Motion Integration in Visual Attention Models for Predicting Simple Dynamic Scenes

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

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

ABSTRACT

Visual attention models mimic the ability of a visual system, to detect potentially relevant parts of a scene. This process of attentional selection is a prerequisite for higher level tasks such as object recognition. Given the high relevance of temporal aspects in human visual attention, dynamic information as well as static information must be considered in computer models of visual attention. While some models have been proposed for extending to motion the classical static model, a comparison of the performances of models integrating motion in different manners is still not available. In this article, we present a comparative study of various visual attention models combining both static and dynamic features. The considered models are compared by measuring their respective performance with respect to the eye movement patterns of human subjects. Simple synthetic video sequences, containing static and moving objects, are used to assess the model suitability. Qualitative and quantitative results provide a ranking of the different models.

Keywords: visual attention, computer model, motion, bottom-up, eye movement, saliency

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, dynamic information as well as static information must be considered in the computer model of 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 investigations are found that focus on visual attention models integrating motion. In Ref. 8, the authors confirm the relevance of motion in computational modeling of visual attention. In Ref. 9, the authors investigate the contribution of low-level saliency to human eye movements in complex dynamic scenes. The results show that a proposed model which includes motion, temporal change, color, intensity and orientation channels, fits better to human eye movement behavior than any model comprising a single channel only. The authors also conclude that single motion and single temporal change are stronger predictors of human fixations than any other single channel. In Ref. 10, the authors study the contribution of motion in visual attention models, by comparing in a similar experimental frame the performances of the classical static model with an extended model integrating motion as additional cue. Surprisingly, the proposed extended model shows no improvement compared to the classical one.

So far, several ways have been proposed for extending the classical model of visual attention or, to state it differently, for combining the static and dynamic contributions. In Refs. 9 and 10, the motion channel is integrated in a straightforward manner, by adding the motion channel with the other static channels at the same level. In Refs. 11 and 12, other ways of motion integration are proposed. However, a comparison of the performances of the different models is still not available. The purpose of this study is to understand how to integrate motion in the computer model of visual attention, such that the model prediction correlates well with the average visual behavior of a population of human subjects. The idea is to evaluate the model performance in situations of increasing complexity: first in simple synthetic scenarios, then in simple real scenes and later in complex real scenes.

In this article, a description of five existing models is provided and two additional models are proposed, resulting in seven considered models combining static and dynamic channels. In an experimental frame, the model performances are then compared with respect to the eye movement patterns of a population of human subjects, while viewing a set of video sequences. Here, the study concentrates on the model prediction in presence of simple synthetic dynamic scenes. The qualitative and quantitative evaluation provide a performance ranking of the seven considered models.

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

2. DYNAMIC VISUAL ATTENTION MODELS

This section describes the seven considered models of visual attention. The first one is the classical model of visual attention for still images, described in Sect. 2.1 as the static model. The second one considers motion only and is defined in Sect. 2.2 as the motion model. Section 2.3 provides the description of five other models combining static and motion channels.

2.1. Static Model

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 Ref. 13):
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.[14]
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 saliency map according to model 1, *the static model* is defined as:

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

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. Refs. 13 and 15 describe and compare linear versus non-linear functions. A comparison with human vision concluded to the superiority of the non-linear methods. The advantage of non-linear integration is to suppress the noise of the map, while promoting isolated high responses in the same map. In this work, a non-linear exponential normalization function defined in Ref. 13 is used.

2.2. Motion Model

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.[16] 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 Ref. 11. 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. $\mathcal{N}()$ is the same normalization function as used in the static model. Finally, the saliency map according to the model 2, *the motion model* is defined as:

$$Model\ 2: \qquad S_{motion} = C_{motion} \tag{4}$$

2.3. Combined models

In this section, five different attention models combining static and motion conspicuities, are described. The first model, used in Refs. 9 and 10, integrates the motion in a cue competition scheme that proceeds by combining all the cues at the same level. The saliency map of the model 3, *the cue competition model* is thus defined as:

$$Model\ 3: \qquad S_{cuecomp} = \mathcal{N}(C_{color}) + \mathcal{N}(C_{int}) + \mathcal{N}(C_{orient}) + \mathcal{N}(C_{motion}) \tag{5}$$

The next model, already proposed in Ref. 11, is a combination of the static map S_{static} with the motion map S_{motion}. The saliency map according to model 4, *the static&dynamic model* is defined as:

$$Model\ 4: \qquad S_{static\&dyn} = \mathcal{N}(S_{static}) + \mathcal{N}(S_{motion}) \tag{6}$$

The next model, also previously proposed in Ref. 11, combines S_{static} and S_{motion} in a motion-conditioned scheme: only moving objects compete for saliency and in a proportion equal to their static conspicuity. The saliency map according to the model 5, *the motion-conditioned model* is defined as:

$$Model\ 5: \qquad S_{cond}(\mathbf{x}, t) = \begin{cases} S_{static}(\mathbf{x}, t) \quad if \quad S_{motion}(\mathbf{x}, t) > T \qquad with \qquad T = p \cdot max(S_{motion}) \\ 0 \quad otherwise \end{cases} \tag{7}$$

where T is a threshold, defined as a given percentage ($p = 25\%$) of the global maximum value $max(S_{motion})$.

Similarly to a model proposed in Ref. 12, we propose two additional models that combine S_{static} and S_{motion} in a motion priority scheme: in presence of motion, the saliency map is computed by suppressing the static channels, the motion has the priority. In absence of motion, the saliency map is computed by the classical static model. The saliency map according to model 6, *the motion priority 1 model* is defined as:

$$Model\ 6: \qquad S_{priority1}(\mathbf{x}, t) = \begin{cases} S_{motion}(\mathbf{x}, t) \quad if \quad max(S_{motion}) > T_1 \\ S_{static} \quad otherwise \end{cases} \tag{8}$$

Accordingly, the saliency map $S_{priority1}$ corresponds to the motion map S_{motion} if the global maximum value $max(S_{motion})$ is higher than a given threshold T_1, otherwise, it corresponds to the static map S_{static}. The last model is similar to the model defined above. The saliency map according to model 7, *the motion priority 2 model* is defined as:

$$Model\ 7: \qquad S_{priority2}(\mathbf{x}, t) = \begin{cases} S_{cond}(\mathbf{x}, t) \quad if \quad max(S_{motion}) > T_1 \\ S_{static} \quad otherwise \end{cases} \tag{9}$$

Accordingly, the saliency map $S_{priority2}$ corresponds to the motion-conditioned map S_{cond} if the global maximum value $max(S_{motion})$ is higher than a threshold T_1, 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 their eye movements are recorded. The data are segmented into saccade, blink, fixation and smooth-pursuit periods. Then blink and saccade periods are discarded in order to take only into account fixations and smooth-pursuits in the analysis.[10] 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 mean 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:

$$S_{human} = H(\mathbf{x}, t) = \frac{1}{K}\sum_{k=1}^{K} h(\mathbf{x}_k, t) \tag{10}$$

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 Refs. 9 and 10. In this article, a similarity score s, defined in Ref. 13, 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:

$$s = \frac{\overline{s}_{fix} - \overline{s}}{\overline{s}}, \quad with \quad \overline{s}_{fix} = \frac{1}{K}\sum_{k=1}^{K} S(\mathbf{x}_k) \tag{11}$$

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, 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. As example, Fig. 4 shows the score evolution of the seven considered models for the sequence 1 and 11.

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. Figure 1 illustrates a few examples of the scenarios. The duration of the sequences is 10 seconds.

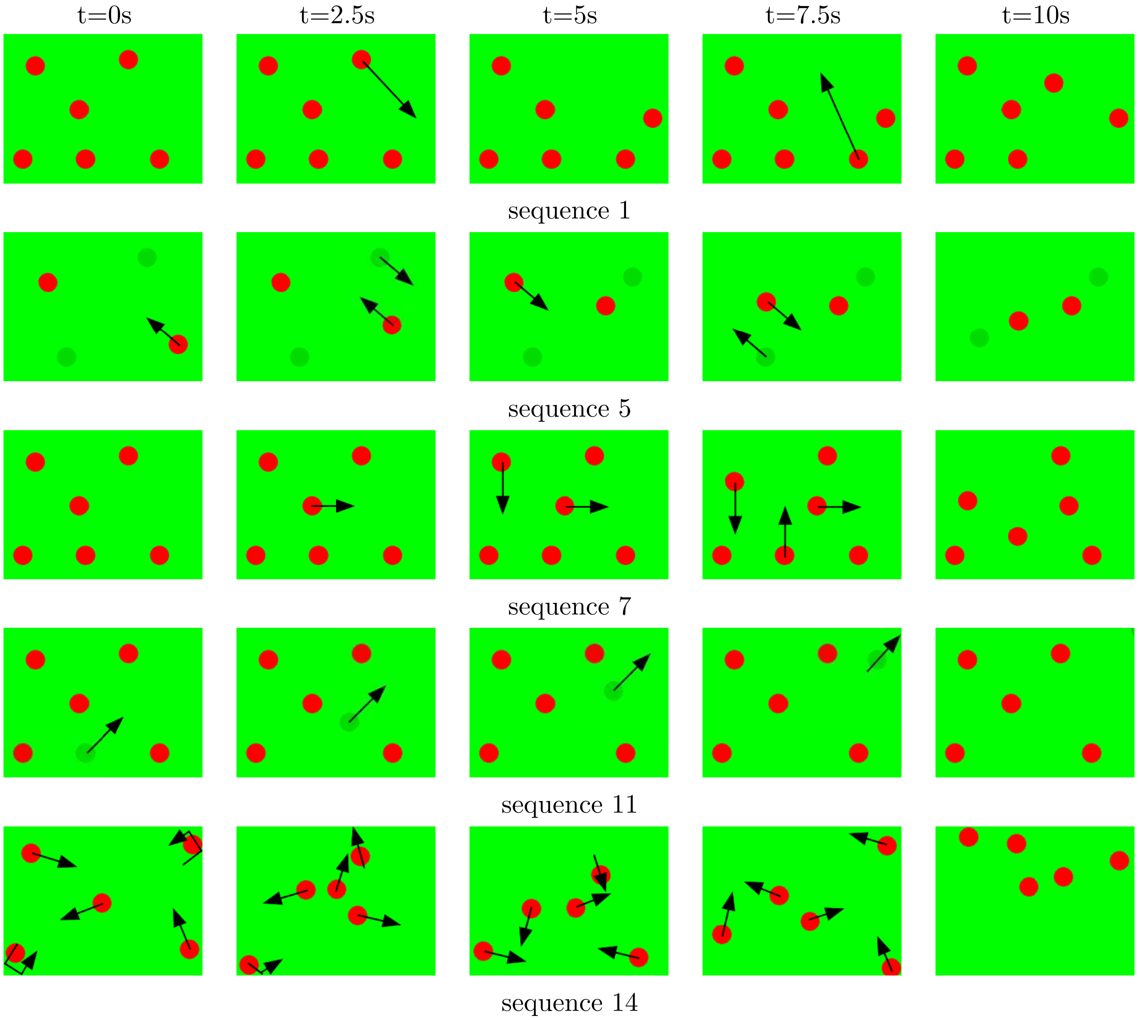


Figure 1. Some examples of synthetic video sequences represented by their respective frames at time 0.0,2.5,5.0,7.5,10.0 [s]. The arrows represents the motion of the spots. In this work, various scenarios containing static and moving objects are used: in sequence 1, all the spots stand still from 0.0 to 2.5 and from 5.0 to 7.5 [s]. From 2.5 to 5.0 [s] and from 7.5 to 10.0 [s], only one spot is moving and the other ones stand still; in sequence 5, static, moving, high-color-contrasted and low-color-contrasted spots are in competition; in sequence 7, the number of moving spots increases over the time; in sequence 11, a moving low color-contrasted spot competes with several static high color-contrasted spots; in sequence 14, all the spots are moving.

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".

4.3. Qualitative Evaluation

Figure 2 shows an example of qualitative evaluation of the seven considered models. Here, the model comparison is performed in the 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 seven models 1 to 7 for both situations. In the first situation, the subjects spread their attention on the static spots. The models 1, 3, 4, 6 and 7 have the same saliency map and are equivalent in term of similarity to the human saliency. As expected, the pure-motion models(model 2 and 5) have a null saliency map and are inadequate in the static situation. In the second situation, all the subjects concentrate their attention on the moving spot. Here, The models 2, 5, 6 and 7 are the most suitable for predicting the human attention. Then follow the model 3 and 4, which are more similar to the human saliency than the static model 1.

Table 1. The suitability of the seven considered models in several situations: static situations, motion situations and combined situations. The models with a motion priority scheme are the most suitable in every situations.

	in static situations	in dynamic situations	in combined situations
1. static	√	x	x
2. motion	x	√√	√√
3. cue competition	√	√	√
4. static&dynamic	√	√	√
5. motion-conditioned	x	√√	√√
6. motion priority 1	√	√√	√√
7. motion priority 2	√	√√	√√

In the frame of these experiments, the suitability of the models is evaluated by analyzing visually the human attention over all the sequences. Table 1 shows an overview of the suitability of the models in several sequence scenarios: static situations, motion situations and combined situations. As observed in Fig. 2, the static model (model 1) fails in motion situations, while the pure-motion models (model 1 and 5) fail in static situations. The models 3, 4, 6 and 7 are suitable in any situation and further, models 6 and 7 are the most similar with respect to the human eye movement patterns. Thus, we can firstly state that the majority of the human subjects concentrates on the moving stimuli. Secondly, by comparing qualitatively all the models over all the sequences, we conclude that the most suitable models are the models that give the priority to the motion (model 6 and 7).

4.4. Quantitative Evaluation

Figure 4 shows the score evolution over the time for the seven considered models in sequence 1 (A) and 11 (B). In the scenario (A), all the spots stand still from the frame #0 to #50 and from #100 to #150 (static situation). One spot is moving while the other ones stand still from #50 to #100 and from #150 to #200 (motion situation). In the scenario (B), one low-color-contrasted spot is moving all along the sequence while the other high-color-contrasted spots stand still (motion situation). In static situations, the scores are identical for all models, except for the motion and motion-conditioned models with negative score $s(t) = -1$, due to a

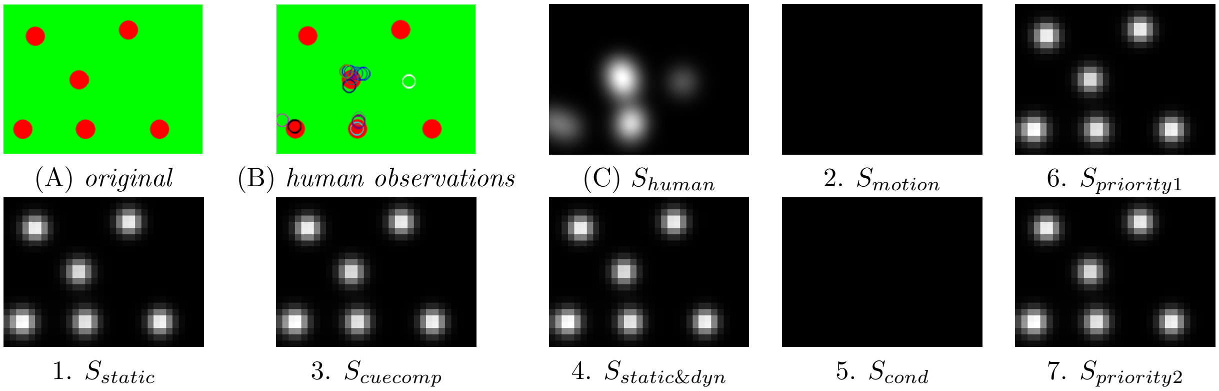


Figure 2a: sequence 1, frame #46: all spots stand still.

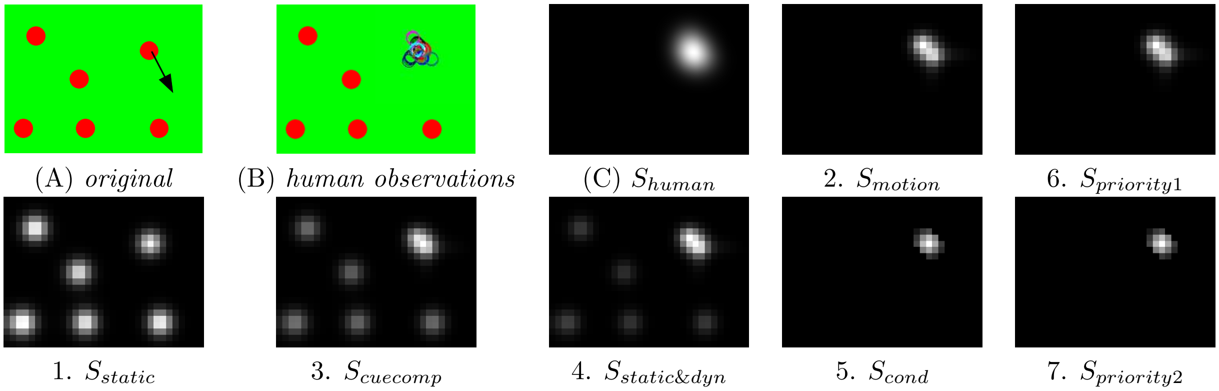


Figure 2b: sequence 1, frame #70: one spot is moving, the other ones stand still.

Figure 2. Comparison of the human saliency map issued from the human recording with the computer saliency maps issued from the seven considered models. Figure 2a shows the comparison for the frame #46 of the sequence 1, where all the red spots stand still. Figure 2b shows the comparison for the frame #70, where one spot is moving(represented by the arrow) and the other ones stand still. (A) represents the original frame. (B) the human observations, each subjects represented by a colored circle. (C) the human saliency map issue from the fixation and smooth pursuit periods. 1. to 7. the saliency maps issue from the seven considered models.

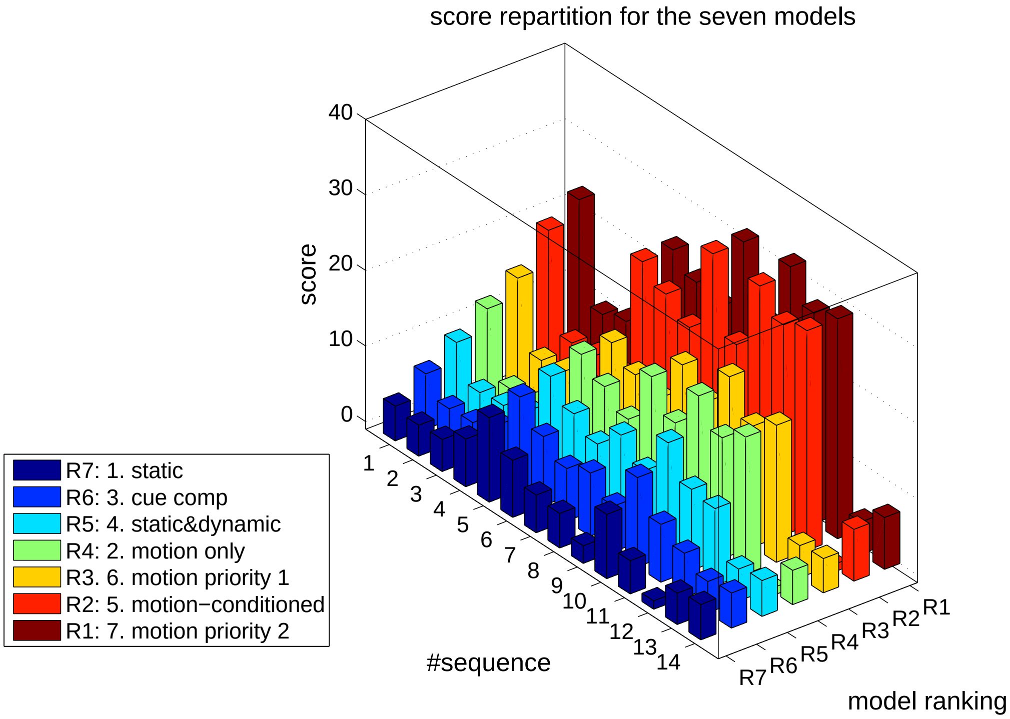


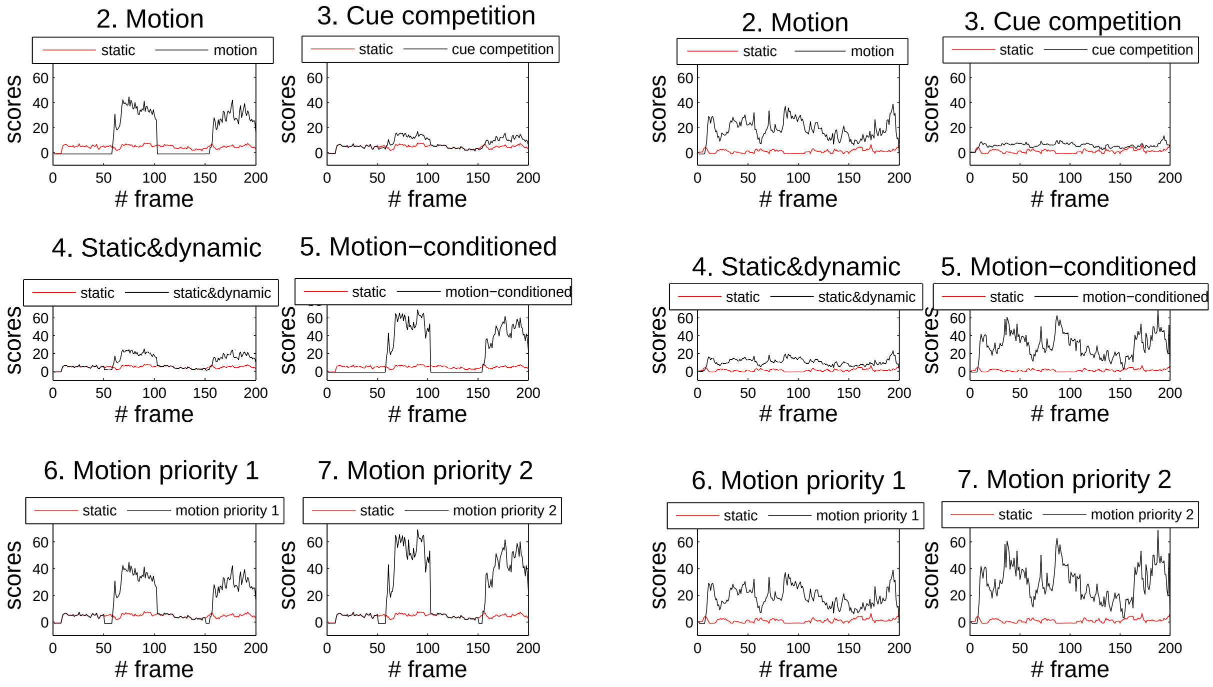
Figure 3. The score repartion evaluated for 14 synthetic video sequences for seven considered models.

null saliency map. In motion situations, the motion priority models (model 6 and 7), the motion and motion-conditioned models (model 2 and 5) give very high scores up to $s(t) = 60$. It shows that human attention is catched in priority by the motion. In other words, motion stimuli have a pop-out effect that strongly attracts the human attention.

Finally, by comparing the model performances in the sequences 1 and 11, we conclude that the motion priority 1 and 2 are the most suitable models in static and motion situations. Further, the motion priority 2 model gives the highest scores.

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 models. Thus, 14 scores represent the performance of a given model. Figure 3 shows the score repartition for each sequence for each model. Table 2 shows an overview of the model performances. First we notice that the average scores for all the models are high. For example, the average score for the static model is 5.18. That means that the average saliency value sampled on the human fixation is 6.18 times higher than the average saliency value sampled randomly. Secondly, all the models integrating motion have higher scores than the static model. Third, even the high scores of the motion and motion-conditioned models (10.87 and 18.09), those models are not suitable in static situations, as observed in Fig. 2. Finally, over all the considered models, the motion priority 2 model gives the highest scores.

To summarize, the experimental qualitative and quantitative evaluation shows that human attention concentrates in priority on moving objects in the case of simple synthetic dynamic scenes. It is thus clear that an attention model integrating a motion priority scheme is the most suitable to predict the attention of human subjects in simple video scenarios.



(A) Sequence 1

(B) Sequence 11

Figure 4. Performances of the seven considered models in sequences 1 and 11. The plots show the score evolution over frame number for models 2 to 7, each compared to model 1.

Table 2. Ranking of the seven considered models: an average score for each model is computed from the 14 sequences

visual attention models	Mean Score	Standard Deviation	Ranking
1. static model	5.18	2.53	R7
2. motion model	10.87	6.51	R4
3. cue competition model	7.16	2.54	R6
4. static&dynamic model	9.01	3.29	R5
5. motion-conditioned model	18.09	10.56	R2
6. motion priority 1 model	12.24	5.44	R3
7. motion priority 2 model	19.47	9.44	R1

5. CONCLUSION

This article provides a comparison of computer visual attention models combining static and dynamic features. Seven models are considered: a static model, a motion model and five models combining static and dynamic channels in different ways. The models are compared by measuring their respective performance with respect to the eye movement patterns of human subjects, while viewing simple synthetic sequences. Qualitative and quantitative results provide a ranking of the different models.

The qualitative evaluation showed that the static model fails in motion situations, while both motion and motion-conditioned models fail in static situations. Much better are the cue competition and static&dynamic models presented in the literature, as they are suitable in any situations, static or dynamic. These models combine static and motion channels in a competitive combination scheme. Finally, among all models, the motion priority 1 and 2 models, proposed in this article and inspired from previous models, are the most suitable. These models differ from previous models by a motion priority scheme that computes the saliency map by suppressing the static channels in presence of motion, and in absence of motion, uses the classical static model.

In the context of simple synthetic scenarios, this comparative study shows that the motion priority scheme is more suitable than the competitive combination scheme for the motion integration in visual attention. 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**, pp. 2218–2221, 1998.
2. A. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cognitive Psychology* **12**, pp. 97–136, 1980.
3. C. Koch and S. Ullman, “Shifts in selective visual attention: Towards the underlying neural circuitry,” *Human Neurobiology* **4**, pp. 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*, *Lecture Notes in Computer Science (LNCS)* **2695**, pp. 537–549, 2003.
5. D. Walther, U. Rutishauser, C. Koch, and P. Perona, “Selective visual attention enables learning and recognition of multiple objects in cluttered scenes,” in *Computer Vision and Image Understanding*, **100**, pp. 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*, *Lecture Notes in Computer Science* **2686**, pp. 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**, pp. 1304–1318, 2004.
8. J. Tsotsos, M. Pomplun, Y. Liu, J. Martinez-Trujillo, and E. Simine, “Attending to motion: Localizing and classifying motion patterns in image sequences,” *Second International Workshop on Biologically Motivated Computer Vision (BMCV'02)* , pp. 439–452, 2002.
9. L. Itti, “Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes,” *Visual Cognition* **12**, pp. 1093–1123, 2005.
10. T. Williams and B. Draper, “An evaluation of motion in artificial selective attention,” in *Computer Vision and Pattern Recognition Workshop (CVPRW'05)*, **3**, p. 85, 2005.
11. N. Ouerhani, *Visual Attention: from bio-inspired Modeling to Real-Time Implementation (PhD Thesis pp.42-52)*, http://www-imt.unine.ch/parlab/, 2004.

12. G. Somma, "Dynamic foveation model for video compression," in *The 18th International Conference on Pattern Recognition*, pp. 339–342, 2006.
13. N. Ouerhani, A. Bur, and H. Hügli, "Linear vs. nonlinear feature combination for saliency computation: A comparison with human vision," *Lecture Notes in Computer Science* **4174**, pp. 314–323, Springer, 2006.
14. L. Itti, C. 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**, pp. 1254–1259, 1998.
15. L. Itti and C. Koch, "A comparison of feature combination strategies for saliency-based visual attention systems," in *Human Vision and Electronic Imaging IV*, *Proc. SPIE* **3644**, pp. 373–382, 1999.
16. J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision* **12**, pp. 1–9, 1994.