Dynamic Visual Attention: Motion Direction versus Motion Magnitude

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

Defined as an attentive process in the context of visual sequences, dynamic visual attention refers to the selection of the most informative parts of video sequence. This paper investigates the contribution of motion in dynamic visual attention, and specifically compares computer models designed with the motion component expressed either as the speed magnitude or as the speed vector. Several computer models, including static features (color, intensity and orientation) and motion features (magnitude and vector) are considered. Qualitative and quantitative evaluations are performed by comparing the computer model output with human saliency maps obtained experimentally from eye movement recordings. The model suitability is evaluated in various situations (synthetic and real sequences, acquired with fixed and moving camera perspective), showing advantages and inconveniences of each method as well as preferred domain of application.

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 clearly involved in visual attention, where rapid detection of moving objects is essential for adequate interaction with the environment.[1] Defined as attentive process in the context of visual sequences, dynamic visual attention refers to the selection of the most informative parts of video sequence, on which higher level task can then focus. Potential computer applications of dynamic visual attention are considerable, including video surveillance, video quality assessment, video compression[2] and object tracking.[3]

Over the last decade, several investigations focused on the architecture of computer model of dynamic visual attention. In order to deal with video sequences, current dynamic models generally integrate additional motion components to the classical saliency-based model proposed by Koch and Ullman,[4] which applies to still images. A brief description of the related state of the art is given below. In Ref. 5, the authors consider a dynamic model combining static features (color, intensity, orientations) and dynamic features (temporal changes, four motion directions (0°, 45°, 90° and 135°) and a comparison with human vision is performed experimentally, by comparing the models with respect to the eye movement patterns of human subjects. In Ref. 6, dynamic models combining in different ways color, intensity, orientation and motion magnitude are compared and the evaluation is performed on simple synthetic sequences. In Ref. 7, the authors propose a dynamic model that is based on the motion contrast, computed as the difference between local (hierarchical block matching) and dominant motion (2D affine motion model with M-estimators). Evaluation and model comparison in real video sequences conduct to higher suitability of the proposed model compared to other models, including Ref. 5. One general conclusion of the above mentioned articles confirms that motion contrast is much more relevant than other features for predicting human attentional behavior.

While Ref. 8 attempts to guide the human scanpath (i.e. the pattern of eye movements) to improve visual communication, our work focused alternatively on predicting the scanpath. We intend to integrate efficiently motion information 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. This paper investigates the contribution of motion in dynamic visual attention, and specifically compares the relevance of motion contrast of different nature, the speed magnitude on one hand and the speed vector on the other hand.

In this article, five computer models are considered: a static model which includes static features (color, intensity and orientation), two motion models designed with the motion component expressed as the speed magnitude on one hand and the speed vector on the other hand, and two dynamic models combining both static and motion channels. A qualitative and quantitative evaluation is performed by comparing the computer model output with human saliency maps obtained experimentally from eye movement recordings of 20 human subjects. The model suitability is evaluated in 84 video sequences, classified in four categories: synthetic scene (respectively real scene) with fixed camera perspective, synthetic scene (respectively real scene) with moving camera perspective. The discussion of the results provides advantages and inconveniences of each model as well as preferred category of application.

The rest of the paper is structured as follows: Section 2 describes the considered models of visual attention. Section 3 provides the methodology for the model evaluation, Section 4 describes the experiments and in Section 5 the results are discussed. Finally, a conclusion is given in Sect. 6.

2. DYNAMIC VISUAL ATTENTION MODELS

This section describes the five considered models of visual attention. The first one is the static model of visual attention, described in Sect. 2.1. The second and third one are the motion models designed with the motion component expressed as the speed magnitude on one hand and the speed vector on the other hand (Sect. 2.2). Section 2.3 presents the dynamic models combining both static and motion channels and finally Section 2.4 describes the normalization process in the map integration.

2.1. Static Model

The saliency-based model of visual attention[4] relies 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 model is described in details in Ref. 9. We briefly remind it here:
1) Seven features F_j 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 $0°, 45°, 90°, 135°$.
2) Each feature map is transformed in its conspicuity map. Each conspicuity map C_j highlights the parts of the scene that strongly differ, according to a specific feature F_j, from their surrounding. This is usually achieved by using a multiscale *center-surround* mechanism.[10] The conspicuity operator $C()$ is defined as follow:

$$C_j = C(F_j) = \sum_{k=1}^{K} \mathcal{N}_{exp}(\mathcal{M}_{j,k}) \tag{1}$$

where $\mathcal{M}(j,k)$ corresponds to the intermediate conspicuity map that highlights the center-surround contrast at a given scale k. Described in Sect. 2.4, $\mathcal{N}()$ is a normalization function used in the map integration process.
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 competitively integrated, into the *saliency map* $\mathcal{S}$. Formally the saliency map according to model 1, *the static model* is defined as:

$$1^{st}\ Model: \qquad S_{static} = \mathcal{N}_{exp/LT}(C_{color}) + \mathcal{N}_{exp/LT}(C_{int}) + \mathcal{N}_{exp/LT}(C_{orient}) \tag{2}$$

2.2. Motion Models

This section defines the motion models designed with the speed magnitude on one hand and the speed vector on the other hand, both resulting from the 2D motion vector field $\overrightarrow{v}$, described below. Here, the considered motion pattern is pure translation and motion estimation is performed on two successive frames (t, t-1) by block matching algorithm (BMA). The proposed algorithm is based on a multi-scale approach in order to detect motion at different scales, by varying the block size. A coarse scale is obtained by applying BMA using a large block size, while a fine scale is obtained using a small block size. In this article, three scales are used and thus three

intermediate 2D vector fields $\overrightarrow{v}_1, \overrightarrow{v}_2$ and $\overrightarrow{v}_3$ are computed. Finally, the 2D motion field $\overrightarrow{v}$ is computed by combining the intermediate motion fields at the highest resolution.

The 2D motion field $\overrightarrow{v}$ having been described above, here are the definition of both motion models. The second model is the motion magnitude model, designed with the motion component expressed as the speed magnitude of the motion field $\overrightarrow{v}$. It is computed in two steps: first, a scalar map v_{norm} is computed as the euclidian norm of $\overrightarrow{v}$, second, a center-surround filtering is applied to v_{norm}. The saliency map S_{magn} according to the *motion magnitude model* is formally defined as:

$$v_{norm} = \|\overrightarrow{v}(i,j)\| = \sqrt{v_x^2 + v_y^2} \tag{3}$$

$$2^{nd}\ Model: \qquad S_{magn} = C(v_{norm}) \tag{4}$$

where C() refers to the conspicuity operator defined in Eq. 1.

The third model is the motion vector model, designed with the motion component expressed as the speed vector. Inspired from Ref. 5 and 11, the model is composed of eight principal direction maps v_Θ, each one corresponding to an activation map for a given direction. It is computed in three steps: each vector of the 2D motion field $\overrightarrow{v}$ is projected into both nearest principal direction maps v_Θ ($\Theta = \frac{\Pi}{4}i,\ i\ \epsilon\ \{0,1,2,...,7\}$), using a parallel projection. Then the center-surround filtering is applied to each direction map v_Θ and finally all maps are combined in a competitive way into the saliency map S_{vector}. According to *motion vector model*, S_{vector} is formally defined as:

$$3^{rd}\ Model: \qquad S_{vector} = \sum_\Theta \mathcal{N}_{exp}(C(v_\Theta)) \tag{5}$$

2.3. Dynamic models

This section defines both dynamic models, the former combining the motion magnitude saliency with the static saliency, the latter combining the motion vector saliency with the static saliency. Due to the different nature of the features to be fused, each map is previously normalized to the same dynamic range by applying a peak to peak normalization, which eliminates across-modality amplitude difference due to dissimilar extraction mechanism. the saliency map according to the dynamic magnitude model and respectively to the dynamic vector model are defined by the following equations:

$$4^{th}\ Model: \qquad S_{dyn\ magn} = \mathcal{N}_{exp/PP}(S_{static}) + \mathcal{N}_{exp/PP}(S_{magn}) \tag{6}$$

$$5^{th}\ Model: \qquad S_{dyn\ vector} = \mathcal{N}_{exp/PP}(S_{static}) + \mathcal{N}_{exp/PP}(S_{vector}) \tag{7}$$

To simplify the notation, *the magnitude models* refer to the 2nd and 4th computer models, while *the vector models* refer to the 3rd and 5th computer models.

2.4. Normalization process in the map integration

This section briefly describes the normalization process that simulates intra-map competition and inter-map competition in the map integration process. Several normalization methods are described in the literature. Refs. 9 and 12 compare functions of linear and non-linear nature, and report a superiority of the non-linear methods. We use the normalization processes described in details in Ref. 9. The normalization clearly depends on the nature of the maps to be integrated. $\mathcal{N}_{exp}(.)$ refers to the non-linear normalization process used in presence of maps of similar value ranges, which is the case for a set of intermediate multiscale maps $\mathcal{M}_{j,k}$ or a set of direction maps $C(v_\Theta)$. Here the non-linear exponential normalization is used:

$$\mathcal{N}_{exp}(C) = C' \cdot \left(\frac{C'}{C'_{max}}\right)^\gamma \quad with \quad C' = w(C) \cdot C \quad and \quad w(C) = \frac{C_{max}}{\overline{C}} \tag{8}$$

where $w(C)$ refers to a weighting scheme that simulates the inter-map competition, C_{max} and $\overline{C}$ are the maximum value and the mean value of the conspicuity map C.

$\mathcal{N}_{exp/LT}(.)$ refers to the normalization process used for integrating the static cues. Here, the values of considered maps have different ranges due to dissimilar extraction mechanisms. The idea is to apply a long-term normalization $\mathcal{N}_{LT}(.)$, which scales each map with respect to a maximum value, corresponding to a long-term specific maximum $\overline{M}_{cue}$ computed experimentally from a large set of images. Thus, $\mathcal{N}_{exp/LT}(.)$ applies first a $\mathcal{N}_{LT}(.)$ followed by a non-linear $\mathcal{N}_{exp}(.)$:

$$\mathcal{N}_{exp/LT}(C) = \mathcal{N}_{exp}(\mathcal{N}_{LT}(C)) \tag{9}$$

Finally, $\mathcal{N}_{exp/PP}(.)$ refers to the normalization process used for integrating the static and motion map. Due to the different value ranges of the map to be fused, each map is first normalized to the same dynamic range by applying a peak to peak normalization. Thus, $\mathcal{N}_{exp/PP}(.)$ applies first a a peak-to-peak normalization $\mathcal{N}_{PP}(.)$ followed by a non-linear $\mathcal{N}_{exp}(.)$:

$$\mathcal{N}_{exp/PP}(C) = \mathcal{N}_{exp}(\mathcal{N}_{PP}(C)) \tag{10}$$

This process has the undesirable drawback to map each channel to its full range, regardless of the effective amplitude of the map. An alternative would be to perform a long term normalization $\mathcal{N}_{LT}(.)$, the computation of the maximum value $\overline{M}_{motion}$ remaining an open issue.

3. MODEL EVALUATION

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

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.[13] 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 different scores that provide a quantitative measure of the similarity between computer saliency and the set of fixation and pursuit points $\{\mathbf{x}(t)\}$.

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:

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

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. Scores

For quantifying the correspondence of human eye movement patterns with a given computer saliency map, three scores which are commonly used in the literature are considered: the correlation coefficient,[7] the fixation-to-chance distance[6] and the Kullback Leibler divergence.[7]

The correlation coefficient s_{cc} quantifies the similarity between the human saliency map S_h and the computer saliency map S, according to the following equation:

$$s_{cc} = \frac{cov(S_h, S)}{\sigma_{S_h}\sigma_S} \tag{12}$$

where $cov()$ refers to the covariance value between the human saliency S_h and computer saliency S. High similarity provides a score s_{cc} close to 1 while low similarity provides a score s_{cc} close to 0.

The fixation-to-chance distance s_{ftc} 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_{ftc} is thus defined as:

$$s_{ftc} = \frac{\overline{s}_{fix} - \overline{s}}{\overline{s}}, \quad with \quad \overline{s}_{fix} = \frac{1}{K}\sum_{k=1}^{K} S(\mathbf{x}_k) \tag{13}$$

A high score s_{ftc} 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 Kullback-Leibler divergence, noted s_{KL}, estimates the dissimilarity between two probability density functions. Formally, the score s_{KL} is defined as:

$$s_{KL} = \sum_x s(x) \cdot ln\left(\frac{s(x)}{s_h(x)}\right) \tag{14}$$

where $s(x)$ and $s_h(x)$ are respectively, the probability densities deduced from the computer saliency map S and from the human saliency map S_h. As it is a measure of dissimilarity, a score s_{KL} close to zero indicates that the map S is almost identical to the map S_h.

The quantitative evaluation is performed as follows: for each model, for each sequence, for each frame t, the considered scores are computed by comparing the saliency map at the frame t with the fixations and pursuits that occur at that time.

4. EXPERIMENTS

4.1. Video sequences

The set of video clips is composed of 84 short sequences (10 sec. duration), representing four categories. The first and second categories include synthetic scenes, while the third and fourth natural real scenes. some examples are illustrated in Figure 1. The first category contains 15 videos with fixed camera perspective, combining static, moving, high color-contrasted and low-color-contrasted spots in a uniform background. The second category contains 19 videos with moving camera perspective. The considered motion pattern is pure-translation in the camera plane. The video content is composed of one or several moving dots (targets supposed to be salient) among a background of moving dots (distractors). The background is either composed of random dots or either of a grid of dots. Various configurations of motion contrast are considered: target dot(s) moving either faster or slower in the same direction relatively to the background; target dot(s) moving slower, at the same speed, or faster in another direction than the background. The third (30 videos in fixed camera) and fourth categories (20 videos in moving camera) include videos of natural real scenes in outdoor and indoor environment (city, traffic road, street, football field, train station, stores).

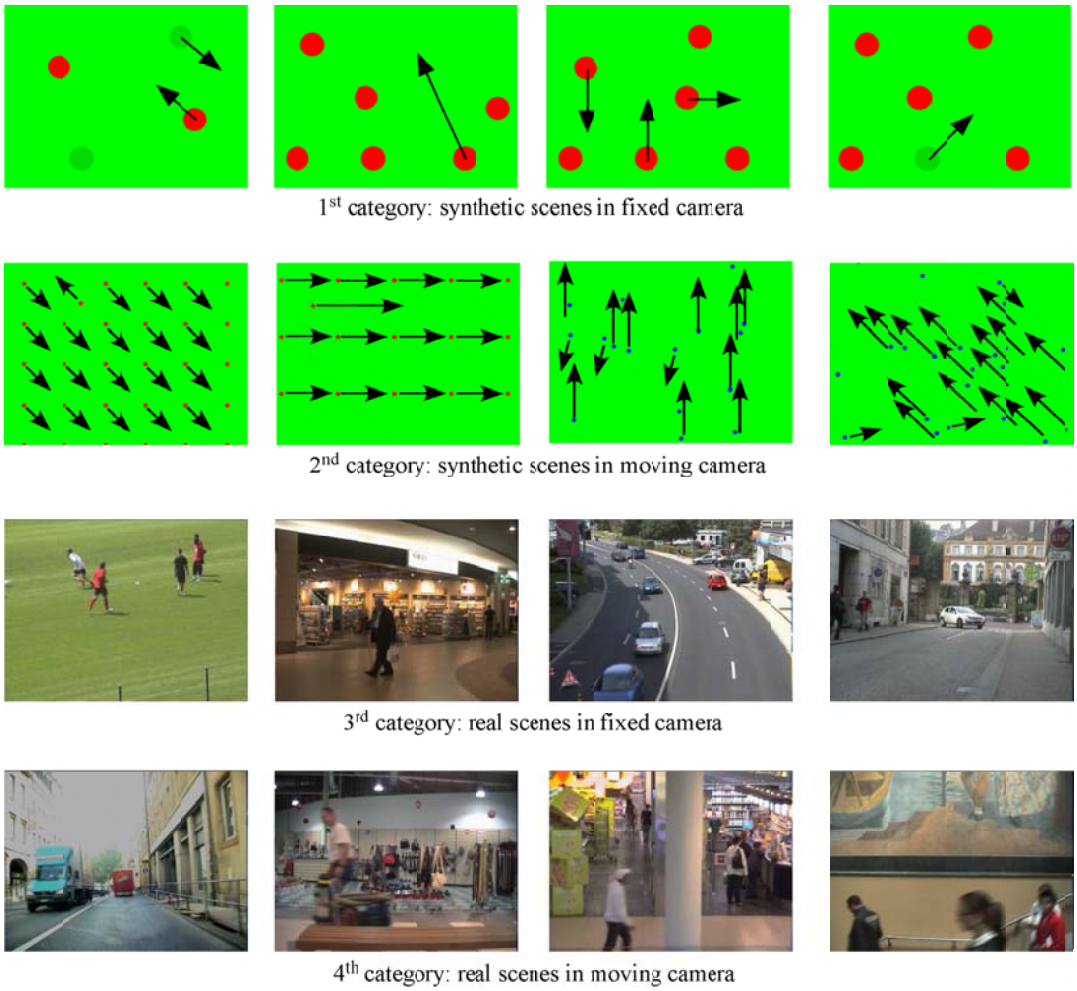


Figure 1. Some examples of the video clips used in the experiments. 84 sequences of synthetic and natural real scenes are considered, classified in four categories: synthetic scenes with fixed camera, synthetic scenes with moving camera, real scenes with fixed camera and real scenes with moving camera. The arrows indicate the direction of motion for the synthetic scenes. The first category contains sequences combining static, moving, high-color-contrasted and low-color-contrasted spots in a uniform background. The second category contains videos composed of one or several moving dots (targets supposed to be salient) among a background of moving dots (distractors). Various configurations of motion contrast are considered: target dot(s) moving either faster or slower and in the same direction relatively to the background; target dot(s) moving slower, at the same speed, or faster and in another direction relatively to the background. The third and fourth categories include videos of natural real scenes in outdoor and indoor environment (city, traffic road, street, football field, train station, stores).

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. 20 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 was preceded by a central fixation cross for 2 seconds. The subjects were instructed to freely look at the sequences with no specific task.

5. RESULTS

This section discusses the suitability of the considered models, according to the different sequence categories (defined in Sect. 4.1). Sections 5.2 and 5.3 discuss the model evaluation in the context of sequences with fixed and moving camera perspective. Qualitative and quantitative evaluations highlight the advantages and inconveniences of each model, as well as preferred situations of application. Finally, Sect. 5.4 compare the performances between synthetic and real scenes. Before entering in the details of the models evaluation, general observations concerning the average human visual behavior over all sequences are given in the next section.

5.1. Motion Contrast and Human Visual Behavior

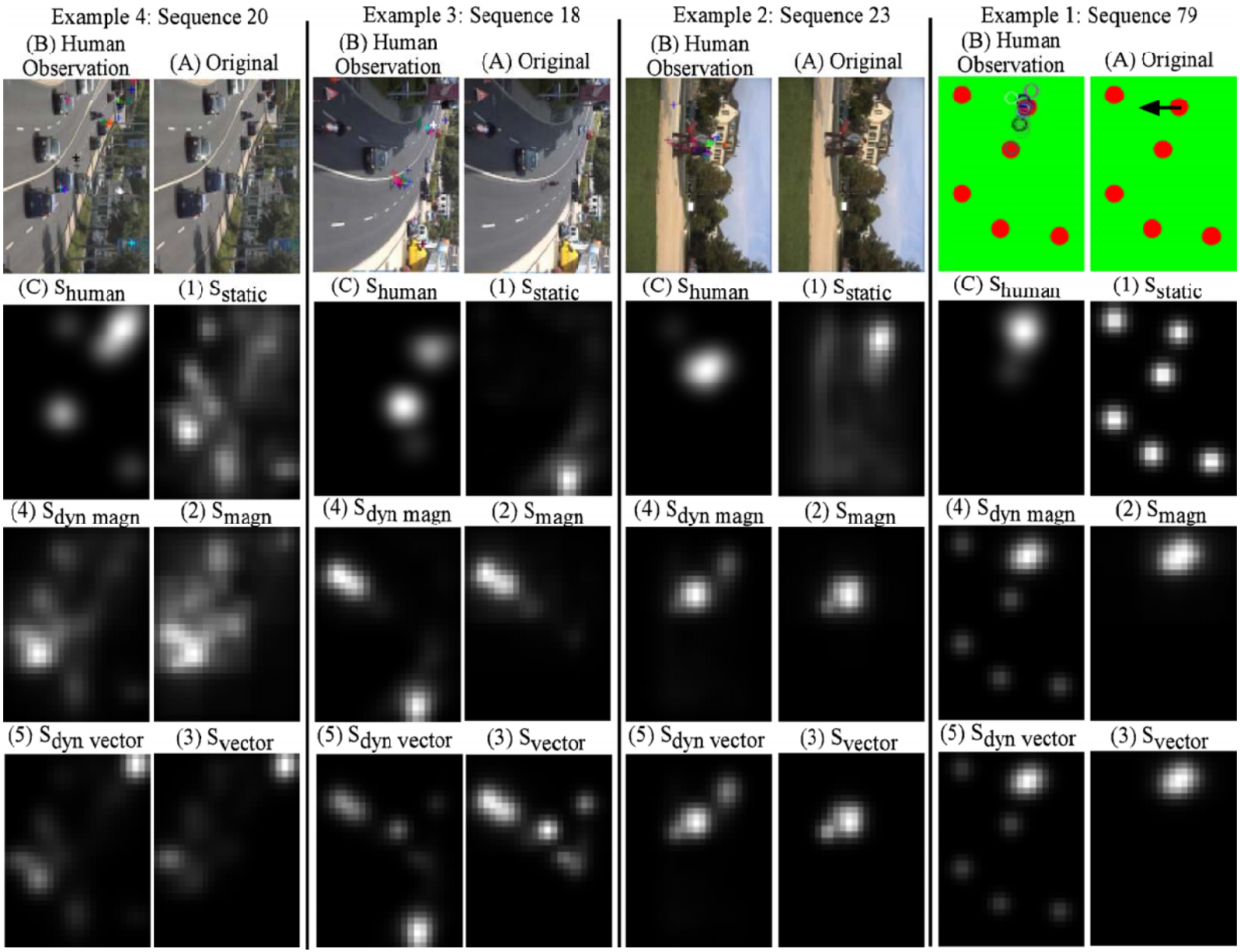
Over all the sequences, most human subjects tend to focus as expected on the image regions containing a strong motion contrast, which can be seen as a motion difference between a centered and a surround region. In this study, the content of the synthetic sequences has been chosen according to various configurations of motion contrast (described in Fig. 1). In all considered configurations, most human observations focused on the target supposed to be salient: target dot(s) moving faster or slower and in the same direction relatively to the background; target dot(s) moving slower, at the same speed, or faster and in another direction relatively to the background. Theses general observations confirm previous studies, performed on complex dynamic scenes, having shown that motion contrast is one of the most important visual attractor.[5,7]

As reminder, *the magnitude models* refer to the 2nd and 4th computer models, while *the vector models* refer to the 3rd and 5th computer models defined in Sect. 2.

5.2. Model Evaluation: Fixed Camera Perspective

Model evaluation of synthetic and real scenes with fixed camera perspective allows to illustrate the variation of the model performance, depending on the sequences content. In a first situation involving a limited number of moving objects (typically 1 to 3), *the vector models* perform identically compared to *the magnitude models*. Examples 1 and 2 in Fig. 2 illustrate this situation. Each example shows the original frame (A), the human observations (B), the human saliency map (C) and (1) to (5) the computer saliency maps according to their respective computer models. In both examples, most human observations focused on the moving object (see (B) and (C)) and the computer saliency maps according to *the magnitude models* ((2) S_{magn} and (4) $S_{dyn\ magn}$) as well as those according to *the vector models* ((3) S_{vector} and (5) $S_{dyn\ vector}$) are very similar compared to the human saliency map (C). Furthermore, a visual comparison of (2) and (3) (respectively (4) and (5)) shows similar performances.

In a second situation, performances are discussed for sequences content including a large set of objects moving in a given direction while another object is moving in a different direction. This situation is illustrated in Example 3 (traffic scene with cars driving from top to bottom while a pedestrian (in the center of the image) is crossing the road from right to left) and in Example 4 (traffic scene with cars driving in two main directions while a truck (top left of the image) is driving from left to right. Most human observations focused on the pedestrian in Example 3 and on the truck in Example 4. Furthermore, the computer saliency maps S_{vector} and $S_{dyn\ vector}$ are more similar to the human saliency map (C) compared to S_{magn} and $S_{dyn\ magn}$. Thus, both examples show a situation in which *the vector models* are more suitable than *the magnitude models*.


Example 4: Sequence 20
(B) Human Observation
(A) Original
(C) S_{human}
(1) S_{static}
(4) $S_{dyn\ magn}$
(2) S_{magn}
(5) $S_{dyn\ vector}$
(3) S_{vector}
Example 3: Sequence 18
(B) Human Observation
(A) Original
(C) S_{human}
(1) S_{static}
(4) $S_{dyn\ magn}$
(2) S_{magn}
(5) $S_{dyn\ vector}$
(3) S_{vector}
Example 2: Sequence 23
(B) Human Observation
(A) Original
(C) S_{human}
(1) S_{static}
(4) $S_{dyn\ magn}$
(2) S_{magn}
(5) $S_{dyn\ vector}$
(3) S_{vector}
Example 1: Sequence 79
(B) Human Observation
(A) Original
(C) S_{human}
(1) S_{static}
(4) $S_{dyn\ magn}$
(2) S_{magn}
(5) $S_{dyn\ vector}$
(3) S_{vector}

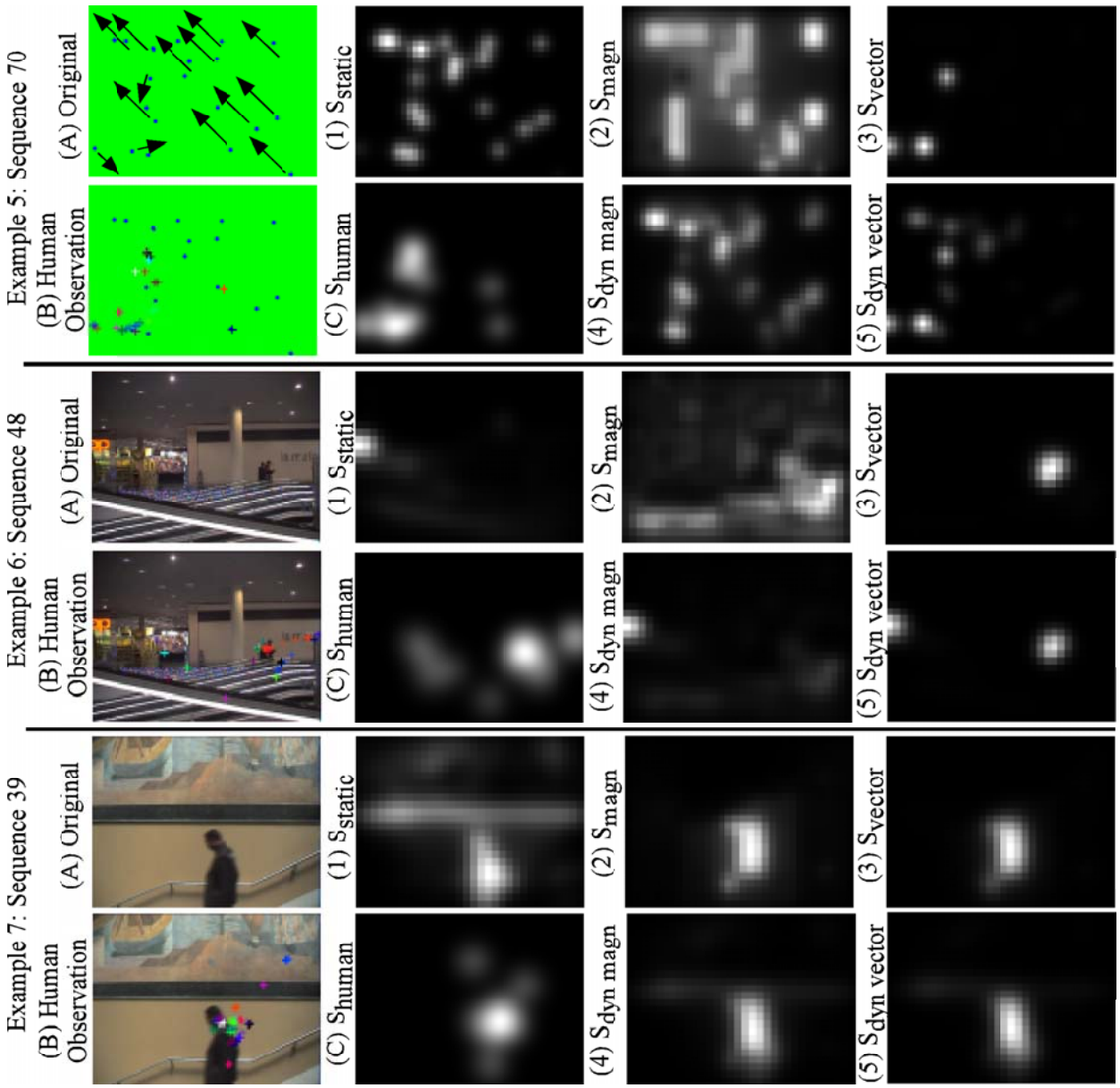


Figure 2. Comparison of the human saliency map issued from the human recordings with the computer saliency maps issued from the five considered models. (A) represents the original frame; (B) the human observations; (C) the human saliency map issued from the fixation and smooth pursuit periods; (1) to (5) the saliency maps issued from the five considered models. Sequences in fixed camera are presented in Examples 1 to 4, while sequences in moving camera in Examples 5 to 7. Example 1: one dot is moving while the others stand still. Example 2: two persons are walking on the path. Example 3: traffic scene with cars driving while a pedestrian (in the center of the image) is crossing the road. Example 4: traffic scene with cars driving in two main directions while a truck (top left of the image) is driving from left to right. Example 5: a background of moving dots (motion direction indicated by the arrows) while three dots are moving in a unique direction. Example 6: scene acquired from a moving escalator, the background is moving from right to left while two persons are walking in the opposite direction. Example 7: scene acquired from a moving escalator, the background is moving from bottom right to top left while a person is moving in the opposite direction.

Over all sequences with fixed camera perspective, the performances of *the vector models* are generally similar compared to *the magnitude models*. The Quantitative evaluation (top table of Table 1) confirms this statement. For both categories (synthetic and real scenes) and for each computer model, the table shows the averages of the different scores and their respective standard deviations. The scores calculated for *the magnitude models* and *the vector models* are very close ((2) S_{magn} compared to (3) S_{vector} and (4) $S_{dyn\ magn}$ compared to $S_{dyn\ vector}$) and thus reflect the qualitative evaluation. However, the specific situation discussed above (sequence content including a large set of objects moving in a given direction while an object is moving in a different direction) suggests that *the vector models* are more suitable. A careful interpretation may suggest that, even in sequences with fixed camera perspective, human attentional behavior may be influenced by motion direction. Moreover, this interpretation may be reinforced by the example of the ballet: spectators tend to look at the main dancer who is moving alone in a given direction, while all other dancers are moving in another direction. However, this interpretation still remains hypothetic, since the content of most considered sequences includes a limited number of moving objects. Globally, the proposed experiments with fixed camera perspective show similar performances between *the vector models* and *the magnitude models*.

Table 1. Quantitative evaluation of the five considered computer models using three different scores: the correlation coefficient s_{cc}, the fixation-to-chance distance s_{ftc} and the Kullback Leibler divergence s_{KL}. For each of the four categories, the average scores are given with their respective standard deviation. As reminder, s_{cc} and s_{ftc} are scores of similarity, while s_{KL} is a score of dissimilarity.

	1st category (15 video): synthetic in fixed camera			3rd category (30 video): real in fixed camera		
computer model	s_{cc}	s_{ftc}	s_{KL}	s_{cc}	s_{ftc}	s_{KL}
(1) S_{static}	0.29 ± 0.10	5.56 ± 2.69	5.66 ± 1.16	0.14 ± 0.15	1.08 ± 1.73	4.59 ± 1.01
(2) S_{magn}	0.47 ± 0.21	6.57 ± 5.12	2.91 ± 1.24	0.39 ± 0.17	5.18 ± 3.55	2.78 ± 1.20
(3) S_{vector}	0.47 ± 0.22	7.49 ± 5.01	2.77 ± 1.18	0.39 ± 0.14	5.85 ± 3.58	2.79 ± 1.12
(4) $S_{dyn\ magn}$	0.51 ± 0.13	7.91 ± 3.09	3.83 ± 0.65	0.38 ± 0.16	3.85 ± 2.37	3.37 ± 0.98
(5) $S_{dyn\ vector}$	0.52 ± 0.13	8.43 ± 2.98	3.82 ± 0.67	0.38 ± 0.14	4.09 ± 2.35	3.36 ±0.90

	2nd category (19 video): synthetic in moving camera			4th category (20 video): real in moving camera		
computer model	s_{cc}	s_{ftc}	s_{KL}	s_{cc}	s_{ftc}	s_{KL}
(1) S_{static}	0.31 ± 0.11	4.40 ± 3.12	4.69 ± 1.08	0.20 ± 0.16	1.30 ± 2.06	4.26 ± 0.98
(2) S_{magn}	0.26 ± 0.15	1.92 ± 2.11	4.45 ± 0.92	0.22 ± 0.16	1.95 ± 2.79	4.14 ± 1.16
(3) S_{vector}	0.45 ± 0.10	9.82 ± 5.67	3.05 ± 0.53	0.21 ± 0.15	2.88 ± 3.35	4.37 ± 1.19
(4) $S_{dyn\ magn}$	0.34 ± 0.10	3.53 ± 2.2	4.25 ± 0.95	0.26 ± 0.14	2.10 ± 2.05	4.05 ± 0.99
(5) $S_{dyn\ vector}$	0.47 ± 0.07	8.59 ± 3.65	3.39 ± 0.49	0.24 ± 0.14	2.67 ± 2.44	4.37 ± 1.00

5.3. Model Evaluation: Moving Camera Perspective

Model evaluation in the context of real and synthetic sequences with moving camera perspective is discussed below. Sequences generally contain a dominant motion due to the moving camera and differ considerably from sequences with fixed camera perspective. Sequence content is thus composed of a moving background, while other objects are moving relatively to the background. Examples 5, 6 and 7 in Fig. 2 illustrate this situation. In Example 5, a background of dots is moving while three other dots are moving in different directions relative to the background. Examples 6 and 7 show indoor scenes acquired from a moving escalator while a few persons are walking in the opposite direction relatively to the escalator. In these examples, most human observations are located in regions containing a strong contrast of motion. In examples 5 and 6, motion contrast is due to relative differences in direction and thus *the vector models* are clearly more suitable for predicting the human observations than *the magnitude models*. Example 7 illustrates a situation in which the motion contrast in magnitude is high (person moving faster than the background). Thus, *the magnitude models* perform equally well as *the vector models* in this example.

Regarding the models suitability in the context of sequences with moving camera perspective, these specific three examples are representative of the overall evaluation in synthetic scenes as well as in real scenes. From these experiments, it can be concluded that *the magnitude models* are limited for detecting a strong motion contrast in magnitude (high local motion relative to low global motion or the inverse), typically an object moving faster than the background motion. Most important point, *the vector models*, which are also suitable for detecting such motion contrast, have the additional ability to highlight a motion contrast due to differences of directions (phase of the motion vector), typically an object moving at the same speed range but in a different direction compared to the motion background.

The quantitative evaluation is presented in the bottom table of Table 1, the left side and the right side respectively showing the average scores for the 2st category (synthetic scenes) and for the 4th category (real scenes). In the context of synthetic scenes, *the vector models* clearly outperforms *the magnitude models*. This is expected, since the content of most synthetic sequences is dominated by motion contrast in direction. In the context of real scenes, *the vector models* and *the magnitude models* show no significative differences. This surprising result is in fact explained by the content of most real sequences, which include more frequently strong contrast in magnitude than in direction. For predicting human visual behavior in the context of video sequences with moving camera perspective, we can conclude that finally *the vector models* are more suitable than *the magnitude models*

5.4. Performance Comparison for Synthetic and Real scenes

The remaining discussion compares the models performances depending on the nature of the scene. The quantitative evaluation of Table 1 illustrates the differences of model performances obtained between synthetic and natural real scenes. Globally, the scores computed for synthetic scenes are higher than those computed for the real scenes. This suggests that the nature of the scene influences the models performance. Furthermore, these experiments suggest a stronger top-down influence for the real scenes compared to the synthetic scenes. This make sense since the content of the real scenes is much more complex. Here are two examples of top-down influences observed in the real scenes: first, human observations focused more frequently on the head of a moving person, while dynamic models highlight stronger motion contrast of the body, typically the motion of the legs. Second, human observations focussed on lettering.

6. CONCLUSION

This article provides a comparison of dynamic computer models of visual attention, designed with the motion component expressed as the speed vector and the speed magnitude. Five computer models are considered: a static model, two motion models designed with the speed magnitude on one hand and with the speed vector on the other hand, and two dynamic models combining both static and motion channels. The models are compared by measuring their respective performances with the eye movement patterns of human subjects, while viewing video sequences. The model suitability is evaluated in four categories (synthetic and real sequences, acquired with fixed and moving camera perspective), showing advantages and inconveniences of each method as well as preferred situation of application.

The model evaluation with fixed camera perspective concluded to comparable performances between the motion models, while with moving camera perspective, *the vector models* perform better than *the magnitude models*.

The magnitude models are limited for detecting motion contrast in magnitude, typically an object moving faster than the background motion. When the background is not moving, these models highlight efficiently the motion contrast and thus fixed camera perspective correspond to the preferred situation of application for *the magnitude models*. Contrary to *the magnitude models*, *the vector models* have the additional ability to highlight a motion contrast due to differences of directions (phase of the motion vector), typically an object moving at the same speed range but in a different direction compared to the motion background. *The vector models* are suitable in both situations of application, in fixed as well as in moving camera perspective.

In conclusion, *the vector models* have the advantage to perform well in both situations, with moving and fixed camera perspective, while *the magnitude models* performs well only in the situation with fixed camera perspective. On the other hand, *the magnitude models* have the advantage in term of computational cost.

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. L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transaction on Image Processing* **13**, pp. 1304–1318, 2004.
3. 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.
4. C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology* **4**, pp. 219–227, 1985.
5. L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Visual Cognition* **12**, pp. 1093–1123, 2005.
6. 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* **6492**, To be published in february 2007.
7. O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research* **47**, pp. 2483–2498, 2007.
8. E. Barth, M. Dorr, M. Böhme, K. R. Gegenfurtner, and T. Martinetz, "Guiding the mind's eye: improving communication and vision by external control of the scanpath," in *Human Vision and Electronic Imaging*, B. E. Rogowitz, T. N. Pappas, and S. J. Daly, eds., *Proc. SPIE* **6057**, 2006.
9. A. Bur and H. Hügli, "Optimal cue combination for saliency computation: A comparison with human vision," in *Second International Work-Conference on the Interplay Between Natural and Artificial Computation (IWINAC 2007)*, *Lecture Notes in Computer Science* **LNCS 4528**, pp. 109–118, 2007.
10. 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.
11. J. K. Tsotsos, Y. Liu, J. C. Martinez-Trujillo, M. Pomplun, E. Simine, and K. Zhou, "Attending to visual motion," *Comput. Vis. Image Underst.* **100**(1-2), pp. 3–40, 2005.
12. 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.
13. 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.