Prediction of the Inter-Observer Visual Congruency (IOVC) and application to image ranking

Olivier Le Meur
University of Rennes 1
Campus de Beaulieu
35042 Rennes, France
olemeur@irisa.fr

Thierry Baccino
University of Paris VIII / LUTIN
30, avenue Corentin Cariou
75930 Paris, France
thierry.baccino@univ-paris8.fr

Aline Roumy
INRIA
Campus de Beaulieu
35042 Rennes, France
aline.roumy@inria.fr

ABSTRACT

This paper proposes an automatic method for predicting the inter-observer visual congruency (IOVC). The IOVC reflects the congruence or the variability among different subjects looking at the same image. Predicting this congruence is of interest for image processing applications where the visual perception of a picture matters such as website design, advertisement, etc. This paper makes several new contributions. First, a computational model of the IOVC is proposed. This new model is a mixture of low-level visual features extracted from the input picture and model's parameters are learned by using a large eye-tracking database. Once the parameters haves been learned, it can be used for any new picture. Second, regarding low-level visual feature extraction, we propose a new scheme to compute the depth of field of a picture. Finally, once the training and the feature extraction have been carried out, a score ranging from 0 (minimal congruency) to 1 (maximal congruency) is computed. A value of 1 indicates that observers would focus on the same locations and suggests that the picture presents strong locations of interest. A second database of eye movements is used to assess the performance of the proposed model. We also show that the proposed method outperforms an existing method. Results show that our IOVC criterion outperforms the Feature Congestion measure [32]. To illustrate the interest of the proposed model, we have used it to rank personalized photograph in an automatic manner.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Experimentation, Human Factors

Keywords

visual dispersion, congruency, photo ranking eye tracking

1. INTRODUCTION

Idiosyncrasy is defined as an individualizing quality or characteristic of a person or group, and is often used to express peculiarity (from Wikipedia). Therefore idiosyncratic eye movements refer to as the difference between the visual scanpaths of observers viewing the same stimulus. More precisely, these differences concern the intrinsic features of visual fixations. For instance, there is a strong variability of fixation durations between and within observers as shown by [30]. In this paper, we define inter-observer visual congruency (IOVC) and we propose a new method to automatically compute it.

The causes explaining the visual dispersion are usually classified into either stimulus-dependent (or bottom-up) or observer-dependent features (or top down).

A first observer-dependent cause is the cultural difference as suggested by [26, 5]. These authors compared the visual scan pattern of two different populations, an American and an Asian one. The conclusion was Asian people tend to look more at the background and spend relatively less time on focal objects than American people. However, a recent study casts doubt on the influence of cultural differences on oculomotor behavior [31].

A second observer-dependent factor of variability is due to our prior knowledge or our prior experience as illustrated by [1]. They showed that the variability differs when observers looked at famous versus non-famous faces. Famous faces tend to decrease the dispersion of visual fixations just after the stimulus onset. The visual scanning is then affected by higher level information such as the eye-movement-based memory effect.

On the opposite, stimulus-dependent features influencing the dispersion between observers is related to the properties of the stimuli itself. For instance, Rousselet et al. [35] shows that human faces as well as animals attract our attention leading to a decrease of inter-observer variability. Contrary to the two previous factors that are observer-dependent, the properties of the stimuli and their characteristics can be easily extracted and analyzed automatically from the picture.

It is also important to note that the difference between observers' scanpath is time-dependent [28, 37]. They indeed showed that the consistency in fixation locations between observers decreases with prolonged viewing. To explain that, two hypothesis have been formulated: [28] shows

that the influence of bottom-up mechanisms decreases with the viewing time and is progressively overridden after several seconds of viewing by top-down mechanisms. On the contrary [37] conjectures that bottom-up mechanisms are not time-dependent and that low-level visual features might keep their ability to attract our visual attention throughout the viewing. They [37] therefore explained that the increase of inter-observer variability would be due to the growing influence of top-down mechanisms over time. Even if both papers propose different explanations, they all agree on the time dependency of IOVC and on the fact that bottom-up (or stimulus-dependent) mechanisms occur first.

In this study, we predict the inter-observer visual congruency that occurs in the first seconds of a picture observation. From a given picture, a score indicating the degree of visual congruence is computed. The computational model we propose, combines stimulus-dependent features which are solely inferred from the low-level visual features of the incoming picture. We train the model by using a large eye-tracking database. In this database, we consider the first eye movements tracked in the first seconds of a picture observation. Observer-dependent features such as prior knowledge are not taken into account since they may vary from one observer to another. In order to get a general model that can apply to any picture, we propose a stimulus-dependent model. Therefore, our method predicts the dispersion based on the influence of stimulus characteristics, rather than individual observer characteristics.

There is almost no study dealing with the computational modeling of the inter-observer visual congruency. The closest method work concerns a method to measure visual clutter. Note that IOVC and visual clutter defers since the former is the dispersion that exists between locations focused by observers whereas the latter is related to the amount of visual information in a scene. However both might be strongly correlated. The most popular method to measure the visual clutter has been proposed by Rosenholtz et al. [32]. The idea is to measure the visual clutter of a scene in order to avoid confusion and to speed up the visual processing of information. For instance, a possible application is to help people to find important information on a web site or simply on a screen. Rosenholtz et al.'s solution is based on a set of low-level visual features. Some of them will be reused in this study. Rosenholtz et al. assessed the performance of their algorithm by comparing the amount of clutter for a scene to the reaction time required to find a target in the same scene. In this paper as in [32], we use low-level visual features. However, our approach differs from [32] since the goal is different and since we not only use low-level visual features but also eye tracking measurements. More precisely, we use the visual scanpaths of observers in order to train a model. Our approach is supported by a number of studies suggesting that the degree of clutter present in the scene affects the deployment of our visual attention [17]. For instance, Ehinger et al. [9] measured the dispersion between observers to define an upper bound of model-predicted saliency. Eye movements were collected when 14 observers performed a task of pedestrian detection in 912 outdoor scenes. The dispersion between observers was computed by a one-against-all method as proposed by [41]. The same approach will be used in this paper to compute our ground truth. However, the comparison between Ehinger et al. [9] stops there because Ehinger et al. proposed a com-putational model to predict where observers look at, while searching for pedestrians. To infer the salient locations in this visual search context, they combined different guidance sources such as low-level saliency (by using Torralba's model [41]), target features (by using a person detector) and scene context. In our method, we do not predict where people look at. We predict the dispersion between observers indicating whether observers look at similar locations or not. Ehinger et al. used the dispersion between observers as an an upper bound for comparing performance of computational models with human fixations. As stated by [46], no saliency prediction can perform better than the dispersion between observers.

Eye movements are also often used to examine the usability of an interface. Golberg and Kotval [13] defined a number of metrics based on the analysis of fixation durations and saccade amplitudes. Cowen et al. [8] featured the efficiency of a visual search by examining the fixation distributions. Fixations concentrated in a small area indicate focused and efficient searching whereas sparse fixations would indicate lower search efficiency.

In this paper, we present a computational model of inter-observer visual congruency (IOVC). From a training set composed of eye tracking data, we build a prediction model by using a supervised approach. Our goal aims to give a score to a given picture that indicates whether the visual strategy of human observers is similar or not. From the eye tracking data, we set up a ground truth by computing for each picture the inter-observer congruency. By using a limited number of low-level visual features, a model is trained and is able to predict efficiently and automatically the dispersion between observers. This predicted score can be used in applications where the visual perception of a picture matters such as website design, advertisement. Indeed, it can help to rank images based on their capacity to attract our attention or help to measure the relevance of web design.

The paper is composed as follows. Section II gives an overview of the proposed approach. Section III describes how the IOVC is measured. A large database of eye tracking data is used for this purpose. Section IV is related to the extraction of visual features that are supposed to influence the attentional allocation. Section V concerns the learning and its performance. Section VI presents an application for ranking personalized pictures based on their interestingness. Finally, we conclude the paper.

In summary, our contributions include:

- a solution for predicting the inter-observer congruency. We use an eye tracking database to train a computational model;

- a novel scheme to estimate the depth of field for a scene is proposed which is simpler than existing methods;

- a new method to rank automatically personalized photographs. The ranking is based on pictures' interestingness.

2. SYSTEM OVERVIEW

Figure 1 illustrates the proposed approach. First, an image database with its corresponding eye tracking data is set up. The feature extraction step extracts different visual attributes for each picture of the training dataset. After the feature extraction, the training set along with eye tracking data is used to train a cluster-weighted model. The trained
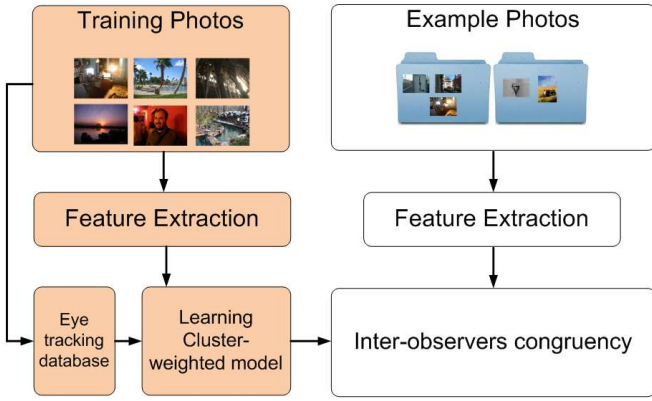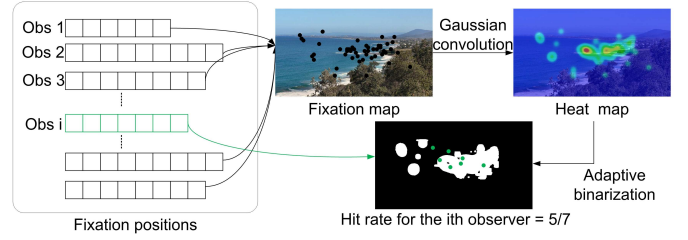
Figure 1: System overview.

Figure 2: Measure of the inter-observer congruency. On the left, the spatial coordinates of visual fixations for each observer are given. By considering all fixations except those stemming from the ith observer, a heat map is computed (on the right). After an adaptive binarization, we count the number of fixations of the ith observer that fall into salient regions (white region on the bottom).

model is then used to predict the inter-observer congruency of a picture taken from a new data set.

Once the estimation of model's parameters has been performed, personalized photograph can be ranked according to their interestingness. The interestingness of an image is related to its ability to attract and to hold our attention. The interestingness of a picture is thus similar to the inter-observer congruency.

3. MEASURING THE INTER-OBSERVER CONGRUENCY

3.1 Eye-tracking database

Judd et al.'s database [20] is used in this study. Let us first review the experimental protocol and the characteristics of this dataset. The database is composed of 1003 images of various contents. Images had different resolutions and were in a landscape or portrait orientation. Fifteen viewers were involved in the eye tracking tests. The users were males and females between the ages of 18 and 35. Two of the viewers were researchers and the others were naive viewers. They viewed each image for 3 seconds in a free-viewing task. Participants sat at a distance of approximately two feet from a 19 inch computer screen of resolution 1280×1024. The images subtended approximately 45 horizontally and 37 vertically of the observer's field of view. The number of pixel per degree is then about 23. This database can be downloaded from `http://people.csail.mit.edu/tjudd/index.html`.

3.2 Inter-observer congruency

To assess the inter-observer congruency, a one-against-all approach (also called leave one out) is used as in [41]. The first step consists in computing a 2D fixation distribution from the fixation data of all observers except one for a given picture. The fixation distributions were then convolved with a two-dimensional Gaussian. Each pixel of this map represents the probability to be fixated. The standard deviation of the Gaussian kernel is set at one degree to reflect estimates of foveal size. This map is then thresholded to select an image area having the highest probability of being fixated. The threshold is adaptively set in order to keep 25% of the image. The goal is now to compute the percentage of the visual fixations of the remaining observer that fall within

salient parts of the threshold saliency map. This process was iterated for all observers. For a given picture, the variability between observers is the average of the aforementioned percentage over all subjects. As most of the dispersion values are in the range of 0.5 to 1, the scale has been stretch from 0 to 1. A value of 1 indicates that observers fixate the same areas. Conversely, a low value would suggest that the scan patterns are uncorrelated meaning a strong variability between subjects. Figure 2 illustrates the method for the ith observer.

Over the whole dataset, the average dispersion is of 72%, the median dispersion is of 76%. Figure 3 shows the distribution of the inter-observer congruency over the whole dataset. It is interesting to notice that, for a number of pictures, the congruency is maximal. This is due to the fact that a fixation point is defined by its spatial coordinates and by its neighborhood, representing one degree of visual angle (representing fovea's size). Figure 4 shows for different pictures the experimental congruency between observers. Results suggest that the congruency is small when there is nothing in the scene that catches our attention. In this context, areas that stand out the background are rare and the scene consistency is strong. As expected, the presence of human faces tends to increase the inter-observer congruency. It is indeed known that human faces attract in an effortlessly manner our attention.

4. EXTRACTION OF VISUAL ATTRIBUTES IMPACTING THE INTER-OBSERVER VARIABILITY

In this section, the visual features used to predict the inter-observer congruency are presented.

4.1 Face detection

As the human faces significantly impact our visual deployment, it is of importance to detect human faces. The face detector we use is the one proposed by OpenCV library. The face detector is based on Haar feature-based cascade classifier for object detection. This kind of detection has been initially proposed by [43] and improved by [22].
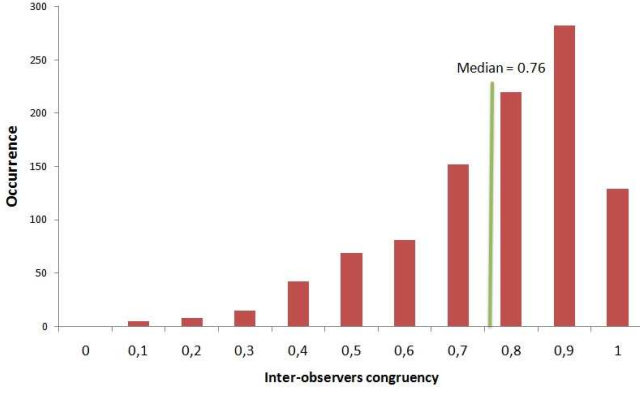
4.2 Color Harmony

Figure 3: Distribution of the inter-observer congruency over Judd et al.'s dataset [20].

Several studies showed that scene incongruency or inconsistency are factors influencing the inspection of an image [23, 14, 42]. Among the scene inconsistency factors (objects, size, etc), the color might be an important factor. For instance, Frey et al. [11] showed that overt attention is significantly influenced by the presence of color. The basic assumption was that the color presence might systematically increase the congruency. The conclusion of [11] is not so straightforward. Indeed, the influence of the color might depend on the picture's category.

In this study, the color inconsistency refers to the color harmony of the scene. We speculate that a scene with a strong consistent color harmony would be less visually disruptive than a scene with a poor color harmony. To measure the color harmony, we propose to follow the process of [6].

In [6], the notion of color harmony is based on the work of Matsuda [25, 38]. Figure 5 illustrates the height harmonic templates on the hue wheel. These templates, that may be rotated by an arbitratry angle, can be used to measure how aesthetically pleasing an image is.

The color harmony of an input color picture I, called $CH(I)$ is computed by equation 1. We use similar notations as [6]. They are briefly reminded below:

$$CH(I) = \min_m F(I, (m, \alpha^*)) \tag{1}$$

where, $m \in \{i, I, L, T, V, X, Y\}$. The function F is defined as follows:

$$F(I, (m, \alpha^*)) = \sum_{i \in I} \left\| H(i) - E_{T_m(\alpha^*)}(i) \right\| \times S(i) \tag{2}$$

where H and S denote the hue and the saturation channels, respectively. The hue distance $\|.\|$ refers to the arc-length distance on the hue wheel (measured in radians). Hues that are enclosed in the sector of T_m are considered to have zero distance from the template. α^* defines the orientation of the template T_m that minimizes the distance F. As in [6], we use Brent's algorithm [29] to find the best orientation α^*. This kind of algorithm seeks a local minimum of a function in a given interval. $E_{T_m(\alpha^*)}(i)$ is the sector border hue of template T_m with orientation α^* that is the closest to the hue of pixel i. Figure 6 gives an example of two pictures extracted from [6]: one is the original whereas the second presents an optimized color harmony. The value F is given

Figure 4: Examples of pictures associated with their corresponding inter-observer congruency. The variability is in the range of 0 (strongest variability) to 1 (no variability).

for each template.

4.3 Depth of Field

The Depth of Field (DoF) is defined as the distance between the nearest and farthest objects in a scene that appear acceptably sharp in an image. A shallow DoF is often used to emphasize the region of interest in a picture. It is for instance used for portraiture photography. All background details are blurred whereas the nearest person (or object) is sharp, attracting our attention. An example is given figure 7 (a). When a large DoF is used, the opposite effect is achieved. The entire picture is sharp so that all the details of the scene are preserved. Picture of figure 7 (c) was taken with a large DoF.

Estimating the DoF is then of importance. As photographers can steer our visual attention towards a particular areas by controlling the DoF, the inter-observer variability might be contingent upon this artistic effect.
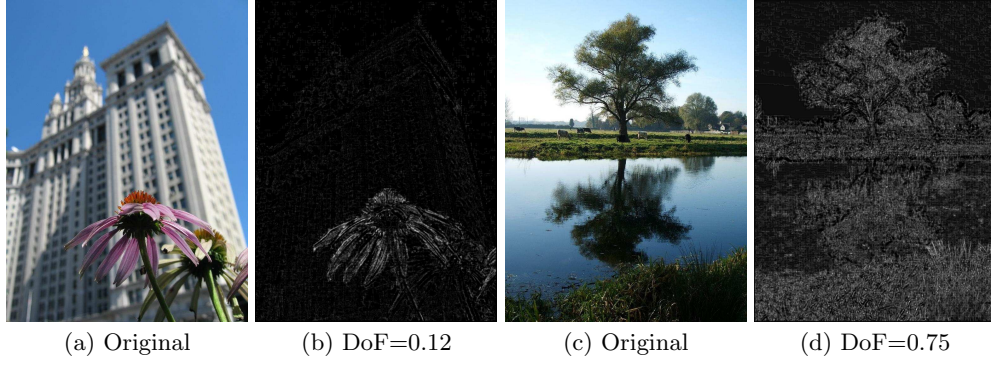
(a) Original   (b) DoF=0.12   (c) Original   (d) DoF=0.75

Figure 7: (a) and (c) two original pictures. (b) and (d) indicates areas sensitive to blur in dark. The brigth areas correspond to unfocuss areas. DoF, standing for Depth of Field, indicates whether the picture is sensitive to blur (deep DoF) or not (shallow DoF).
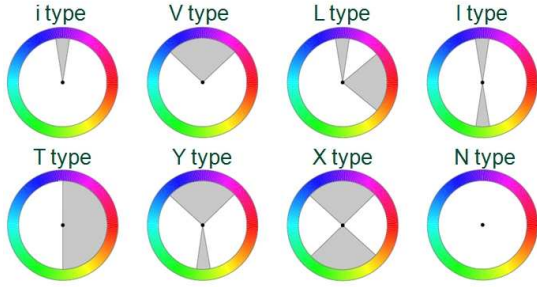

Figure 5: The seven color templates defined by [25] on the hue wheel. The templates may be rotated by an arbitrary angle (extracted from [6]). The template N is used for gray-scale images.

To determine the depth of field, the proposed algorithm relies on the fact that the shape of the horizontal/vertical derivatives histogram is modified after a blurring operation [21, 24]. The proposed scheme to compute the DoF of a picture is described below.

Let I the input picture and f_k the bluring kernel of size $k \times k$ ($k = \{3, 5, 7\}$). The blurring kernels are first applied on the luminance L of I and the vertical and horizontal derivatives are then computed. The vertical and horizontal derivatives are given by:

$$p_{xk} \quad \alpha \quad hist(I * f_k * d_x) \quad (3)$$
$$p_{yk} \quad \alpha \quad hist(I * f_k * d_y) \quad (4)$$

where $d_x = [1 \ -1]$ and $d_y = [1 \ -1]^T$.
For a pixel (i, j) and for a kernel k, we compute the KL-divergence between the distributions p_{xk} and p_{yk} and the original distributions p_{x1} and p_{y1}:

$$D_k(i,j) = \sum_{(n,m) \in W_{ij}} KL(p_{xk}|p_{x1})(n,m) + KL(p_{yk}|p_{y1})(n,m)$$

$$(5)$$

where, W_{ij} is a window centered on the pixel (i, j). In this study, all the experiments were performed using uniform kernels.

The KL-divergence for a given pixel located at (i, j) is

Figure 6: Original (top) and harmonized (bottom) picture extracted from [6]. On the right-hand side, the value $F(I, (m, \alpha^*))$ for each template m is given. The red value is the lowest one, called here $CH(I)$. The harmonized picture presents a lower CH value than the original one.

given by the following formula:

$$KL(p|q)(i,j) = p_{ij} log(\frac{p_{ij}}{q_{ij}}) \quad (6)$$

The KL-divergence involves two probability density functions p and q. They both sum to 1. The KL-divergence is only defined when p_{ij} and q_{ij} are greater than zero. The quantity $0 log 0$ is considered as zero.

The use of the KL-divergence is especially interesting in the equation 7 because of its similarity with the DoF values. Indeed, D_k tends to zero when the distributions p_{xk} and p_{yk} are close to p_{x1} and p_{y1}, respectively. In this case, it means that the incoming picture is not sensitive to blur indicating that the picture is already blurred. The DoF is

then low. When the value D_k increases, it suggests that the areas under analyis is rather sharp (DoF is probably high).

The DoF value is finally computed as follows:

$$DoF = \sum_{(i,j) \in I} \sum_k D_k(i,j) \qquad (7)$$

Figure 7 (b) and (d) give the value of DoF for two examples. For the first one, the DoF is of 0.12 suggesting that the picture is composed of large blurred areas. As the DoF is greater than zero the picture probably presents a sharp areas, sensitive to a blurring operation. For the second picture, the DoF is of 0.75. Unlike the previous one, this picture is more sensitive to blurring operations, suggesting that most of the pictures are sharp. Figure 7 (b) and (d) illustrate in bright areas regions that are sensitive to blur. For the sake of visibility, the two pictures have been normalized in the range of 0 to 255 by using their own global maximum (3.56 and 4.68, respectively). This kind of map might be used to extract the region of interest when the DoF value is rather low, as proposed by [24].

4.4 Scene complexity

The amount of visual information as well as the visual clutter in a picture might contribute to explain the observers' variability [32]. Oliva et al. in [27] determined a list of factors that correlates with our representation of the visual complexity of a scene. Among them, the most important would be the quantity and the variety of objects, detail and color. To assess the visual complexity, three computational measures are used: the entropy, the number of regions and the amount of contours.

4.4.1 Entropy-based scene complexity

To compute the entropy of the incoming picture I, we follow a similar approach of the one described in [32]. The complexity of the scene is the sum of the entropies of wavelet subbands. The procedure is described below:

1. The input picture is first converted into the Lab color space;

2. Each component is transformed by using a 2D dyadic wavelet transform. The level number is set to 2;

3. A non parametric method is used to compute the probability distribution of wavelet coefficients. The entropy for each subband is computed as $E = -\sum_i p_i log(p_i)$, where, p_i is the probability distribution of wavelet coefficients for a given subband.

4. We sum the subband entropies for each component $\{L, a, b\}$;

5. The final complexity C is obtained by using the pooling of [32]: $C = 0.84 \times E_L + 0.08 \times E_a + 0.08 \times E_b$, where E_L, E_a and E_b represent the entropy of the component L, a and b, respectively.

   Figure 8 gives two complexity values C for two pictures.

4.4.2 Color mean shift segmentation

The color mean-shift segmentation has been proposed by [7]. Based on a bilateral filtering, the color mean-shift segmentation associates each pixel of the incoming picture with a significant mode of the joint domain density located in the neighborhood of the considered pixel. The software designed by [4] and available at `http://coewww.rutgers.edu/riul/research/code/EDISON/index.html` is used to perform the segmentation.

The sizes of the filtering kernel (h_s, h_r) (by using the notations of [7]) are both set to 5. M is the minimum number of pixels enclosed by a region (equal to 1 percent of the input resolution).

Figure 8 shows two segmented pictures ((b) and (d)). The number of regions of the segmented picture is also given. This value is used to feature the visual complexity of the scene.

4.4.3 Amount of contours

Edges play an important role in our perception. For instance, Baddeley and Tatler [2] showed that edges correlate with fixation location in real-world scenes better than luminance contrast.

To measure the amount of contours in an image, Sobel edge detectors are used to detect horizontal, vertical and diagonal edges. The kernel has a size of 3×3. These kernels are applied on each level of a Gaussian pyramid. The number of levels is equal to 3. For each level, we compute the average energy of the Sobel detectors (by averaging over the level the squared output of Sobel filter). This process is performed for each level of the pyramid and for the three components $\{L, a, b\}$. For a given component, we combine the energy across scale by taking the average. Finally, to get a single measure, the average energy value is computed across the three components.

5. LEARNING: DESCRIPTION AND PERFORMANCE

5.1 Learning

Each image is then represented by a features vector, having a dimension of 6. The dimensionality of the features vector is not reduced as the number of dimension is low.

The estimation of the inter-observer congruency is equivalent to the estimation of the joint probability density function $p(IOVC, \mathbf{v})$. The random variable $IOVC$ represents the inter-observer visual congruency whereas $\mathbf{v}$ is the feature vector containing the six indicators. To infer the relationship between these two random variables, a learning algorithm is used. We follow the same procedure described in [39, 33] and use software kindly provided by [33]. We just remind the main aspects of this learning procedure.

The learning consists in estimating the relationship between a measure of congruency and the extracted visual features described in the previous section. A cluster-weighted model (CWM) initially proposed by [12] is used. This is a generalization of Gaussian mixture, in which each Gaussian function expressed a part of the relationship between the input and the output distributions. The joint PDF $p(IOVC, \mathbf{v})$ is given by:

$$p_\theta(IOVC, \mathbf{v}) = \sum_{i=1}^{N} p(c_i)p(\mathbf{v}|c_i)p(IOVC|\mathbf{v}, c_i) \qquad (8)$$

where $IOVC$ is the inter-observer congruency and $\mathbf{v}$ refers to the image features. N is the number of clusters. Each cluster is decomposed in three factors:

(a) Original   (b)   Segmented   (c) Original   (d)   Segmented
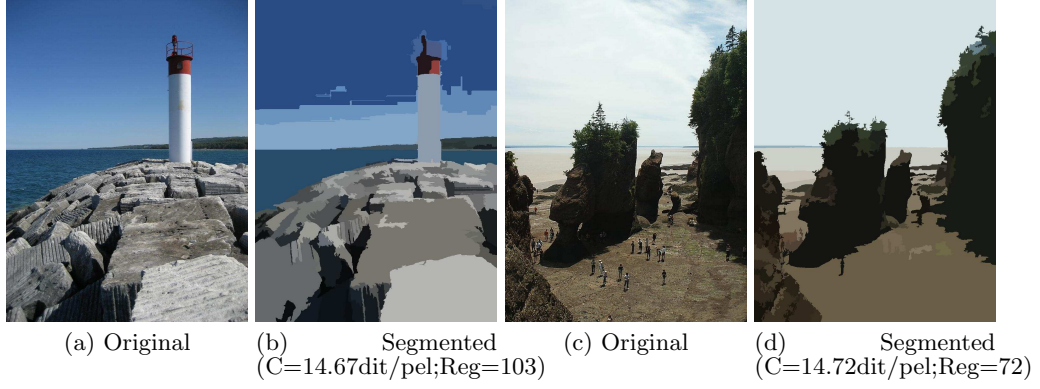(C=14.67dit/pel;Reg=103)   (C=14.72dit/pel;Reg=72)

Figure 8: (a) and (c) two original pictures. (b) and (d) are the segmented pictures. C and Reg stand for the complexity measure and the number of regions, respectively.

- $p(c_i)$ is the weight of the cluster c_i;

- $p(\mathbf{v}|c_i)$ is a multivariate Gaussian with mean μ_i and covariance matrix $\sum_i$:

$$p(\mathbf{v}|c_i) = \frac{exp\left[-\frac{1}{2}(\mathbf{v} - \mu_i)^T (\sum_i)^{-1}(\mathbf{v} - \mu_i)\right]}{(2\pi)^{L/2} \left|\sum_i\right|^{1/2}} \quad (9)$$

- $p(IOVC|\mathbf{v}, c_i)$ is the probability of the inter-observer congruency $IOVC$ given the input data in the cluster i:

$$p(IOVC|\mathbf{v}, c_i) = \frac{exp\left[-\frac{1}{2}(IOVC - w_i^T \mathbf{v}^*)^2\right]}{\sqrt{2\pi}\sigma_i} \quad (10)$$

This is a Gaussian function with a variance equal to σ_i^2 and a mean dependent on the input feature $\mathbf{v}^*$ (same as $\mathbf{v}$ with a value 1 concatenated to its end) and a weight vector w_i. This vector indicates the weight of each input data.

Parameters of the model θ, $(p(c_i)$, μ_i, $\sum_i$, σ_i^2, w_i, with $i = 1...N)$ are estimated using the Expectation-Maximization algorithm [18].

As explained in [15], in data-rich situation, it would be possible to split the data into three parts (a training set, a validation set and a test set). As this is not the case here (1000 pictures), we use the Bayesian Information Criterion (BIC) to define the model complexity. The BIC is given by:

$$BIC = -2 \times loglik + d \times \log S \quad (11)$$

where d is the number of free parameters depending on the number of clusters, S is the size of the dataset and $loglik$ is the maximized log-likelihood:

$$loglik = \sum_{n=1}^{S} \log p_{\widehat{\theta}}(IOVC, \mathbf{v}) \quad (12)$$

where $p_{\widehat{\theta}}(IOVC, \mathbf{v})$ is defined in equation 10. $\widehat{\theta}$ are the estimated parameters of the model.

Figure 9 presents the BIC values in function of N (the number of clusters). $N = 9$ is a good trade-off between complexity and quality of prediction. This value allows to predict quite efficiently the inter-observer congruency without over fitting the training data. Indeed, over fitting the data would lead to an almost perfect prediction but the risk

is to loss the generalization property. As mentioned by [10], it is important to accept error to make less error. By using $N = 9$, we respect this first point. Concerning the quality of prediction, the ground truth and the predicted values of $IOVC$ are correlated $r(2004) = .34$, $p < .001$ (Pearson coefficient) and $r(2004) = .28$, $p < .001$ (Spearman coefficient).
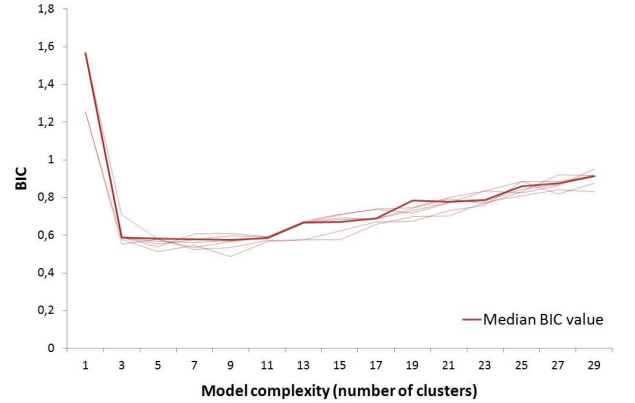

Figure 9: BIC in function of the model complexity. Several trials have been performed (light red curves). The dark red curve gives the median BIC values.

Remark: during the learning phase, we didn't use the face detector in order to limit the impact of false alarms on the estimated parameters. Instead, hand-label data are used indicating for each picture of the dataset the number of faces present.

5.2 Performance

A qualitative and quantitative evaluation of the proposed approach has been performed. Figure 11 presents some qualitative results. Figure 11 (a) gives five pictures having a high IOVC whereas figure 11 (b) illustrates five pictures with a small IOVC. These results are relevant with our own subjective ranking. Indeed, pictures of figure 11 (a) are much more interestingness than those of figure 11 (b). In other words, it would be difficult to predict where an observer would focus on this kind of pictures.

A quantitative evaluation is also performed by using an-

other eye tracking database. This database is composed of 27 pictures and can be downloaded from `http://www.irisa.fr/temics/staff/lemeur/visualAttention/`. We compute the Pearson correlation coefficient between IOVC stemming from this new ground truth and our prediction. Both are correlated $r(54) = .27$, $p < .17$. The correlation is not significant due to the small number of pictures in this database. In addition, the face detector fails to detect the human faces on 5 pictures due to the varying face poses. This lack of accuracy in the detection lowers the correlation coefficient.

The proposed method is compared to the Feature Congestion measure of Rosenholtz et al. [32]. This measure aims to evaluate the visual clutter of a scene. The software available on Rosenholtz's web page is used. We run the Feature Congestion measure on the aforementioned dataset. The correlation coefficient between the Feature Congestion measure and IOVC of this dataset is $r(54) = -0.15$, $p < .43$. The correlation is negative since a high visual clutter might be interpreted as a weak congruency.

5.3 Limitations

The proposed model is relevant in order to predict the dispersion of observers only in free-viewing task. In the introduction, we have dressed a list of factors influencing the dispersion between observers. One factor that was not mention is the task to perform. For instance, if we measure the inter-observer congruency when the task is to detect pedestrians, the inter-observer congruency is very high, indicating that observers share the same strategy to perform the task. To illustrate this point, we compute the inter-observer congruency over the whole eye tracking database of Ehinger [9] (some details are given in the introduction). The average dispersion is of 82%, the median dispersion is of 88%. Compared to the dispersion measured on Judd's database, there is a significant difference (unpaired t-test, $F(1, 1356) = 8.28$, $p < .001$).

Another limitation concerns the influence of the viewing time on the dispersion. It has been shown that the dispersion is time-dependent and increases with the time viewing. This feature is here not taken into account. For the targeted application, this feature was not judged as fundamental.

The last limitation concerns the limited accuracy of the detector we use. More specifically, as the presence of face plays an important role, the face detector has to be as efficient as possible.

6. IMAGE RANKING BASED ON INTERESTINGNESS

The interestingness of an image is related to its ability to attract and to hold our attention. For instance, to give a score of interestingness, Flickr (`http://www.flickr.com`) uses a combination of several parameters such as comments, annotations, favorites, etc. This is an excellent indicator but it requires a feedback or an effort of the users. An indicator based on the content analysis, such as the proposed method, might help evaluating the immediate interest of an image.

The proposed method can then be used in a context of photos browsing and automatic photograph organization. As in [34, 44, 24, 36, 45], we propose to organize a large set of photograph. However the proposed ranking is based on the picture interestingness. This is different from state-of-the-

art methods. For instance, Luo and Tang [24] proposed to rank images according to their quality. This score is based on composition, lighting, focus controlling and color. Although there are some similarities among the extracted features (such as the DoF), better photo quality does not mean more relevant or interestingness, as mentioned in [24]. For instance, Judd et al. [19] show that the dispersion between observers depends on image complexity and that fixations from lower-resolution images (low quality) can predict fixations on higher-resolution images (high quality).

To illustrate the proposed method, we propose to sort forty nine images. We run the proposed model on these pictures in order to estimate their interestingness. Figure 10 illustrates the results by showing the pictures ranked according to their interestingness. The first picture (top-left) has the most important IOVC whereas the picture having the lowest IOVC appears at bottom-right. On the last pictures, we can notice that there is nothing that stands out the background. In other words, it would be very difficult to predict for this kind of picture where an observer would focus on.

7. CONCLUSION

In this paper we proposed a new criterion to automatically estimate the visual congruence between observers. We have evaluated our method qualitatively and quantitatively. We showed that our IOVC criteria outperforms the Feature Congestion measure of [32] since the absolute value of the correlation between the ground truth IOVC and our criterion is larger that with the Feature Congestion measure of [32]. The predicted IOVC can be used in image processing applications where the visual perception of a picture matters such as website design, advertisement. For instance, we considered ranking personalized photograph: pictures are sorted out in function of their predicted IOVC.

However, the proposed method is still an approximation of the 'true' IOVC. It can best estimate short-term IOVC, that is the IOVC experienced in the first instant of a picture observation. In order to improve this method, it would be necessary to consider both higher level factors such as those proposed by [40] and higher level cognitive factors like the scene coherence [16]. Taking into account these factors is difficult because of their complexity.

In future work it is planned to build a new eye tracking database in order to improve the training and the performance assessment. The influence of higher level information such as the type of the scene (indoor, outdoor...) will also be examined. Last but not the least, we will consider more IOVC-based applications. One of them is related to photo-quality assessment as presented in [3, 24]. The use of the predicted IOVC might be combined with other image ranking work.

8. REFERENCES

[1] R. Althoff and N. Cohen. Eye-movement-based memory effect: a reprocessing effect in face perception. *Jounral Of Experimental Psychology-Learning Memory and Cognition*, 25(4):997–1010, 1999.

[2] R. Baddeley and B. Tatler. High frequency edges (but not contrast predict where we fixate: A bayesian system identification analysis. *Vision Research*, 46:2824–2833, 2006.
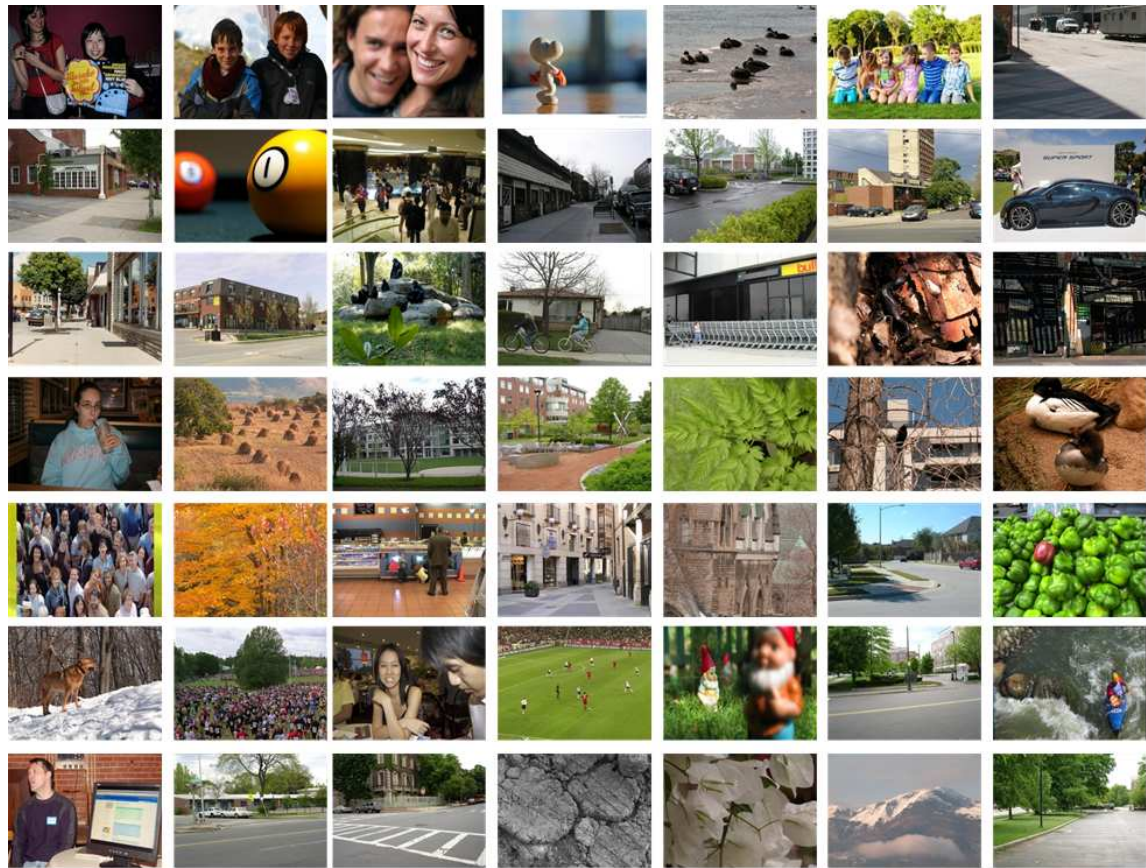
Figure 10: 49 pictures of various contents sorted out in function of their interestingness (from top-left (highest congruency) to bottom-right (lowest congruency)).

[3] S. Bhattacharya, R. Sukthankar, and M. Shah. A coherent framework for photo-quality assessment and enhancement based on visual aesthetics. In *in ACM Multimedia International conference*, 2010.

[4] C. Christoudias, B. Georgescu, and P. Meer. Synergism in low-level vision. In *16th International Conference on Pattern Recognition*, volume IV, pages 105–155, 2002.

[5] H. Chua, J. Boland, and R. Nisbett. Cultural variation in eye movements during scene perception. In *Proceedings of the National Academy of Sciences*, volume 102, pages 12629–12633, 2005.

[6] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y. Xu. Color harmonization. In *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, volume 56, pages 624–630, 2006.

[7] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24:603–619, 2002.

[8] L. Cowen, L. Ball, and J. Delin. An eye-movement analysis of web-page usability. In L. S. V. Ltd, editor, *People and Computers XVI-Memorable yet invisible: Proceedings of HCI 2002*, pages 317–335, 2002.

[9] K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modeling search for people in 900 scenes. *Visual Cognition*, 17:945–978, 2009.

[10] H. Einhorn. Accepting erro to make less error. *Journal of Personality Assessment*, 50(3):387–395, 1986.

[11] H. Frey, C. Honey, and P. Konig. What's color got to do with it? the influence of color on visual attention in different categories. *Journal of Vision*, 8(14), October 2008.

[12] Gershnfel. *The nature of mathematical modelling*. Cambridge, Univ. Press, 1999.

[13] H. Golberg and X. Kotval. Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, 24:631–645, 1999.

[14] R. Gordon. Attentional allocation during the perception of scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 30:760–777, 2004.

[15] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics, 2001.

[16] J. Henderson. Regarding scenes. *Current Directions in Psychological Science*, 16:219–222, 2007.

[17] J. Henderson, M. Chanceaux, and T. Smith. The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, 9, 2009.
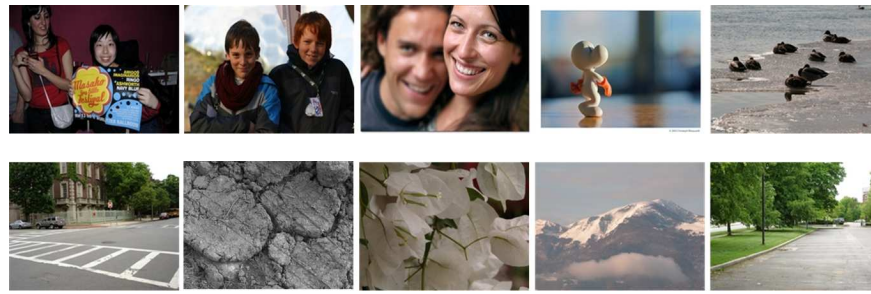
[18] M. Jordan and R. Jacobs. Hierarchical mixtures of

Figure 11: Top: pictures having high IOVC; bottom: pictures having low IOVC.

experts and the em algorihtm. *Neural Computation,* 6:181–214, 1994.

[19] T. Judd, F. Durand, and A. Torralba. Fixations on low-resolution images. *Journal of Vision,* 11(4), 2011.

[20] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where people look. In *ICCV,* 2009.

[21] A. Levin. Blind motion deblurring using image statistics. In *NIPS,* 2006.

[22] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *ICIP,* volume 1, pages 900–903, 2002.

[23] G. Loftus and N. Mackworth. Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performances,* 4:565–572, 1978.

[24] Y. Luo and X. Tang. Photo and video quality evaluation: focussing on the subject. In *ECCV,* pages 386–399, 2008.

[25] Y. Matsuda. Coor design. In *Asakura Shoten,* 1995.

[26] R. Nisbett. *The geography of thought: how Asians and Westerners think differently... and why.* New York: Free Press, 2003.

[27] A. Oliva, M. Mack, M. Shrestha, and A. Peeper. Identifying the perceptual dimensions of visual complexity of scenes. In *26th annual meeting of the Cognitive Science Society Meeting,* 2004.

[28] D. Parkhurst, K. Law, and E. Niebur. Modelling the role of salience in the allocation of overt visual attention. *Vision Research,* 42:107–123, 2002.

[29] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C: the art of Scientific Computing.* Cambridge University Press, New York, NY, USA, 1992.

[30] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin,* 124(3):372–422, 1998.

[31] K. Rayner, M. Catelhano, and J. Yang. Eye movements when looking at unusual-weird scenes: are there cultural differences? *Journal of Experimental psychology: learning, Memory and cognition,* 35(1):154–259, 2009.

[32] R. Rosenholtz, Y. Li, and L. Nakano. Measuring visual clutter. *Journal of Vision,* 7(2):1–22, 2007.

[33] M. Ross and A. Oliva. Estimating perception of scene layout properties from global image features. *Journal Of Vision,* 1:1–25, 2010.

[34] C. Rother, L. Bordeaux, Y. Hamadi, and A. Black. Autocollage. In *in ACM Transactions on Graphics (SIGGRAPH),* 2006.

[35] G. A. Rousselet, M. J.-M. Macé, and M. Fabre-Thorpe. Is it an animal? is it a human face? fast processing in upright and inverted natural scenes. *Journal of Vision,* 3:440–455, 2003.

[36] X. Sun, H. Yao, R. Ji, and S. Liu. Photo assessment based on computatinal visual attention model. In *ACM Multimedia,* pages 541–544, 2009.

[37] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research,* 45:643–659, 2005.

[38] M. Tokumaru, N. Muranaka, and S. Imanishi. Color design support system considering coor harmony. In *IEEE International Conference on Fuzzy Systems,* pages 378–383, 2002.

[39] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE Pattern Analysis and Machine Intelligence,* 24(9):1226–1238, 2002.

[40] A. Torralba and A. Oliva. Statistics of natural image catagories. *network,* 14:391–421, 2003.

[41] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review,* 113(4):766–786, 2006.

[42] G. Underwood and T. Foulsham. Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *The Quarterly journal of experimental psychology,* 59(11):1931–1949, 2006.

[43] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR,* 2001.

[44] M. Wallick. *Automatic Organization of Large Photograph Collections.* PhD thesis, University of Wisconsin-Madison, 2007.

[45] C.-G. Yeh, Y. Ho, B. Barsky, and M. Ouhyoung. Personalized photograph ranking and selection system. In *ACM Multimedia,* 2010.

[46] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision,* 11(3):1–15, 2011.