# A human visual model-based approach of the visual attention and performance evaluation

O. Le Meur[1,2] , D. Barba[2] , P. Le Callet[2], D. Thoreau[1]
[1] THOMSON R&D France, 1 Avenue de Belle Fontaine, 35511 Cesson-Sévigné cedex, FRANCE.
[2] IRCCyN UMR n°6597 CNRS, Ecole Polytechnique de l'Université de Nantes, rue Christian Pauc, La chantrerie, BP50609, 44306 Nantes cedex, FRANCE.

## ABSTRACT

In this paper, a coherent computational model of visual selective attention for color pictures is described its performances are precisely evaluated . The model based on some important behaviours of the human visual system is composed of four parts: visibility, perception, perceptual grouping and saliency map construction. This paper focuses mainly on its performances assessment by achieving extended subjective and objective comparisons with real fixation points captured by an eye-tracking system used by the observers in a task-free viewing mode. From the knowledge of the ground truth, qualitatively and quantitatively comparisons have been made in terms of the measurement of the linear correlation coefficient (CC) and of the Kulback Liebler divergence (KL). On a set of 10 natural color images, the results show that the linear correlation coefficient and the Kullback-Leibler divergence are of about 0.71 and 0.46, respectively. CC and Kl measures with this model are respectively improved by about 4% and 7% compared to the best model proposed by L.Itti. Moreover, by comparing the ability of our model to predict eye movements produced by an average observer, we can conclude that our model succeeds quite well in predicting the spatial locations of the most important areas of the image content.

**Keywords :** Visual Attention, Point of Fixation, Human Visual Model, Saliency Map, Performance Evaluation

## 1. INTRODUCTION

Visual attention is an essential function in the set of mechanisms used by the Human Visual System (HVS) to look at his 3D environment. It is a rather complex process which selects the most relevant objects in a scene – more precisely, the most relevant areas in the image of the scene, these areas covering only a part or the entire view of an object - according to some specific features they behave and to the importance they show (cognitive interest) at the moment of perception. Visual attention is of great importance in many applications, firstly in visual pattern recognition and scene analysis as it contributes to extract and select interesting visual features which will be further associated in the recognition process [1] , secondly in image and video communications with lossy data compression [2]. In this later application, since a decade now many methods have been proposed for adapting the available bit budget to the degree of interest of the different parts (block-based or region –based ) of the image or video: more bits are used in the coding of visually interesting image locations than in the others and so the image/video quality is interest-dependant [3].

It is well established that visual attention is driven at least by two mechanisms working in conjunction: a top-down control mechanism which is task-dependent, a bottom-up mechanism which is visual stimulus-dependent [4]. Most of the researchs in our image/video community concerns only the last one as it is common to all the applications, contrary to the former. Another important reason for that is one needs to know and modelize many aspects of the control of human perception, which is in close relationship with natural (visual) intelligence [5].The different processing involved in it are not still clear up to now. Therefore, as L. Itti did in designing his model [6], we have limited the objective of our model to simulate only the bottom-up visual attention process performed in the HVS. This model is based on some important properties of this later that provides noticeable advantages compared to other classical published approaches [7]. In this paper, first we recall the coherent approach for modeling the visual attention we recently proposed [8]. Then we focus the paper mainly on the performances assessment of this model by achieving extended subjective and objective comparisons with real fixation points captured by an eye-tracking system used by human observers looking at still color pictures. Comparison with others models will be also made before concluding.

## 2. VISUAL ATTENTION MODELING

We developped recently a coherent approach to model the bottom-up visual attention process [8] (the reader can access to the mathematical description of the model in this reference). Contrary to others classical approaches, it takes into account the main properties of the human visual system involved in this process. This functional and computational model has been derived from a series of psychophysic experiments. It is decomposed mainly into four parts (see figure 1).

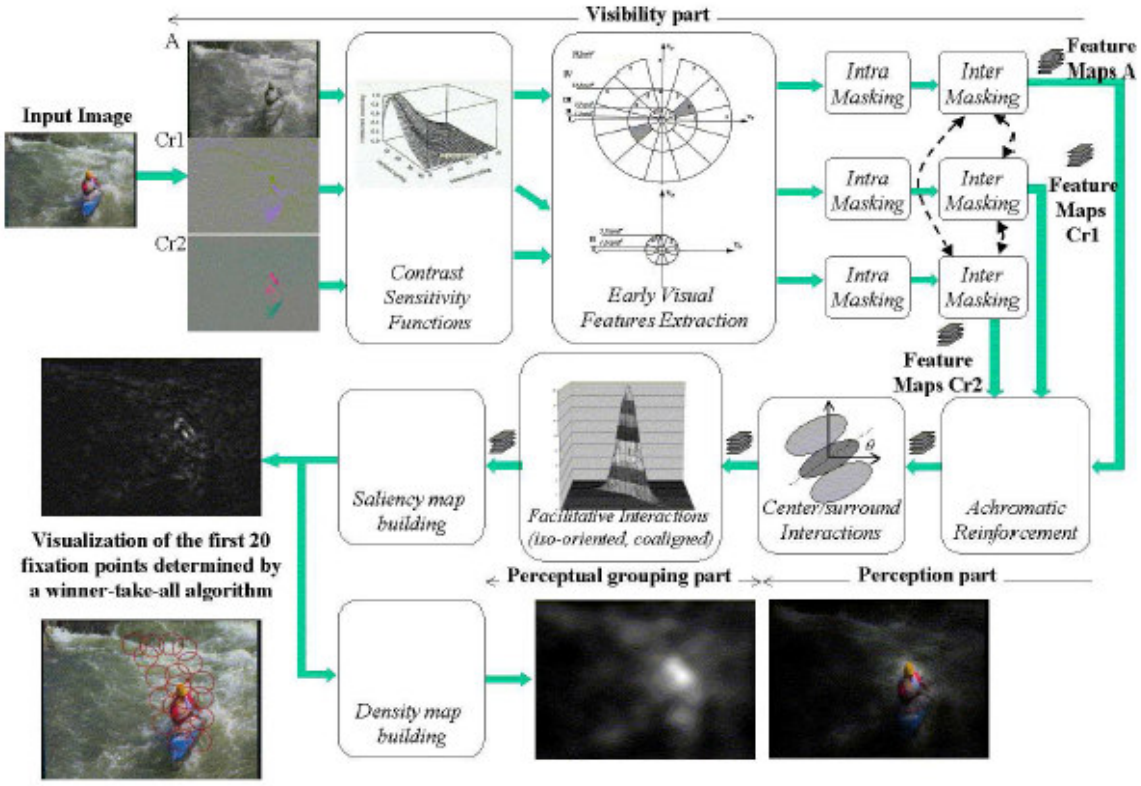


Figure 1 : Block diagram of our computational model of bottom-up visual selective attention. It is composed of four main parts: visibility, perception, perceptual grouping, saliency map building.

❶ **Visibility part**. It implies a set of transformations which modelize the progressive transformation of RGB color pixels (digital data in computer) into multiresolution visual signals found in the visual primary cortex (V1 area) :

- two successive non linear transformations of RGB digital data into first physical luminances ($L_R$ , $L_G$ , $L_B$ ) then perceptive luminance due to the 3-type cones absorption (L, M, S);
- a color normalization of the LMS signals;
- the projection onto a perceptual color space which makes separation between a pure brightness and two opponent color components (Krauskopf's color space: A, CR1, CR2);
- a perceptual subband decomposition which corresponds to a partitionning of the spatial frequency space both in radial spatial frequency and in orientation (non dyadic radial frequency bandwidths and orientation selectivity depending of the radial frequency band). Actually, the subband decomposition modelizes the behaviour of the different types of cells in the primary visual cortex showing a classical receptive field (CRF). The frequency

decomposition leads for SDTV with normalized viewing conditions (d = 6H) to 17 subbands for the achromatic component A and only to 5 subbands for the two chromatic components CR1 and CR2, respectively. The contrast sensitivity functions (one CSF for each of the 3 components) are also used to weight the signals in the different subbands of each of the three components ;

- finally, intra and inter subband masking effects are modelized for each of the three component. As masking effects correspond to a modulation of the differential visibility threshold of a visual signal (stimuli) due to others rather strong visual signals in its surroundings (masking signals), the outputs of the different subbands are adaptively weighted again according to the value of the masking. This later can be in some cases lower than one (facilitation effects) but very often higher than one due to really what is strictly so-called masking effect.

The different psychovisual signals obtained at the end of this first part can be view as visual band-limited signals whose magnitudes have been normalized: each unit of magnitude corresponds at this image location to value of the local differential visibility threshold for the considered band-limited visual component.

❷ **Perception part**. This part provides a structural description of the achromatic component. When a signal is visible, the perception we have of it is not linearly dependant of the normalized visual signal. Even more important is that perception can be strongly modified by non-adjacent contextual surroundings. It is due to the effects on cortical cells activity in V2 area of contextual stimuli outside although not far from the receptive field of the cortical cells. On an another way, only for sake of simplicity, we also incorporated in this part the impact of the two chromatic components on the achromatic component, through the way of a reinforcement of this later by the chromatic context. So, in summary, two successive processings are performed:

- a possible reinforcement of the achromatic band-limited normalized signals by the presence of contrasted band-limited color components in a direction orthogonal to the direction associated with the achromatic subband [5] but within the same subband. Reinforcement is dependant on the type of the contrasted chromatic component and is realized in a multiplicative manner;
- modeling the center-surround suppressive interactions in the neurons of the primary visual cortex which expresses the so-called anisotropic non-CRF inhibition. The inhibition term is computed in an inhitory area adjacent to the CRF. To do it, a difference of 2D gaussians kernel filtering is used whose spatial extension (related to 2 standard-deviations ) is depending on the orientation. Clipping is finally performed to restrict perceptive signals to non negative amplitudes.

❸ **Perceptual grouping**. This part refers to the properties that the human visual system (HVS) has in grouping local perceptive features having the same properties and which are organized into more or less regular 1D or 2D structures and in binding these local features into a structural feature of higher level which can be further interpreting by human at a semantic level. One typical example of that is the capability of binding local perceptual edges into elongated contours ( called *contour integration*) even if some texture elements can disturb at some spots the perception of edges, and so their detection. This can be extended to the easy and well perception of corner by binding two half straight contours showing a difference in angular directions at their very closed ends.

So we simulate the contour integration function in the early visual processing (area V2) by a sequence of two processings:

- a facilitative interaction based on the properties of iso-orientation and interaction between co-aligned cells: to compute it, a linear filtering is performed with the use of two half butterfly filters having a proper point spread function size each other;
- a reinforcement (facilitative interaction) due to interaction between achromatic subbands of higher radial frequencies but with the same orientation.

❹ **Saliency map**. A saliency map is a representation of the visual field, coding the degree of interest or of attractiveness of each site or pixel. So a unique 2D saliency map S should be computed from the different band-limited signals obtained at the outputs of the perceptual grouping. As in all the previous steps, a coherent normalization has been used, we can fuse all the perceptive signals by a simple summation. We have to take care of other effects in producing a saliency map. A first effect is that when a human looks at a spot in a scene, the exact location varies along the time and from one test to

the other with the same observer. It depends also of the observer. In summary, we modelized all these effects by filtering with a 2D gaussian point spread function the outputs of the perceptual grouping part, in order to re-inforce regions showing a high density of points of interest and, conversely, to attenuate the other regions (with a low density of points). One important second effect can be easily taken in account. Usually in a view of a scene, the objects of interest are more likely located in the central part than nearby the borders in relation with the manner to film. So the output of the filtering step is at the end multiplied by an anisotropic 2D gaussian function using a different standard deviation in horizontal and vertical direction (these two values have been optimized), respectively, scaled at a unitary amplitude at the center of the picture.

## 3. PERFORMANCE EVALUATION OF THE SALIENCY MODEL

Evaluation of performances of a computational bottom-up visual attention model cannot readily be achieved as there is not a real consensus on a particular method. A first rather simple but necessary method consists in subjective comparisons. This qualitative method refers to subjective assessments in which human observers can judge the more or less ressemblance between the ground truth saliency map and the prediction of it. A second class of performance evaluation concerns purely objective methods in which some specific parameters measure the accordance or dissemblance between them.

### *A. Subjective evaluation*

Subjective performance evaluations can provide insights into the effectiveness of the proposed model. However, one has to not get too strong conclusions from these kind of evaluations since different observers can judge the dissemblance differently. As an example, we illustrate the application of our saliency map construction on three pictures: “LightHouse2”, “Rapids”, “Patin_Couleur0”. Figure 2 displays for each of them: the original color picture (left) , the human Region Of Interest (ROI) (middle) , the saliency map (right).


Figure 2. Saliency maps. Results from human observers and from our method on pictures “LightHouse2”, “Rapids”, “Patin_Couleur0” : original color pictures (left) , human Region Of Interest (middle), saliency map (right).

To improve the visibility of these maps, they are used through a logarithm law in an amplitude modulation of the pixels in the original color picture (the log transformation is used only for visualization purposes). For each picture, the map of human ROI has been computed with data collected from experiments made on a set of 20 natural color images (in fact 10 were presented with the full color and the other with their luminance component only) presented to 40 observers in a task-free viewing mode, a viewing duration of 14 seconds, using an eye tracker apparatus from Cambridge Research Corporation. The experiments has been already described in [8]. We can observe on average a rather good correspondence between human ROI and predictive ROI but with a widening of the predictive Region Of Interest for the two pictures "LightHouse2" and "Rapids" and, at the opposite, a contraction of the predictive Region Of Interest for picture "Patin_Couleur0". The difference can be, in part, explained by the fact that the model does not take care of the semantic effects which are always included in the human strategy.
In conclusion from this subjective evaluation on our picture databse, it seems that our model detects more conspicuous areas than human observer does but most of human ROI are well covered by the model.

## *B. Objective evaluation*

Two objective methods have been used to evaluate the effectiveness of the proposed model. They are based on the measure of the linear correlation coefficient and on the Kullback-Liebler divergence, respectively.

### *1)* Measure by the linear correlation coefficient

The linear correlation coefficient is widely used for comparing two sequences of data (1D or multi dimensional fields) for different purposes such as image registration, object recognition, and disparity measurement. To assess the model, we measured the correlation coefficient CC between the predicted saliency map (noted PS in the following) and the real one deduced from eye-tracking experiments (noted HS). The linear correlation coefficient is given by:

$$CC(HS, PS) = [Cov(HS_c, PS_c)] / [\sigma_H * \sigma_P]$$

with: $HS_c$ , $PS_c$ the centered human and predictive density maps, respectively (average values of density maps is set to zero);
Cov (. , .) is the experimental covariance function;
$\sigma_H$ and $\sigma_P$ are the experimental standard deviations of the human and predictive density map, respectively.

The correlation coefficient measurements CC for a set of the 10 tested color images are given in Table I for our model and also, for comparison, for the Itti's model which we consider as the reference model. Both models have been evaluated on the same color image database. All the experimental data belonging to the real periods of fixation have been used in the full viewing time (14s): we rejected eye tracking coordinates considered as member of saccade periods. For the Itti's model, two options have been tested: one with and one without the application of an anisotropic 2D Gaussian function which attenuates the importance of a fixation point with its excentricity. For most of the color pictures tested, the correlation between human data and our model is higher than the correlation with Itti's model. If we consider this last model only with the use of the eccentricity weighting function as such weighting is also included in our model, the differences in correlation vary from – 0.09 to + 0.26. There are two pictures (*Kayak_Couleur0*, *Sailing1*) for which the reference model with the weighting function shows better performances than our model does. The two models give the same correlation on one image (*Zebres_Couleur797*), and for the other, our model provides higher correlation. A particular case concerns image *Manfishing* for which our model achieves a rather high correlation ( CC = 0.87) contrary to the other model.
Taken as a whole, our model outperforms the reference model. The average correlations cofficient are improved by 0.04 and 0.31 with the reference model' with the use of a weighting function and without it, respectively.

Another interesting results have been analysed. It concerns the effect of the viewing duration. We expect that, a priori, it may have an impact in the selection of the ROI. As the time is running, the strategy for human to look at a picture and fix it may be changing. The correlation measurements for 3 viewing durations ( 2s, 5s ,14s) are given in Table II for our model and for the reference model (Itti's model) too. The standard deviation in the eccentricity weighting has been

adapted and optimized for each of the 3 durations ( 2°, 2.25° , 2.5°) and used in the two models. At a glance, one notices that the correlation decreases a slight when the viewing duration moves from 2s to 5 s, then the correlation increases when the duration goes up to 14s and overcomes even slightly the value obtained with 2s. This qualitative behaviour is roughly the same with the two models and the gap in the performances remains almost the same whatever the viewing duration is: our model provides always ( along the time) better performances than the reference model. One explanation of the dependance with the wiewing duration is the following. With a short duration (2s), the effect of eccentricity is rather important and human observers looks more directly to salient features closer to the image center. During the next duration (from 2s to 5s), humans looks also at features which can be located also on the sides of the picture, provided that they have a sufficient saliency, in order to get an overview of the image content. Then, during the last part of the viewing time (from 5s to 14s), human focus of attention is again concentrated on the more interesting and more salient parts, more or less already scanned at the beginning, to detail them in deeper.

**TABLE I**

LINEAR CORRELATION COEFFICIENT FOR DIFFERENT PICTURES, FOR OUR MODEL AND THE REFERENCE MODEL (ITTI'S MODEL) WITHOUT AND WITH ECCENTRICITY FUNCTION (WEIGHTING)

| **Pictures** | **Our model** | **Itti's model + weighting** | **Itti's model** |
|---|---|---|---|
| Kayak_Couleur0 | 0,68 | 0,37 | 0,79 |
| Manfishing | 0,87 | 0,25 | 0,61 |
| Church&Capitol | 0,62 | 0,32 | 0,52 |
| Vautour_Couleur538 | 0,76 | 0,25 | 0,67 |
| Zebres_Couleur797 | 0,53 | 0,26 | 0,53 |
| Patin_Couleur0 | 0,84 | 0,28 | 0,83 |
| Light_House2 | 0,74 | 0,52 | 0,73 |
| Rapids | 0,61 | 0,45 | 0,58 |
| Dancers2 | 0,72 | 0,26 | 0,69 |
| Sailing1 | 0,69 | 0,44 | 0,75 |
| **Average** | **0,71** | **0,40** | **0,67** |

**TABLE II**

LINEAR CORRELATION COEFFICIENT FOR DIFFERENT PICTURES, FOR OUR MODEL AND THE REFERENCE MODEL (ITTI'S MODEL) WITH ECCENTRICITY WEIGHTING, FOR DIFFERENT VIEWING DURATIONS

| **Pictures** | **Our model** | | | **Itti's model + weighting** | | |
|---|---|---|---|---|---|---|
| **Viewing duration** | **2s** | **5s** | **14s** | **2s** | **5s** | **14s** |
| Kayak_Couleur0 | 0,685 | 0,660 | 0,683 | 0,779 | 0,764 | 0,775 |
| Manfishing | 0,828 | 0,870 | 0,871 | 0,537 | 0,603 | 0,615 |
| Church&Capitol2 | 0,762 | 0,659 | 0,624 | 0,662 | 0,581 | 0,532 |
| Vautour_Couleur538 | 0,768 | 0,732 | 0,764 | 0,686 | 0,661 | 0,683 |
| Zebres_Couleur797 | 0,558 | 0,493 | 0,537 | 0,535 | 0,486 | 0,531 |
| Patin_Couleur0 | 0,798 | 0,854 | 0,846 | 0,800 | 0,837 | 0,826 |
| LightHouse2 | 0,702 | 0,696 | 0,740 | 0,708 | 0,690 | 0,724 |
| Rapids | 0,619 | 0,555 | 0,617 | 0,535 | 0,497 | 0,552 |
| Dancers2 | 0,830 | 0,655 | 0,729 | 0,739 | 0,590 | 0,645 |
| Sailing1 | 0,478 | 0,608 | 0,695 | 0,625 | 0,717 | 0,746 |
| **Average** | **0,703** | **0,679** | **0,711** | **0,661** | **0,643** | **0,663** |

### *2)* Measure by the Kulback-Liebler divergence

To provide an objective measure of the dissimilarity degree between our prediction and the ground truth, we may also use the Kulback-Liebler divergence. This measure consists in considering a density fixation map as a probability density function. The degree of dissimilarity between two probability functions h and p is given by the Kullback-Liebler divergence measure *KL*:

$$KL\ (\ p/h) = \sum_x\ p(x)\ \mathrm{Log}\ [\ p(x)\ /\ h(x)\ ]$$

with: h the human density probability function,
p the predicted density probability function,
x are the sites where the probabilities are defined.

When the two probability densities are identical, *KL* is null.
The performance evaluation mainly consists in comparing our saliency model with the reference model with and without a weighting function. In order to determine and interpret the gain in the divergence reduction, we also used an other simple reference: a uniform model, considered as a naïve model. It has been set by assigning to every site (pixel) the same saliency value in the probability map. The Kulback divergence measures are given in Table III on the same set of color images.

**TABLE III**

KL DIVERGENCES FOR DIFFERENT COLOR PICTURES : FOR OUR MODEL, THE REFERENCE MODEL (ITTI'S MODEL) AND A UNIFORM MODEL

| **Pictures** | **Our model** | **Itti's model** | **Itti's model + weighting** | **Uniform model** |
|---|---|---|---|---|
| Kayak_Couleur0 | 0,64 | 1,71 | 0,49 | 2,21 |
| Manfishing | 0,20 | 1,27 | 0,52 | 1,49 |
| Church&Capitol | 0,47 | 0,92 | 0,61 | 1,65 |
| Vautour_Couleur538 | 0,49 | 3,1 | 0,58 | 3,97 |
| Zebres_Couleur797 | 0,52 | 1,39 | 0,56 | 1,97 |
| Patin_Couleur0 | 0,24 | 2,2 | 0,26 | 3,02 |
| LightHouse2 | 0,43 | 1,06 | 0,45 | 2,63 |
| Rapids | 0,68 | 1,89 | 0,74 | 3,03 |
| Dancers2 | 0,49 | 1,56 | 0,55 | 1,90 |
| Sailing1 | 0,48 | 1,32 | 0,43 | 2,50 |
| **Average** | **0,46** | **1,64** | **0,52** | **2,44** |

Not surprisingly, the results show that the uniform model gives the worst results for each of the different pictures. The interest of this model relies in the divergence it yields which can be considered in practice as an apper bound (even if it don't give the maximum value). On average, our model provides the greatest performances. The gain (in absolute value) over the reference model (Itti) with an eccentricity weighing function is about 0.06 on average and around 11.5% in relative value. Notwithstanding the difference in average performance between these two approaches, the performance of our model is not always higher than those issued from the reference model. As noticed previously with the correlation measurement, for picture *Kayak_Couleur0*, the performance of Itti'model with weighting outperforms our model by 0.15 and by 0.05 for picture *Kayak_Couleur0.* On the opposite, the gain in quite impressive for picture *Manfishing* with a reduction in the KL divergence of more than 0.32, which goes down from 0.52 to 0.20 (a gain of 150% in relative value !).
We may also evaluate the performances as a function of the viewing duration. Table IV displays the KL divergence measures for the two models and for 3 durations (2s, 5s, 14s). In that case, the gain obtained in the KL divergence (divergence reduction) when the viewing duration increases, is much more visible than with the measure by correlation.

Moreover, the decreasing is monotonic with the duration, contrary to what has been observed with the correlation measurement: the two models perform better with the increasing of the viewing time, at least in the range 2s to 14s.

**TABLE IV**

KL DIVERGENCES MESEARE FOR DIFFERENT COLOR PICTURES : FOR OUR MODEL AND THE REFERENCE MODEL (ITTI'S MODEL), FOR DIFFERENT VIEWING DURATIONS.

| **Pictures** | | **Our model** | | | **Itti's model + weighting** | |
|---|---|---|---|---|---|---|
| **Viewing duration** | **2s** | **5s** | **14s** | **2s** | **5s** | **14s** |
| Kayak_Couleur0 | 1,152 | 0,968 | 0,646 | 1,05 | 0,792 | 0,494 |
| Manfishing | 0,657 | 0,249 | 0,202 | 1,14 | 0,512 | 0,520 |
| Church&Capitol2 | 0,526 | 0,495 | 0,472 | 0,651 | 0,598 | 0,611 |
| Vautour_Couleur538 | 0,812 | 0,594 | 0,497 | 0,893 | 0,635 | 0,582 |
| Zebres_Couleur797 | 0,763 | 0,647 | 0,523 | 0,846 | 0,667 | 0,560 |
| Patin_Couleur0 | 0,471 | 0,239 | 0,248 | 0,623 | 0,300 | 0,266 |
| LightHouse2 | 0,547 | 0,492 | 0,437 | 0,545 | 0,502 | 0,453 |
| Rapids | 1,24 | 1,110 | 0,686 | 1,910 | 1,580 | 0,740 |
| Dancers2 | 0,589 | 0,710 | 0,490 | 0,874 | 1,030 | 0,557 |
| Sailing1 | 0,869 | 0,602 | 0,482 | 0,723 | 0,460 | 0,437 |
| **Average** | **0,763** | **0,611** | **0,468** | **0,926** | **0,708** | **0,522** |

Another interesting way to assess the confidence of our model consists in computing over all the observers the average dissimilarity. This could be obtained in the following. Firstly, we compute the Kulback-Liebler divergence between the probability density function for one observer and the probability density function for the global human (average human observer). Secondly, we iterate this computation over the set of all observers. Finally, we take the average. The average observers's behaviour is given by *Klavg*. A high value means that the visual strategy of all observers is different. In others words, the deviation among the observers is high. On the opposite, a low value means that the visual attention strategies of all observers are quite similar. The lower bound is zero and will be obtained only if all the observers look at the same set of locations during the same viewing duration (independantly of the order).

**TABLE V**

COMPARISONS BETWEEN THE *KL* DIVERGENCE FOR OUR MODEL AND $Kl_{avg}$, FOR TWO DIFFERENT VIEWING TIMES

| **Pictures** | **Our model** | $KL_{avg}$ | **Our model** | $KL_{avg}$ |
|---|---|---|---|---|
| **Viewing time** | **4s** | **4s** | **14s** | **14s** |
| Kayak_Couleur0 | 1,18 | 0,72 | 0,64 | 0,60 |
| Manfishing | 0,41 | 0,67 | 0,20 | 0,52 |
| Church&Capitol | 0,16 | 0,58 | 0,47 | 0,30 |
| Vautour_Couleur538 | 0,82 | 0,70 | 0,49 | 0,47 |
| Zebres_Couleur797 | 0,79 | 0,93 | 0,52 | 0,73 |
| Patin_Couleur0 | 0,33 | 0,48 | 0,24 | 0,58 |
| LightHouse2 | 0,48 | 0,80 | 0,43 | 0,56 |
| Rapids | 1,23 | 0,73 | 0,68 | 0,43 |
| Dancers2 | 0,82 | 0,73 | 0,49 | 0,41 |
| Sailing1 | 0,64 | 0,85 | 0,48 | 0,55 |
| **Average** | **0,69** | **0,76** | **0,46** | **0,52** |

From Table V, several remarks can be drawn. First, the temporal variation of the divergence $Klavg$ is interesting since when the viewing time increases, divergence *KLavg* tends to decrease which means that the visual strategies of the observers are the closest: they converge in time. A possible explanation could refer to a property of the

human visual strategy previously observed : rather than scanning the whole scene, humans pursue to fixate areas of interest. Therefore, when the viewing duration increases, the contribution of spatial locations that are looked at with a very low probability decreases.
A second remark consists in comparing the divergence value (*KL(p|h)*) in the comparison between our model and the global human probability density function with the *KLavg* value.
We have to consider three cases :
• *KL(p|h)≈ Klavg* . There is a good pairing between the predicted density functions and the set of density functions obtained for each observer. For examples, for a viewing time of 14s, pictures *Kayak_Couleur0* and *Vautour_Couleur538* are in this case.
• *KL(p|h) < Klavg* . The most important part of the predicted density is well paired with the set of density functions obtained for each observer. In that case, the most conspicuous areas of the picture are well predicted. This is the case for most of the pictures in ths database.
• *KL(p|h) > KLavg* . There is a weak, or even a bad pairing between the most important part of the predicted density and the set of density functions obtained for each observer. Pictures *Church&Capitol* and *Rapids* are examples of an insufficient matching. Differences stem from the spatial locations of the most important areas in the two density functions.
From all these results, in summary we can say that, in most situations, our model succeeds in predicting the locations of the most important areas of interest in color images from a perceptual point of view. It is in accordance with the previous results given in Tables I and II, on one side, and Tables III and IV, on the other side.

## 4. CONCLUSION

In this paper, a coherent model of visual selective attention that computes a saliency map indicating the most salient spatial locations of a still color image, is briefly recalled. This model has been designed from some important behaviours of the human visual system. In this model, only a pure bottom-up mechanism found in early stages of the human visual system is modelized It takes into account rather precisely the different process which produce visibility effects, perception effects and perceptual feature grouping. To test its performances, eye tracking experiments were conducted in order to track and record real observers eye movements looking at images for capturing the content. From the knowledge of the ground truth, in terms of interest map, qualitative and quantitative comparisons have been developped. On a set of 10 natural color images presented to 40 observers in a task-free viewing mode ( viewing duration of up to 14 seconds), the results show that, on average, the linear correlation coefficient achieves 0.71 and the Kullback-Leibler divergence measure takes about 0.46. Over this database, the linear coefficient correlation and the Kullback-Liebler divergence are respectively improved by about 4% and 7% for our model compared to the model proposed by L.Itti. Measurements performed at intermediate viewing duration (2s and 5s) show the same difference in performances. Moreover, by comparing the ability of our model to predict eye movements produced by an average observer, we can conclude that our model succeeds quite well in predicting the spatial locations of the most important areas in a picture.

## ACKNOWLEGMENT

We are grateful to Laurent Itti for his help and for his agreement concerning this performance comparison.

## REFERENCES

1. J. M. Wolf, *Visual Search*, pp. 13-74, in Pashler, H. editor, Attention, Psychology Press, 1998.

2. E. Nguyen, C.Labit, J. M. Odobez, *A ROI approach for hybrid image sequence coding*, Proceedings ICIP-94, Vol. 3, pp. 245-249, Nov. 1994.

3. Z. Wang, L. Lu, A. C. Bovik, *Foveation Scalable Video Coding with Automatic Fixation Selection*, IEEE Trans. On Image Processing, Vol. , pp. , 2003.

4. C. Kock, S. Ullman, *Shifts in selective visual attention: towards the underlying neural circuity*, Human Neurobiology, Vol. 4(4), pp. 219-2237, 1985.

5. A. Oliva, A. Torralba, *Top-Down Control of Visual Attention in Object Detection*, Proceedings ICIP-03, Barcelone, 2003.

6. L. Itti, C. Kock, and E. Niebur, *A model of saliency-based visual attention for rapid scene analysis*, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), vol. 20, N11, pp. 1254-1259, 1998.

7. U. Rajashekar, L.K. Cormack, A.C. Bovik, *Images features that draw fixations,* Proc. ICIP'03, Barcelone, 2003.

8. O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, and E. Francois, *From low level perception to high level perception, a coherent approach for visual attention modeling*, Proceedings of SPIE Human Vision and Electronic Imaging IX (HVEI'04), San Jose, CA, (B. Rogowitz, T. N. Pappas Ed.), 2004.