Published in
Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach.
(Series: Lecture Notes in Computer Science) 3562, 469-478, 2006
which should be used for any reference to this work

Model Performance for Visual Attention in Real 3D Color Scenes

Heinz Hügli, Timothée Jost, and Nabil Ouerhani

Institute of Microtechnology,
University of Neuchâtel,
Rue A.-L. Breguet 2,
CH-2000 Neuchâtel, Switzerland
{Heinz.Hugli, Nabil.Ouerhani}@unine.ch

Abstract. Visual attention is the ability of a vision system, be it biological or artificial, to rapidly detect potentially relevant parts of a visual scene. The saliency-based model of visual attention is widely used to simulate this visual mechanism on computers. Though biologically inspired, this model has been only partially assessed in comparison with human behavior. The research described in this paper aims at assessing its performance in the case of natural scenes, i.e. real 3D color scenes. The evaluation is based on the comparison of computer saliency maps with human visual attention derived from fixation patterns while subjects are looking at the scenes. The paper presents a number of experiments involving natural scenes and computer models differing by their capacity to deal with color and depth. The results point on the large range of scene specific performance variations and provide typical quantitative performance values for models of different complexity.

1 Introduction

Visual attention is the ability of a system, be it biological or artificial, to analyze a visual scene and rapidly detect potentially relevant parts on which higher level vision tasks, such as object recognition, can focus. On one hand, artificial visual attention exists as the implementation of a model on the computer. On the other hand, biological visual attention can be read from human eye movements. Therefore, the research presented in this paper aims at assessing the performance of various models of visual attention by comparing the human and computer behaviors.

It is generally agreed nowadays that under normal circumstances human eye movements are tightly coupled to visual attention. This can be partially explained by the anatomical structure of the human retina. Thanks to the availability of sophisticated eye tracking technologies, several recent works have confirmed this link between visual attention and eye movements [1, 2, 3]. Thus, eye movement recording is a suitable means for studying the temporal and spatial deployment of visual attention in most situations.

In artificial vision, the paradigm of visual attention has been widely investigated during the last two decades, and numerous computational models of visual attention have been suggested. A review on existing computational models of visual attention is available in [4]. The saliency-based model proposed in [5] is now widely used in numerous software and hardware implementations [6, 7] and applied in various fields.

However, and despite the fact that it is inspired by psychophysical studies, only few works have addressed the biological plausibility of the saliency-based model [8]. Parkhurst et al [9] presented for the first time a quantitative comparison between the computational model and human visual attention. Using eye movement recording techniques to measure human visual attention, the authors report a relatively high correlation between human attention and the saliency map, especially when the images are presented for a relatively short time of few seconds. Jost et al [10] run similar experiments on a much larger number of test persons and could measure the quantitative improvement of the model when chromaticity channels are added to the conventional monochrome video channels. Visual attention in 3D scenes was first considered in [11] and recently, a visual attention model for 3D was quantitatively analyzed in presence of various synthetic and natural scenes [12].

This paper presents a more global analysis, where the performance of a family of visual attention models in presence of 3D color scenes is evaluated. The basic motivation is to get insight into the contribution of the various channels like color and depth. Another motivation is to assess possible improvements when artificial visual attention is made more complex.

The remainder of this paper is organized as follows. Chapter 2 recalls basics of the saliency models. Chapter 3 presents the methods for acquiring the human fixation patterns and comparing them to the saliency map. Chapter 4 details the experiments and obtained results. A general conclusion follows in Chapter 5.

2 Saliency Models

The saliency-based visual attention [5] operates on the input image and starts with extracting a number of features from the scene, such as intensity, orientation chromaticity, and range. Each of the extracted features gives rise to a conspicuity map which highlights conspicuous parts of the image according to this specific feature. The conspicuity maps are then combined into a final map of attention named saliency map, which topographically encodes stimulus saliency at every location of the scene. Note that the model is purely data-driven and does not require any a priori knowledge of the scene.

2.1 Feature and Conspicuity Maps

From a scene defined by a color image (R, G, B) and a range image Z, a number of features F_j are extracted as follows:
Intensity feature $F_1 = I = 0.3 \cdot R + 0.59 \cdot G + 0.11 \cdot B$.

Four features F_2, F_3, F_4, F_5 for the local orientation according to the angles $\theta \in \{0^o, 45^o, 90^o, 135^o\}$.
Two chromaticity features F_6, F_7 based on the two color opponency components R^+G^- and B^+Y^- defined with the help of the yellow component Y as follows:

$$Y = \frac{R+G}{2} \qquad F_6 = \frac{R-G}{I} \qquad F_7 = \frac{B-Y}{I} \tag{1}$$

Depth feature represented by a depth map $F_8 = Z$.
Each feature map is then transformed into its conspicuity map C_j which highlights the parts of the scene that strongly differ, according to the feature specificity, from their surroundings. The computation of the conspicuity maps noted $C_j = T(F_j)$ relies on the center-surround mechanism, a multiscale approach and a normalization and summation step during which, the maps from each scale are combined, in a competitive way, into the feature-related conspicuity map C_j.

2.2 Cue Maps

Given the nature of the different features, the model groups together conspicuities belonging to the same category and we define cue conspicuity maps for intensity (int), orientation (orient), chromaticity (chrom.) and range as follows:

$$\hat{C}_{int} = C_1; \;\; \hat{C}_{orient} = \sum_{j \in \{2,3,4,5\}} N(C_j); \;\; \hat{C}_{chrom} = \sum_{j \in \{6,7\}} N(C_j); \;\; \hat{C}_{range} = C_8 \tag{2}$$

where $N(.)$ is a normalization operator which simulates the competition between the different channels. A detailed description of the normalization strategy is given in [6].

2.3 Saliency Map

Finally, the cue maps are integrated, in a competitive manner, into a universal saliency map S as follows:

$$S = \sum_{cue} N(\hat{C}_{cue}) \tag{3}$$

More specifically, in this study we work with three alternative saliency maps of in-creasing complexity, namely:

- A greyscale saliency map S_{grey} that includes intensity and orientation: $S_{grey} = N(\hat{C}_{int}) + N(\hat{C}_{orient})$.
- A color saliency map S_{color} that includes intensity, orientation and chromaticity: $S_{color} = N(\hat{C}_{int}) + N(\hat{C}_{orient}) + N(\hat{C}_{chrom})$.
- A depth saliency map S_{depth} that includes intensity, orientation, chromaticity and range: $S_{depth} = N(\hat{C}_{int}) + N(\hat{C}_{orient}) + N(\hat{C}_{chrom}) + N(\hat{C}_{range})$.

3 Comparing Computer and Human Visual Attention

The evaluation principle illustrated in figure 1 is based on the comparison of the computed saliency map with human visual attention. Under the assumption that under most circumstances, human visual attention and eye movements are tightly coupled, the deployment of visual attention is experimentally derived from the spatial pattern of fixations.

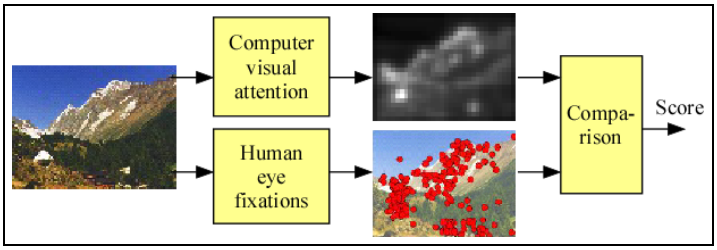

Fig. 1. Comparison of computer and human visual attention

3.1 Eye Movement and Fixation Pattern Recording

Eye movements were recorded with an infrared video-based tracking system (Eye-Link^{TM}). It has a temporal resolution of 250 Hz, a spatial resolution of 0.01^o, and a gaze-position accuracy relative to the stimulus position of $0.5^o - 1.0^o$, largely dependent on subjects' fixation accuracy during calibration. As the system incorporates a head movement compensation, a chin rest was sufficient to reduce head movements and ensure constant viewing distance.

A considerable challenge of this research has been to record eye movements while a subject is watching a stereo image. It was made possible with the use of an autostereoscopic display. It avoids using glasses on the subject, which would prevent eye movement tracking. The images were presented in blocks of 10. Each image block was preceded by a 3×3 point grid calibration scheme. The images were presented in a dimly lit room on the autostereoscopic 18.1" CRT display (DTI 2018XLQ) with a resolution (in stereo mode) of 640×1024, 24 bit color depth, and a refresh rate of 85 Hz. Active screen size was 36×28.5 cm and viewing distance 75 cm, resulting in a viewing angle of 29×22^o. Every image was shown for 5 seconds, preceded by a center fixation display of 1.5 seconds. Image viewing was embedded in a recognition task.

Eye monitoring was conducted on-line throughout the blocks. The eye tracking data was parsed for fixations and saccades in real time, using parsing parameters proven to be useful for cognitive research thanks to the reduction of detected microsaccades and short fixations (< 100 ms). Remaining saccades with amplitudes less than 20 pixels (0.75^o visual angle) as well as fixations shorter than 120 ms were discarded after-wards [10].

For every image and each subject i, the measurements yielded an eye trajectory T^i composed of the coordinates of the successive fixations f_k, expressed as image coordinates (x_k, y_k):

$$T^i = (f_1^i, f_2^i, f_3^i, ...) \tag{4}$$

3.2 Score *s*

The score s is used as a metric to compare human fixations and computer saliency maps. Also called chance-adjusted saliency by Parkhurst et al. [9], the score s corresponds to the difference of average values of two sets of samples from the computer saliency map $S(x)$. Formally:

$$s = \frac{1}{N} \sum_{f_k \in T} S(f_k) - \mu_S \tag{5}$$

The first term corresponds to the average value of N fixations f_k from an eye trajectory T^i . The second term μ_S is the saliency map average value. Thus the score measures the excess of salience found at the fixation points with respect to arbitrary points. If the human fixations are focused on the more salient points in the saliency map, which we expect, the score should be positive. Furthermore, the better the model, the higher the probability to reach the points with highest saliency and the higher this score should be.

4 Experiments and Results

The experimental process was divided into two parts. A first part is devoted to the measurement of visual attention induced by 2D images. A second part compares human visual attention in presence of 3D color scenes.

4.1 Dataset 2D

This dataset consists of 41 color images containing a mix of natural scenes, fractals, and abstract art images (see figure 2). Most of the images (36) were shown to 20 subjects. As stated above, these images were presented to the subjects for 5 seconds apiece, resulting in an average of 290 fixations per image.

4.2 Dataset 3D

This dataset consists of 12 3D scenes representing quite general natural scenes. Each scene is represented by a stereo image pair. Figure 3 presents sample images from this dataset. These image pairs were presented to 20 different subjects for 5 seconds apiece, resulting in an average of 290 fixations per image.

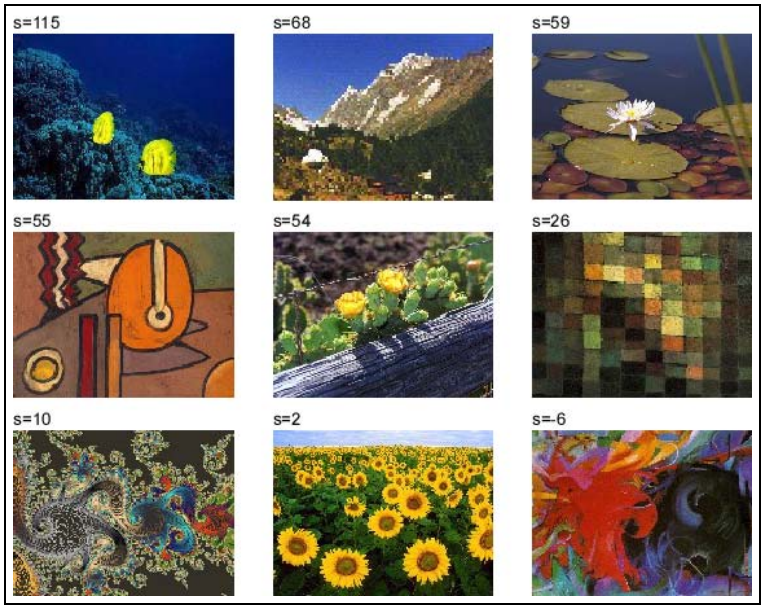

Fig. 2. images from the dataset 2D, ranked by score for the color model

Fig. 3. Sample scenes from the dataset 3D

4.3 Performance in Presence of 2D Images

For all images of dataset 2D, we created a greyscale saliency map S_{grey} and a color saliency map S_{color}, both normalized to the same dynamic range. Then, a comparison of these two models with the whole set of human fixation patterns was performed in order to obtain the respective scores. Note that the score s was computed taking the first 5 fixations of each subject into account, since it has been suggested that, with regard to human observers, initial fixations are controlled mainly in a bottom-up manner [10].

Figure 4 shows the scores for the different individual images. The main observation here is that the resulting scores are widely spread in their value, covering the range [-7 .. 115]. The values show the model performance depends in a strong way on the kind of image. To illustrate these results and explain somehow these

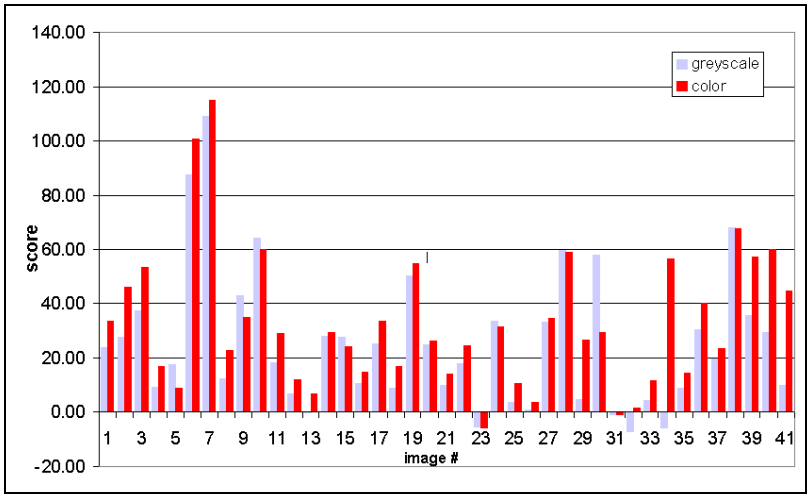

Fig. 4. Individual scores for images of dataset 2D, both for the greyscale and color models

strong variations, we refer to figure 2 showing sample images from the dataset 2D. There, the images are ordered according to the score S_{color} obtained by each image. The image yielding the best results is top left. The score decreases from left to right and top to bottom. It is apparent that the images found on the top row generally contain few and strong salient features, such as the fish, the small house or the water lily. They yield the best results. On the other hand, images that lack highly salient features, such as the abstract art or the fractal images on the bottom row, result in much lower scores. Here, the model loses its effectiveness in the single image (out of 41) yielding a negative score.

Referring to performance of the models, it is expected that the color model performs better because it includes the additional chromaticity cue. We therefore expect the score for the color model to be at least as good as the score of the greyscale model. Although this is not true for all images it is the case for a majority of about 85% of the cases.

A general comparison is given in table 1 showing the estimated average model scores. The standard error was computed using the variance from both random picks and human fixations means. The main observation is that the color model fares better than the greyscale one. More specifically, the color model yields an average score 25.8% higher than the greyscale model. This underlines the

Table 1. Scores of the greyscale and color models

	score s
greyscale model S_{grey}	24.8 ± 1.2
color model S_{color}	31.2 ± 1.1

usefulness of the chromaticity cue in the model and goes toward assessing that this cue has a considerable influence on visual attention.

4.4 Performance in Presence of 3D Scenes

For all scenes of dataset 3D, we created a color saliency map S_{color} and a depth saliency map S_{depth} , both normalized to the same dynamic range. Then, a comparison of these two models with the whole set of human fixation patterns was performed in order to obtain the respective scores. The score s was computed as in previous experiments.

Figure 5 shows the scores for the 12 individual images. The main observation here is that the resulting scores are widely spread in their value [5 .. 76]. The effect is the same as in previous experiments and above comments keep their full validity here. It shows again that the model performance depends in a strong way on the kind of scene.

Referring to the model performance, table 2 presents the average scores s over the whole dataset, for both the color and the depth models. The standard error was computed as above. The main observation is that the depth model fares better than the color one. More specifically, the depth model yields an average score s that is 11.8% better than the color model. This general result underlines

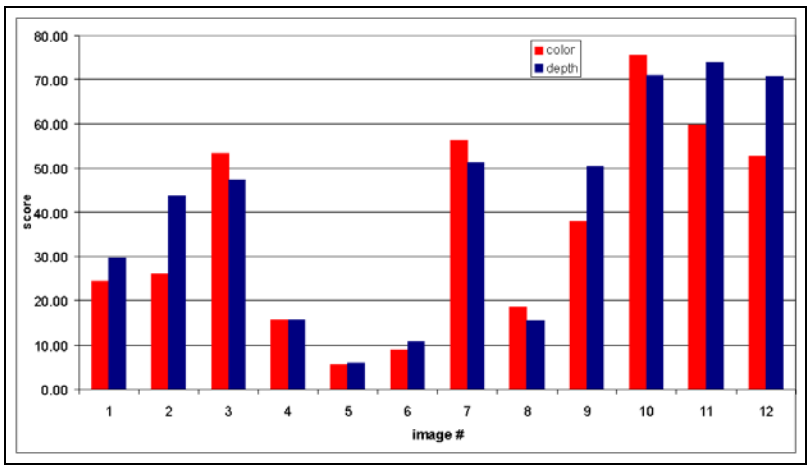

Fig. 5. Individual scores for images of dataset 3D, both for the color and depth models

Table 2. Scores of the color and depth models

	score s
color model S_{color}	36.2 ± 2.1
depth model S_{depth}	40.5 ± 2.1

the usefulness of the depth channel in the model and goes toward assessing that depth contributes to the visual attention process.

5 Conclusion

The research described in this paper aims at assessing the performance of the saliency model of visual attention in the case of natural scenes. The evaluation is based on the comparison of computer saliency maps with human visual attention derived from fixation patterns while subjects are looking at the scenes of interest.

A new aspect of this research is the application to 3D color scenes. In this respect, this study provides for the first time quantitative performance comparisons of different models of visual attention, giving new insights to the contribution of some of its components, namely color and depth.

The experiments involved test persons watching at 3D scenes generated by stereo-vision. An autostereoscopic display was used so that stereo image pairs could be shown to the subjects while recording their eye movements. A first series of experiments refers to the performance in presence of color images. It involves 40 images of different kinds and nature. A second series refers to the performance in presence of color 3D scenes. It involves 12 scenes of different kinds and nature. The number of test persons is 20 in each case.

The eye saccade patterns were then compared to the saliency map generated by the computer. The comparison provides a score (s), i.e. a scalar that measures the similarity of the responses. The higher the score, the better the similarity and the better the performance of the attention model for predicting human attention.

The experiments provide scores covering a wide range of values, i.e. the range is [-5 .. 120] for the images and [5...75] for the 3D scenes. These large score variations illustrate the strong dependence on the kind of scenes: Visual attention of some scenes is very well predicted by the model, while the prediction is quite poor in some cases. These results confirm previous understanding of the model capability and earlier measurements on smaller datasets.

Beyond these large variations, the study shows significant performance differences between the three investigated models. The model performance increases with the model complexity. The performance is first increased when passing from the basic greyscale model to the color model. This is quantitatively assessed by a score increase of 25%. A further performance increase, assessed by a score increase of 11%, characterizes the model extension to depth.

The study therefore confirms the feasibility of a quantitative approach to performance evaluation and provides a first quantitative evaluation of specific models differing by their capacity to deal with color and depth.

Acknowledgements

This work was partially supported by the Swiss National Science Foundation under grant FN 64894. We acknowledge the valuable contribution of René Müri

and Roman von Wartburg from University of Berne, Switzerland, who provided the eye saccade data.

References

1. A. A. Kustov and D.L. Robinson. Shared neural control of attentional shifts and eye movements. *Nature, Vol. 384, pp. 74-77*, 1997.
2. D.D. Salvucci. A model of eye movements and visual attention. *Third International Conference on Cognitive Modeling, pp. 252-259*, 2000.
3. C. Privitera and L. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *Pattern Analysis and Machine Intelligence (PAMI), Vol. 22, No. 9, pp. 970-981*, 2000.
4. D. Heinke and G.W. Humphreys. Computational models of visual selective attention: A review. *In Houghton, G., editor, Connectionist Models in Psychology*, in press.
5. Ch. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, Vol. 4, pp. 219-227*, 1985.
6. L. Itti, Ch. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 20, No. 11, pp. 1254-1259*, 1998.
7. N. Ouerhani and H. Hugli. Real-time visual attention on a massively parallel SIMD architecture. *International Journal of Real Time Imaging, Vol. 9, No. 3, pp. 189-196*, 2003.
8. N. Ouerhani, R. von Wartburg, H. Hugli, and R. Mueri. Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis (ELCVIA), Vol. 3 (1), pp. 13-24*, 2004.
9. D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research, Vol. 42, No. 1, pp. 107-123*, 2002.
10. T. Jost, N. Ouerhani, r. von Wartburg, R. Muri, and H. Hugli. Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding Journal (CVIU)*, to appear.
11. N. Ouerhani and H. Hugli. Computing visual attention from scene depth. *International Conference on Pattern Recognition (ICPR'00), IEEE Computer Society Press, Vol. 1, pp. 375-378*, 2000.
12. T. Jost, N. Ouerhani, r. von Wartburg, R. Muri, and H. Hugli. Contribution of depth to visual attention: comparison of a computer model and human. *Early cognitive vision workshop, Isle of Skye, Scotland, 28.5. - 1.6*, 2004.