PREDICTING SALIENCY USING TWO CONTEXTUAL PRIORS: THE DOMINANT DEPTH AND THE HORIZON LINE

O. Le Meur (olemeur@irisa.fr)

University of Rennes 1 (IRISA-TEMICS)
Campus de Beaulieu 35042 Rennes, France.

ABSTRACT

A computational model of visual attention using visual inferences is proposed. The dominant depth and the horizon line position are inferred from low-level visual features. This prior knowledge helps to find salient areas on still color pictures. Regarding the dominant depth, the idea is to favor the lowest spatial frequencies on close-up scenes whereas the highest spatial frequencies are used to predict salient areas on panoramic view. Some studies showed that the horizon line is a natural attractor of our gaze. Horizon detection is then used to improve the saliency prediction. Results show that the proposed model outperforms existing approaches. However, the dominant depth does not bring any gain in the saliency prediction.

Index Terms— visual attention, contextual guidance, saliency, dominant depth, horizon line.

1. INTRODUCTION

Since 1998 with the publication of the influential work of Itti, Kock and Niebur [1], the computational modelling of the visual attention has known an increasing interest. Several computational models have been proposed and may be grouped into three categories. The first one may concern computational models that are purely bottom-up; they are solely based on the low-level visual features without taking into account neither visual inferences nor top-down mechanisms. Two seminal studies are at the origin of these models: the biologically plausible architecture for controlling bottom-up attention proposed by Koch and Ullman [2] and the Feature Integration Theory of Treisman and Gelade [3] positing that the visual processing is able to encode in a parallel manner visual features such as color, form, orientation, and others. As mentioned previously, the first model was proposed by Itti et al. [1]. Several features maps are extracted from different visual dimensions at different levels of resolution. A center-surround filter was then applied to determine in each feature map the most salient locations. The final saliency map was then obtained by combining the filtered feature maps. In the same vein, Le Meur et al. [4] proposed a model based on contrast sensitivity function and visual masking. Some parts of this model will be re-used in this study (see section 2.1). Another solution to compute the saliency is to use a probabilistic framework, assuming that a rare event attracts our visual attention. An elegant solution to simulate this behavior consists in using the self information or the mutual information of Shannon's information theory [5, 6, 7, 8].

A second category may concern more recent modelling that endeavor to take into account some visual inferences. Most of researchers are now convinced that the low-level visual features are not sufficient to predict accurately the gaze location. Higher-level factors, such as the visual inferences, play a determinant role in the gaze allocation [9]. As Rao et al. [10] proposed, we believe that prior knowledge is a key concept that can shape the bottom-up saliency. The most recent contribution related to this work has been proposed by Judd et al. [11]. In this work, several visual features are combined by using a learning algorithm. The list of visual features involved in the learning is composed of Felzenszwalb car and person detectors, Viola Jones Face detector, horizon line detector and features used by Itti's model.

Finally, the last category could be related to visual attention models dedicated to the solving of a particular visual task [12, 13]. These models are not described here since they are not in the scope of the study.

In this paper, a computational model of visual attention belonging to the second aforementionned category is proposed. As mentioned before, prior knowledge can shape the bottom-up saliency. Among the different factors influencing our gaze, we are interested in two cues: the dominant depth and the horizon line position. In Section II, the proposed bottom-up model is presented. The method to determine the dominant depth information as well as the position of the horizon line is briefly described in section III. Section IV presents how we propose to use these prior knowledge in the bottom-up model. Section V presents the performance and section VI concludes the paper.

2. BOTTOM-UP MODEL OF VISUAL ATTENTION

The proposed model is close to the one proposed in [4]. There are some minor modifications and one major concerning the computation of the final saliency map (pooling of intermediate maps). The synoptic is shown on figure 1.

2.1. Early visual features extraction

The feed-forward model is inspired by Le Meur's model [4]. The input picture with a 256×256 resolution is first projected into an opponent color space (Lab). A Fourier transform is applied on these components. The amplitude Fourier spectrum of the achromatic component is split into seventeen subbands spread over 4 crowns. For the chromatic decomposition, only two crowns (5 subbands) are used. Each crown has its own spatial frequency range and a given angular selectivity orientation. More details are given in [4]. The same process, described below, is applied on each component c ($c \in \{L, a, b\}$). A center-surround filter is applied on each subband in order to simulate responses of retinal ganglion cells and lateral geniculate cells. A Difference-Of-Gaussian (DoG) filter [14] (with a center On or Off) can be used for this purpose. A filtered subband $\widehat{SB}_{(r,\theta)}$ at the spatial position (x, y) of the component c is then given by:

$$\widehat{SB}^{c}_{(r,\theta)}(x,y) = \max(SB^{On-Off}_{(r,\theta)}(x,y), SB^{Off-On}_{(r,\theta)}(x,y)) \tag{1}$$

where,

$$SB^{On-Off}_{(r,\theta)}(x,y) = \max(0, SB_{(r,\theta)}(x,y) * DoG_{\sigma_c,\sigma_s}(x,y)) \tag{2}$$

$$SB^{Off-On}_{(r,\theta)}(x,y) = \max(0, SB_{(r,\theta)}(x,y) * DoG_{\sigma_s,\sigma_c}(x,y)) \tag{3}$$

The Difference-Of-Gaussian DoG is given by:

$$DoG_{\sigma_1,\sigma_2}(x,y) = G_{\sigma_1}(x,y) - G_{\sigma_2}(x,y) \tag{4}$$

where G_σ is a bi-dimensional Gaussian function with a standard deviation σ. σ_1 and σ_2 represent the standard deviations of the center and surround Gaussian functions. As proposed by Marr and Hildret [14], a ratio of 1.6 is used between both standard deviation ($\sigma_2 = 1.6 \times \sigma_1$).

2.2. Computation of the saliency map

From all these filtered subbands, a final unique saliency map has to be deduced. In a context of free-viewing, severals approaches have been proposed [1, 15, 16]. All of them use the low-level visual features without taking into account prior knowledge. Two pooling methods, a simple or a more complex, are proposed. Before giving details about these pooling methods, it is important to mention that all subbands have been upsampled and normalized in order to have homogeneous spatial resolutions and range of greyscale values. The simple pooling, called SP, is given by:

$$\begin{aligned} S(x,y) = & \max_{c=\{L,a,b\}} (SM^c(x,y)) \\ & + \beta \times log(1 + \prod_{c=\{L,a,b\}} (1 + SM^c(x,y)) \end{aligned} \tag{5}$$

where β is a constant ($\beta = 20$ for this study). The intermediate saliency maps SM^c are given by

$$SM^c(x,y) = \sum_i \alpha_i \times \max_\theta \left(\widehat{SB}^{c}_{i,\theta} \right) \tag{6}$$

with $SB^c_{i,\theta}$ is a subband belonging to the i^{th} crown with an angular selectivity θ of the component c. α_i are the weighing coefficients, set to 1 by default.
The complex pooling, called CP, is given by the following equation. This is inspired by Harel et al.'s method [17]. The idea is to compute the dissimilarity between a current point (x, y) and all others points. The sum of these dissimilarities indicates the saliency of the point (x, y):

$$S(x,y) = \sum_{c=\{L,a,b\}} \sum_{p=0}^{M-1} \sum_{m=0}^{N-1} d\left(SM^c(x,y), SM^c(p,m)\right) \tag{7}$$

Where $d(.)$ is the Euclidean distance. M and N are the size of the picture.

3. VISUAL INFERENCES

Our visual perception is a complex process that results from the combination of prior beliefs and inferences with low-level visual features stemming from the environment. Most of the visual attention models focus on visual data we gather from the visual field. To go further, we propose to infer from the low-level visual features contextual information such that the dominant depth and position of the horizon line (if any):

Dominant depth: several studies support the hypothesis that there are separate neural pathways for processing information about different visual properties [3]. These properties would be processed very quickly and unconsciously. The depth feature is one of them. We are indeed able to perceive the depth from our visual field in an effortlessly manner. The most striking is that, even when we look at a picture, we are able to extrapolate it. As depth information is quickly available, this prior knowledge might affect eye movements. For instance, depth might contribute to an early recognition of the scene layout. In addition, from the knowlegde of the dominant depth value, the average size of salient areas might be inferred. This property is used in the final pooling in other to favor a spatial frequency range. The tested assumption is that salient features are more likely to be present in low spatial frequencies for close-up scenes. For panoramic scenes, it might be more interesting to consider high frequencies than low spatial ones.

Horizon line: in a recent study, Foulsham et al. [18] provided evidence that the natural horizon line systematically attracts our visual attention. In this study, authors recorded the gaze allocation for five stimuli presenting the same visual scene. The difference is that these stimuli are the result of a rotation of 0, 45, 90,135 and 180 degrees. The distribution

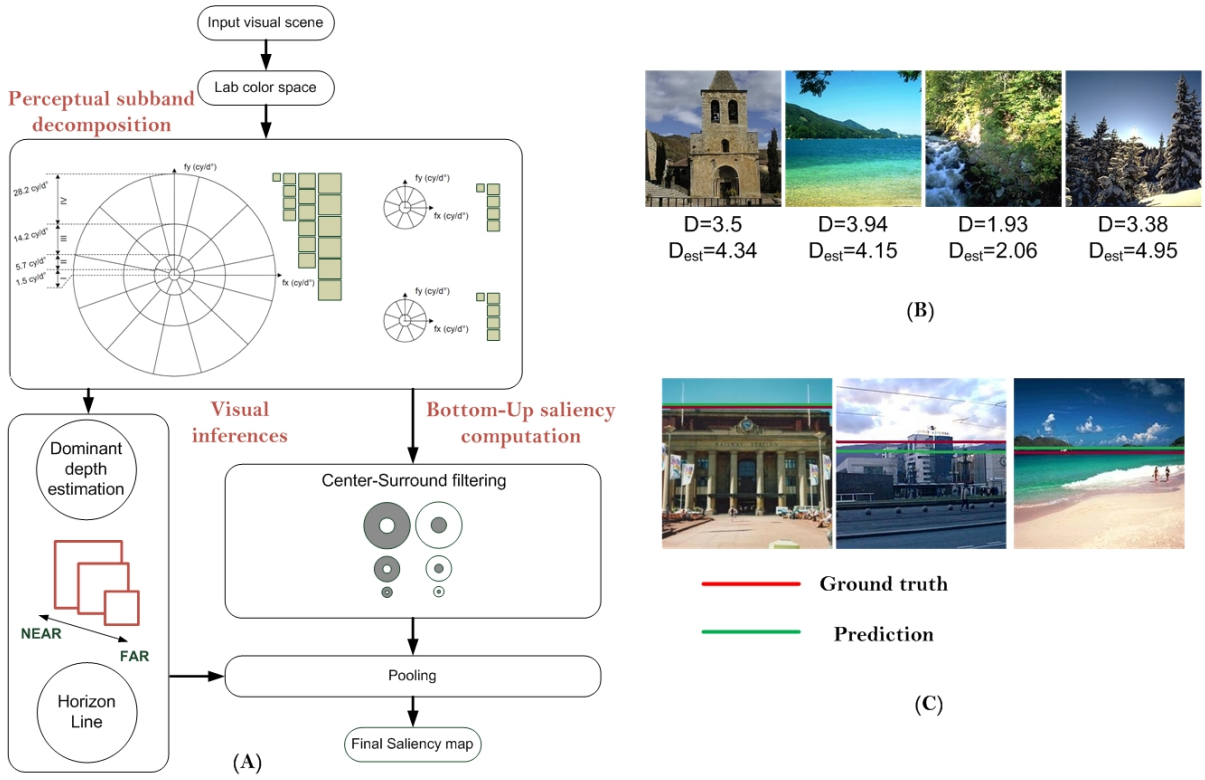

Fig. 1. (A) Synoptic of the proposed approach. The input visual scene is projected into the Lab color space. A hierarchical decomposition is performed in the Fourier domain. Depth and the position of the horizon line (if any) are used to adapt the saliency computation. (B) Inference of the dominant depth based on a machine learning (D and D_{est} represent the ground truth and the prediction of the dominant depth, respectively). (C) Inference of the position of the horizon line.

of saccade directions is elongated along the axis of the natural horizon, whatever the rotation angle. The detection of the horizon line will then be used as a contextual prior in order to bias the computation of the saliency map.

3.1. Dominant depth

As in [19], the mean depth is inferred from the low-level visual features. These features are extracted from the hierarchical decomposition of the achromatic component presented in the section 2.1. Each spatial subband is split into 9 blocks. For each block, the average spatial energy is computed. This is defined as follows:

$$E_{(r,\theta)}(k) = \frac{1}{card(\mathcal{B}_{r,k})} \sum_{(i,j)\in\mathcal{B}_{r,k}} SB^{L}_{(r,\theta)}(x,y) \qquad (8)$$

where $\mathcal{B}_{r,k}$ is the considered block indexed by k for the radial spatial frequency r. Note that as the subbands SB don't have the same resolution, the local energy is computed over blocks having different sizes. The more the spatial frequency r, the more the block size is important. For a picture having a resolution of 256×256 pixels, the block sizes range from 4×4, 17×17, 42×42 and 85×85 for the first, the second, the third and the fourth frequential crowns, respectively. This choice is rather arbitrary and far to be biologically plausible. However, this blockification is very convenient from a computational point of view. In addition, past studies have successfuly used this approach in a context of scene recognition [20] or depth estimation [19, 21]. Each scene is then represented by a feature vector, having a dimension of 153 (17 subbands multiplied by 9). Principal components analysis (PCA) was used to reduce the dimensionality of the features vector while preserving 90% of the variance. The new feature vector $\mathbf{v}$ has 30 dimensions and is composed of the projection of the local energies onto the first 30 eigenvectors.

The estimation of the dominant depth is equivalent to the estimation of the joint probability density function $p(D, \mathbf{v})$. The random variable D represents the depth whereas $\mathbf{v}$ stands for the information about the spatial layout of the scene. To infer the relationship between these two random variables, a learning algorithm is used. We followed the same procedure described in [19, 21] and used data provided by [21][1]. The main aspects of the learning procedure is reminded below.

The learning consists in estimating the relationship between

[1] See `http://cvcl.mit.edu/layout/`

a subjective rating (the depth) and objectives measurement (local energies). A cluster-weighted model (CWM) initially proposed by [22] 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 ouptut distributions. The joint PDF $p(D, \mathbf{v})$ is given by:

$$p(D, \mathbf{v}) = \sum_{i=1}^{N} p(cl_i)p(\mathbf{v}|cl_i)p(D|\mathbf{v}, cl_i) \tag{9}$$

where D is the mean depth and $\mathbf{v}$ refers to the image features. N is the number of clusters. Each cluster i is decomposed into three factors:

- $p(cl_i)$ is the weight of the cluster cl_i;
- $p(\mathbf{v}|cl_i)$ is a multivariate Gaussian with mean μ_i and covariance matrix $\sum_i$:

$$p(\mathbf{v}|cl_i) = \frac{exp\left[-\frac{1}{2}(\mathbf{v}-\mu_i)^T(\sum_i)^{-1}(\mathbf{v}-\mu_i)\right]}{(2\pi)^{L/2}\left|\sum_i\right|^{1/2}} \tag{10}$$

- $p(D|\mathbf{v}, cl_i)$ is the probability to have the depth D given the input data in the cluster i:

$$p(D|\mathbf{v}, cl_i) = \frac{exp\left[-\frac{1}{2}(D - w_i^T\mathbf{v}^*)^2\right]}{\sqrt{2\pi}\sigma_i} \tag{11}$$

  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, $p(cl_i)$, μ_i, $\sum_i$, σ_i^2, w_i, with $i = 1...N$ are estimated using the Expectation-Maximization algorithm [23]. The training data set is composed of the 1380 pictures stemming in part from the work of [21] and in other part from personal pictures. For each picture, an average dominant depth score going from 1 to 6 (from near to far) is given [21]. The pictures of the training data set are uniformly distributed on the continuous scale. The dominant depth scores were used to train the model described above. For the training, $N = 20$ clusters and the first 30 eigenvectors of the PCA were used. The number of clusters was chosen based on a trade-off between complexity, mean squared error and linear correlation value. We choose twenty clusters to achieve a good trade-off between the quality of prediction and the complexity. Figure 1 (B) gives the predicted depth for some images.

3.2. Horizon line

As for the previous detector, a CWM learning is perfomed to infer the horizon line position. The five lowest subbands of the component L and the five subbands of the blue component (the negative values of the component b) are used in the learning. These subbands are again split into 9 blocks. The average spatial energy, given by equation (8), is computed for each block. A PCA was used to reduce the number of features from 90 to the first 50 eigenvectors (97% of the variance). A CWM learning was performed with 5 clusters. The learning involved 213 natural outdoor visual scenes. For each scene, the position of the horizon line was manually set. A scalar value indicates the row index where the horizon line is located: zero means that the line is at the top of the scene. Figure 1 (C) illustrates this point.

4. VISUAL ATTENTION MODEL BASED ON VISUAL INFERENCES

4.1. Depth-based pooling

The dominant depth gives an information about the size of the salient area. Indeed, when the dominant depth is high (panoramic view), we can assume that the salient areas will be small. It is then more appropriate to favor the medium to high frequencies subbands to the detriment to the lowest spatial frequencies. When the dominant depth is small, the picture is likely a close-up and it is appropriate to use the lowest spatial frequencies in order to compute the saliency map. The dominant depth is then used to favor the scales of the achromatic decomposition for which it is likely to find a salient area. Equation (6) is simply modified by adjusting the coefficient α_i to the dominant depth.

4.2. Facilitation based on the spatial position of the horizon line

Previous studies [18] demonstrated that there is a strong systematic tendency to look at the natural horizon. Therefore, the use of horizon detection in a context of visual attention modelling is then important to provide contextual priors to bias the saliency map to certain locations. We simply propose to weight the final saliency map in function of the spatial position of the horizon line. The weighting function is given by:

$$W(x, y) = exp\left(-\frac{(y-h)^2}{2\sigma_h^2}\right) \tag{12}$$

where, h is the predicted position of the horizon line and σ_h a parameter to control the spread of the weighting. For pictures having a resolution of 256×256, σ_h is equal to 75.

5. PERFORMANCES

To measure how well the proposed model predicts fixation locations on a given image, a ROC analysis is performed. In this kind of analysis, the saliency maps, whether it be predicted or not, are considered as a binary classifier. Each pixel

of the map is then labeled as being salient or not. Two sets of threshold are required, a first for the human saliency maps and a second for the predicted ones. The former set of threshold is defined in order to obtain 10, 20, 30 and 40 percent of salient areas. To threshold the predicted saliency maps, 128 thresholds, uniformly distributed, are used. For each pair of thresholds, the true positive and the false positive rates are computed. A ROC curve is obtained by varying the different thresholds. The Area Under Curve (AUC) is a good metric indicating the degree of similarity between the human and the predicted saliency maps. In this paper, Bruce and Tsotsos's database of visual fixations (120 natural color pictures, see [5]) is used. Results are given in figure 2. We make the following observations:

The best results are achieved when luminance and chrominance components are used. This is consistent with previous studies (such as [24]) indicating that the best model is the model combining all visual features.

However, the pooling of the intermediate feature maps plays a fundamental role. The complex pooling (called CP) largely outperforms the simple pooling (called SP). It is not surprising since the complex pooling is based on both local information (the saliency at a given location) and global information (dissimilarity between local and global information (see equation (7))). For example, when images are thresholded at 20% salient, the model with luminance and chrominance (with a complex pooling) performs at 76% while the model using a simple pooling is at 62%.

Without using the dominant depth or horizon detection, the best results are given by the model called *CP(Uniform Weighting)Luma-Chroma*. The uniform weighting means that the coefficients α_i (used to compute the achromatic saliency map) are uniformly distributed (see equation 6). When the dominant depth is used in order to adapt these coefficients, AUC values systematically decrease. A number of coefficient set has been tested without success. These results are not given on figure 2 for the sake of visibility. The assumption that a given spatial frequency range can be favored in function of the dominant depth seems to be a wrong idea. It might indicate that the different subbands containing complementary and redundant information are important in the computation of final saliency map. However, it is important to underline, before drawing a definitive conclusion, that Bruce's database might not be well adapted. Indeed, 75% of the pictures have a dominant depth between 2 and 4. This fact might explain why an uniform weighting gives the best performance.

Contrary to the dominant depth, results are improved by the use of the facilitation induced by the position of the horizon line. The overall performance of the model with horizon detection is indeed slighty above the model called *CP(Uniform Weighting)Luma-Chroma*. When horizon line is detected in an image (51 detections over 120 pictures), the median gain (see the bottom-right of figure 2) is equal to 2%.

Compared to existing approaches, the best proposed model performs better than Itti's and Bruce's model. For instance, at the 30% salient location threshold, the model with the complex pooling and the horizon detection performs at 0.76 (0.75 without horizon detection) whereas the AUC values is of 0.68 and 0.71 for Itti's[2] and Bruce's model, respectively.

6. CONCLUSION

In the proposed model, saliency is based on low-level visual features combined with the extraction of global features. They provide layout information and contextual priors to bias the saliency map to certain locations.

We use machine learning to train models to infer the dominant depth and the position of the horizon line. These inferences are based on the low-level visual features. The proposed model is compared to both purely bottom-up model and existing models. We found that the dominant depth doesn't bring any improvement when compared to a naive model. Regarding the horizon line, the median gain is of 2% on pictures for which there is an horizon line. Compared to existing approaches, the proposed model performs at 0.76 while Itti's and Bruce's model are at 0.68 and 0.71, respectively. Supplementary materials are available on `http://www.irisa.fr/temics/staff/lemeur/`.

Future studies will focus on other visual inferences. The contextual guidance plays an important role in the gaze allocation. By using the same approach and framework, the type of the scene and the gist are the next prior knowledge that we would like to use in order to improve the relevance of this kind of model.

7. ACKNOWLEDGEMENT

This work is supported by the French national program CONTINT through the project entitled PERSEE.

8. REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model for saliency-based visual attention for rapid scene analysis," *IEEE Trans. on PAMI*, vol. 20, pp. 1254–1259, 1998.

[2] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[3] A. Treisman and G. Gelade, "A feature integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[4] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model the bottom-up visual attention," *IEEE Trans. on PAMI*, vol. 28, pp. 802–817, 2006.

[5] N.D.B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*, 2006, vol. 18, pp. 155–162.

[2] SaliencyToolbox 2.2 `www.saliencytoolbox.net`

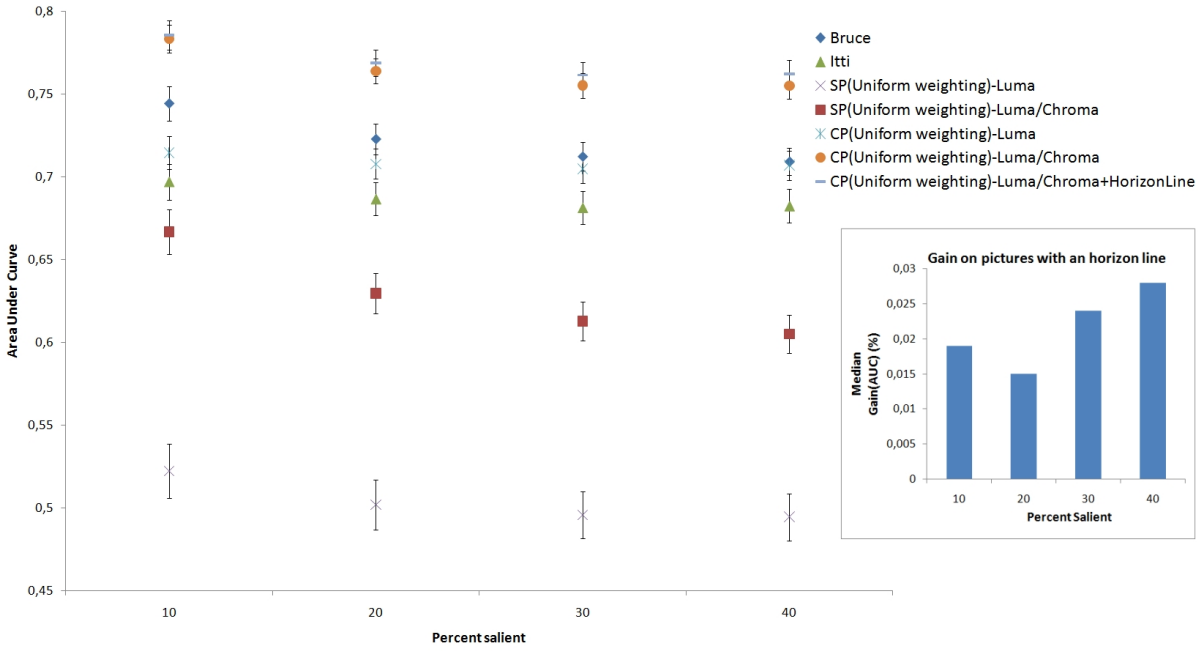

Fig. 2. AUC value for the different versions of the proposed model. The performance of Itti's and Bruce's model is also given. On the bottom-up right, the median gain brings by the use of the horizon line is shown. For pictures having an horizon line, the gain is about 2%.

[6] N.D.B. Bruce and J.K. Tsotsos, "Saliency, attention and visual search: an information theoretic approach," *Journal of Vision*, vol. 9, pp. 1–24, 2009.

[7] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G. W. Cottrell, "Sun: a bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, pp. 1–20, 2008.

[8] D.Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, pp. 1–13, 2008.

[9] J.M. Henderson, "Regarding scenes," *Current Directions in Psychological Science*, vol. 16, no. 4, pp. 219–222, 2007.

[10] R. P. N. Rao, G.J. Zelinsky, and M.M. Hayhoe D.H. Ballard, "Eye movements in iconic visual search," *Vision Research*, pp. 1447–1463, 2002.

[11] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *ICCV*, 2009.

[12] A. Oliva, A. Torralba, M.S. Castelhano, and J.M. Henderson, "Top-down control of visual attention in object detection," in *IEEE ICIP*, 2003.

[13] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, pp. 205–231, 2005.

[14] D. Marr and E. Hildreth, "Theory of edge detection," in *Proceedings of the Royal Society of London. Series B, Biological Sciences (The Royal Society)*, 1980.

[15] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, pp. 2483–2498, 2007.

[16] A. Bur and H. Hugli, "Optimal cue combination for saliency computation: a comparison with human vision," *in Lecture Notes in computer science, Springer-Verlag GmbH*, vol. 4528, pp. 109–118, 2007.

[17] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2006.

[18] T. Foulsham, A. Kingstone, and G. Underwood, "Turning the world around: Patterns in saccade direction vary with picture orientation," *Vision Research*, vol. 48, pp. 1777–1790, 2008.

[19] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Trans. On PAMI*, vol. 24, no. 9, pp. 1226–1238, 2002.

[20] A. Torralba and A. Oliva, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[21] M.G. Ross and A. Oliva, "Estimating perception of scene layout properties from global image features," *Journal Of Vision*, vol. 10, no. 1, pp. 1–25, 2010.

[22] N. Gershnfeld, *The nature of mathematical modelling*, Cambridge, Univ. Press, 1999.

[23] M.I. Jordan and R.A. Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

[24] O. Le Meur and J.C. Chevet, "Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks," *IEEE Trans. On Image Processing*, vol. 19, no. 11, pp. 2801–2813, 2010.