Extraction d'indices spatiaux et temporels dans des séquences vidéo couleur

Sébastien Lefèvre
LSIIT – Université Louis Pasteur (Strasbourg I)
Parc d'Innovation, Bd Brant, BP 10413, 67412 Illkirch Cedex
lefevre@lsiit.u-strasbg.fr

Nicole Vincent
CRIP5 – Université René Descartes (Paris V)
45 rue des Saints Pères, 75270 Paris Cedex 06
nicole.vincent@math-info.univ-paris5.fr

Résumé

Dans cet article, nous considérons les séquences vidéo couleur comme des données complexes. Notre contribution porte sur deux méthodes adaptées à ce type de données et dont l'objectif est d'extraire respectivement des indices spatiaux et temporels. Nous pensons que ces méthodes d'extraction d'indices peuvent être intégrées avec succès dans un processus plus complexe de fouille de données multimédia, aspect qui ne sera pas abordé ici. Les méthodes que nous présentons ici sont basées sur l'espace Teinte Saturation Luminance, réputé pour sa meilleure représentation de la vision humaine, mais caractérisé par des composantes particulières (une composante angulaire, la teinte, et deux composantes scalaires, la saturation et la luminance) qui nécessitent des méthodes de calcul spécifiques. L'extraction d'indices spatiaux est assimilée au problème de la séparation du fond et des objets pour lesquels nous proposons une approche multirésolution ne nécessitant qu'une seule image. L'extraction d'indices temporels correspond à la détection des changements de plans dans une séquence d'images, et la méthode que nous proposons pour l'obtenir se base sur l'utilisation de mesures de distances indépendantes du contexte. Les caractéristiques communes de nos deux méthodes (extraction d'indices spatiaux et temporels) sont l'utilisation de l'espace TSL, l'efficacité (temps de traitement respectant la cadence vidéo), et la robustesse (notamment aux changements d'illumination). Nous illustrons ces deux approches par des résultats obtenus sur des séquences vidéo sportives.

1 Introduction

Dans notre société de l'information et de la communication, les données numériques occupent une place de plus en plus importante et il devient nécessaire de disposer d'outils adaptés pour les traiter, les synthétiser, les fouiller. En particulier, les séquences vidéo issues des canaux télévisuels fournissent des volumes de données dont la taille ne permet plus aujourd'hui un parcours linéaire. Afin d'analyser ces données et d'en extraire les éléments pertinents, des méthodes permettant l'extraction d'indices doivent être élaborées.

Nous nous intéressons ici au problème de l'extraction d'indices dans les séquences vidéo. Puisque celles-ci sont le plus souvent composées d'images couleur, nous proposons d'utiliser l'espace couleur Teinte Saturation Luminance qui fournit des caractéristiques intéressantes. En se

basant sur cet espace, nous cherchons tout d'abord à extraire des indices spatiaux, que nous assimilons aux différents éléments contenus dans les images : les objets et l'arrière-plan de la scène. Puis nous nous focalisons sur l'extraction d'indices spatiaux représentant les limites des différents plans d'une séquence. Notre article sera donc organisé de la manière suivante : après avoir présenté l'espace TSL, nous décrirons nos méthodes d'extraction d'indices spatiaux et temporels, et enfin nous commenterons les résultats obtenus dans le contexte de séquences vidéo sportives.

2 L'espace Teinte Saturation Luminance

Le codage de la couleur dans des images numériques peut être effectué en utilisant différents espaces de représentation, appelés traditionnellement espaces couleur. Pour plus d'informations, le lecteur pourra se référer au récent ouvrage de Trémeau *et al.* [6] ou au panorama de Chang *et al.* [2] sur la segmentation couleur. Dans cet article, nous nous focalisons sur l'espace Teinte Saturation Luminance (TSL) que nous présentons ici.

L'espace TSL est représenté par trois composantes : la teinte, la saturation, et la luminance. Tandis que la saturation et la luminance sont codées "de manière classique" sous forme scalaire, la teinte est pour sa part une valeur angulaire. Ces composantes peuvent être interprétées de la manière suivante :

Teinte : représente la couleur perçue (rouge, jaune, vert, *etc.*),

Saturation : mesure la pureté de la couleur (par exemple pour une teinte rouge, le rose se caractérise par une saturation plus faible que le rouge, tandis que le noir, le blanc, et le gris sont caractérisés par une saturation nulle),

Luminance : représente le niveau de gris, de "sombre" pour une valeur faible à "clair" pour une valeur élevée.

L'espace TSL fournit, au travers de ses 3 composantes, des informations complémentaires. Il a l'avantage de permettre l'élaboration de méthodes robustes aux changements d'illumination. En effet, ces artefacts affectent principalement la composante de luminance. En ne tenant compte que des composantes de chrominance (teinte et saturation), il est donc possible de diminuer la sensibilité aux changements d'illumination. Nous avons également observé que la teinte était une composante plus robuste que la saturation ou la luminance dans un cadre multirésolution, où les données peuvent être analysées à une échelle plus ou moins fine. En effet, la teinte est moins sensible aux artefacts dus à des moyennages successifs des valeurs des pixels, nécessaires dans le cas d'une représentation multirésolution pyramidale. Nous avons ainsi observé une indépendance vis-à-vis de la résolution de certaines mesures calculées sur les teintes (comme la plage de valeur ou l'écart-type).

La teinte est donc une composante intéressante, invariante aux changements d'illuminations et aux cadres multirésolutions. Cependant elle doit être analysée avec précaution. En effet, sa fiabilité dépend du niveau de saturation et la teinte n'est significative que si la saturation est élevée. Les méthodes d'analyse basées sur la teinte des pixels doivent donc vérifier que ceux-ci ne sont pas achromatiques. Une autre contrainte de la teinte provient de sa nature mathématique (mesure angulaire) qui nécessite l'utilisation de mesures statistiques spécifiques. Ainsi, lorsqu'on cherche à mesurer la similarité entre deux valeurs, on est logiquement amené à calculer la différence absolue entre ces deux valeurs. Dans le cas de valeurs angulaires, la transposition est relativement triviale :

$$|\theta_i - \theta_j|_\angle = \min\left(|\theta_i - \theta_j|, 2\pi - |\theta_i - \theta_j|\right) \quad (1)$$

en notant $|\theta_i - \theta_j|_\angle$ la différence absolue de deux valeurs angulaires θ_i et θ_j. En cas de fusion ou de combinaison d'informations, la moyenne est fréquemment utilisée en analyse d'images. Nous

avons choisi d'utiliser la définition donnée dans [5] pour le calcul de la moyenne $\bar{\theta}$ d'un ensemble de mesures d'angles $\{\theta_i\}_{i\in[1,\Theta]}$. Elle s'obtient par la formule :

$$\bar{\theta} = \begin{cases} \text{Arctan}\left(\frac{\sum_{i=1}^{\Theta}\sin(\theta_i)}{\sum_{i=1}^{\Theta}\cos(\theta_i)}\right) & \text{si } \sum_{i=1}^{\Theta}\cos(\theta_i) \geq 0, \\ \text{Arctan}\left(\frac{\sum_{i=1}^{\Theta}\sin(\theta_i)}{\sum_{i=1}^{\Theta}\cos(\theta_i)}\right) + \pi & \text{sinon} \end{cases} \tag{2}$$

Dans [5] est donné également un algorithme permettant le calcul de l'amplitude de variation de l'ensemble $\{\theta_i\}_{i\in[1,\Theta]}$. Cette méthode nécessite un tri croissant préalable de tous les angles considérés, et est donc caractérisée par une complexité algorithmique relativement élevée. Nous proposons ici une méthode applicable dans le cas où l'angle moyen $\bar{\theta}$ a déjà été calculé. Cette méthode nécessite, en plus de la moyenne, la connaissance des angles minimum et maximum dans l'intervalle de longueur 2π choisi, respectivement notés θ_{min} et θ_{max}. Si la moyenne appartient à l'intervalle limité par les deux angles, c'est-à-dire si $\theta_{\text{min}} \leq \bar{\theta} \leq \theta_{\text{max}}$, alors l'amplitude de variation est égale à $\theta_{\text{max}} - \theta_{\text{min}}$. Dans le cas contraire, l'angle complémentaire doit être considéré et l'amplitude de variation est égale à $2\pi - (\theta_{\text{max}} - \theta_{\text{min}})$.

Par la suite, nous utiliserons les différentes définitions introduites ici pour mesurer et comparer des valeurs angulaires, la teinte dans notre contexte. Nous avons décrit ici l'espace TSL et ses caractéristiques. Nous allons maintenant décrire deux méthodes d'extraction d'indices spatiaux et temporels, et basés sur cet espace.

3 Extraction d'indices spatiaux

Nous avons précédemment décrit l'espace TSL qui nous semble opportun pour la fouille de séquences vidéo couleur. En se basant sur cet espace, nous proposons d'extraire des indices spatiaux à partir des différentes trames des séquences vidéo.

Au sein d'une image, il est possible d'extraire différents types d'indices au cours d'un processus de fouille de données. Nous avons choisi de considérer comme indices les régions (position, taille, *etc*) appartenant soit à des objets soit à l'arrière-plan de la scène. Ce choix nous semble opportun dans la mesure où la séparation fond/objets est cruciale dans de nombreuses applications, telles que le suivi d'objet, l'interprétation du contenu des images et des séquences d'images, ou encore la compression. En effet, la norme de compression MPEG-4 décrit une scène par les différents objets qui la composent et par son arrière-plan [3].

Afin d'extraire des indices relatifs aux objets ou à l'arrière-plan d'une scène, il est généralement nécessaire de disposer de plusieurs images et de les comparer : ainsi, les objets sont délimités par les zones de l'image dont le contenu a évolué au cours du temps. Si une image de référence (sans objet) est disponible, les objets correspondent aux zones de l'image analysée qui diffèrent de l'image de référence. Ce principe n'est malheureusement valable que si la caméra est statique. Dans le cas d'une caméra mobile, une étape coûteuse d'estimation/compensation de mouvement est nécessaire : le temps de calcul du processus de fouille croît alors considérablement. Nous présentons donc ici une méthode d'extraction d'indices spatiaux relatifs aux éléments d'une scène (objets et arrière-plan) qui ne requiert pas d'information de mouvement et qui peut être appliquée sur chaque image de manière indépendante. Notre méthode est adaptée aux scènes où l'arrière-plan occupe une partie du champ de vision plus importante que les objets (statiques ou en mouvement). Elle est basée sur le constat suivant : au fur et à mesure que la résolution d'une image diminue,

les détails disparaîssent et le contenu de l'image tend à représenter exclusivement l'arrière-plan au détriment des objets présents dans la scène.

Cette réflexion nous amène à proposer une approche multirésolution : en considérant une image originale $\underset{0}{I}$, il est possible de diminuer fortement sa résolution pour obtenir une image à très faible résolution $\underset{r_{\max}}{I}$. Aucun objet n'est plus perceptible dans cette image qui n'est composée que de l'arrière-plan. Un modèle du fond peut donc être calculé à partir de cette image. En augmentant la résolution r de manière itérative, il est alors possible de comparer les différentes régions de l'image $\underset{r}{I}$ avec le modèle de l'arrière-plan suivant un critère donné. Cette comparaison permet de déterminer si chaque région correspond ou non à l'arrière-plan. De plus, le principe de multirésolution permet d'analyser des données à un degré de précision plus ou moins fin selon le résultat attendu. Une analyse multirésolution a aussi généralement l'avantage de limiter le nombre de calculs nécessaires par rapport à son homologue monorésolution. L'inconvénient principal est lié à la difficulté du choix des résolutions initiale et finale de l'analyse.

Le traitement décrit précédemment s'effectue donc en quatre étapes successives : création de la représentation multirésolution, estimation du modèle de l'arrière-plan, segmentation itérative aux différentes résolutions, et enfin segmentation finale et obtention des indices relatifs aux objets et à l'arrière-plan. La première étape consiste en la création de la représentation multirésolution de l'image, par décomposition pyramidale, où la valeur d'un pixel $P(x, y)$ à la résolution $r + 1$ est calculée comme la moyenne d'un ensemble de p^2 pixels à la résolution r. La taille de l'image dépend alors de la résolution. La moyenne a été préférée à d'autres mesures nécessitant des calculs plus importants, comme par exemple la valeur médiane. Le calcul est effectué itérativement, à partir de la résolution originale $r = 0$, et jusqu'à obtenir la résolution voulue $r = r_{\max}$.

Il est alors possible de modéliser l'arrière-plan à l'aide de l'image $\underset{r_{\max}}{I}$, sommet de la pyramide. Cette considération n'est bien sûr valable que si l'arrière-plan occupe une partie importante de l'image, s'il diffère suffisamment des objets présents dans la scène, et s'il est assez homogène. Afin de garantir une faible complexité tout en assurant une robustesse aux changements d'illuminations, nous avons décidé d'utiliser comme modèle la moyenne des teintes d'une région, puisque la teinte nous semblait une composante relativement robuste aux moyennages successifs nécessaires pour construire la représentation multirésolution de l'image. Le modèle de l'image I est noté $\varphi(I)$ et l'arrière-plan sera donc caractérisé par $\varphi\left(\underset{r_{\max}}{I}\right)$. Ce choix requiert la validité de deux hypothèses : il est nécessaire que les pixels ne soient pas achromatiques (saturation non nulle) pour que la valeur de leur teinte soit fiable, et l'arrière-plan doit être de couleur (ou plutôt de teinte) homogène.

La caractérisation de chaque région (et donc la génération des indices correspondants) peut ensuite être effectuée et améliorée de manière itérative (à la manière d'un *quadtree* incomplet) depuis la résolution $r_{\max} - 1$ jusqu'à la résolution initiale $r = 0$, base de la pyramide. A une résolution donnée r' avec $r_{\max} > r' > 0$, l'image $\underset{r'}{I}$ doit être analysée. Cette image est comparable à l'image initiale $\underset{0}{I}$ qui aurait été découpée en $K = (K_0)^{r_{\max} - r'}$ régions, avec K_0 une constante dont la valeur pourrait être en toute logique égale à p^2. Chacune de ces régions est alors comparée avec le modèle de l'arrière-plan. Cet appariement entre deux régions I^m et I^n s'effectue en calculant sur les moyennes respectives des teintes la mesure de similarité $\delta(\varphi(I^m), \varphi(I^n) = |\varphi(I^m) - \varphi(I^n)|_{\angle}$. Une fois cette mesure δ calculée entre le modèle d'une région donnée et le modèle de l'arrière-plan, elle est comparée à un seuil S_1 afin d'effectuer ou non la reconnaissance. Une valeur inférieure au seuil signifie que la région est considérée comme l'arrière-plan.

Cependant, un second test est nécessaire pour vérifier la cohérence de la région étudiée et éviter les artefacts liés à l'utilisation de la moyenne. En effet, si une région est composée de deux pixels ayant des teintes opposées, la moyenne ne reflétera pas correctement le contenu de

la région. Nous analysons donc la cohérence de chaque région appariée avec l'arrière-plan. Nous avons préféré l'amplitude de variation à d'autres mesures de dispersion telles que la variance pour évaluer ici la cohérence d'une région : plus la plage de valeurs d'une région est faible, plus celle-ci est homogène. Pour calculer cette plage de valeurs angulaires, nous n'utilisons pas la méthode donnée dans [5] mais l'approche originale présentée précédemment. Une fois la plage de valeurs calculée pour une région $\underset{r}{I^k}$, nous comparons cette mesure de dispersion avec un second seuil S_2. Une plage inférieure au seuil assure l'homogénéité de la région concernée. Celle-ci est alors étiquetée en fond ou arrière-plan. Dans le cas contraire, l'hétérogénéité de la région candidate à l'étiquetage implique son rejet.

Si une région fournit une réponse positive à ces deux tests sucessifs, elle est affectée à l'arrière-plan. Dans ce cas la région n'est plus analysée à de meilleures résolutions. A l'opposé, une région sans étiquette sera analysée plus en détail à la résolution $r' - 1$. Cette segmentation est effectuée si nécessaire jusqu'à la résolution initiale $r = 0$. Dans le cas d'applications nécessitant une segmentation et des indices très précis, les régions correspondant aux objets peuvent être analysées par la suite afin d'affiner les contours des objets. Au contraire, si la précision des contours des objets n'est pas nécessaire, le processus peut être arrêté à une résolution r_{final} avec $r_{\text{max}} > r_{\text{final}} \geq 0$. Dans ce cas, nous considérons que les régions sans étiquette représentent les objets.

Afin d'améliorer la qualité du modèle de l'arrière-plan, il est possible de recalculer celui-ci au cours du processus de segmentation. Dans ce cas, le modèle obtenu à la résolution r_{max} ne représente que l'état initial de l'arrière-plan. A mesure que la résolution devient plus fine, les résultats sont plus précis, et il est possible d'obtenir un modèle de l'arrière-plan plus fiable en ne considérant que les parties de l'image déjà étiquetées comme le fond.

Nous avons décrit ici une méthode d'extraction d'indices spatiaux dans des séquences vidéo couleur utilisant l'espace TSL dans un cadre multirésolution. Nous allons montrer maintenant que cet espace peut également être employé pour fournir des indices temporels.

4 Extraction d'indices temporels

Après avoir montré comment les séquences vidéo couleur pouvaient être analysé dans l'espace TSL pour fournir des indices de nature spatiale, nous allons étudier maintenant l'extraction d'indices temporels. Plus précisément, les indices que nous cherchons à extraire sont les frontières des différents plans de la séquence (ou changements de plans). Nous rappelerons briévement les différents types de changements de plans et présenterons ensuite notre approche en détaillant le nécessaire prétraitement des données, la définition de la mesure de distance utilisée, et enfin la solution globale proposée.

Un plan est défini comme une suite d'images issues d'une acquisition continue d'une caméra donnée. Ainsi, toutes les images d'un plan ont été acquises avec la même caméra. Le plan est souvent l'unité temporelle la plus petite pour une séquence vidéo si l'on ne prend pas en compte l'image pour laquelle la notion de temps a disparu. Chaque plan est séparé du précédent et du suivant par une transition, qui peut être brusque ou progressive. Lors d'une transition brusque (appelée *cut*), la dernière image du premier plan est directement suivie par la première image du second plan. Dans le cas où les deux plans sont connectés en utilisant un effet particulier, on parle de transition progressive : fondu, volet, *etc*. On distingue le fondu du noir vers un plan, d'un plan vers le noir, ou d'un plan vers un autre plan. Au cours d'un fondu, le niveau de chaque pixel des images intermédiaires (appartenant à la transition progressive) est calculé en fonction des niveaux des pixels de la dernière image du premier plan et de la première image du second plan. La proportion varie au cours de la transition de 0 à 1 pour la première image du second plan et de 1 à

0 pour la dernière image du premier plan. Lors d'un volet, chaque pixel des images intermédiaires a un niveau égal à celui du pixel de mêmes coordonnées spatiales soit dans la dernière image du premier plan soit dans la première image du second plan. Les images appartenant à un volet vont donc contenir de plus en plus de pixels extraits de la première image du second plan et de moins en moins de pixels extraits de la dernière image du premier plan.

La plupart des méthodes proposées pour résoudre le problème de la détection des changements de plans fonctionnent en deux étapes : le calcul d'une mesure de dissimilarité entre deux trames successives d'une séquence vidéo, puis la comparaison de la valeur obtenue avec un seuil, afin de déterminer ou non la présence d'un changement de plans. Suivant ce principe, la détection d'un changement de plans est effective si la condition $d(I_t, I_{t-1}) > S$ est respectée. Nous rappelons que I_t représente l'image de la séquence vidéo obtenue à l'instant t, d une distance, et S un seuil. On trouvera dans [4] un panoramas approfondi des méthodes adaptées aux données non-compressées.

Afin de garantir un temps de calcul relativement faible et d'assurer une certaine robustesse au bruit, nous proposons d'introduire un prétraitement des données. Celui-ci consistera à diminuer la résolution spatiale et sera obtenu par l'approche multirésolution décrite dans la section précédente, en considérant des blocs de 8×8 pixels (choix qui nous permet de traiter également des données compressées JPEG ou MPEG à l'aide des coefficents DC). De manière à accroître la robustesse aux changements d'illumination et aussi à réduire les temps de calcul nous avons choisi de limiter la représentation des pixels à un espace de dimension 2 composé de la teinte et de la saturation. Pour toutes les scènes d'extérieur qui sont fréquentes dans les séquences vidéo, on peut noter une amélioration par rapport aux résultats obtenus dans l'espace RVB.

Comme [1], nous mesurons la différence entre deux images dans le sous-espace TS. Cette différence est obtenue par la distance d définie comme :

$$d_k(I_{t_1}, I_{t_2}) = \sum_{x=1}^{X} \sum_{y=1}^{Y} I_{t_1}(x,y) \ominus I_{t_2}(x,y) \tag{3}$$

avec $\ominus$ un opérateur algébrique utilisé pour comparer deux pixels et défini par :

$$\begin{aligned} I_{t_1}(x,y) \ominus I_{t_2}(x,y) = \alpha_{T,S}(I_{t_1}(x,y,T) - I_{t_2}(x,y,T)) \pmod{2\pi} \\ + (1 - \alpha_{T,S}) \left| I_{t_1}(x,y,S) - I_{t_2}(x,y,S) \right| \end{aligned} \tag{4}$$

où $\alpha_{T,S}$ est un coefficient permettant de donner plus ou moins d'influence aux composantes T et S. En effet, dans le cas de pixels achromatiques, il est important de ne pas donner d'importance à la composante T qui n'est alors pas fiable.

La mesure de distance d, quoique relativement simple, permet d'estimer correctement la différence entre deux images en assurant une invariance à l'illumination. L'utilisation d'une mesure de distance plus complexe pourrait apporter une quantité d'information supplémentaire mais engendrerait également un surcoût en terme de temps de calcul. Cependant, l'utilisation directe de cette mesure de dissimilarité entre deux images successives nécessiterait la comparaison à un seuil S. Le seuil utilisé doit souvent être fixé de manière empirique, et dépend du domaine vidéo étudié (sport, bulletin d'informations, *etc.*) ou du type de plans présents dans la séquence. Ainsi, un plan éloigné, où les objets en mouvement sont petits, sera caractérisé par une valeur d relativement faible tandis qu'un plan proche ou serré, où les objets en mouvement occupent une portion importante de l'image, sera caractérisé par une valeur d plus élevée. Le seuil S devra donc être ajusté en conséquence afin d'éviter les fausses détections ou l'absence de détection. Comme certaines séquences vidéo, notamment les retransmissions télévisées d'événements sportifs, contiennent des plans éloignés et des plans proches, il est nécessaire d'utiliser une méthode plus générale qui puisse s'adapter à ces différents types de plans. Nous proposons donc d'introduire un seuil adaptatif, noté S_d, qui dépend du temps. La valeur du seuil est mise à jour pour chaque nouvelle image

de la séquence, soit $S_d(t) = \alpha_{S_d} S_d(t-1) + (1-\alpha_{S_d}) d(I_t, I_{t-1})$ où $S_d(t)$ représente la valeur du seuil S_d à l'instant t. De cette façon, il s'adapte automatiquement avec une certaine inertie (représentée par le coefficient α_{S_d}) au contenu de la vidéo étudiée, sa valeur étant modifiée en fonction des valeurs de $d(I_t, I_{t-1})$. Les mesures de précision et de rappel dépendront évidemment de α_{S_d}.

L'utilisation directe d'une mesure d entre deux images successives est très sensible au bruit et au mouvement présent dans la séquence étudiée. L'introduction d'un seuil adaptatif permet de limiter cette sensibilité dans une certaine mesure, mais évidemment pas dans sa totalité. Nous proposons donc de considérer une mesure relative et non une mesure absolue. Cette mesure relative, notée d', permet d'accroître la robustesse au bruit et aux mouvements importants présents dans la séquence, et est définie par $d'(I_t) = |d(I_t, I_{t-1}) - d(I_{t-1}, I_{t-2})|$. Contrairement à la mesure d, la mesure d' est définie de manière relative et son ordre de grandeur dépend donc moins du type de plan ou de vidéo étudié. Afin de détecter un changement de plans, cette mesure peut donc être comparée à un seuil $S_{d'}$ fixé empiriquement au début de la séquence. La valeur de $S_{d'}$ pourra évoluer en fonction du type de vidéo ou de plan analysé.

Comme il a été précisé précédemment, un changement de plans peut être brusque ou progressif. En considérant *une transition progressive comme une transition brusque dont les effets sont étalés sur plusieurs images*, nous proposons ici une approche permettant de détecter les transitions brusques ou progressives de manière relativement similaire. Pour détecter un changement brusque, nous comparons directement la valeur d' avec un seuil $S_{d'}$. En effet, si la valeur de d' est élevée, c'est-à-dire si la différence absolue entre $d(I_t, I_{t-1})$ et $d(I_{t-1}, I_{t-2})$ est significative, alors l'évolution du contenu de la séquence entre les images I_{t-2} et I_{t-1} n'est pas cohérente avec celle observée entre les images I_{t-1} et I_t. Un changement brusque existe donc à l'instant $t-1$.

Si aucun changement brusque n'a été détecté, il est encore possible de se trouver en présence d'une transition progressive. La valeur d' ne peut être utilisée directement dans le cas d'une transition progressive puisqu'elle ne reflète l'évolution de la mesure de distance d qu'à un instant donné. Il est donc nécessaire de cumuler les valeurs d' obtenues pour toutes les trames composant une transition progressive afin d'obtenir une mesure qui sera du même ordre de grandeur que la valeur du seuil considéré dans le cas d'une transition brusque. La détection des transitions progressives s'effectue donc en deux étapes successives. La première étape consiste en la détection des trames susceptibles d'être les frontières d'une transition progressive. Pour détecter celles-ci, nous analysons l'évolution de la mesure de distance d et nous comparons à chaque instant t la valeur $d(I_t, I_{t-1})$ avec le seuil adaptatif $S_d(t)$ défini précédemment. L'utilisation de ce seuil adaptatif nous permet de gérer tout type de situation (plan proche ou éloigné, mouvement important ou pas, *etc.*). Si la condition $d(I_{t_1}, I_{t_1-1}) > S_d(t_1)$ est vérifiée, alors il est possible qu'une transition soit présente dans la séquence à partir de l'instant t_1. L'instant t_2 de fin de cette transition correspondrait à la première trame vérifiant la condition $d(I_{t_2}, I_{t_2-1}) < S_d(t_2)$ avec $t_2 > t_1$. Une fois les frontières t_1 et t_2 d'une possible transition déterminées, il est nécessaire d'analyser les trames t de cet intervalle de temps. Pour cela, nous calculons la valeur cumulée de d' sur l'ensemble des trames $[t_1, t_2]$ notée $d'_{\text{cumul}}(t_1, t_2)$, soit $d'_{\text{cumul}}(t_1, t_2) = \sum_{t=t_1}^{t_2} d'(I_t)$. La comparaison de $d'_{\text{cumul}}(t_1, t_2)$ avec le seuil $S_{d'}$ permet alors de valider ou non la présence d'un changement entre les trames I_{t_1} et I_{t_2}.

Les deux méthodes présentées ici ont été testées dans le contexte de l'analyse de séquences vidéo de football. Les résultats obtenus vont maintenant être présentés.

5 Résultats

Afin d'illustrer l'intérêt de l'espace TSL dans les problèmes de segmentation, nous avons décrit précédemment deux méthodes basées sur cet espace couleur et permettant respectivement la segmentation spatiale et la segmentation temporelle de données image. Nous présentons ici quelques résultats obtenus avec ces méthodes, nous permettant ainsi d'illustrer l'intérêt de l'espace TSL et les capacités des méthodes proposées. Le contexte considéré est celui de l'analyse temps-réel de séquences vidéo de match de football. Les images analysées représentent des scènes d'exterieur où l'illumination n'est pas constante. Les mesures de temps de calcul ont été obtenues en considérant une architecture PC Pentium 4 candencé 1700 MHz.

5.1 Indices spatiaux

La méthode d'extraction des indices spatiaux a pour but de séparer les objets et le fond. Dans le contexte proposé, l'objectif est d'identifier les joueurs par rapport au terrain de football. La figure 1 montre les résultats obtenus sur deux images différentes où les tailles des objets varient considérablement. Les objets sont correctement détectés, indépendamment de l'aire qu'ils occupent dans l'image. Les temps de calcul observés dépendent de la taille de l'image : de 30 millisecondes pour une image de taille 192×128 pixels à 350 millisecondes pour une image de taille 704×576 pixels. La segmentation a été obtenue en utilisant les seuils $S_1 = \pi/18$ et $S_2 = \pi/4$. La structure de la pyramide est régulière, nous avons donc p^2 et K_0 égaux à 4. Les résolutions ont été fixées à $r_{\text{max}} = 7$ et $r_{\text{final}} = 5$.

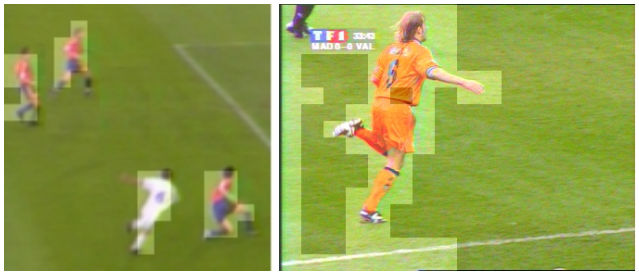

FIG. 1 – Résultats obtenus en considérant des objets de différentes tailles. Les régions étiquetées comme "objet" sont représentées plus claires que le reste de l'image.

Le processus de segmentation est itératif. La figure 2 illustre la diminution progressive de la résolution pour obtenir le modèle initial de l'arrière-plan, ainsi que le résultat de la segmentation aux différentes résolutions. On peut observer que le choix de la résolution finale r_{final} influe directement sur la précision du résultat. Cependant, même en considérant une résolution finale similaire à la résolution originale (*i.e.* $r_{\text{final}} = 0$), les contours des objets détectés resteront grossiers et parallèles aux côtés de l'image. En supposant que la séparation des objets et du fond n'est qu'une étape intermédiaire, la précision obtenue est néanmoins suffisante.

Nous avons également observé que l'utilisation de la teinte fournit un résultat de meilleure qualité que l'espace RVB. Aucune des composantes R, V ou B ne contient l'information suffisante pour atteindre des résultats aussi précis qu'avec la teinte (l'information pertinente étant dispersée dans les trois canaux de base). De plus, nous avons évalué la robustesse de la composante couleur utilisée au réglage des paramètres. Là encore, la teinte fournit les résultats les plus intéressants et les plus robustes.

Les principales limites de la méthode proposée ont été identifiées *a priori* dans la section 3. D'une part, les pixels ne doivent pas être achromatiques (puisque seule la teinte est utilisée dans le

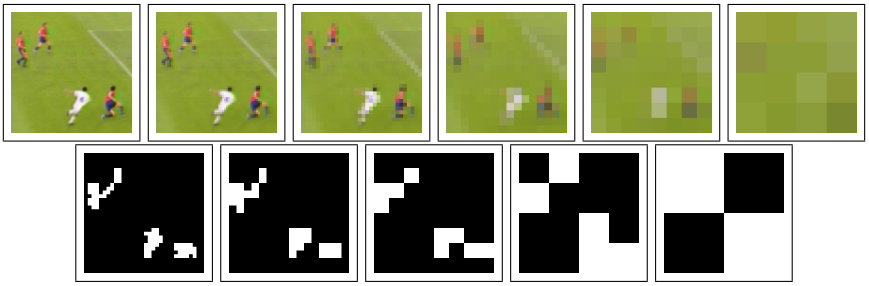

FIG. 2 – En haut : représentation d'une image à différentes résolutions, de la résolution originale $r = 0$ (à gauche) à $r = r_{\text{max}} = 5$ (à droite). En bas : résultat obtenu aux différentes résolutions, de $r_{\text{final}} = r_{\text{max}} - 1 = 4$ (à droite) à $r_{\text{final}} = 0$ (à gauche).

processus de segmentation), et d'autre part l'arrière-plan doit être relativement uniforme (puisqu'il est modélisé ensuite par une moyenne).

5.2 Indices temporels

La méthode proposée ici a été testée sur des séquences vidéo de différents domaines. Afin de souligner ses capacités, nous avons choisi de limiter notre présentation à des résultats obtenus à partir de retransmissions de matchs de football et considérons ce contexte comme plus difficile que celui des journaux télévisés par exemple. En effet, les séquences analysées sont caractérisées par des plans proches ou éloignés, un grand nombre d'effets différents utilisés, une variation importante des mouvements de la caméra et des objets présents dans la scène, des changements d'illumination fréquents liés au système d'éclairage, et enfin une certaine homogénéité des images de la séquence vidéo (même si elles appartiennent à différents plans). De plus, un ensemble de tests nous a permis de vérifier expérimentalement que les résultats obtenus avec cette méthode étaient meilleurs que ceux obtenus avec d'autres méthodes de la littérature.

Après l'étape de réduction, les images à traiter contiennent 20×15 pixels. Les séquences vidéo contiennent différents effets (cut, volet, fondu, mais aussi certains effets combinant volet et fondu). Elles incluent aussi du bruit dû au mouvement global de la caméra, aux objets en mouvement, ainsi qu'aux effets d'illumination. Pour la taille d'image considérée (160×120 pixels), le temps de calcul nécessaire est égal à 4 millisecondes par image en considérant la même architecture matérielle que précédemment.

La figure 3 montre l'évolution des mesures d, d', et d'_{cumul} pour des séquences vidéo contenant différents types de transition. Les paramètres utilisés sont $\alpha_{T,S} = 0.3$ et $\alpha_{S_d} = 0.25$. Nous avons constaté que le contraste des valeurs au niveau des extremums locaux est beaucoup plus marqué dans l'espace TSL que dans l'espace RVB. Ce choix d'espace assure donc à la méthode proposée une plus grande robustesse et confirme notre hypothèse théorique de départ.

La qualité d'une méthode de détection de changements de plans peut être évaluée grâce aux mesures de rappel et de précision, qui tiennent compte respectivement des détections manquées et des fausses détections. Des tests effectués sur une vingtaine de séquences composées chacune de 500 images et contenant chacune entre un et trois changements de plans (brusques ou progressifs) ont permis d'obtenir des mesures de rappel et de précision respectivement égales à 80 % (100 % dans le cas de cuts) et 100 %.

La principale limite de l'approche proposée ici est sa sensibilité au mouvement présent dans la séquence. Ce mouvement apparent peut être provoqué par une accéleration brusque de la caméra ou par le mouvement d'un objet occupant quasiment toute l'image dans un plan rapproché.

FIG. 3 – Evolution temporelle des mesures d (en vert), d' (en rouge), et d'_{cumul} (en bleu) pour une séquence contenant des transitions brusques (à gauche) et des transitions progressives (à droite) indiquées par les flèches.

6 Conclusion

De nos jours, les données complexes telles que des séquences vidéo sont la plupart du temps représentées en couleur. Cependant, l'espace de représentation généralement considéré est l'espace Rouge Vert Bleu prévu pour l'affichage des images à l'écran. Nous avons montré ici comment un autre espace de représentation de la couleur, l'espace Teinte Saturation Luminance, pouvait apporter une amélioration en analyse d'images dans l'optique d'un processus de fouille. Pour celà nous avons cherché à extraire deux types d'indices, spatiaux et temporels, à l'aide de l'espace TSL. Tandis que l'extraction des indices spatiaux ne nécessite qu'une seule image grâce à un cadre d'analyse multirésolution, l'obtention des indices temporels utilise une méthode s'adaptant au contenu, notamment par l'usage d'une mesure de distance intertrames différentielle. Notre contribution a donc porté sur ces trois aspects : caractéristiques et calculs dans l'espace TSL, extraction d'indices spatiaux par séparation des objets et du fond, extraction d'indices temporels par détection des changements de plans.

Outre la validation de nos approches par leur intégration à un processus de fouille, les perspectives des travaux présentés ici sont de deux ordres. D'une part, il nous paraît important d'approfondir les résultats présentés, en continuant d'identifier les caractéristiques de l'espace TSL et de proposer des modes de calcul appropriés à cet espace. D'autre part, nous souhaitons atténuer les limites des méthodes proposées : meilleure prise en compte des données achromatiques, et considération de scènes au contenu plus complexe.

Références

[1] T. CARRON, *Segmentations d'images couleur dans la base Teinte-Luminance-Saturation : approche numérique et symbolique*, thèse de doctorat, Université de Savoie, Décembre 1995.

[2] H. CHANG, X. JIANG, Y. SUN, AND J. WANG, Décembre 2001, *Color image segmentation : Advances and prospects*, Pattern Recognition, 34, pp. 2259–2281.

[3] B. HASKELL ET AL., Novembre 1998, *Image and video coding — emerging standards and beyond*, IEEE Transactions on Circuits and Systems for Video Technology, 8, pp. 814–837.

[4] S. LEFÈVRE, J. HOLLER, AND N. VINCENT, (2003), *A study of real-time segmentation of uncompressed video sequences for content-based search and retrieval*, Real-Time Imaging, 9, pp. 73–98.

[5] K. MARDIA AND P. JUPP, *Directional Statistics*, Wiley & Sons Ltd., (2000).

[6] A. TRÉMEAU, C. FERNANDEZ-MALOIGNE, AND P. BONTON, *Imagerie Numérique Couleur*, Dunod, (2004).