Dynamic visual attention: competitive versus motion priority scheme

Bur A.[1], Wurtz P.[2], Müri R.M.[2] and Hügli H.[1]

[1]Institute of Microtechnology, University of Neuchâtel, Neuchâtel, Switzerland
[2]Perception and Eye Movement Laboratory, Departments of Neurology and Clinical Research, University of Bern, Bern, Switzerland

Abstract. Defined as attentive process in presence of visual sequences, dynamic visual attention responds to static and motion features as well. For a computer model, a straightforward way to integrate these features is to combine all features in a competitive scheme: the saliency map contains a contribution of each feature, static and motion. Another way of integration is to combine the features in a motion priority scheme: in presence of motion, the saliency map is computed as the motion map, and in absence of motion, as the static map. In this paper, four models are considered: two models based on a competitive scheme and two models based on a motion priority scheme. The models are evaluated experimentally by comparing them with respect to the eye movement patterns of human subjects, while viewing a set of video sequences. Qualitative and quantitative evaluations, performed in the context of simple synthetic video sequences, show the highest performance of the motion priority scheme, compared to the competitive scheme.

1 INTRODUCTION

Motion is of fundamental importance in biological vision systems. Specifically, motion is involved in visual attention, where rapid detection of moving objects is essential for adequate interaction with the environment [1]. Given the high relevance of temporal aspects in visual attention mechanisms, motion information as well as static information must be considered in the computer model of dynamic visual attention.

During the two last decades, computer models simulating human visual attention have been widely investigated. Most of them rely on the feature integration theory [2]. Many models used today stem from the classical saliency-based model proposed by Koch and Ullman [3], apply to still images and are used for detecting the most informative parts of an image, on which higher level tasks can then focus. This paradigm is used in various applications including color image segmentation [4] and object recognition [5]. Dynamic scene analysis is another field of interest where computer visual attention is applicable [6, 7].

In the literature, several ways have been proposed for extending the classical model of visual attention or, to state it differently, for combining the static and motion contributions. In [8], the motion channel is integrated with the other static channels at the same level, as additional channel. In [9] and [10], other ways of motion integration are proposed. Over all the proposed models in the literature, they can be classified in two distinct map integration schemes: (1) the competitive scheme and (2) the motion priority scheme.

In this article, both schemes are described and discussed. Then, four dynamic models, issued from both schemes, are considered. Their performance is evaluated experimentally by comparing the models with respect to the eye movement patterns of a population of human subjects, while viewing a set of video sequences. Qualitative and quantitative results conclude to the superiority of the motion priority scheme.

The rest of the paper is structured as follow. Section 2 describes both integration schemes and the four specific dynamic visual attention models. Section 3 provides the methodology for the model evaluation and Section 4, the description and results of the experiments. Finally, a conclusion is given in Section 5.

2 DYNAMIC VISUAL ATTENTION MODELS

Section 2.1 provides a description of the computation of the static map. In Section 2.2, two pure motion models are presented. Finally, Section 2.3 provides a description of the different dynamic models considered for computing the saliency map of dynamic sequences.

2.1 Static map

The saliency-based model of visual attention [3] is based on three major principles: visual attention acts on a multi-featured input; local saliency is influenced by the surrounding context; the saliency is represented on a scalar saliency map. In this article, three cues namely, color, intensity and orientation are used and the cues stem from seven features. The different steps of the model are briefly described here (more details are available in [11]):
1) Seven features are extracted from the scene by computing the so-called features from an RGB image color: one intensity feature; two chromatic features based on the two color opponency blue-yellow and red-green; four local orientation features according to the angles $\theta \epsilon \{0°, 45°, 90°, 135°\}$.
2) Each feature map is transformed in its conspicuity map. Each conspicuity map highlights the parts of the scene that strongly differ, according to a specific feature, from their surrounding. This is usually achieved by using a multiscale *center-surround*-mechanism [12].
3) The seven features are then grouped, according to their nature into three conspicuity cues of intensity C_{int}, color C_{color} and orientation C_{orient}.
4) Finally, the cue conspicuity maps are integrated together, in a competitive way, into the *saliency map* $\mathcal{S}$. Formally the static saliency map is defined as:

$$S_{static} = \mathcal{N}(C_{color}) + \mathcal{N}(C_{int}) + \mathcal{N}(C_{orient}) \tag{1}$$

where $\mathcal{N}()$ is a normalization function that simulates intra-map competition and inter-map competition in the map integration process. Several normalization methods exist in the literature. [11] and [13] describe and compare linear versus non-linear functions. A comparison with human vision concluded to the superiority of the non-linear methods, which tend to suppress the low level noise of the map, while promoting isolated high level responses. In this work, a non-linear exponential normalization function defined in [11] is used.

2.2 Motion maps

A. The motion map

The general idea is to have a channel acting as a motion component in the model. Among various possibilities for detecting motion, here we consider the absolute value of the local speed computed with a gradient-based optical flow method [14]. Based on the brightness conservation, the optical flow is computed from the temporal and spatial derivatives of the image intensity. Formally, the absolute value of normal velocity s is given by:

$$s(\mathbf{x}, t) = \frac{|I_t(\mathbf{x}, t)|}{\|\nabla I(\mathbf{x}, t)\|} \tag{2}$$

where ∇I refers to the spatial gradient and I_t is the temporal derivative of the image intensity I. In order to deal with displacement of variable amplitude, a multi-scale approach is used. The details of the implementation are given in [15]. Formally, the motion conspicuity is defined as:

$$C_{motion} = \sum_{i=1}^{4} \mathcal{N}(\mathcal{M}_i) \tag{3}$$

where $\mathcal{M}_i$ refers to the multi-scale motion map s at the scale i and $\mathcal{N}()$ is the same normalization function as used in the static model. Finally, the motion map is defined as:

$$S_{motion} = C_{motion} \tag{4}$$

B. The motion-conditioned map

Proposed in [15], the motion-conditioned map computes motion differently. Here, the motion map defined in Eq. 4 is conditioned by the static map: only moving objects compete for saliency and in a proportion equal to their static conspicuity. Formally, the motion-conditioned map is defined as:

$$S_{cond}(\mathbf{x}) = \begin{cases} S_{static}(\mathbf{x}) & if \quad S_{motion}(\mathbf{x}) > T_\varepsilon \\ 0 & otherwise \end{cases} \tag{5}$$

where T_ε is a threshold, corresponding to the minimum value, for which motion response is considered as significant.

2.3 Dynamic maps

A. The competitive scheme

Given a set of feature maps $\mathcal{F}$ to be integrated, the competitive scheme combines all the maps additively. The resulting map S contains a contribution of each feature. Formally, the competitive scheme is defined as:

$$S = \sum_{i=1}^{n} \mathcal{N}(\mathcal{F}_i) \tag{6}$$

where $\mathcal{F}_i$ refers to one of the n feature maps and $\mathcal{N}()$ is the same normalization function as defined in Eq.1. We notice that the considered scheme is identical to the feature integration process used in the static model.

In this paper, two models using this competitive scheme are considered. The first model integrates the motion in the static model as an additional cue. All the cues (color, intensity, orientation and motion) are integrated into the saliency map in a competitive way [8] [16]. The saliency map of the model 1, named *the cue competition model* is thus defined as:

$$Model\ 1: \quad S_{cuecomp} = \mathcal{N}(C_{color}) + \mathcal{N}(C_{int}) + \mathcal{N}(C_{orient}) + \mathcal{N}(C_{motion}) \tag{7}$$

The second model, proposed in [15], integrates motion at a higher level. The motion map is directly combined to the static map in a competitive scheme. Formally, the saliency of model 2, *the static&dynamic model*, is defined as:

$$Model\ 2: \quad S_{static\&dyn} = \mathcal{N}(S_{static}) + \mathcal{N}(S_{motion}) \tag{8}$$

Compared to the first model, this results to a higher motion contribution in the saliency map.

B. The motion priority scheme

Proposed in [9], the motion priority scheme combines the static map and motion map by prioritizing motion: in presence of any motion, the saliency map is computed by suppressing the static channels, the motion has the priority. In absence of any motion, the saliency map is computed by the classical static model. This integration scheme acts like a switch between the static and motion map. The third model, *the motion priority model* is defined as:

$$Model\ 3: \quad S_{priority1} = \begin{cases} S_{motion} & if \quad max_x(S_{motion}(\mathbf{x})) > T_\varepsilon \\ S_{static} & \quad otherwise \end{cases} \tag{9}$$

Accordingly, the saliency map $S_{priority1}$ corresponds either to the motion map S_{motion} if its global maximum value is higher than the threshold T_ε, or otherwise, it corresponds to the static map S_{static}.

The fourth model combines in a similar way the static map with the motion-conditioned map of Eq. 5. The saliency map according to model 4, *the motion-conditioned priority model* is defined as:

$$Model\ 4: \quad S_{priority2} = \begin{cases} S_{cond} & if \quad max_x(S_{motion}(\mathbf{x})) > T_\varepsilon \\ S_{static} & \quad otherwise \end{cases} \tag{10}$$

Accordingly, the saliency map $S_{priority2}$ corresponds either to the motion-conditioned map S_{cond} if the global maximum value of S_{motion} is higher than the threshold T_{ε}, or otherwise, it corresponds to the static map S_{static}.

3 MODEL EVALUATION

This section describes the method used to evaluate the performance of the models of visual attention in comparison with human vision. The basic idea consists in measuring, for a given set of video sequences, the correspondences between the computed saliency sequences and the corresponding human eye movement patterns.

Video sequences are used as visual source. On one hand, the computer operates according to a selected model and produces saliency maps for each video frame and therefore a saliency sequence corresponding to a video source sequence. On the other hand, the same video sequence is shown to human subjects while recording their eye movements. The data are segmented into saccade, blink, fixation and smooth-pursuit periods. Then blink and saccade periods are discarded in order to take into account only fixations and smooth-pursuits in the analysis [16]. We end up with a set of fixation and pursuit points $\{\mathbf{x}(t)\}$.

For the purpose of a qualitative comparison of human and computer results, we present next a means to transform the set $\{\mathbf{x}(t)\}$ into a so called human saliency map that provides the possibility to visually compare the computer saliency and human saliency sequences.

For the purpose of a quantitative comparison, we present next the definition of a score that provide a quantitative measure of the similarity between computer saliency and the set $\{\mathbf{x}(t)\}$.

3.1 Human saliency

The human saliency map $H(\mathbf{x}, t)$ is computed under the assumption that it is an integral of gaussian point spread functions $h(\mathbf{x}_k)$ sampled in time and space at the locations of the fixation and pursuit points $\{\mathbf{x}(t)\}$. The width of the gaussian is chosen to approximate the size of the fovea. Formally, the human saliency map $H(\mathbf{x}, t)$ computed at a given frame t is:

$$S_{human} = H(\mathbf{x}, t) = \frac{1}{K} \sum_{k=1}^{K} h(\mathbf{x}_k, t) \tag{11}$$

where $\mathbf{x}_k$ refers to the position of one of the K fixation and pursuit points that occur at the time t.

3.2 Score

For quantifying the correspondence of human eye movement patterns with a given saliency map, an analysis of the saliency value located at the human observation points is performed. Several approaches are defined in [8] and [16]. In this article, a similarity score s, defined in [11], is computed for evaluating the suitability of the seven considered models. The score s quantifies the similarity of a given saliency map S with respect to a set of fixation and pursuit points $\{\mathbf{x}(t)\}$.

The idea is to define the score as the difference of average saliency $\overline{s}_{fix}$ obtained when sampling the saliency map S at the fixation and pursuit points with respect to the average $\overline{s}$ obtained by a random sampling of S. In addition, the score used here is normalized and thus independent of the scale of the saliency map. Formally, the score s is thus defined as:

$$s = \frac{\overline{s}_{fix} - \overline{s}}{\overline{s}}, \quad with \quad \overline{s}_{fix} = \frac{1}{K}\sum_{k=1}^{K} S(\mathbf{x}_k) \tag{12}$$

A high score s means high saliency values at the fixation and pursuit points, in comparison to the average value of the saliency map S. The score represents simply the ratio $\frac{\overline{s}_{fix}}{\overline{s}}$ shifted with an offset of -1.

The quantitative evaluation is performed as follows: for each model, for each sequence and for each frame t, a score $s(t)$ is computed by comparing the saliency map at the frame t with respect to the fixations and pursuits that occur at that time.

4 EXPERIMENTS

4.1 Video sequences

The set of video clips is composed of 14 short synthetic video sequences, containing either static objects, moving objects or both. In the experiments, various scenarios are used, alternating moving situations and still situations, combining high color-contrasted, low color-contrasted, moving and standing spots. The duration of the sequences is 10 seconds.

4.2 Eye Movement Recording

Eye movements were recorded using an infrared-video-based eye tracker (HiSpeedTM, SensoMotoric Instruments GmbH, Teltow, Germany, 240Hz), tracking the pupil and the corneal reflection to compensate the head movements. 10 human subjects observed the video sequences on a 20" color monitor with a refresh rate of 60 Hz. The viewing distance was 71.5 cm and the video sequences were displayed full screen, resulting to a visual angle of approximately 32° by 24°. Each synthetic sequence was displayed randomly in alternation with a real video sequence in order to keep a close attention of the subject throughout the viewing session. Each video sequence lasted 10 seconds and was preceded by a central fixation cross for 2 seconds. The instruction given to the subjects was "just look at the screen and relax".

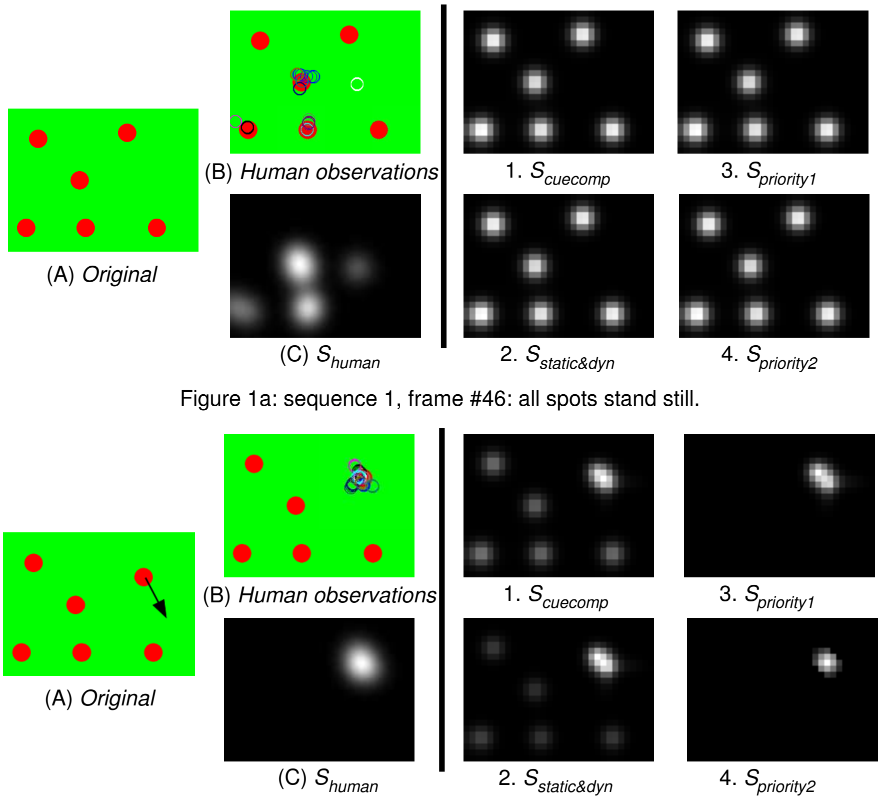

Figure 1a: sequence 1, frame #46: all spots stand still.

Figure 1b: sequence 1, frame #70: one spot is moving, the other ones stand still.

Fig. 1. A comparison of the human saliency map issued from the human recording with the computer saliency maps issued from the four considered models (1. to 4.). (A) the original frame, (B) the human observations and (C) the human saliency map.

4.3 Qualitative Evaluation

Figure 1 shows an example of qualitative evaluation of the four considered models. Here, the model comparison is performed for sequence 1 in two situations: all the spots stand still (frame #46); one spot is moving while the other ones stand still (frame #70). The human saliency (C) is compared with the four models for both situations. In the first situation, the subjects spread their attention on the static spots. All the models have the same saliency map and are equivalent in term of similarity to the human saliency. In the second situation, all the subjects concentrate their attention on the moving spot. Here, the models based on the motion priority schemes are more suitable for predicting the human attention, compared to the competitive-based models.

In the frame of the experiments, we observe over all the sequences that the human saliency map highlights most of the time moving objects. Thus we can state that most human subjects concentrate their attention on moving stimuli. In other word, motion stimuli have a pop-out effect that strongly attracts the human attention. This is an explanation why the motion priority is more suitable, compared to a competitive scheme.

4.4 Quantitative Evaluation

The next paragraph discusses the overall model performances based on the set of 14 sequences. For each sequence, an average score is computed for each model. Thus, 14 scores represent the performance of a given model. Figure 2 shows the score repartition for each sequence for each model. Over all the sequences, models 3 and 4 have higher scores compared to the models 1 and 2. This results to the superiority of the motion priority scheme compared to the competitive scheme in the dynamic visual attention model.

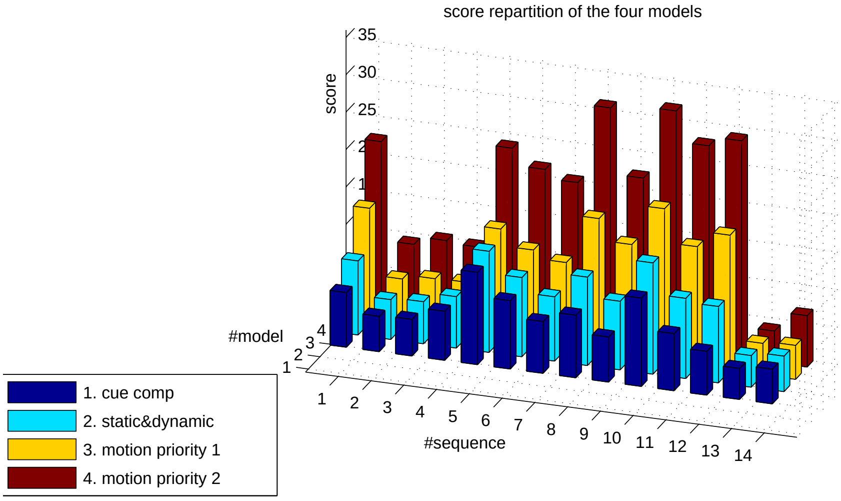

Fig. 2. The score repartion evaluated for 14 synthetic video sequences for the four considered models.

Table 1 shows an overview of the model performances. First we notice that all scores are quite high. For example, the average score for the cue competition model is 7.16, which means that the average saliency value sampled on the human fixations is 8.16 times higher than the average saliency value sampled randomly. Finally, a comparison of the average scores confirms the superiority of the motion priority scheme.

Table 1. Performance evaluation for the four considered models: an average score for each model is computed from the 14 sequences

visual attention models	Mean Score	Standard Deviation
1. cue competition model	7.16	2.54
2. static&dynamic model	9.01	3.29
3. motion priority 1 model	12.24	5.44
4. motion priority 2 model	19.47	9.44

To summarize, the experimental qualitative and quantitative evaluation show that the motion priority scheme is more suitable than the competitive scheme in the architecture of dynamic visual attention models. The motion priority scheme acts like a switch: in presence of any motion, the saliency map is computed by suppressing the static channels, the motion has the priority. In absence of any motion, the saliency map is computed as the static model.

During the experiments, most human subjects concentrate their attention on moving stimuli, which induce a pop-out effect that strongly attracts their attention. This is an explanation of the higher suitability of the motion priority scheme, compared to the competitive scheme. We should keep in mind a limitation: the motion priority scheme does not allow to detect a high salient static object in presence of motion, while the competitive scheme allows it. We notice also that the results are limited to the analysis of synthetic scenes. Future research will extend the analysis to natural scenes.

5 CONCLUSION

This article compared two alternative schemes of map integration for combining static and motion features in the field of a computer visual attention model for dynamic scenes. Four models, belonging to the considered integration schemes, are compared by measuring their respective performance with respect to the eye movement patterns of human subjects, while viewing simple synthetic sequences.

In the context of the simple synthetic scenarios provided, this comparative study shows that the motion priority scheme is more suitable than the competitive scheme, for integrating motion in visual attention. Both qualitative and quantitative evaluations show the superiority of the motion priority scheme. The motion priority scheme acts like a switch: in presence of any motion, the saliency map is computed by suppressing the static channels, the motion has the priority. In absence of any motion, the saliency map is computed as the static model. An interpretation in human vision suggests that attentional behavior is best explained by the motion priority scheme. Future research will investigate this interpretation in the general context of real natural scenes.

ACKNOWLEDGMENTS

The presented work was supported by the Swiss National Science Foundation under project number FN-108060.

References

1. T. Watanabe *et al.* Attention-regulated activity in human primary visual cortex. *Journal of Neurophysiology*, 79:2218–2221, 1998.
2. A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.
3. Ch. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.
4. N. Ouerhani and H. Hügli. MAPS: Multiscale attention-based presegmentation of color images. In *4th International Conference on Scale-Space theories in Computer Vision*, volume 2695 of *LNCS*, pages 537–549, 2003.
5. D. Walther, U. Rutishauser, Ch. Koch, and P. Perona. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. In *Computer Vision and Image Understanding*, volume 100, pages 41–63, 2005.
6. N. Ouerhani and H. Hügli. A model of dynamic visual attention for object tracking in natural image sequences. In *International Conference on Artificial and Natural Neural Network*, volume 2686 of *LNCS*, pages 702–709, Springer, 2003.
7. L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transaction on Image Processing*, 13:1304–1318, 2004.
8. L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12:1093–1123, 2005.
9. A. Bur and H. Hügli. Motion integration in visual attention models for predicting simple dynamic scenes. In *Human Vision and Electronic Imaging XII*, Proc. SPIE, To be published in february 2007.
10. G. Somma. Dynamic foveation model for video compression. In *The 18th International Conference on Pattern Recognition*, pages 339–342, 2006.
11. N. Ouerhani, A. Bur, and H. Hügli. Linear vs. nonlinear feature combination for saliency computation: A comparison with human vision. volume 4174 of *Lecture Notes in Computer Science*, pages 314–323, Springer, 2006.
12. 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)*, 20:1254–1259, 1998.
13. L. Itti and Ch. Koch. A comparison of feature combination strategies for saliency-based visual attention systems. In *Human Vision and Electronic Imaging IV*, volume 3644 of *Proc. SPIE*, pages 373–382, 1999.
14. J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12:1–9, 1994.
15. N. Ouerhani. *Visual Attention: from bio-inspired Modeling to Real-Time Implementation (PhD Thesis pp.42-52)*. http://www-imt.unine.ch/parlab/, 2004.
16. T. Williams and B. Draper. An evaluation of motion in artificial selective attention. In *Computer Vision and Pattern Recognition Workshop (CVPRW'05)*, volume 3, page 85, 2005.