whose onset does not vary with RT*. This is in keeping with a recent study showing that the response latencies of neurons from the anterior infero-temporal cortex are completely independent from reaction times (DiCarlo & Maunsell, 2005). Such data demonstrate that the presence of a link between RT and onset latency of ERP effects cannot be used as the hallmark for high-level processing, as proposed by Johnson & Olshausen (2003, 2005).

In this context, we would like to argue that the differential activities recorded for humans and animals as early as 120-130 ms do not necessarily reflect irrelevant low-level physical differences, but might in fact be the signature of the early activation of high-level units coding for diagnostic properties in the image. The best evidence that very complex process can occur at such short latencies comes from a study by Kirchner & Thorpe (in press). In this experiment, subjects had to perform a forced-choice animal categorization task using ocular movements. Two natural scenes were flashed simultaneously on each side of a fixation cross and subjects had to move their eyes as fast as they could towards the image that contained an animal. The median reaction time was 230 ms with 90% of correct responses, and the minimal reaction time was as low as 120-130 ms. Such very short behavioral response latencies do not seem compatible with complex processing of natural scenes and might suggest that some sort of low-level image properties were used to perform the task. However, Kirchner & Thorpe were unable to find any statistical parameter in the images that could be used to distinguish reliably between target and non-target images, arguing against the idea of a simple analysis of image features by the visual system in this experiment.

Another point that must be considered in the present discussion is that "high-level" categorization does not necessarily imply that high-level representations are used to perform the task. It has been shown that "mid-level" representations can perfectly be used to perform this kind of classification, like the detection of faces in natural scenes (Ullman, Vidal-Naquet, & Sali, 2002). Such "mid-level" representations might be used as diagnostic features in our task, allowing subjects to respond for the presence of complex objects (Schyns, 1998). As this kind of feature might well be processed in areas V4-TEO of the ventral pathway and activated by a feedforward wave of activation, this strengthens the hypothesis of an early "high-level" process of objects in natural scenes. Indeed, as suggested by panel A of Figure 11, it may be possible to initiate category-specific behavioral responses without explicitly using categorical coding. The solution would be to use top-down biasing to pre-activate large numbers of units selective to high-level features combinations diagnostic for stimuli

characteristic of the target category. In this case, a decision mechanism that simply monitor the amount of activation in this population could trigger a category-specific behavioral response without the need for explicit categorization.

Furthermore, it has been suggested that the earliest evidence for coarse face processing might be found at around 120 ms (Itier & Taylor, 2002; Linkenkaer-Hansen et al., 1998). In keeping with this hypothesis, recent source analysis on ERP data has revealed that the fusiform gyrus, an area of the ventral pathway involved in high-level object recognition, can be activated under 110 ms after stimulus onset (Di Russo et al., 2002; Martinez et al., 2001). It has also been suggested that such early activities might not be as "early" as generally thought because visual mechanisms in this time window might well be influenced by feedback from prefrontal cortex (Foxe & Simpson, 2002).

However, following this line of thinking, we do not mean that object categorization in natural scenes is achieved in 120-130 ms. Indeed, a significant difference between two ERP waveforms is not synonymous with the completion of the task by the visual system. What we mean is that by 120-130 ms after stimulus onset, it might well be that some objects are coarsely categorized, or more generally speaking, that at the neuronal population level the categorization process has already started. A similar idea has been proposed, in which a coarse processing of the objects in the scene is used to create a saliency map that could orient more detailed visual processing (Bar, 2004; Macé, Thorpe, & Fabre-Thorpe, 2005).

In addition, the current data also revealed that the fast categorization of objects in natural scenes is relatively unaffected by inversion. The shallower slope of differential activity recorded for inverted stimuli compared to upright ones reinforces the model of accumulation of evidence (Perrett, Oram, & Ashbridge, 1998) used previously in Rousselet et al. (2003) to explain how slightly performance was affected by inversion in these tasks. This small effect of inversion suggests that the neuronal representations used to perform the task are relatively coarse and view invariant, but this issue remains to be investigated more deeply. The data also suggest that stimuli like faces and humans form a very specific class of objects which are by default processed to a larger extent than other objects, for example animals in the present study (see discussion in Rousselet, Macé, & Fabre-Thorpe, 2004).

Finally, we would like to emphasize the main message of this paper, namely that high-level visual categorization can be done without strong task-related ERP differences.

Acknowledgments

We acknowledge support from the CNRS and the Integrative and Computational Neuroscience ACI. The two first authors were supported by PhD grants from the French government. GAR is currently supported by a CIHR fellowship program. The CNRS Research Ethical Board (COPE) approved this work. Caitlin R. Sternberg and Anne-Sophie Paroissien are acknowledged for their help in running subjects in experiment 1 and 2 respectively. Thanks to Nadège M. Bacon for programming stimulus presentation in experiment 2. We also thank Rufin VanRullen for several brainstorming sessions about these data.

Commercial relationships: none.

Corresponding author: Guillaume A. Rousselet. Email: rousseg@mcmaster.ca.
Address: McMaster University, 1280 Main street west, Hamilton, ON, L8S4K1.

References

Allison, T., Puce, A., Spencer, D. D., & McCarthy, G. (1999). Electrophysiological studies of human face perception. I: Potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cereb Cortex, 9*(5), 415-430.

Antal, A., Keri, S., Kovacs, G., Janka, Z., & Benedek, G. (2000). Early and late components of visual categorization: an event-related potential study. *Brain Res Cogn Brain Res, 9*(1), 117-119.

Bar, M. (2004). Visual objects in context. *Nat Rev Neurosci, 5*(8), 617-629.

Chelazzi, L., Duncan, J., Miller, E. K., & Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *J Neurophysiol, 80*(6), 2918-2940.

Delorme, A., Rousselet, G. A., Mace, M. J., & Fabre-Thorpe, M. (2004). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Brain Res Cogn Brain Res, 19*(2), 103-113.

Di Russo, F., Martinez, A., Sereno, M. I., Pitzalis, S., & Hillyard, S. A. (2002). Cortical sources of the early components of the visual evoked potential. *Hum Brain Mapp, 15*(2), 95-111.

DiCarlo, J. J., & Maunsell, J. H. (2005). Using neuronal latency to determine sensory-motor processing pathways in reaction time tasks. *J Neurophysiol, 93*(5), 2974-2986.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*(2), 171-180.

Fabre-Thorpe, M., Richard, G., & Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport, 9*(2), 303-308.

Fize, D., Fabre-Thorpe, M., Richard, G., Doyon, B., & Thorpe, S. (in press). Foveal vision is not necessary for rapid categorisation of natural images: a behavioural and ERP study. *Brain & Cognition*.

Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans. A framework for defining "early" visual processing. *Exp Brain Res, 142*(1), 139-150.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J Neurosci, 23*(12), 5235-5246.

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science, 182*(108), 177-180.

Hillyard, S. A., & Münte, T. F. (1984). Selective attention to color and location: an analysis with event- related brain potentials. *Percept Psychophys, 36*(2), 185-198.

Itier, R. J., & Taylor, M. J. (2002). Inversion and Contrast Polarity Reversal Affect both Encoding and Recognition Processes of Unfamiliar Faces: A Repetition Study Using ERPs. *Neuroimage, 15*(2), 353-372.

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. Journal of Vision, 3(7), 499-512, http://journalofvision.org/3/7/4/, doi:10.1167/3.7.4.

Johnson, J. S., & Olshausen, B. A. (2005). The earliest EEG signatures of object recognition in a cued-target task are postsensory. Journal of Vision, 5(4), 299-312, http://journalofvision.org/5/4/2/, doi:10.1167/5.4.2.

Kirchner, H., & Thorpe, S. J. (in press). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vis Res*.

Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat Neurosci, 3*(9), 946-953.

Linkenkaer-Hansen, K., Palva, J. M., Sams, M., Hietanen, J. K., Aronen, H. J., & Ilmoniemi, R. J. (1998). Face-selective processing in human extrastriate cortex around 120 ms after stimulus onset revealed by magneto- and electroencephalography. *Neurosci Lett, 253*(3), 147-150.

Macé, M. J.-M., Thorpe, S. J., & Fabre-Thorpe, M. (2005). Rapid categorization of achromatic natural scenes: how robust at very low contrasts? *Eur J Neurosci*, 21(7), 2007-2018.

Martinez, A., DiRusso, F., Anllo-Vento, L., Sereno, M. I., Buxton, R. B., & Hillyard, S. A. (2001). Putting spatial attention on the map: timing and localization of stimulus selection processes in striate and extrastriate visual areas. *Vision Res, 41*(10-11), 1437-1457.

Perrett, D. I., Hietanen, J. K., Oram, M. W., & Benson, P. J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philos Trans R Soc Lond B Biol Sci, 335*(1273), 23-30.

Perrett, D. I., Oram, M. W., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition, 67*(1-2), 111-145.

Puce, A., Allison, T., & McCarthy, G. (1999). Electrophysiological studies of human face perception. III: Effects of top-down processing on face-specific potentials. *Cereb Cortex, 9*(5), 445-458.

Rossion, B., & Gauthier, I. (2002). How does the brain process upright and inverted faces? *Behavioral and Cognitive Neuroscience Reviews, 1*(1), 62-74.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nat Neurosci, 5*(7), 629-630.

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2004). Animal and human faces in natural scenes: How specific to human faces is the N170 ERP component? *Journal of Vision*, 4(1), 13-21, http://journalofvision.org/4/1/2/, doi:10.1167/4.1.2.

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision*, 3(6), 440-455, http://journalofvision.org/3/6/5/, doi:10.1167/3.6.5.

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2004). Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vision Res, 44*(9), 877-894.

Schyns, P. G. (1998). Diagnostic recognition: task constraints, object information, and their interactions. *Cognition, 67*(1-2), 147-179.

Sheinberg, D. L., & Logothetis, N. K. (2001). Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *J Neurosci, 21*(4), 1340-1350.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu Rev Neurosci, 19*, 109-139.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat Neurosci, 5*(7), 682-687.

VanRullen, R., & Thorpe, S. J. (2001a). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception, 30*(6), 655-668.

VanRullen, R., & Thorpe, S. J. (2001b). The time course of visual processing: from early perception to decision-making. *J Cogn Neurosci, 13*(4), 454-461.

3.5 - Les activités différentielles précoces dans une double tâche

Ces nouvelles données apportent des éléments intéressants quant à l'origine des activités différentielles précoces entre essais-cibles et essais-distracteurs autour de 80 à 120 ms. Nous avons montré grâce aux résultats obtenus à l'issue des expériences dans lesquelles le contraste et la luminance des images étaient profondément modifiés que l'amplitude des différences précoces enregistrées entre les signaux associés aux essais-cibles et aux essais-distracteurs pouvait être réduite ou augmentée selon que les différences physiques entre les images étaient diminuées ou augmentées.
Inversement, on peut s'affranchir des contaminations dues aux différences physiques en comparant pour les mêmes images le signal EEG obtenu lorsqu'elles sont traitées en tant que cible ou en tant que distracteur dans une tâche de catégorisation. Les deux dernières expériences décrites ci-dessus et dans lesquelles les sujets doivent catégoriser des visages d'humains parmi des visages d'animaux (ou l'inverse) permettent d'effectuer ce type d'analyse sur l'activité cérébrale enregistrée.

Dans l'article n°7 la comparaison du signal EEG entre images cibles et distracteurs de l'expérience 1 fait apparaître des activités cérébrales différentielles précoces (figure n°3 A et B). Cette même comparaison réalisée dans l'expérience 2 pour laquelle n'ont été sélectionnés que des gros plans de visages d'humains et d'animaux (figure n°3 E et F) montre une forte diminution de ces activités différentielles précoces. En fait, ces activités différentielles précoces ne se retrouvaient plus que sur les électrodes médianes sur les animaux. Cependant, la méthode la plus efficace pour supprimer totalement les influences des différences physiques entre les images sur le signal reste la différentielle calculée sur le signal EEG enregistré sur les mêmes images traitées soit comme cibles, soit comme distracteurs (par des sujets différents puisqu'une image donnée n'est vue qu'une seule fois par chaque sujet). En procédant ainsi dans les deux expériences ont constate que jusqu'à une latence de 150-200 ms, il n'existe plus aucune différence entre les potentiels évoqués par les images vues en tant que cibles ou en tant que distracteurs (figure n°3 C et D + G et H). Ainsi, lorsque l'on supprime les différences physiques entre les cibles et les distracteurs, on supprime également les différences précoces dans le signal EEG.

L'ensemble des données qui permettent de mieux interpréter l'origine de ces activités différentielles précoces et qui ont été présentées dans cette thèse au fil des différents chapitre fera l'objet d'un article de synthèse actuellement en préparation. La figure N°9 permet de rassembler les résultats concernant les activités différentielles précoces obtenues dans nos différentes expériences.

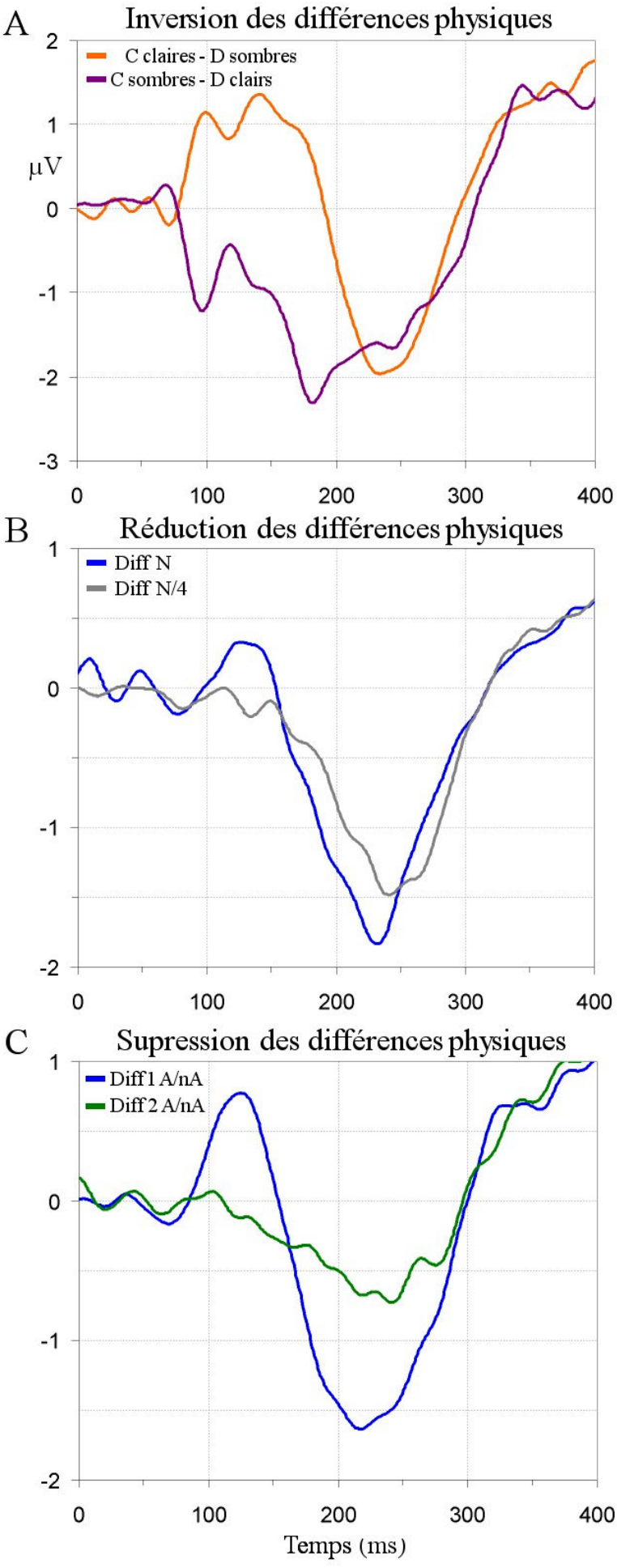

A
Inversion des différences physiques
C claires - D sombres
C sombres - D clairs
µV
2
1
0
-1
-2
-3
0
100
200
300
400
B
Réduction des différences physiques
Diff N
Diff N/4
1
0
-1
-2
0
100
200
300
400
C
Supression des différences physiques
Diff1 A/nA
Diff2 A/nA
1
0
-1
-2
0
100
200
300
400
Temps (ms)

Figure 9 : Les 3 sections de cette figure font la synthèse des différents effets que nous avons observés sur les activités différentielles précoces tout au long de ce mémoire. A : La luminance des images est modifiée et l'on maximise les différences physiques entre les images en soustrayant les distracteurs sombres aux cibles claires (orange) ou bien en soustrayant les distracteurs clairs aux cibles sombres (violet). L'augmentation des différences physique induit de très fortes activités différentielles précoces avant 100 ms. Le signe de ces activités différentielles précoces est inversé entre les deux conditions jusqu'à environ 150 ms, latence à partir de laquelle les deux courbes évoluent dans le même sens. B : Le contraste des images est réduit, ce qui a pour effet de diminuer les différences physiques entre les cibles et les distracteurs puisque la quantité d'informations qu'elles contiennent est également réduite. Seule la courbe correspondant au contraste original des images (en bleu) reflète une activité différentielle précoce entre 100 et 150 ms. Dans la condition en gris, le contraste est divisé par 4 et l'activité différentielle précoce a disparue. C : L'utilisation de deux tâches de catégorisation successives permet de s'affranchir des différences physiques qui existent toujours entre deux groupes d'images en catégorisant les images soit comme cibles, soit comme distracteurs. La courbe bleue correspond à l'activité différentielle classique lors de la tâche animal/non animal (animaux cibles - distracteurs visages et distracteurs neutres) et la courbe verte à la différentielle de type 2 (animaux cibles - animaux distracteurs). La suppression des différences physiques fait disparaître les différences précoces avant 150 ms mais préserve l'activité différentielle après 150 ms (juste avant le croisement des deux courbes).

3.6 - Conclusion générale

Les deux premières études sur les niveaux de catégorisation nous ont permis de montrer que l'idée très répandue d'une plus grande rapidité d'exécution lors d'une catégorisation au niveau de base comparée à une catégorisation au niveau superordonné trouve très probablement son explication dans les tâches utilisées lors de ces anciennes études et qui toutes font appel à des processus lexicaux. Lorsque le système visuel est le seul impliqué dans une tâche de catégorisation, c'est l'accès au niveau superordonné qui est le plus rapide et non l'accès au niveau de base. Ces données bousculent la hiérarchie traditionnelle des catégories et semblent montrer que le système visuel construit dans un premier temps une représentation grossière de la scène visuelle pour la détailler ensuite au cours du temps, lorsque des informations plus fines lui parviennent (Figure 10). Non seulement cette idée est en accord avec la majorité des modèles de reconnaissance d'objets développés dans l'introduction de ce chapitre, mais elle correspond aussi particulièrement bien à un modèle de traitement de l'information de type "coarse to fine".

Un autre point majeur soulevé dans cette série d'études concerne la notion d'éléments diagnostiques utilisés pour répondre à une tâche de catégorisation donnée. Le nombre et la pertinence de ces éléments dépendent des distances inter- et intra- catégories exploitables au cours d'une tâche donnée (Figure 6). Déterminer quels sont ces éléments dans des tâches complexes utilisant des scènes naturelles est une question fondamentale si l'on souhaite

comprendre et modéliser le fonctionnement du système visuel. Il faut aussi noter que la définition d'un élément diagnostique n'est pas figée et qu'elle est en interaction avec l'expertise que possède un sujet avec une classe de stimuli. L'expertise permet de dégager des éléments diagnostiques même dans des situations pour lesquelles les indices sont très subtils à distinguer (on peut reprendre l'exemple de la catégorisation des visages humains et des têtes d'animaux que les sujets humains n'ont aucun mal à accomplir).

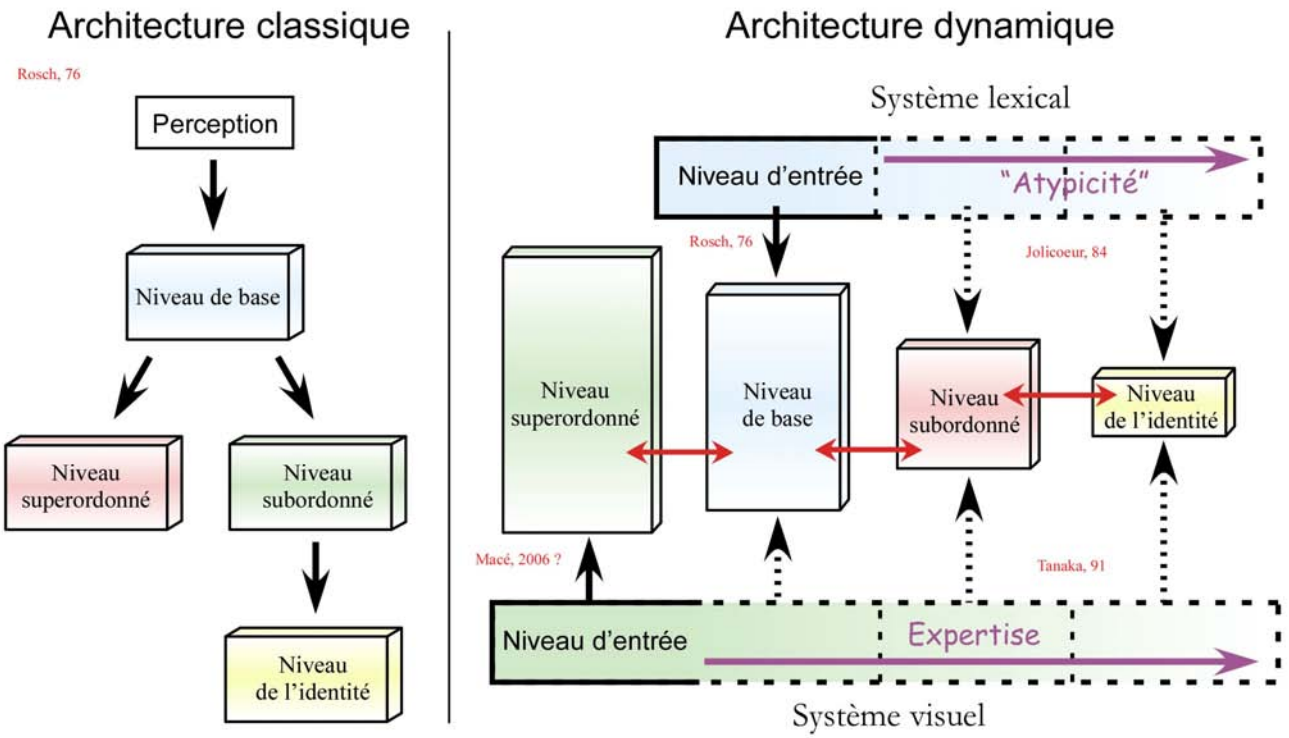


Figure 10 : Les expériences menées par Rosch dans les années 70 ont abouti à l'architecture de la catégorisation classiquement acceptée avec un niveau de traitement de l'information privilégié : le niveau de base. Ce niveau correspond à un optimum entre la profondeur de traitement et la quantité d'information obtenue. Les éléments plus abstraits (niveau superordonné) ou plus figuratifs (niveau subordonné puis niveau de l'identité) sont atteints en partant de ce niveau. Cette vision a été peu à peu enrichie par l'introduction de la typicité (Jolicoeur, 1984) et de l'expertise (Tanaka, 1991) qui peuvent chacun modifier le niveau d'entrée de la catégorisation qui n'est donc plus nécessairement située au niveau de base. Dans l'architecture dynamique de la catégorisation que nous proposons ici, le système lexical et le système visuel peuvent chacun amorcer l'accès aux catégories. Le système lexical possède un niveau d'entrée principalement au niveau de base, qui peut être décalé vers les niveaux inférieurs lorsque les stimuli sont atypiques. Le système visuel possède un niveau d'entrée préférentiellement au niveau superordonné de par la nature des informations magnocellulaires et l'architecture du traitement visuel. Ce niveau d'entrée peut être déplacé vers des niveaux plus figuratifs lorsque le système visuel possède une expertise pour les stimuli considérés. Dans notre schéma, la taille des blocs représente la taille des niveaux de catégorie.

En ce qui concerne la dernière étude sur le traitement spécifique (ou non) des visages, on peut observer une dissociation entre les résultats comportementaux et électrophysiologiques. Alors que les visages humains et les têtes d'animaux sont catégorisés avec des temps de réaction similaires, la comparaison de l'activité cérébrale enregistrée sur les cibles et les distracteurs donne des résultats très différents. Ces divergences prennent probablement leur origine dans l'expertise qu'ont les humains dans la reconnaissance des visages et le besoin qu'ils ont dans leur vie quotidienne de toujours considérer ces stimuli comme significatifs et donc de les traiter en détails même si la tâche ne le demande pas de façon explicite.

Synthèse et perspectives

"Vous devez relâcher le doigt dès que vous apercevez un animal". Une tâche très simple, presque ludique (du moins pendant les mille premiers essais !). Elle peut même être effectuée par des animaux dépourvus de langage. Mais cette simplicité apparente cache des processus d'une grande complexité. Si nous sommes capables d'extraire le sens d'une dizaine d'images présentées en une seule seconde ([Potter, 1976 #267]), c'est que le système visuel peut calculer plusieurs fois par seconde les contours présents dans l'image véhiculée par le nerf optique. Cette simple extraction des bords présents dans une image constituée par un million de fibres en provenance de la rétine requiert une puissance de calcul brute probablement supérieure à 1 GigaFlops ([Moravec, 1998 #268] : 100 instructions élémentaires par pixel x 1 million de pixels x 10 hertz). Et ceci ne représente qu'une portion très restreinte du système visuel, sans parler du cerveau dans son ensemble ! Les puissances de calcul nécessaires pour reproduire ces capacités (si l'on omet pour l'instant l'architecture sophistiquée et la souplesse du système visuel), ne sont accessibles que depuis très récemment aux spécialistes de la vision par machine. Ce simple fait explique au moins en partie pourquoi les tentatives pour reproduire les capacités visuelles des primates dans des systèmes de vision artificielle se sont jusqu'à maintenant soldées par des échecs retentissants. Le temps joue cependant en faveur des modélisateurs et des roboticiens puisque la puissance de calcul des microprocesseurs continue de progresser de manière exponentielle. Mais la puissance brute ne suffit pas et produire des systèmes artificiels capables d'effectuer une tâche d'apparence aussi simple pour les humains et les singes que celle de trier des images selon leur contenu sera grandement facilitée par une compréhension plus approfondie du fonctionnement du système visuel.
Les travaux présentés dans cette thèse sont inscrits dans cette démarche de compréhension des mécanismes biologiques de la vision qui permettra à terme de reproduire dans des systèmes artificiels les capacités de traitement de l'information des systèmes vivants.

Pour réduire la complexité inhérente au problème de l'interprétation d'une scène, le système visuel a recours à des artifices de calcul qui lui permettent d'être à la fois plus efficace et plus rapide. L'une des idées les plus importantes qui émerge dans le travail présenté ici est celle d'un système qui ne traite pas toute l'information visuelle en une seule fois, mais qui utilise dans un premier temps une version en "basse résolution" de la scène (à l'instar de ce qui est effectué par certains logiciels de traitement d'images) avant d'analyser toute l'image en détail. Nous avons mis en évidence cette caractéristique du système visuel dans le 1er chapitre avec les expériences de réduction du contraste et de la luminance qui montrent que les premières informations disponibles, qui sont de type magnocellulaire, sont suffisantes pour effectuer une

tâche de catégorisation animal/non animal. L'analyse fine de la scène visuelle dans son ensemble, avec tous ses détails, demande plus de temps et de calculs et nécessite surtout d'attendre l'arrivée des informations parvocellulaires. Le système visuel construirait donc dans un premier temps une représentation approximative de la scène qui servirait à la fois à déclencher des réponses motrices rapides et le plus souvent adaptées et à recevoir les informations détaillées pour qu'émerge une représentation de la scène en "haute résolution".

Les représentations précoces de la scène visuelle sont intéressantes parce qu'elles sont construites en une fraction de seconde à partir d'informations parcellaires et permettent néanmoins aux sujets d'effectuer certaines tâches visuelles complexes. Le décours temporel de la catégorisation visuelle est par exemple si rapide qu'il impose de fortes contraintes temporelles aux modèles de reconnaissance d'objets qui peuvent être proposés, comme nous l'avons vu dans le 2ème chapitre.
Les résultats des expériences détaillées dans le 3ème chapitre permettent d'explorer les limites de ces représentations précoces puisqu'elles semblent insuffisamment détaillées pour permettre de catégoriser les chiens et les oiseaux lorsqu'ils sont présentés parmi d'autres images d'animaux. Les caractéristiques visuelles qui différencient tous ces animaux ne sont peut être pas suffisamment distinctes sur la base des seules informations fournies par le système magnocellulaire et c'est après un cours délai que des informations supplémentaires (parvocellulaires ?) peuvent aider à lever l'ambiguïté sur la nature du stimulus. En poursuivant cette logique, on pourra observer comment va évoluer le temps de réaction et la précision des sujets dans une tâche de catégorisation qui requiert l'extraction de caractéristiques visuelles encore plus fines, par exemple lorsque les sujets doivent trouver des lévriers parmi d'autres chiens ou des mésanges parmi d'autres oiseaux. Notre prédiction est que le système visuel devra attendre encore plus de temps pour intégrer des informations très détaillées sur l'image et que la performance devrait encore se dégrader.

Dans les expériences décrites ci-dessus, seul l'espace de recherche des cibles est manipulé. On peut obtenir des résultats ayant une plus large portée en faisant varier les trois principaux paramètres de la catégorisation : la variabilité des cibles, la variabilité des distracteurs et la distance cibles-distracteurs. C'est ce que nous avons en partie réalisé dans l'expérience de catégorisation chien/non chien dans laquelle la proportion de distracteurs animaux était variable. L'homogénéité des cibles est augmentée lors du passage de la catégorisation animal/non animal à la catégorisation chien/non chien. Au sein de la tâche chien/non chien, l'augmentation de la proportion d'animaux parmi les distracteurs a pour conséquence l'augmentation de l'homogénéité des distracteurs et la diminution de la distance cibles-distracteurs. Beaucoup de travail reste cependant à faire pour mieux contrôler et caractériser

tous ces paramètres avant de pouvoir tenter de modéliser leur influence respective et leurs interactions dans des tâches de catégorisation.

Les facteurs propres aux stimuli, comme ceux que nous venons d'énoncer, ne sont pas les seuls à influencer la performance dans une tâche de catégorisation : l'expertise des sujets avec les stimuli est un facteur interne qui peut aussi avoir un rôle très important. Nous avons observé cet effet dans les expériences de catégorisation des visages du 3ème chapitre. Les sujets sont des experts dans le traitement des visages et leurs performances sont supérieures à celles qu'ils atteignent dans la tâche animal/non animal malgré le niveau de détail plus poussé nécessaire pour séparer correctement les visages d'humains et les visages d'animaux. L'expertise interviendrait ici en augmentant les capacités de discrimination du système visuel, peut-être grâce à une sur-représentation neuronale des objets concernés.

Il est également possible de modifier les contraintes temporelles qui pèsent sur la tâche en manipulant la manière dont le sujet fournit sa réponse. Une réponse qui nécessite l'utilisation des deux mains (choix oui/non) plutôt qu'une seule (go/no-go) prend plus de temps, probablement à causes d'interactions entre les deux programmes moteurs préparés. Nous avons également vu qu'une tâche de discrimination dans laquelle la réponse est donnée au moyen des yeux donne lieu à des TR beaucoup plus courts qu'une tâche dans laquelle la réponse est donnée manuellement. J'aimerai m'intéresser à cette variabilité très importante imputée au système moteur et je vais pour cela effectuer mon stage postdoctoral dans un laboratoire spécialisé dans la motricité et les interactions visuo-motrices. Je vais donc pouvoir au cours des années qui viennent explorer ces questions relatives à la chronométrie des réactions motrices.

Finalement comme nous l'avons vu tout au long de ce mémoire, la compréhension des mécanismes de catégorisation passe par la définition et l'analyse d'un ensemble de paramètres influençant le contexte de la tâche étudiée, faisant intervenir l'espace de recherche des images cibles et distracteurs, l'expertise des sujets, le degré de pré-activation du système ou encore le mode de réponse utilisé. L'ensemble de ces paramètres ont pu être évalué par des protocoles comportementaux. Le développement de l'EEG chez le singe et de la stimulation magnétique transcrânienne (TMS) chez l'homme, alliées ou non à des techniques de masquage, devrait apporter de nouvelles indications tout aussi précieuses sur la chronométrie de la catégorisation visuelle. Les singes présenteront très probablement une activité cérébrale différentielle plus précoce que celle des hommes (leurs latences sont toujours inférieures) et l'utilisation de la TMS devrait également permettre d'explorer le décours temporel des traitements visuels à travers les différentes aires en faisant varier la localisation de la stimulation magnétique et le délai entre l'apparition de l'image et cette stimulation. C'est l'un des autres objectifs de mon

post-doctorat que de me familiariser avec cette technique de stimulation à travers l'étude des interactions visuo-motrices.

La plate-forme technique de l'IFR Sciences du Cerveau de Toulouse devrait permettre à terme de mener des expériences à la fois chez l'homme et le singe. Il sera particulièrement intéressant d'observer les aires activées lorsque les sujets sont engagés dans des tâches à différents niveaux de catégorisation (animaux, chiens, lévriers) ou à différents niveaux d'expertise (visages humains ou animaux, greebles chez des naïfs ou des experts, etc...)
Un grand enjeu pour les expériences à venir est également d'arriver à déterminer quelles sont les informations qui sont les plus diagnostiques ou les plus utilisées par les sujets pour répondre dans la tâche demandée. Cette analyse est par exemple possible en faisant appel à des techniques proches de la "reverse correlation", comme les bubbles ([Gosselin, 2001 #269]) ou en effectuant des transformations sur la phase ou le spectre d'amplitude des images. La grande variabilité des images naturelles est cependant un obstacle pour déterminer les indices utilisés par les sujets et il sera probablement nécessaire de travailler sur les algorithmes de convergence de ces techniques pour parvenir à des images de classification fiables avec un nombre d'essais raisonnable.

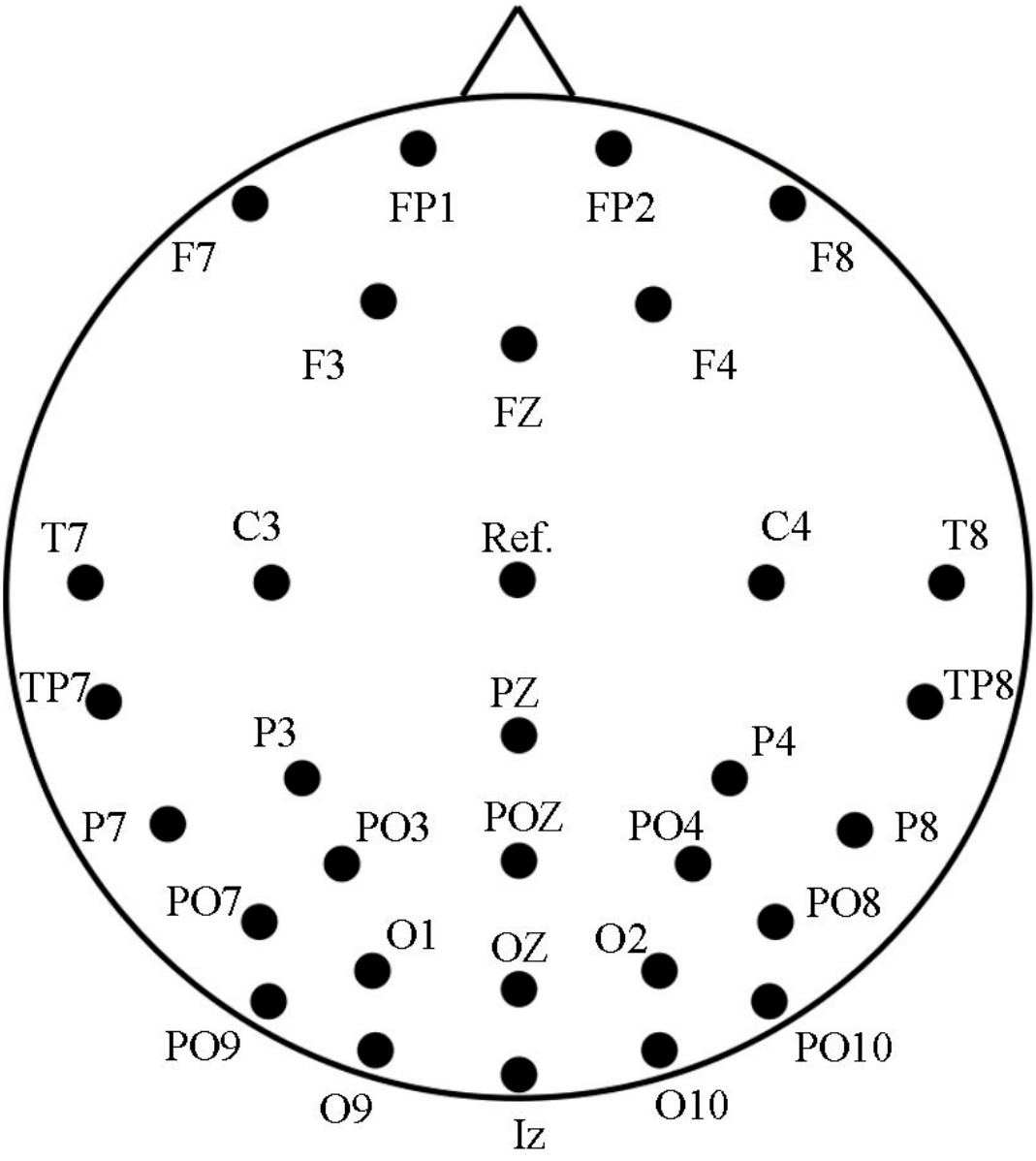
Enfin pour terminer, la tâche de catégorisation visuelle rapide se déroule à des latences si courtes qu'elle pourrait bien précéder l'accès conscient à l'information si celui-ci repose par exemple sur des oscillations corticales induites qui ne peuvent se mettre en place instantanément (Tallon-Baudry & Bertrand, 1999). C'est particulièrement le cas dans la tâche de discrimination oculaire dans laquelle chacun constate avec étonnement que ses yeux se sont déplacés sur la bonne image *avant* que celle-ci ne soit apparue à l'écran (subjectivement bien sûr, puisqu'il ne s'agit pas d'anticipations et que la performance est supérieure au niveau chance). Il est cependant difficile de prouver que les processus de catégorisation mis en jeu sont effectués de manière totalement inconsciente et de nouveaux protocoles à base de mouvements oculaires, de masquage et de TMS restent à développer.

Conclusion :

Les expériences réalisées pendant ma thèse ont permis d'aborder différents aspects des traitements visuels précoces en terme de chronométrie et de mieux caractériser les tâches qui peuvent être réalisées grâce aux représentations rudimentaires construites à partir des premières informations visuelles. Il paraît cependant indispensable de se tourner à présent vers d'autres outils d'analyse, complémentaires de l'EEG et de la psychophysique, tels que l'IRMf et la TMS pour mieux comprendre les mécanismes impliqués dans la catégorisation des scènes naturelles.

Annexe A - Répartition des électrodes

Répartition des électrodes à la surface de la tête. Ce bonnet est adapté du système classique 10-20, avec des électrodes surnuméraires en région occipitale, en regard des aires corticales spécialisées dans la vision.



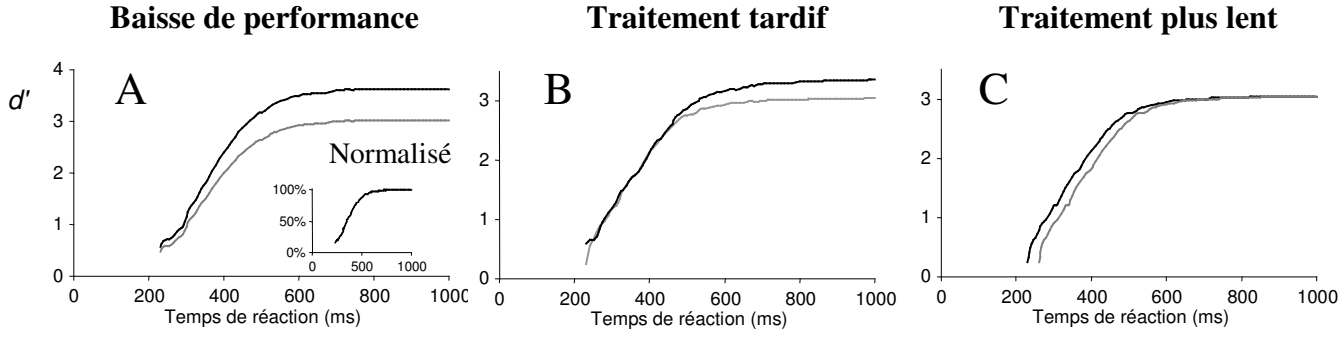
Annexe B - Précision en fonction du temps : d'

La précision d'un sujet s'évalue en calculant et en comparant le nombre de cibles et de distracteurs correctement catégorisés. Mais comment peut-on estimer la performance en ne prenant que les TR inférieurs à 500 ms ? Il n'est plus possible dans ce cas de comparer le pourcentage de réussite des sujets sur les cibles et les distracteurs puisqu'ils disposent encore de 500 ms pour répondre. Le *d'* est un indice qui permet de faire cette estimation de la précision des réponses au cours du temps[1]. Le calcul du *d'* peut se faire sur le nombre de réponses par unité de temps ou sur le nombre de réponses cumulées. Dans le cas d'un *d'* cumulé, le niveau du plateau final des courbes de performance dépend directement de la précision des sujets dans la tâche.

La figure ci-dessous illustre trois comparaisons de deux courbes de *d'* et la façon dont elles doivent être interprétées :

- Tout d'abord, si les courbes sont proportionnelles, c'est-à-dire si les courbes normalisées sont confondues, il s'agit d'une simple variation de précision indépendante du temps de réaction : les précisions cumulées sont en effet proportionnelles à chaque pas de temps (A).
- Ensuite, si seul le plateau final diffère, cela signifie que le traitement rapide est identique pour les deux catégories - parties initiales des courbes - et diverge pour les réponses tardives, la catégorisation devenant plus précise pour l'une des deux catégories (B).
- Enfin, si l'une des courbes semble retardée par rapport à l'autre, cela signifie que l'une des catégories est traitée plus rapidement que l'autre et que ce gain est indépendant du temps de réaction. Dans ce cas, on peut observer également des distributions de TR similaires mais décalées l'une par rapport à l'autre (C).

Dans les analyses des résultats, les cas sont rares où l'interprétation est aussi claire. Souvent, les différents types d'effets sont mélangés, ce qui complique l'interprétation.



[1] À partir du pourcentage de réponses sur l'ensemble des cibles, on détermine à l'aide d'une table de correspondance un Z-score à un pas de temps donné : *Zc*. On effectue de même pour les distracteurs : *Zd*. La valeur du *d'* est la soustraction de ces deux valeurs : *d'=Zc-Zd*

Acronymes :

Aires cérébrales :

CGL : Corps genouillé latéral

V1 : Aire visuelle primaire

V2, V3, V4 : Aires visuelles extrastriées

STS : Sulcus temporal supérieur

IT : Cortex inféro-temporal (antérieur : AIT, postérieur : PIT)

TE : Aire temporale (AIT)

TEO : Aire temporo-occipitale (PIT)

MT : Aire médiane temporale

MST : Aire médiane temporale supérieure

LIP : Aire Latérale intrapariétale

FEF : Frontal Eye Field

Techniques d'analyse :

MEG : Magnéto-encéphalographie

IRMf : Imagerie par résonance magnétique fonctionnelle

TMS : Stimulation magnétique transcrânienne

EEG : Electro-encéphalographie

PEV : Potentiels évoqués visuels

P1, N1, (P100, N170) : Premières ondes positives et négatives des potentiels évoqués

DA : Activité différentielle

Divers :

TR : Temps de réaction

SOA : Stimulus onset asynchrony (délai entre l'apparition de l'image et du masque)

An/nAn (ou A/nA) : Animal/non Animal

N/2 (N/X) : contraste divisé par 2 (par X)

RBF : Radial Basis Function

IFR : Institut Fédératif de Recherche

Bibliographie

Albrecht, D.G., Geisler, W.S., Frazor, R.A. & Crane, A.M. (2002). Visual cortex neurons of monkeys and cats: temporal dynamics of the contrast response function. *J Neurophysiol*, **88**, 888-913.

Allison, J.D., Melzer, P., Ding, Y., Bonds, A.B. & Casagrande, V.A. (2000). Differential contributions of magnocellular and parvocellular pathways to the contrast response of neurons in bush baby primary visual cortex (V1). *Vis Neurosci*, **17**, 71-76.

Allison, T., Ginter, H., McCarthy, G., Nobre, A.C., Puce, A., Luby, M. & Spencer, D.D. (1994). Face recognition in human extrastriate cortex. *J Neurophysiol*, **71**, 821-825.

Allison, T., Puce, A., Spencer, D.D. & McCarthy, G. (1999). Electrophysiological studies of human face perception. I: Potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cereb Cortex*, **9**, 415-430.

Antal, A., Keri, S., Kovacs, G., Janka, Z. & Benedek, G. (2000). Early and late components of visual categorization: an event-related potential study. *Brain Res Cogn Brain Res*, **9**, 117-119.

Aubertin, A., Fabre Thorpe, M., Fabre, N. & Geraud, G. (1999). Catégorisation visuelle rapide et vitesse de traitement chez le migraineux. *C R Acad Sci III*, **322**, 695-704.

Bacon-Macé, N., Kirchner, H., Fabre-Thorpe, M. & Thorpe, S.J. (in revision). Effects of task requirements on rapid natural scene processing: From common sensory encoding to distinct decisional mechanisms. *J Exp Psychol Hum Learn*.

Bacon-Macé, N., Macé, M.J.-M., Fabre-Thorpe, M. & Thorpe, S.J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Res*, **45**, 1459-1469.

Bar, M. (2004). Visual objects in context. *Nat Rev Neurosci*, **5**, 617-629.

Barcelo, F., Suwazono, S. & Knight, R.T. (2000). Prefrontal modulation of visual processing in humans. *Nat Neurosci*, **3**, 399-403.

Barlow, H.B. (1953). Summation and inhibition in the frog's retina. *J Physiol*, **119**, 69-88.

Barsalou, L.W. & Sewell, D.R. (1985). Contrasting the representation of scripts and categories. *J Mem Lang*, **24**, 646-665.

Baylis, G.C. & Rolls, E.T. (1987). Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. *Exp Brain Res*, **65**, 614-622.

Baylis, G.C., Rolls, E.T. & Leonard, C.M. (1987). Functional subdivisions of the temporal lobe neocortex. *J Neurosci*, **7**, 330-342.

Benardete, E.A. & Kaplan, E. (1997). The receptive field of the primate P retinal ganglion cell, II: Nonlinear dynamics. *Vis Neurosci*, **14**, 187-205.

Benardete, E.A. & Kaplan, E. (1999). Dynamics of primate P retinal ganglion cells: responses to chromatic and achromatic stimuli. *J Physiol (Lond)*, **519 Pt 3**, 775-790.

Bentin, S., Allison, T., Puce, A., Perez, E. & Mccarthy, G. (1996). Electrophysiological Studies of Face Perception in Humans. *J Cogn Neurosci*, **8**, 551-565.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol Rev*, **94**, 115-147.

Biederman, I. & Gerhardstein, P.C. (1993). Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance [published erratum appears in J Exp Psychol Hum Percept Perform 1994 Feb;20(1):80] [see comments]. *J Exp Psychol Hum Percept Perform*, **19**, 1162-1182.

Booth, M.C. & Rolls, E.T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb Cortex*, **8**, 510-523.

Braeutigam, S., Bailey, A.J. & Swithenby, S.J. (2001). Task-dependent early latency (30-60 ms) visual processing of human faces and other objects. *Neuroreport*, **12**, 1531-1536.

Bringuier, V., Chavane, F., Glaeser, L. & Fregnac, Y. (1999). Horizontal propagation of visual activity in the synaptic integration field of area 17 neurons. *Science*, **283**, 695-699.

Buckley, M.J., Gaffan, D. & Murray, E.A. (1997). Functional double dissociation between two inferior temporal cortical areas: perirhinal cortex versus middle temporal gyrus. *J Neurophysiol*, **77**, 587-598.

Buffalo, E.A., Ramus, S.J., Clark, R.E., Teng, E., Squire, L.R. & Zola, S.M. (1999). Dissociation between the effects of damage to perirhinal cortex and area TE. *Learn Mem*, **6**, 572-599.

Bullier, J. (2001). Integrated model of visual processing. *Brain Res Brain Res Rev*, **36**, 96-107.

Bullier, J. & Kennedy, H. (1983). Projection of the lateral geniculate nucleus onto cortical area V2 in the macaque monkey. *Exp Brain Res*, **53**, 168-172.

Bulthoff, H.H. & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc Natl Acad Sci U S A*, **89**, 60-64.

Bulthoff, H.H., Edelman, S.Y. & Tarr, M.J. (1995). How are three-dimensional objects represented in the brain? *Cereb Cortex*, **5**, 247-260.

Busettini, C., Masson, G.S. & Miles, F.A. (1997). Radial optic flow induces vergence eye movements with ultra-short latencies. *Nature*, **390**, 512-515.

Carson, C., Belongie, S., Greenspan, H. & Malik, J. (1997). Region-based image querying. *Proc. IEEEW. on Content-Based Access of Image and Video Libraries*, pp. 42-49.

Clark, V.P. & Hillyard, S.A. (1996). Spatial selective attention affects early extrastriate but not striate components of the visual evoked potential. *J Cogn Neurosci*, **8**, 387-402.

Cowey, A. & Gross, C.G. (1970). Effects of foveal prestriate and inferotemporal lesions on visual discrimination by rhesus monkeys. *Exp Brain Res*, **11**, 128-144.

Dacey, D.M. & Petersen, M.R. (1992). Dendritic field size and morphology of midget and parasol ganglion cells of the human retina. *Proc Natl Acad Sci U S A*, **89**, 9666-9670.

Damasio, A.R., Damasio, H. & Van Hoesen, G.W. (1982). Prosopagnosia: anatomic basis and behavioral mechanisms. *Neurology*, **32**, 331-341.

de Gelder, B., Bachoud-Levi, A.C. & Degos, J.D. (1998). Inversion superiority in visual agnosia may be common to a variety of orientation polarised objects besides faces. *Vision Res*, **38**, 2855-2861.

De Valois, R.L., Morgan, H. & Snodderly, D.M. (1974). Psychophysical studies of monkey vision. III. Spatial luminance contrast sensitivity tests of macaque and human observers. *Vision Res*, **14**, 75-81.

De Valois, R.L., Morgan, H.C., Polson, M.C., Mead, W.R. & Hull, E.M. (1974). Psychophysical studies of monkey vision. I. Macaque luminosity and color vision tests. *Vision Res*, **14**, 53-67.

De Valois, R.L., Smith, C.J., Kitai, S.T. & Karoly, A.J. (1958). Response of single cells in monkey lateral geniculate nucleus to monochromatic light. *Science*, **127**, 238-239.

Debruille, J.B., Guillem, F. & Renault, B. (1998). ERPs and chronometry of face recognition: following-up Seeck et al. and George et al. *Neuroreport*, **9**, 3349-3353.

Deco, G. & Zihl, J. (2001). A neurodynamical model of visual attention: feedback enhancement of spatial resolution in a hierarchical system. *J Comput Neurosci*, **10**, 231-253.

Delorme, A. (2000). Traitement visuel rapide des scènes naturelles chez le singe, l'homme et la machine. UPS, pp. 293.

Delorme, A., Richard, G. & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. *Vision Res*, **40**, 2187-2200.

Delorme, A., Rousselet, G.A., Macé, M.J.-M. & Fabre-Thorpe, M. (2004). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Brain Res Cogn Brain Res*, **19**, 103-113.

Derrington, A.M. & Lennie, P. (1984). Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *J Physiol*, **357**, 219-240.

Desimone, R., Albright, T.D., Gross, C.G. & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J Neurosci*, **4**, 2051-2062.

Desimone, R., Schein, S.J., Moran, J. & Ungerleider, L.G. (1985). Contour, color and shape analysis beyond the striate cortex. *Vision Res*, **25**, 441-452.

DeYoe, E.A. & Van Essen, D.C. (1985). Segregation of efferent connections and receptive field properties in visual area V2 of the macaque. *Nature*, **317**, 58-61.

Di Russo, F., Martinez, A., Sereno, M.I., Pitzalis, S. & Hillyard, S.A. (2001). Cortical sources of the early components of the visual evoked potential. *Hum Brain Mapp*, **15**, 95-111.

Di Russo, F., Pitzalis, S., Spitoni, G., Aprile, T., Patria, F., Spinelli, D. & Hillyard, S.A. (2005). Identification of the neural sources of the pattern-reversal VEP. *Neuroimage*, **24**, 874-886.

DiCarlo, J.J. & Maunsell, J.H. (2005). Using neuronal latency to determine sensory-motor processing pathways in reaction time tasks. *J Neurophysiol*, **93**, 2974-2986.

Duncan, J. & Humphreys, G.W. (1989). Visual search and stimulus similarity. *Psychol Rev*, **96**, 433-458.

Edelman, S. & Bulthoff, H.H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Res*, **32**, 2385-2400.

Edelman, S. & Duvdevani-Bar, S. (1997). A model of visual recognition and categorization. *Philos Trans R Soc Lond B Biol Sci*, **352**, 1191-1202.

Enroth-Cugell, C. & Robson, J.G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. *J Physiol*, **187**, 517-552.

Fabre-Thorpe, M., Delorme, A., Marlot, C. & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *J Cogn Neurosci*, **13**, 171-180.

Fabre-Thorpe, M., Richard, G. & Thorpe, S.J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport*, **9**, 303-308.

Farah, M.J. (1996). Is face recognition 'special'? Evidence from neuropsychology. *Behav Brain Res*, **76**, 181-189.

Farah, M.J. & Aguirre, G.K. (1999). Imaging visual recognition: PET and fMRI studies of the functional anatomy of human visual recognition. *Trends Cogn Sci*, **3**, 179-186.

Farah, M.J., Levinson, K.L. & Klein, K.L. (1995). Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia*, **33**, 661-674.

Ferrera, V.P., Nealey, T.A. & Maunsell, J.H. (1992). Mixed parvocellular and magnocellular geniculate signals in visual area V4. *Nature*, **358**, 756-761.

Fitzpatrick, D., Lund, J.S. & Blasdel, G.G. (1985). Intrinsic connections of macaque striate cortex: afferent and efferent connections of lamina 4C. *J Neurosci*, **5**, 3329-3349.

Fize, D., Boulanouar, K., Chatel, Y., Ranjeva, J.P., Fabre-Thorpe, M. & Thorpe, S. (2000). Brain areas involved in rapid categorization of natural images: an event-related fMRI study. *Neuroimage*, **11**, 634-643.

Fize, D., Fabre-Thorpe, M., Richard, G., Doyon, B. & Thorpe, S.J. (2005). Rapid categorization of foveal and extrafoveal natural images: Associated ERPs and effects of lateralization. *Brain Cogn*, **59**, 145-158.

Foxe, J.J. & Simpson, G.V. (2002). Flow of activation from V1 to frontal cortex in humansA framework for defining "early" visual processing. *Exp Brain Res*, **142**, 139-150.

Freedman, D.J., Riesenhuber, M., Poggio, T. & Miller, E.K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, **291**, 312-316.

Freedman, D.J., Riesenhuber, M., Poggio, T. & Miller, E.K. (2002). Visual categorization and the primate prefrontal cortex: neurophysiology and behavior. *J Neurophysiol*, **88**, 929-941.

Galifret, Y. & Pieron, H. (1949). Vitesse de réaction et intensité de sensation : données expérimentales sur le problème d'une courbe sigmoïde des vitesses. *L'ann. Psychol.*, **51**, 1-16.

Gauthier, I., Anderson, A.W., Tarr, M.J., Skudlarski, P. & Gore, J.C. (1997). Levels of categorization in visual recognition studied using functional magnetic resonance imaging. *Curr Biol*, **7**, 645-651.

Gauthier, I. & Tarr, M.J. (1997). Becoming a "Greeble" expert: exploring mechanisms for face recognition. *Vision Res*, **37**, 1673-1682.

Gauthier, I., Tarr, M.J., Moylan, J., Anderson, A.W., Skudlarski, P. & Gore, J.C. (2000). Does subordinate-level categorization engage the functionally-defined face area? *Cognitive Neuropsychology*, **17**, 143-163.

Gauthier, I., Tarr, M.J., Moylan, J., Skudlarski, P., Gore, J.C. & Anderson, A.W. (2000). The fusiform "face area" is part of a network that processes faces at the individual level. *J Cogn Neurosci*, **12**, 495-504.

Gautrais, J. & Thorpe, S. (1998). Rate coding versus temporal order coding: a theoretical approach. *Biosystems*, **48**, 57-65.

Gemba, H. & Sasaki, K. (1989). Potential related to no-go reaction of go/no-go hand movement task with color discrimination in human. *Neurosci Lett*, **101**, 263-268.

George, N., Dolan, R.J., Fink, G.R., Baylis, G.C., Russell, C. & Driver, J. (1999). Contrast polarity and face recognition in the human fusiform gyrus. *Nat Neurosci*, **2**, 574-580.

George, N., Jemel, B., Fiori, N. & Renault, B. (1997). Face and shape repetition effects in humans: a spatio-temporal ERP study. *Neuroreport*, **8**, 1417-1423.

Gosselin, F. & Schyns, P.G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Res*, **41**, 2261-2271.

Grieve, K.L., Acuna, C. & Cudeiro, J. (2000). The primate pulvinar nuclei: vision and action. *Trends Neurosci*, **23**, 35-39.

Grill-Spector, K., Knouf, N. & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nat Neurosci*, **7**, 555-562.

Grill-Spector, K., Kourtzi, Z. & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Res*, **41**, 1409-1422.

Gross, C.G. (1992). Representation of visual stimuli in inferior temporal cortex. *Philos Trans R Soc Lond B Biol Sci*, **335**, 3-10.

Gross, C.G., Rocha-Miranda, C.E. & Bender, D.B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. *J Neurophysiol*, **35**, 96-111.

Grossberg, S. (1997). Cortical dynamics of three-dimensional figure-ground perception of two-dimensional pictures. *Psychol Rev*, **104**, 618-658.

Halgren, E., Baudena, P., Heit, G., Clarke, J.M., Marinkovic, K., Chauvel, P. & Clarke, M. (1994). Spatio-temporal stages in face and word processing. 2. Depth-recorded potentials in the human frontal and Rolandic cortices. *J Physiol Paris*, **88**, 51-80.

Halgren, E., Baudena, P., Heit, G., Clarke, J.M., Marinkovic, K. & Clarke, M. (1994). Spatio-temporal stages in face and word processing. I. Depth-recorded potentials in the human occipital, temporal and parietal lobes [corrected]. *J Physiol Paris*, **88**, 1-50.

Halgren, E., Raij, T., Marinkovic, K., Jousmaki, V. & Hari, R. (2000). Cognitive response profile of the human fusiform face area as determined by MEG. *Cereb Cortex*, **10**, 69-81.

Hamker, F.H. & Worcester, J. (2002). Object detection in natural scenes by feedback *Biologically Motivated Computer Vision Workshop*. Springer-verlag, Tübingen, pp. 398-407.

Hartline, H.K. (1940). The receptive field of optic nerve fibers. *Am J Physiol*, **130**, 690-699.

Haxby, J.V., Hoffman, E.A. & Gobbini, M.I. (2000). The distributed human neural system for face perception. *Trends Cogn Sci*, **4**, 223-233.

Hendrickson, A.E., Wilson, J.R. & Ogren, M.P. (1978). The neuroanatomical organization of pathways between the dorsal lateral geniculate nucleus and visual cortex in Old World and New World primates. *J Comp Neurol*, **182**, 123-136.

Hess, C.W., Mills, K.R. & Murray, N.M. (1987). Responses in small hand muscles from magnetic stimulation of the human brain. *J Physiol*, **388**, 397-419.

Horton, J.C. & Hubel, D.H. (1981). Regular patchy distribution of cytochrome oxidase staining in primary visual cortex of macaque monkey. *Nature*, **292**, 762-764.

Hubel, D.H. (1985) *Eye, Brain and Vision*. W.H. Freeman & Company, New York.

Hubel, D.H. & Wiesel, T.N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J Physiol*, **148**, 574-591.

Hubel, D.H. & Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol*, **160**, 106-154.

Hubel, D.H. & Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J Physiol*, **195**, 215-243.

Huber, L. & Fagot-Joel, e. (1999). Generic perception: open-ended categorization of natural classes. Picture perception in animals. *Cahiers-de-psychologie-cognitive*, **18**, 845-887.

Huber, L., Troje, N., Loidolt, M., Aust, U. & Grass, D. (2000). Natural categorization through multiple feature learning in pigeons. *Quart J of Exp Psychol B: Comp Physiol Psychol*, **53**, 341-357.

Hummel, J.E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychol Rev*, **99**, 480-517.

Humphreys, G.W., Price, C.J. & Riddoch, M.J. (1999). From objects to names: a cognitive neuroscience approach. *Psychol Res*, **62**, 118-130.

Imbert, I. (2000). Comparison of visual cortex areas in the macaques and in humans. *Primatologie*, **2**.

Irvin, G.E., Norton, T.T., Sesma, M.A. & Casagrande, V.A. (1986). W-like response properties of interlaminar zone cells in the lateral geniculate nucleus of a primate (Galago crassicaudatus). *Brain Res*, **362**, 254-270.

Johnson, J.S. & Olshausen, B.A. (2003). Timecourse of neural signatures of object recognition. *J Vis*, **3**, 499-512.

Jolicoeur, P., Gluck, M.A. & Kosslyn, S.M. (1984). Pictures and names: Making the connection. *Cognitive Psychology*, **16**, 243-275.

Kanwisher, N. (2000). Domain specificity in face perception. *Nat Neurosci*, **3**, 759-763.

Kanwisher, N., McDermott, J. & Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*, **17**, 4302-4311.

Kanwisher, N., Tong, F. & Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition*, **68**, B1-11.

Kaplan, E. & Shapley, R.M. (1986). The primate retina contains two types of ganglion cells, with high and low contrast sensitivity. *Proc Natl Acad Sci U S A*, **83**, 2755-2757.

Kessels, R.P., Postma, A. & de Haan, E.H. (1999). P and M channel-specific interference in the what and where pathway. *Neuroreport*, **10**, 3765-3767.

Keysers, C. & Perrett, D.I. (2002). Visual masking and RSVP reveal neural competition. *Trends Cogn Sci*, **6**, 120-125.

Kirchner, H. & Thorpe, S. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Res*, **46**, 1762-1776.

Kreiman, G., Koch, C. & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat Neurosci*, **3**, 946-953.

Kreiter, A.K. & Singer, W. (1996). Stimulus-dependent synchronization of neuronal responses in the visual cortex of the awake macaque monkey. *J Neurosci*, **16**, 2381-2396.

Kuffler, S.W. (1953). Discharge patterns and functional organization of mammalian retina. *J Neurophysiol*, **16**, 37-68.

Lavenex, P., Suzuki, W.A. & Amaral, D.G. (2002). Perirhinal and parahippocampal cortices of the macaque monkey: Projections to the neocortex. *J Comp Neurol*, **447**, 394-420.

Lévi-Strauss, C. (1962) *La pensée sauvage*, Paris.

Liu, J., Harris, A. & Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. *Nat Neurosci*, **5**, 910-916.

Livingstone, M.S. & Hubel, D.H. (1982). Thalamic inputs to cytochrome oxidase-rich regions in monkey visual cortex. *Proc Natl Acad Sci U S A*, **79**, 6098-6101.

Livingstone, M.S. & Hubel, D.H. (1987). Psychophysical evidence for separate channels for the percetion of form, color, movement, and depth. *J of Neuroscience*, **7**, 3416-3468.

Livingstone, M.S. & Hubel, D.H. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, **240**, 740-749.

Logothetis, N.K. & Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cereb Cortex*, **5**, 270-288.

Macé, M.J.-M., Richard, G., Delorme, A. & Fabre-Thorpe, M. (2005). Rapid categorization of natural scenes in monkeys: target predictability and processing speed. *Neuroreport*, **16**, 349-354.

Macé, M.J.-M., Thorpe, S.J. & Fabre-Thorpe, M. (2005). Rapid categorization of achromatic natural scenes: how robust at very low contrasts? *Eur J Neurosci*, **21**, 2007-2018.

Mackworth, A.K. (1972) How to see a simple world: An exegis of some computer programs for scene analysis. In Elcock, E.W., Michie, D. (eds.) *Machine intelligence*. Wiley, New York, pp. 510-537.

Malpeli, J.G. & Baker, F.H. (1975). The representation of the visual field in the lateral geniculate nucleus of Macaca mulatta. *J Comp Neurol*, **161**, 569-594.

Markman, E.M., Horton, M.S. & McLanahan, A.G. (1980). Classes and collections: principles of organization in the learning of hierarchical relations. *Cognition*, **8**, 227-241.

Marr, D. (1982) *Vision*. Freeman.

Martin, P.R., White, A.J., Goodchild, A.K., Wilder, H.D. & Sefton, A.E. (1997). Evidence that blue-on cells are part of the third geniculocortical pathway in primates. *Eur J Neurosci*, **7**, 1536-1541.

Masson, G.S., Rybarczyk, Y., Castet, E. & Mestre, D.R. (2000). Temporal dynamics of motion integration for the initiation of tracking eye movements at ultra-short latencies. *Vis Neurosci*, **17**, 753-767.

Matsumoto, N., Okada, M., Sugase-Miyamoto, Y., Yamane, S. & Kawano, K. (2005). Population Dynamics of Face-responsive Neurons in the Inferior Temporal Cortex. *Cereb Cortex*, **15**, 1103-1112.

Maunsell, J.H. & Gibson, J.R. (1992). Visual response latencies in striate cortex of the macaque monkey. *J Neurophysiol*, **68**, 1332-1344.

Maunsell, J.H., Nealey, T.A. & DePriest, D.D. (1990). Magnocellular and parvocellular contributions to responses in the middle temporal visual area (MT) of the macaque monkey. *J Neurosci*, **10**, 3323-3334.

McCarthy, G., Puce, A., Gore, J.C. & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *J Cogn Neurosci*, **9**, 605-610.

Mel, B.W. (1997). SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comput*, **9**, 777-804.

Mishkin, M., Ungerleider, L.G. & Macko, K.A. (1983). Object vision and spatial vision: two cortical pathways. *Trends Neurosci*, 414-417.

Moore, T. & Armstrong, K.M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, **421**, 370-373.

Moravec, H. (1998). When will computer hardware match the human brain? *Journal of Evolution and Technology*, **1**.

Moscovitch, M., Winocur, G. & Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *J Cogn Neurosci*, **9**, 555-604.

Mouchetant-Rostaing, Y., Giard, M.H., Bentin, S., Aguera, P.E. & Pernier, J. (2000). Neurophysiological correlates of face gender processing in humans. *Eur J Neurosci*, **12**, 303-310.

Mouchetant-Rostaing, Y., Giard, M.H., Delpuech, C., Echallier, J.F. & Pernier, J. (2000). Early signs of visual categorization for biological and non-biological stimuli in humans. *Neuroreport*, **11**, 2521-2525.

Mumby, D.G. & Pinel, J.P. (1994). Rhinal cortex lesions and object recognition in rats. *Behav Neurosci*, **108**, 11-18.

Munk, M.H., Nowak, L.G., Girard, P., Chounlamountri, N. & Bullier, J. (1995). Visual latencies in cytochrome oxidase bands of macaque area V2. *Proc Natl Acad Sci U S A*, **92**, 988-992.

Murphy, G.L. & Wisniewski, E.J. (1989). Categorizing objects in isolation and in scenes: what a superordinate is good for. *J Exp Psychol Learn Mem Cogn*, **15**, 572-586.

Murray, E.A. & Richmond, B.J. (2001). Role of perirhinal cortex in object perception, memory, and associations. *Curr Opin Neurobiol*, **11**, 188-193.

Nakamura, H., Gattass, R., Desimone, R. & Ungerleider, L.G. (1993). The modular organization of projections from areas V1 and V2 to areas V4 and TEO in macaques. *J Neurosci*, **13**, 3681-3691.

Nealey, T.A. & Maunsell, J.H. (1994). Magnocellular and parvocellular contributions to the responses of neurons in macaque striate cortex. *J Neurosci*, **14**, 2069-2079.

Nelson, S.B. & LeVay, S. (1985). Topographic organization of the optic radiation of the cat. *J Comp Neurol*, **240**, 322-330.

Nowak, L.G. & Bullier, J. (1997) The timing of information transfer in the visual system. In Rockland, K.S., Kaas, J.H., Peters, A. (eds.) *Extrastriate visual cortex in primates*. Plenum Press, New York, pp. 205-241.

Nowak, L.G., James, A.C. & Bullier, J. (1997). Corticocortical connections between visual areas 17 and 18a of the rat studied in vitro: spatial and temporal organisation of functional synaptic responses. *Exp Brain Res*, **117**, 219-241.

Nowak, L.G., Munk, M.H.J., Girard, P. & Bullier, J. (1995). Visual latencies in areas V1 and V2 of the macaque monkey. *Visual Neuroscience*, **12**, 371-384.

Oliva, A. & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, **42**, 145-175.

Palmeri, T.J. & Gauthier, I. (2004). Visual object understanding. *Nat Rev Neurosci*, **5**, 291-303.

Perrett, D.I., Hietanen, J.K., Oram, M.W. & Benson, P.J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philos Trans R Soc Lond B Biol Sci*, **335**, 23-30.

Perrett, D.I., Oram, M.W. & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition*, **67**, 111-145.

Perrett, D.I., Oram, M.W., Harries, M.H., Bevan, R., Hietanen, J.K., Benson, P.J. & Thomas, S. (1991). Viewer-centred and object-centred coding of heads in the macaque temporal cortex. *Exp Brain Res*, **86**, 159-173.

Perrett, D.I., Rolls, E.T. & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Exp Brain Res*, **47**, 329-342.

Peyrin, C., Chauvin, A., Chokron, S. & Marendaz, C. (2003). Hemispheric specialization for spatial frequency processing in the analysis of natural scenes. *Brain Cogn*, **53**, 278-282.

Pieron, H. (1914). Recherches sur les lois de variation des temps de latence sensorielle et sur la loi qui relie ce temps à l'intensité de l'exitation. *L'ann. Psychol.*, **22**, 58-142.

Poggio, T. & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, **343**, 263-266.

Poggio, T. & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, **247**, 978-982.

Pohl, W. (1973). Dissociation of spatial discrimination deficits following frontal and parietal lesions in monkeys. *J Comp Physiol Psychol*, **82**, 227-239.

Potter, M.C. (1976). Short-term conceptual memory for pictures. *J Exp Psychol [Hum Learn]*, **2**, 509-522.

Premack, D. (1983). The codes of man and beasts. *Behav Brain Sci*, **6**, 125-167.

Puce, A., Allison, T., Asgari, M., Gore, J.C. & McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *J Neurosci*, **16**, 5205-5215.

Puce, A., Allison, T., Gore, J.C. & McCarthy, G. (1995). Face-sensitive regions in human extrastriate cortex studied by functional MRI. *J Neurophysiol*, **74**, 1192-1199.

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C. & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, **435**, 1102-1107.

Ratcliff, R. & Rouder, J.N. (1998). Modeling response times for two-choice decisions. *Psychol Sci*, **9**, 347-356.

Ratcliff, R. & Smith, P.L. (2004). A comparison of sequential sampling models for two-choice reaction

time. *Psychol Rev*, **111**, 333-367.

Richmond, B.J., Wurtz, R.H. & Sato, T. (1983). Visual responses of inferior temporal neurons in awake rhesus monkey. *J Neurophysiol*, **50**, 1415-1432.

Riesenhuber, M. & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat Neurosci*, **2**, 1019-1025.

Riesenhuber, M. & Poggio, T. (2002). Neural mechanisms of object recognition. *Curr Opin Neurobiol*, **12**, 162-168.

Rolls, E.T., Tovee, M.J. & Panzeri, S. (1999). The neurophysiology of backward visual masking: information analysis. *J Cogn Neurosci*, **11**, 300-311.

Rosch, E. (1973). Natural categories. *Cogn Psychol*, **4**, 328-350.

Rosch, E. (1978) Principles of categorization. In Rosch, E., Lloyds, B.B. (eds.) *Cognition and categorization*. Erlbaum, Hillsdale, NJ, pp. 27-48.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**, 382-439.

Roufs, J.A. (1974). Dynamic properties of vision. V. Perception lag and reaction time in relation to flicker and flash thresholds. *Vision Res*, **14**, 853-869.

Rousselet, G.A., Fabre-Thorpe, M. & Thorpe, S.J. (2002). Parallel processing in high-level categorization of natural images. *Nat Neurosci*, **5**, 629-630.

Rousselet, G.A., Joubert, O.R. & Fabre-Thorpe, M. (2005). How long to get to the “gist” of real-world natural scenes? *Visual Cogn*, **12**, 852-877.

Rousselet, G.A., Macé, M.J.-M. & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *J Vis*, **3**, 440-455.

Rousselet, G.A., Macé, M.J.-M., Thorpe, S. & Fabre Thorpe, M. (Submitted). Temporal course of ERP in fast object categori-zation in natural scenes: a story more compli-cated than expected? *J Vis*.

Rousselet, G.A., Thorpe, S.J. & Fabre-Thorpe, M. (2004). Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vision Res*, **44**, 877-894.

Rugg, M.D., Doyle, M.C. & Wells, T.J. (1995). Word and nonword repetition within-modality and across-modality: an event-relatedpotential study. *J Cogn Neurosci*, **7**, 209-227.

Sasaki, K., Gemba, H., Nambu, A. & Matsuzaki, R. (1993). No-go activity in the frontal association cortex of human subjects. *Neurosci Res*, **18**, 249-252.

Schendan, H.E., Ganis, G. & Kutas, M. (1998). Neurophysiological evidence for visual perceptual categorization of words and faces within 150 ms. *Psychophysiology*, **35**, 240-251.

Schiele, B. & Crowley, J.L. (1996). Object recognition using multimensional receptive field histograms. In Buxton, B., Cipolla, R. (eds.) *Proc. ECCV'96, Lecture Notes in Computer Science*. Berlin: Springer, pp. 610-619.

Schiller, P.H. & Malpeli, J.G. (1978). Functional specificity of lateral geniculate nucleus laminae of the rhesus monkey. *J Neurophysiol*, **41**, 788-797.

Schmolesky, M.T., Wang, Y., Hanes, D.P., Thompson, K.G., Leutgeb, S., Schall, J.D. & Leventhal, A.G. (1998). Signal timing across the macaque visual system. *J Neurophysiol*, **79**, 3272-3278.

Schyns, P.G. (1998). Diagnostic recognition: task constraints, object information, and their interactions. *Cognition*, **67**, 147-179.

Schyns, P.G. & Oliva, A. (1994). From blobs to boundary edges: Evidence for time and spatial scale dependent scene recognition. *Psychol Sci*.

Seeck, M., Michel, C.M., Mainwaring, N., Cosgrove, R., Blume, H., Ives, J., Landis, T. & Schomer, D.L. (1997). Evidence for rapid face recognition from human scalp and intracranial electrodes. *Neuroreport*, **8**, 2749-2754.

Sergent, J., Ohta, S. & MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain*, **115 Pt 1**, 15-36.

Shapley, R., Kaplan, E. & Soodak, R. (1981). Spatial summation and contrast sensitivity of X and Y cells in the lateral geniculate nucleus of the macaque. *Nature*, **292**, 543-545.

Shapley, R. & Perry, V.H. (1986). Cat and monkey retinal ganglion cells and their visual functional roles. *Trends Neurosci*, **May**, 229-235.

Shepard, R.N. & Chang, J.J. (1963). Stimulus generalization in the learning of classifications. *J Exp Psychol*, **65**, 94-102.

Sherman, S.M. (1985). Functional organization of the W-, X- and Y-cell pathways in the cat:a review and hypothesis. *Prog. Psychobiol. Physiol Psychol.*, **11**, 233-314.

Silveira, L.C. & Perry, V.H. (1991). The topography of magnocellular projecting ganglion cells (M-ganglion cells) in the primate retina. *Neuroscience*, **40**, 217-237.

Singer, W. & Gray, C.M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annu Rev Neurosci*, **18**, 555-586.

Strasburger, H. & Rentschler, I. (1996). Contrast-dependent dissociation of visual recognition and detection fields. *Eur J Neurosci*, **8**, 1787-1791.

Sugase, Y., Yamane, S., Ueno, S. & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, **400**, 869-873.

Suzuki, W.A. & Amaral, D.G. (1994). Perirhinal and parahippocampal cortices of the macaque monkey: cortical afferents. *J Comp Neurol*, **350**, 497-533.

Tallon-Baudry, C. & Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends Cogn Sci*, **3**, 151-162.

Tanaka, J.W. & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognit Psychol*, **23**, 457-482.

Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science*, **262**, 685-688.

Tanaka, K. (1997). Mechanisms of visual object recognition: monkey and human studies. *Curr Opin Neurobiol*, **7**, 523-529.

Tanaka, K. (2003). Columns for Complex Visual Object Features in the Inferotemporal Cortex: Clustering of Cells with Similar but Slightly Different Stimulus Selectivities. *Cereb Cortex*, **13**, 90-99.

Tarr, M.J., Williams, P., Hayward, W.G. & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nat Neurosci*, **1**, 275-277.

Thorpe, S., Fize, D. & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, **381**, 520-522.

Thorpe, S.J. (1990) Spike arrival times: A highly efficient coding scheme for neural networks. In R. Eckmiller, G.H.G.H. (ed.) *Parallel processing in neural systems and computers*. Elsevier, pp. 91-94.

Thorpe, S.J., Gegenfurtner, K. R., Fabre-Thorpe, M., Bulthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *Eur J Neurosci*, **14**, 869-876.

Thorpe, S.J. & Fabre-Thorpe, M. (2001). Neuroscience. Seeking categories in the brain. *Science*, **291**, 260-263.

Tolhurst, D.J. & Tadmor, Y. (2000). Discrimination of spectrally blended natural images: optimisation of the human visual system for encoding natural images. *Perception*, **29**, 1087-1100.

Tootell, R.B., Hamilton, S.L. & Switkes, E. (1988). Functional anatomy of macaque striate cortex. Contrast and magno-parvo streams. *J. Neurosci.*, **5**, 1594-1609.

Tovee, M.J. & Rolls, E.T. (1995). Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. *Visual Cogn*, **2**, 35-58.

Troje, N.F., Huber, L., Loidolt, M., Aust, U. & Fieder, M. (1999). Categorical learning in pigeons: the role of texture and shape in complex static stimuli. *Vision Res*, **39**, 353-366.

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition*, **67**, 21-44.

Ungerleider, L.G. (1995). Functional brain imaging studies of cortical mechanisms for memory. *Science*, **270**, 769-775.

Van Essen, D.C. (1979). Visual areas of the mammalian cerebral cortex. *Annu Rev Neurosci*, **2**, 227-

263.

Van Essen, D.C. & Zeki, S.M. (1978). The topographic organization of rhesus monkey prestriate cortex. *J Physiol*, **277**, 193-226.

Van Hoesen, G.W., Yeterian, E.H. & Lavizzo-Mourey, R. (1981). Widespread corticostriate projections from temporal cortex of the rhesus monkey. *J Comp Neurol*, **199**, 205-219.

Vanni, S., Warnking, J., Dojat, M., Delon-Martin, C., Bullier, J. & Segebarth, C. (2004). Sequence of pattern onset responses in the human visual areas: an fMRI constrained VEP source analysis. *Neuroimage*, **21**, 801-817.

VanRullen, R. (2000). Une première vague de potentiels d'action, une première vague idée de la scène visuelle. Paul Sabatier - Toulouse 3, Toulouse, pp. 217.

VanRullen, R. (2003). Visual saliency and spike timing in the ventral visual pathway. *J Physiol Paris*, **97**, 365-377.

VanRullen, R. & Thorpe, S.J. (2001). The time course of visual processing: from early perception to decision- making. *J Cogn Neurosci*, **13**, 454-461.

VanRullen, R. & Thorpe, S.J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, **30**, 655-668.

Vaughan, H.G., Jr., Costa, L.D. & Gilden, L. (1966). The functional relation of visual evoked response and reaction time to stimulus intensity. *Vision Res*, **6**, 645-656.

Vetter, T., Hurlbert, A. & Poggio, T. (1995). View-based models of 3D object recognition: invariance to imaging transformations. *Cereb Cortex*, **5**, 261-269.

Vidyasagar, T.R. (1999). A neuronal model of attentional spotlight: parietal guiding the temporal. *Brain Res Brain Res Rev*, **30**, 66-76.

Vidyasagar, T.R., Kulikowski, J.J., Lipnicki, D.M. & Dreher, B. (2002). Convergence of parvocellular and magnocellular information channels in the primary visual cortex of the macaque. *Eur J Neurosci*, **16**, 945-956.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *Eur J Neurosci*, **11**, 1239-1255.

Vogels, R. & Orban, G.A. (1996). Coding of stimulus invariances by inferior temporal neurons. *Prog Brain Res*, **112**, 195-211.

Wallis, G. & Rolls, E.T. (1997). Invariant face and object recognition in the visual system. *Prog Neurobiol*, **51**, 167-194.

Wilson, P.D., Rowe, M.H. & Stone, J. (1976). Properties of relay cells in cat's lateral geniculate nucleus: a comparison of W-cells with X- and Y-cells. *J Neurophysiol*, **39**, 1193-1209.

Wittgenstein, L. (1953) *Recherches philosophiques*. Gallimard, Paris.

Xiang, J.Z. & Brown, M.W. (1998). Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology*, **37**, 657-676.

Yukie, M. & Iwai, E. (1981). Direct projection from the dorsal lateral geniculate nucleus to the prestriate cortex in macaque monkeys. *J Comp Neurol*, **201**, 81-97.

Zeki, S. & Shipp, S. (1988). The functional logic of cortical connections. *Nature*, **335**, 311-317.

Zeki, S.M. (1978). Functional specialisation in the visual cortex of the rhesus monkey. *Nature*, **274**, 423-428.

Auteur : Marc Macé Directrice de thèse : Michèle Fabre-Thorpe

Résumé :

"Représentations visuelles précoces dans la catégorisation rapide de scènes naturelles chez l'homme et le singe"

Cette thèse porte sur le traitement rapide des scènes naturelles par les hommes et les singes. Elle est composée de trois chapitres, chacun abordant un aspect particulier de la construction des représentations visuelles précoces utilisées pour catégoriser rapidement les objets.
Nous montrons dans le premier chapitre que les informations magnocellulaires sont probablement très impliquées dans la construction des représentations visuelles précoces. Ces représentations rudimentaires de la scène visuelle pourraient servir à guider les traitements effectués sur les informations parvocellulaires accessibles plus tardivement.
Dans le deuxième chapitre, nous nous intéressons à la chronométrie des traitements visuels, en analysant les résultats de tâches conçues pour diminuer le temps de réaction des sujets ainsi que la latence de l'activité différentielle cérébrale. Nous étudions également la dynamique fine de ces traitements grâce à un protocole de masquage dans lequel l'information n'est accessible à l'écran que pendant une période de temps très courte et nous montrons ainsi toute l'importance des 20-40 premières millisecondes de traitement.
Le troisième chapitre traite de la nature des représentations visuelles précoces et des tâches qu'elles permettent de réaliser. Des expériences dans lesquelles les sujets doivent catégoriser des animaux à différents niveaux montrent que le premier niveau auquel le système visuel accède n'est pas le niveau de base mais le niveau superordonné. Ces résultats vont à l'encontre de l'architecture classiquement admise sur la base de travaux utilisant des processus lexicaux et met en évidence l'importance de facteurs comme l'expertise et la diagnosticité des indices visuels pour expliquer la vitesse d'accès aux différents niveaux de catégorie.
Ces différents résultats permettent de caractériser les représentations précoces que le système visuel utilise pour extraire le sens des informations qui lui parviennent et faire émerger la représentation interne du monde telle que nous la percevons.

Mots-clés : perception visuelle, catégorisation, paradigme go/no-go, scènes naturelles, représentations précoces, potentiels évoqués visuels, magno/parvo-cellulaire, comparaison homme/singe.

Summary:

"Early visual representations in rapid visual categorization of natural scenes in humans and monkeys"

Humans and monkeys can process objects in natural scenes very rapidly and the three chapters of this thesis try to define the different aspects of the early representations that underlie rapid visual categorization.
In the first chapter, we show that magnocellular information is probably highly involved in the construction of an early visual representation. The coarse description of the visual scene could be helpful to guide the processing of the delayed parvocellular information.
The second chapter deals with the chronometry of visual processing by analyzing the results of different types of tasks designed to minimize behavioral reaction times and the latency of differential ERP activity. We also use a masking protocol in which a mask is presented shortly after the image to decompose the precise dynamics of visual processing and this technique reveals the great importance of the first 20-40 ms of information processing.
The third chapter explores the nature of early visual representations and the type of tasks we can perform with them. Experiments in which subjects had to categorize animals at different levels show that the first level accessed is not at the basic but rather at the superordinate level. This finding contradicts previous results obtained with lexical experiments and strongly suggests that other components such as the subject's expertise and the diagnosticity of visual features can explain the processing speed when accessing to the different levels of categories.
These results allowed us to characterize the early representations that the visual system is using to extract the meaning of a scene and construct through an emerging process the internal representation of the world that is consciously perceived.

Key-words: visual perception, categorization, go/no-go, natural scenes, early representations, visual evoked potentials, magno/parvo-cellular, human/monkey comparison.