Contribution of depth to visual attention: comparison of a computer model and human behavior

Timothee Jost[1], Nabil Ouerhani[1], Roman von Wartburg[2], René Müri[2], Heinz Hugli[1]

[1] Institute of microtechnology*
University of Neuchâtel
CH-2000 Neuchâtel
Switzerland

[2] Department of Neurology
University of Bern
Inselspital, CH-3010 Bern
Switzerland

Abstract

The research described in this paper aims at assessing the contribution of depth to visual attention. It reports the measurement of depth induced human visual attention derived from fixation patterns and preliminary results of a quantitative comparison with visual attention as modeled by different versions of a computational model. More specifically, a two-cues color model is compared to a three-cues color and depth model. The preliminary conducted experiments give the first hints of the quantitative contribution of depth features in visual attention.

1. Introduction

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 paradigm of computational visual attention has been widely investigated during the last two decades, and numerous computational models of visual attention have been suggested. For a more complete overview on existing computational models of visual attention, the reader is referred to [HeHu].

The saliency-based model of visual attention was proposed by Koch and Ullman in [Koch85]. It is based on four major principles: visual attention acts on a multi-featured input; saliency of locations is influenced by the surrounding context; the saliency of locations is represented on a scalar map (the saliency map); and the winner-take-all and inhibition of return mechanisms are suitable means to provide the locations for consecutive attentional shifts. A brief summary of the saliency based model is found in chapter 2.

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 [Par02, OuWa04]. So far, these works considered only 2D scene features as source of visual attention.

The work presented in the present paper goes further and aims at analyzing the saliency-based model by quantitatively assessing the contribution of depth in visual attention from 3D scenes. Our hypothesis is that depth contrast contributes to visual attention similarly to intensity contrast for example and that, consequently, a model including a depth based feature channel globally fares better in predicting human fixations than a model not including one.

In a first step we measured human fixations while subjects were looking at pure depth images in an attempt to show the relation between depth (or disparity) contrast and human visual attention (sec 5.1).

Then, we compared human fixations derived from eye movement experiments with the computational maps of attention produced by two different versions of the saliency-based model, using the metrics found in section 4. To this end, stereo image pairs were presented to human subjects while their eye movements were recorded, providing information about the spatial locations of foveated image parts, as well as the duration of each fixation, as presented in section 3.

2. Computer visual attention

The model starts with extracting a number of features from the scene, such as color, intensity, depth and orientation. Each of the extracted features gives rise to a conspicuity map which highlights conspicuous parts of the image according to this specific feature. The cue-related conspicuity maps $\hat{C}_{cue}$ are integrated, in a competitive manner, into a saliency map *S*:

$$S = \sum_{cue=1}^{m} N(\hat{C}_{cue}) \qquad \textbf{(1)}$$

where *m* is the number of the considered cues and $N(.)$ the normalization operator.

* name.surname@unine.ch

Several works have dealt with the realization of this model [Mil93, Itt98]. In our work, we used an implementation of the saliency-based model of visual attention that was inspired by these works and which considers also 3D scene features [Oue00].

3. Human visual attention

Under the assumption that under most circumstances, visual attention and eye movements are tightly coupled, the deployment of human visual attention is experimentally derived from the spatial pattern of fixations.

A considerable challenge of the research has been to be able to record eye movements while a subject is watching a stereo image. It was made possible with the use of an autostereoscopic display. Such a screen uses a special illumination plate located behind the LCD that permits to display both halves of a stereo pair simultaneously and to direct them to corresponding eyes. This allows avoiding using glasses on the subject, which would prevent eye movement tracking.

Eye movements were recorded with a infrared video-based tracking system (EyeLink™, SensoMotoric Instruments GmbH, Teltow/Berlin). This system consists of a headset with a pair of infrared cameras tracking the eyes, and a third camera monitoring the screen position in order to compensate for any head movements.

Every stereo image was shown for 5 seconds, preceded by a center fixation display of 1.5 seconds. For every image and each subject i, the measurements yielded an eye trajectory T^{i} composed of the coordinates of the successive fixations $\boldsymbol{f}_k$, expressed as image coordinates (x_k, y_k):

$$\mathrm{T}^{\mathrm{i}} = \left(\boldsymbol{f}_1^i, \boldsymbol{f}_2^i, \boldsymbol{f}_3^i, ..., \boldsymbol{f}_k^i, ...\right) \qquad \textbf{(2)}$$

As a global representation of the set of all fixations $\boldsymbol{f}_k^i$, a human saliency map $H(\boldsymbol{x})$ was computed, under the assumption that this map is an integral of weighted point spread functions $h(\boldsymbol{x})$ located at the positions of the successive fixations. It is assumed that each fixation gives rise to a normally (gaussian) distributed activity. The width σ of the activity patch was chosen to approximate the size of the fovea. Formally, $H(\boldsymbol{x})$ is computed according to:

$$H(\boldsymbol{x}) = H(x, y) = \sum_{i=1}^{N_{subj}} \sum_{\boldsymbol{f}_k \in T^i} \exp\left(\frac{(x_k - x)^2 + (y_k - y)^2}{\sigma^2}\right) \qquad \textbf{(3)}$$

4. Comparison metrics

Two different metrics are considered in order to compare human fixations and computer saliency maps: a correlation and a score measurement.

The first metric is simply defined by the correlation coefficient ρ between the computational saliency map $S(\boldsymbol{x})$ and human saliency map $H(\boldsymbol{x})$ defined above. The value of ρ lies in the [-1, 1] interval. A value of 1 indicates that both maps are exactly similar, a value of 0 indicates that both maps are totally different and a value of -1 indicates that the two maps are anti-correlated, i.e. that a salient feature in one map is not salient in the other one.

The score s, also called chance-adjusted saliency by Parkhurst *et al.* [Par02] is defined by the following.

$$\mathrm{s} = \frac{1}{\mathrm{N}} \sum_{\boldsymbol{f}_k \in T} S\left(\boldsymbol{f}_k\right) - \mu_S \qquad \textbf{(4)}$$

where μ_S is the mean value of the map $S(\boldsymbol{x})$. It basically measures the distance between average random picks and picks made at human fixation positions in a saliency map. 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 similarity and the higher this score should be.

The principal difference between the correlation ρ and the score s is that the first is more “global” than the score s and is independent regarding the scaling of the considered saliency map, while the later avoids relying on parameters (i.e. σ as used to create the human map).

5. Experiments and results

A first part is devoted to the measurement of visual attention induced by pure depth. A second part compares human visual attention with a computer model including, or not, a depth channel.

5.1. Human visual attention from depth

Random dot stereograms (RDSs) were used to assess the saliency based model on depth. These image pairs allow the investigation of the effects of "pure" depth information. This is possible since these stereo image pairs are not defined by patterns, colours, lines etc., but only by stereo disparity information alone in a randomly generated black and white image. Our RDSs consisted of 4 objects each, which were drawn with different disparities between 2 and 16 pixels (i.e. protruding from the screen plane). Figure 1 presents a RDS example with the corresponding human fixations density plot. The density of fixations is clearly higher on the moon object, where the disparity is higher. Also the graphic on the right shows the relationship between the median order of fixations on the objects and their disparity. It is clear that objects with higher disparities attract the earlier (low order) fixations. Generalizing these results suggests that items with larger disparities are easier to be seen and often attract the early fixations, i.e. are more salient.

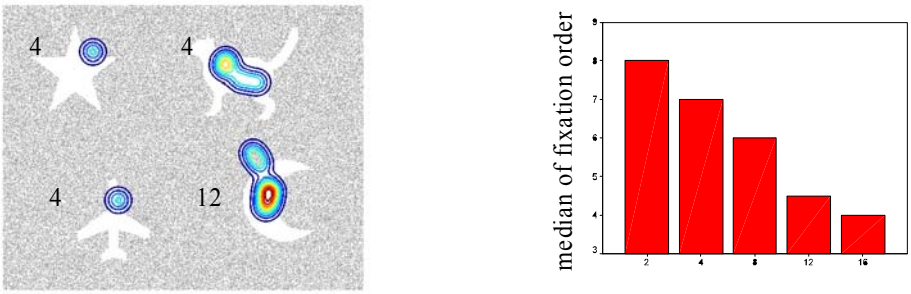

Figure 1. RDS example and relationship between the order of fixations on the object and the disparity

5.2. Color vs. color & depth

4 stereo image pairs of wood pieces compositions were shown to the subjects in order to compare the color model with the color & depth model. For all images, we first created the human map $H(\boldsymbol{x})$ from human fixations and a two-cues color saliency map $S_{col}(\boldsymbol{x}) = N(\hat{C}_{chrom})+N(\hat{C}_{\text{int}})$ and a three-cues color & depth saliency map $S_{depth}(\boldsymbol{x}) = N(\hat{C}_{chrom})+N(\hat{C}_{\text{int}})+N(\hat{C}_{depth})$, according to equation 1. Both saliency maps are also normalized to the same dynamic range. Then, a comparison of each of these 2 models with the human fixations was performed, following the metrics defined in the previous chapter. Figure 2 shows an example of the different measurements and maps involved for an image of the dataset.
Table 1 presents the average scores s and correlation coefficients ρ over the whole preliminary dataset, for both the color and color & depth models. 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, while the human maps $H(\boldsymbol{x})$ included all fixations.

	Score s	ρ
color model $S_{col}(\boldsymbol{x})$	67	.57
color & depth model $S_{depth}(\boldsymbol{x})$	80	.64

Table 1. Similarity measurements of the two computer models $S(x)$ vs. human behavior

The main observation is that, based on both evaluation methods, the color & depth model fares better than the color one. More specifically, the color & depth model yields an average score s and correlation coefficients ρ between 10% and 20% higher than the color model. Even with the limited nature of the dataset, this underlines the usefulness of the depth channel in the model and goes toward assessing that depth contributes to the visual attention process.

6. Conclusion

The work reported in this paper aims at assessing the contribution of depth to visual attention by performing comparisons of computer models with human behavior as measured by recording eye movements of human subjects. A considerable challenge of the research was to consider the depth information, given by stereovision in the human case, as a cue of the model. An autostereoscopic

display was used so that stereo image pairs could be shown to the test subjects while recording their eye movements.

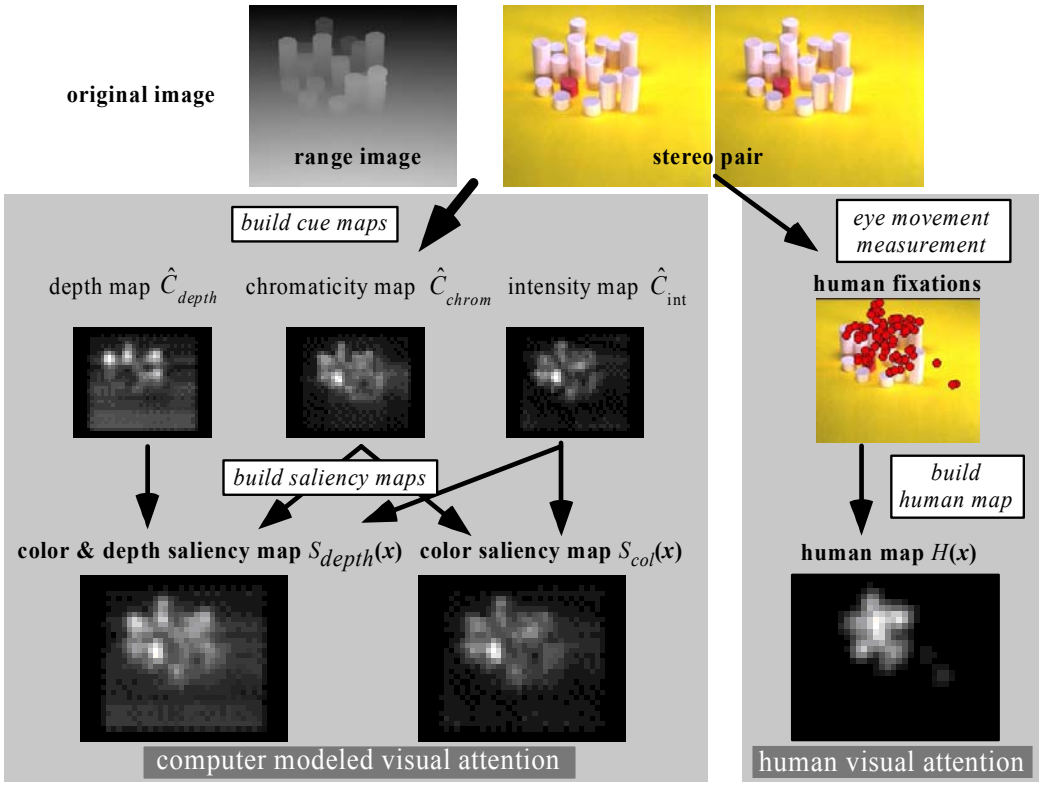

Figure 2. Overview of the different measurements and maps

A first set of experiments permitted to validate the experimental process and to show the link existing between disparity and visual attention, by using random dot stereograms. It was shown that the order of the fixations is related to the depth contrast, i.e. that features with high disparity contrast attract the human attention first.

In a second part, the contribution of depth in visual attention was quantitatively measured as the increase in similarity when the two-cues computer model for color is replaced by the three-cues computer model for color and depth. The similarity improvement, measured as the average on the preliminary dataset, is from ρ= .57 to .64 for the correlation coefficient, and from 67 to 80 for the score s and every tested stereo pairs actually yielded an improvement when the depth cue was considered. The limited nature of the preliminary dataset and the low number of subjects don't permit to draw final conclusions at this time but these results provide the first quantitative measurement of the contribution of the depth channel in a computer model of visual attention.

References

[HeHu] D. Heinke and G.W. Humphreys. "Computational Models of Visual Selective Attention: *A Review." In Houghton, G., editor,* Connectionist Models in Psychology, In press.

[Itt98]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, Vol. 20, No. 11, pp. 1254-1259*, 1998.

[Koc85] Ch. Koch and S. Ullman. "Shifts in selective visual attention: Towards the underlying neural circuitry." *Human Neurobiology, Vol. 4, pp. 219-227*, 1985.

[Mil93] R. Milanese. "Detecting Salient Regions in an Image: from Biological Evidence to Computer implementation." *PhD thesis, Dept. of Computer Science, University of Geneva, Switzerland*, 1993.

[Oue00] N. Ouerhani and H. Hugli. "Computing visual attention from scene depth". *International Conference on Pattern Recognition (ICPR'00), Vol. 1, pp. 375-378*, 2000.

[OuWa] N. Ouerhani, R. von Wartburg, H. Hügli, R.M. Müri, " Empirical validation of the Saliency-based model of visual attention", *Electronic Letters on Computer Vision and Image Analysis (ELCVIA), Vol. 3, pp. 13-24, 200*

[Par02] D. Parkhurst, K. Law, E. Niebur, "Modeling the role of salience in the allocation of overt visual attention", *Vision Research*, vol. 42, pp. 107-123, 2002.

[War03] R. von Wartburg. "Visuo-motor behaviour during complex image viewing: The influence of colour and image type". *Licentiate paper, Dept. of Psychology, University of Bern, Switzerland,* 2003.