ROBUSTNESS AND REPEATABILITY OF SALIENCY MODELS SUBJECTED TO VISUAL DEGRADATIONS

Olivier Le Meur

ROBUSTNESS AND REPEATABILITY OF SALIENCY MODELS SUBJECTED TO VISUAL DEGRADATIONS

Olivier Le Meur

University of Rennes 1
Campus Universitaire de Beaulieu
35042 Rennes Cedex France

ABSTRACT

The present study investigates the sensitivity of computationnal models of visual attention when subjected to visual degradations. One hundred and twenty natural color pictures were degraded using 6 filtering operations. By using different settings, five state-of-the-art models are used to compute 11400 saliency maps. The comparison of these maps to human saliency maps indicates that the tested models are robust to most of the visual degradations they were subjected to. These findings have implications on saliency-based applications, such as quality assessment and coding. A last point concerns the high repeatability of saliency models that might be used in a context of image retrieval.

***Index Terms*—** Saliency, visual degradations, transformations, robustness, repeatability.

1. INTRODUCTION

Visual attention allows people to select and process some particular areas of our visual field. This selection is influenced by a number of factors. In general, we make a distinction between bottom-up and top-down factors. The former rely on the signal whereas the latter concern the task to be performed. Other factors such as the context or the visual inferences should be mentioned. Wolf and Horowitz [1] have dressed a list of attributes that might guide the deployment of visual attention. One factor, the quality of the perceived scene, is not mentioned. A recent study [2] gives evidences and strengthens the hypothesis that the quality of the visual scene does not significantly influence the way we watch a video clip in free-viewing task. Another study previously showed the same trend on still color pictures for a JPEG and JPEG2000 compression [3][1]. The invariance of saliency models subjected to degradation is important especially in the context of quality assessment and compression. Indeed the design of quality metric integrates more and more properties of the human visual system. For instance computational models of visual attention are used to steer the pooling of errors [4]. Errors on salient areas are deemed to be more significant than those located on non-salient areas. However, there are a number of issues. Among them, one concerns the sensitivity of saliency models to degradations. In other words, if the original picture is not available, does it make sense to use the impaired one to compute the saliency map? Concerning the compression, the fact that saliency models would be invariant to degradations might be interesting. Indeed, in this case, it would not be required to transmit the saliency map (or the positions of salient areas) throughout a network. The saliency map could be computed at the decoder side. This invariance property might also be used in a context of saliency-based cropping. Rather than transmitting the spatial coordinates of the cropping window, these data might be re-computed at the user-side. The aim of this paper is then to test whether computation models of visual attention are sensitive to visual degradations. To this end, five state of the art models are put to the test. The basic idea is to compare predicted saliency maps to human saliency maps. Predicted saliency maps are obtained from an original picture and its transformed or impaired version. This paper is organized as follows. Section II presents the database of still pictures as well as the transformation and visual degradations used in this study. Section III presents first the intrinsic performance of the five state-of-the-art models. Then the sensitivity of these models to visual degradation is examined. The last point of Section II concerns the repeatability of saliency models. Section IV will conclude the paper.

(a) Original (b) Human saliency map

Fig. 1. Human saliency map (b) for the picture (a). Bright areas are the most attractive.

2. DATABASE AND LIST OF DEGRADATION

Bruce and Tsotsos's database of visual fixations [5] is used throughout this work. This database is composed of a variety of real color images (120), ranging from natural outdoor scenes to indoor scenes. It can be downloaded at `http://www-sop.inria.fr/members/Neil.Bruce/`. An original picture with its associated human saliency map is depicted on figure 1. Different transformations and image processing filtering are applied on each picture. Figure 2 illustrates some of them on a given picture. The degradation operations are listed below:

- Blur: a Gaussian kernel of size 11×11 is used. Three values of variance are used: 1, 3 and 8. Obviously, the bigger the variance value the greater the smoothing produced;
- Uniform variation of illumination: the RGB components of the pictures are weighted by a fixed coefficient (0.2, 0.6, 1.4,

[1] Available on `http://www.irisa.fr/temics/staff/lemeur/`, see the dedicated page on this topic.

1.8). Coefficients greater than 1 tend to lighten the picture whereas a coefficient less than 1 darkens the picture;

- Gaussian noise: an independent Gaussian noise is added to the original image. The noise is with zero mean and a variance equal to 0.001, 0.01, 0.05 or 0.1. The bigger the variance the more the image is noisy;
- Flip: original pictures are flipped in right/left and up/down directions;
- Rotation: a rotation of the pictures is performed by an angle of 90, 180 and 270 (anti-clockwise). The rotation center is the picture's center. The invariance of models to rotation is interesting to investigate. Indeed Foulsham et al. recently have given evidences of a strong systematic tendency for saccades to occur along the axis of the natural horizon, whatever the picture orientation [6];
- JPEG coding: a JPEG coding is applied on the original picture. The compression is performed by using the software *XnView*. Three quality factors (Q) are used: 40, 10 and 1. A small quality factor indicates a strong compression (or a low quality). For the smallest quality factor, strong block effects appear on the pictures, as illustrated by figure 2 (j).

A total of 2280 pictures (19 kinds of degradations multiplied by 120 pictures) is obtained. 2280×5 saliency maps plus those corresponding to original pictures have been computed, for all 5 models.

3. RESULTS

3.1. State-of-the-art models of visual attention

Five state-of-the-art computational models are used in the test. These models involve different strategies to compute the saliency. The three first models, Itti [7], Le Meur [8] and Bruce [9], rely on two seminal works: the biologically plausible architecture for controlling bottom-up attention proposed by [10] and the Feature Integration Theory [11] positing that the visual processing is able to encode in a parallel manner visual features such as color, form, orientation, and others. The major difference between Bruce's model and the others is that a probabilistic framework is used to derive the saliency. A fourth model [12] is based on a supervised learning of eye tracking data. Compared to the previous ones, this model includes higher-level information such as the position of the horizon line, human face, a detector of cars and pedestrians and a feature indicating the distance to the center for each pixel. Finally, the last one [13] is based on the low pass filtering of the log-spectrum of the input picture.

3.2. Performance assessment

To measure the degree of similarity between saliency maps, a ROC (Receiver Operating Characteristic) 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. One threshold and a set of threshold are required: a first for the human saliency maps and a second for the predicted ones. The former is defined in order to obtain 30 percent of salient areas. Different thresholds allowing to obtain 10, 20 and 40 percent of salient areas have been tested leading to the same conclusion. 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 finally calculated in order to gauge the degree of similarity between the human and the predicted saliency maps.

Table 1. AUC values (average$\pm$SEM and median) for the five models (the human saliency maps were threshold to keep the top 30 percent of salient pixels). (SEM stands for Standard Error of the Mean).

Model	average$\pm$SEM	median
Itti	0.68$\pm$0.009	0.69
Le Meur	0.73$\pm$0.01	0.75
Bruce	0.72$\pm$0.009	0.73
Hou	0.68$\pm$0.01	0.69
Judd	0.76$\pm$0.007	0.77

Table 2. Paired t-test, 95%. (ns) means that results are not statistically different; (*) and (**) mean that the difference is statistically significant ($p < 0.05$ and $p << 0.001$ respectively).

Model	Itti	Le Meur	Bruce	Hou	Judd
Itti	–	**	*	ns	**
Le Meur		–	*	**	**
Bruce			–	**	**
Hou				–	**
Judd					–

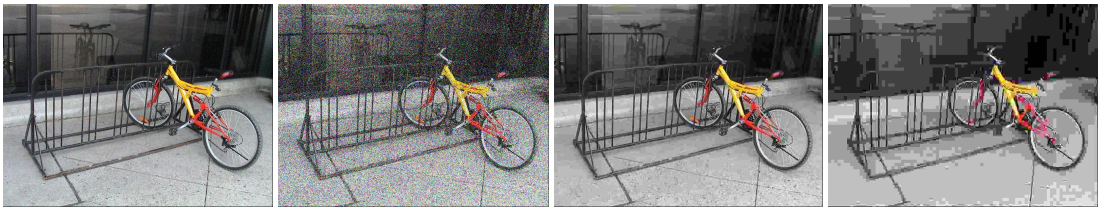
3.3. Performance of state-of-the-art models on Bruce's database

Table 1 gives the performance of the five saliency models in term of AUC values, when human saliency maps are clipped at 30% salient. Judd's model provides the best performance by performing at 76%. As indicated by Table 2, this model significantly outperforms all others. This strong performance is likely due to the specialized detectors (Face, pedestrians, horizon line, car detectors) embedded in the model. This model also uses a central weighing to take central biasing into account[8, 14].
The two next best models are Le Meur and Bruce's models performing at 73% and 72%, respectively whereas Itti and Hou's model perform both at 68%. The same ranking is observed when human saliency maps are clipped at 10, 20 and 40% salient.

3.4. Robustness to visual degradations

Figure 3 gives the AUC values for each kind of degradation. On each subfigure, the performance of the five models on original pictures is also recalled.
Results indicate that the five tested models are invariant to rotation and flip operations. It is however important to underline that other rotation angles such as 45, 135, etc should be tested. These rotation angles will bring new difficulties since it is required either to crop the resulting picture or to enlarge it. It is however possible to speculate on the results. For instance, Le Meur's model uses an anisotropic contrast sensitivity function (CSF), favoring the horizontal and vertical structures. Concerning other models, there are no obvious reasons to be sensitive to these rotation angles. Models are robust to blur and variation of luminance, except for Le Meur's model. Indeed its performance drops down with an increasing blur. The loss of AUC is about 3%. Concerning the luminance variation, Le Meur's model totally fails to predict salient areas for the lowest luminance. It is probably due to the fact that the viewing distance is set to $6H$

(a) Original (b) High blur (c) Under exposure (d) Over exposure



(e) Small noise (f) Strong noise (g) JPEG (Q=10) (h) JPEG (Q=1)

Fig. 2. Examples of some degradations applied on the pictures.

by default (H being the height of the screen). This parameter acts on the CSF. The farther the observer is from the screen the less likely he is to see small features. Conversely, Le Meur's model performs better when the luminance increases. The gain is about 3%. When the input picture is corrupted by Gaussian noise, the performance of Judd, Le Meur, Hou and Bruce's model decreases with the amount of noise. The performance loss is more pronounced for Le Meur's model than others. Whatever the amount of noise, Itti's model performs at 68% showing its strong robustness to noise. Except for high compression ratio, the JPEG compression does not have a significant influence. A compression with a quality factor greater than 10 does not impact the salient prediction. Bruce's model is the less robust for small compression ratio. This could be explained by looking at the architecture of the models. The five models share almost the same architecture by processing separately the low-level visual features in different channels (Note that Hou's model only uses the luminance). Promoting the highest contrast in each channel might be a good solution to be less sensitive to impairments. In addition, as these models are not sensitive to rotation and flip, they all have an isotropic behaviour. However, it would be interesting to test other rotation angle, such as 45, 135...

3.5. Repeatability of saliency models

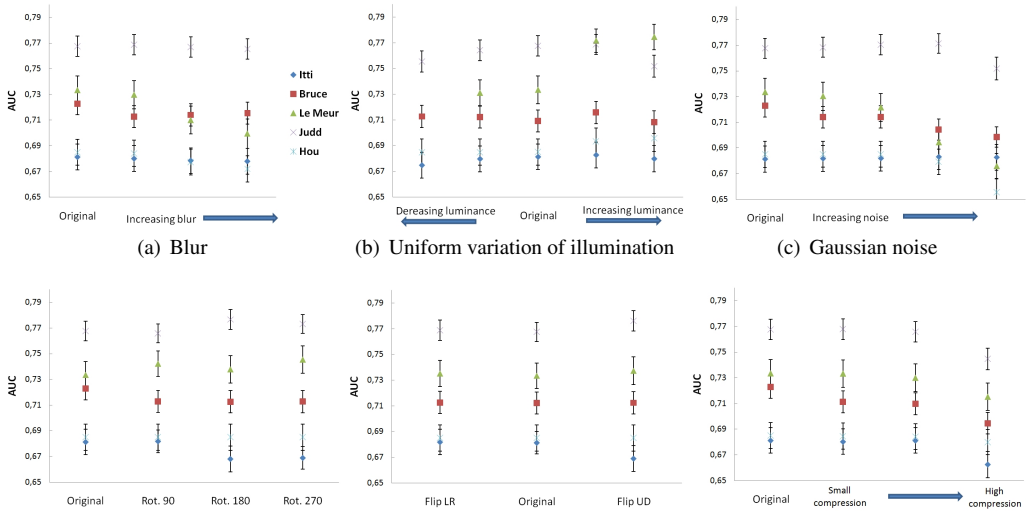
In the previous paragraph, the ability of saliency models to detect salient areas in different impairment conditions has been shown. In this section, the repeatability of saliency models is examined in a similar fashion to [15]. The aim is to establish if the predicted saliency maps depend on visual impairments. From a qualitative point of view, the predicted saliency maps computed on impaired pictures are very similar to the saliency maps computed on unimpaired images. This feeling is confirmed by the ROC analysis as explained below. To quantify this invariance, a ROC analysis is performed between the predicted saliency maps stemming from the original pictures and those stemming from the impaired pictures. Table 3 gives AUC values for each model and for the highest blur, noise, JPEG and luminance degradations. The average AUC value is very high for all models. It indicates that the predicted saliency maps are very similar (almost the same) whatever the visual degradations. It can be concluded that the repeatability of the saliency model is very good. For the highest degradations, the lowest AUC value is equal to 0.82, that is still a good similarity indicator between predicted saliency maps. There is no value of repeatability for Le Meur's model for the same reasons given in the previous section.

Table 3. Repeatability of saliency models for the highest degradations (average AUC values between predicted saliency maps (computed from original and impaired pictures)). Averages over degradations and models are given in the two last lines.

Model	Blur	Luminance		Noise	JPEG
	$\sigma^2 = 8$	Under	Over	$\sigma^2 = 0.1$	$Q = 1$
Itti	0.99	0.94	0.92	0.96	0.94
Le Meur	0.91	–	0.92	0.84	0.91
Bruce	0.97	0.99	0.82	0.95	0.95
Hou	0.99	0.99	0.88	0.88	0.96
Judd	0.97	0.88	0.92	0.86	0.90
Avg/Deg.	0.96	0.95	0.89	0.89	0.93
Avg/Model	0.95	0.89	0.93	0.94	0.90

4. DISCUSSION AND CONCLUSION

Overall, the results indicate that the tested computational models of visual attention are almost invariant to visual degradations. Although some studies [3, 2] have given some evidences of the invariance of the visual attention to visual degradations, it is too early to say whether the invariance of saliency models reflects human behaviors or not. The best and most robust model is Judd's model [12]. Compared to other tested models, this model uses low-level, medium-level and high-level information. Pedestrians, cars, horizon lines, face detection and a central weighting function are at the heart of the model. This is probably why it outperforms the others. The only questionable point about this model concerns the central weighting. It has been shown recently that this central biasing is a laboratory artifact that arises from the onscreen presentation of visual scenes [14]. Therefore, the use of such central weighting would artificially increase the quality of the prediction and does not reveal



(a) Blur (b) Uniform variation of illumination (c) Gaussian noise

(d) Rotation (e) Flip (f) JPEG compression

Fig. 3. AUC values for the different models and degradations. Error bars correspond to SEM.

a real occulomotor behavior. It is also worth noting that the conclusions put forth in this paper have an impact on the use of saliency map in different applications. For instance, in a context of quality assessment (for reduced or no reference metric), the error map can be indifferently weighted by the saliency map stemming from either the original or the impaired one. For applications that would require saliency information at the user-side, it is not necessary to transmit the corresponding saliency map. It can be safely recomputed at the user end, thus saving bandwidth. Another conclusion is related to the repeatability of these models (i.e. their ability to produce the same predicted saliency map whatever the visual degradations). This property of repeatability might be used to choose the points of interest (such as Harris) that are salient. The idea would be to define salient points of interest. Another axis for future studies will focus on video sequences. It could be investigated whether we observe the same invariance on impaired video clips.

5. ACKNOWLEDGEMENT

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

6. REFERENCES

[1] J.M. Wolfe and T.S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?," *Nature Reviews Neuroscience*, vol. 5, pp. 495–501, June 2004.

[2] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Do video coding impairments disturb the visual attention deployment?," *Elsevier, Signal Processing: Image Communication*, vol. 25, no. 8, pp. 597–609, 2010.

[3] O. Le Meur, *Attention sélective en visualisation d'images fixes et animées affichées sur écran: modeles et évaluation de performances - applications*, Ph.D. thesis, University of Nantes, 2005.

[4] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric," in *ICIP*, 2007.

[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.

[6] 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.

[7] 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.

[8] 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, no. 5, pp. 802–817, May 2006.

[9] 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.

[10] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

[11] A.M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[12] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[13] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *CVPR*, 2007.

[14] M. Bindemann, "Scene and screen center bias early eye movements in scene viewing," *Vision Research*, vol. 50, no. 23, pp. 2577–87, November 2010.

[15] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *IJCV*, vol. 60, no. 1, pp. 63–86, 2004.