Cogn Comput manuscript No.
(will be inserted by the editor)

Improving Visual Saliency by Adding 'Face Feature Map' and 'Center Bias'

Sophie Marat* · **Anis Rahman*** · **Denis Pellerin** · **Nathalie Guyader** · **Dominique Houzet**

Received: date / Accepted: date

Abstract Faces play an important role in guiding visual attention, thus the inclusion of face detection into a classical visual attention model can improve eye movement predictions. In this study, we proposed a visual saliency model to predict eye movements during free viewing of videos. The model is inspired by the biology of the visual system, and breaks down each frame of a video database into three saliency maps, each earmarked for a particular visual feature. (i) A 'static' saliency map emphasizes regions that differ from their context in terms of luminance, orientation and spatial frequency. (ii) A 'dynamic' saliency map emphasizes moving regions with values proportional to motion amplitude. (iii) A 'face' saliency map emphasizes areas where a face is detected with a value proportional to the confidence of the detection. In parallel, a behavioral experiment was carried out to record eye movements of participants when viewing the videos. These eye movements were compared with the models' saliency maps to quantify their efficiency. We also examined the influence of center bias on the saliency maps, and incorporated it into the model in a suitable way. Finally, we proposed an efficient fusion method of all these saliency maps. Consequently, the fused master saliency map developed in this research is a good predictor of participants' eye positions.

Keywords visual saliency model · static features · dynamic features · face features · center bias · gaze prediction · video

* Contributed equally as first authors to this work.

S. Marat
University of Southern California,
Los Angeles, CA, USA
E-mail: sophie.marat.ilab@gmail.com

A. Rahman · D. Pellerin · N. Guyader · D; Houzet
Department Images and Signal,
GIPSA-lab, Grenoble, France
E-mail: {firstname}.{lastname}@gipsa-lab.grenoble-inp.fr

The final publication is available at www.springerlink.com.

1 Introduction

Faces play an important role in guiding visual attention, and they immediately attract the eyes when people are looking at static images [1]. Many studies conclude that this natural preference is behind the early-onset responses to face stimuli [2–4]. Moreover, in [5], even when participants were explicitly asked to not attend to faces, they had difficulty doing so. On the contrary, other object-stimuli are easily avoided when faces are the targets of attention. These results indicate a natural preference for faces. Therefore, this preference might lead to a slight delay in the deployment of complete endogenous or voluntary control for attention [6–9]. The bias might be linked to the social and biological importance of faces, or the information conveyed by faces: eye gaze, visual speech, and facial emotions.

Different neuroimaging studies have established that faces activate relatively specific brain areas in the fusiform gyrus [10]. There is also strong evidence of a fast sub-cortical face-processing pathway that operates on low-spatial frequencies and modulates cortical responses [11, 12]. Furthermore, other studies suggest that changes in faces are quickly detected compared to other object stimuli [2]. This implies speedier processing of faces, and automatic shifts of attention without endogenous control; consequently, neglecting other objects presented alongside faces [13–15].

Numerous visual attention models have been proposed to predict eye movements for static images [16–20], or for dynamic images [21–24]. Most of them are based on low-level image features (such as color, orientations and spatial frequencies, motion), despite the fact that high-level stimulus properties (e.g., semantic information) also play an important role in visual perception. A recent paper [25] demonstrates that a combined model of high-level object detection and low-level saliency significantly outperformed a low-level saliency model in predicting eye movements. Other studies work on adding a face detection algorithm to increase the saliency at the location of a face [1,26]. In fact, [27] finds that visual saliency computed through a classical visual attention model, similar to the one proposed by Itti and Koch [18], does not explain human eye fixations when looking at videos with complex social scenes. Moreover, they conclude that observers often direct their initial gaze toward the eyes and heads of people present in the scenes, and these elements are not emphasized using a classical visual attention model.

Based on all these reported researches, it appears important to add a 'face feature' into existing visual attention models to improve their efficiency to predict eye movements. This inclusion of faces in models was already done in [1, 28] using static visual stimuli, and in [26] using dynamic stimuli but for a specific application—video summarization. However, none of the studies incorporate faces in the dynamic visual attention model, and compare model predictions to eye movements recorded during free viewing of videos. The aim of the present research was to study the impact of faces on the recorded eye movements of observers looking at videos with various types of content, and to examine whether the face feature is biased towards the center of the stimuli, as are the other elementary visual features. Our main contributions are (1) a saliency model which combines low-level feature extraction (static features with orientations and spatial frequencies, dynamic features with object motion amplitude) with a higher-level feature: face, and the comparison to eye movements in videos, and (2) the analysis of the impact of the center bias on the 'face feature'.

The organization of the paper is as follows. In Section 2, we present the model with static, dynamic and face features. We emphasize face detection based on Viola-Jones algorithm [29]. Section 3 describes the evaluation of the model, and the contribution of the face pathway.

2 Visual Saliency Model

The presentation of the model in Figure 1 is broken down into in three steps: the initial spatio-temporal saliency model (Section 2.1), the model integrating face features (Section 2.2), and the final model that takes into account the center bias (Section 2.3).

2.1 Spatio-temporal Saliency Model

The spatio-temporal saliency model integrates two pathways (static and dynamic) drawn on the left side of the Figure 1. The model uses two saliency maps: a static and dynamic one. The computation of these maps shares common stages: a retina-like filter and a cortical-like bank of filters. Each pathway is tuned to the processing of a specific type of feature (luminance, orientation and spatial frequency for the static pathway and motion amplitude for the dynamic pathway). The details of this part are given in [23].

Retina-like Filter

The retina model is simply simulated by a cascade of three linear filters. It simulates the two main outputs of the retina that are projected on the magnocellular and parvocellular cells. The first output is the high-spatial frequencies of the scene that is used as input for the static pathway. On the contrary, the second output extracts low-spatial frequencies for the dynamic pathway.

Cortical-like Filters

Visual information is processed into different spatial frequencies, orientations and motion in the primary visual cortex (V1) [30]. The model classically simulates the primary visual cortex complex cells through a bank of Gabor filters with six orientations and four frequency bands in the Fourier domain.

The Static Pathway

As mentioned above, the high-spatial frequency output of the retina is the first stage of the static pathway which gives detailed information about the visual signal. This information is further processed by a bank of Gabor filters, where each filter is sensitive to a specific orientation and a spatial frequency band. The outputs of the filter bank are raw feature maps. Afterwards, the different raw feature maps interact together to reinforce objects belonging to a specific orientation.

The last step of this pathway models the fact that a region is salient if the region is different to its neighbors. Thus, to strengthen intermediate maps that

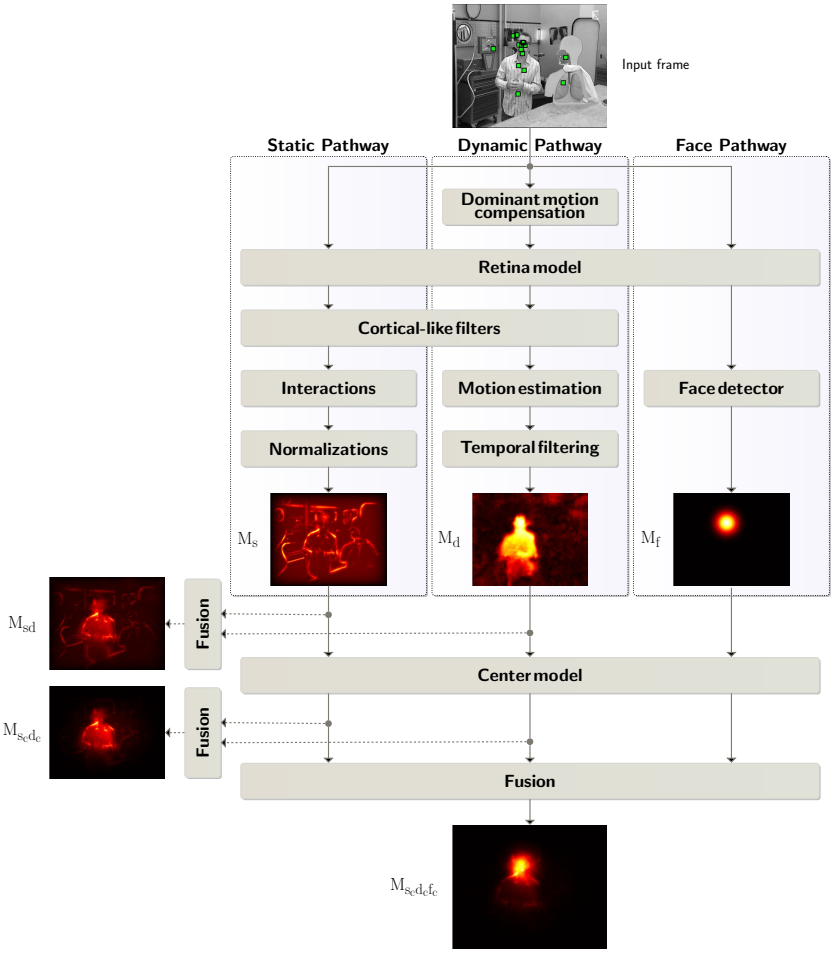


Fig. 1: Block diagram of the proposed visual saliency model with three saliency maps dedicated to specific features: static, dynamic, and face. All these features are computed in parallel pathways, and resultantly each produces a saliency map—such as M_s, M_d, and M_f. The maps may then be fused together either before or after applying the center model to analyze the influence of the center bias. Here, $M_{s_c d_c f}$ is the final saliency model that combines all the three features with center bias.

have spatially distributed maxima, the method developed by Itti [18] was used. Finally, all the intermediate maps were added together to obtain a static saliency map M_s for each frame k (Figure 2b).

The Dynamic Pathway

Humans see stable and moving components in a movie effortlessly. This is true for the case when an object tracked by a camera is seen as moving even if it is

stationary on frames. Therefore, we assumed that the human gaze is attracted by motion contrast, which is the motion of a region against its neighbors. The dynamic pathway starts with the estimation and compensation of relative motion using the 2D motion estimation algorithm developed in [31]. At the output of this algorithm camera motion is compensated, and then the retina and the Gabor filters allow moving objects to be extracted.

Motion Estimation: A differential approach was used for motion estimation that relies on the assumption of luminance constancy. For every frame, the optical flow constraint was applied to each Gabor filter output in the same frequency band. This resulted in an over-determined system of equations, which overcomes the aperture problem [32]. A motion vector was defined (per pixel) by its modulus and angle—corresponding to the motion amplitude and direction respectively. We only used the modulus of the motion vector to define the saliency of an area, assuming that the motion saliency map of a region is proportional to its speed against the background.

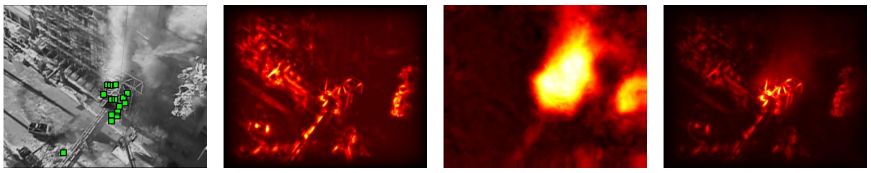
Temporal Filtering: If a pixel moved in one frame but not in the previous ones, it is probably the noise resulting from motion estimation. Hence, a temporal median filter was applied to five successive frames—the current one and the four previous frames—to remove this possible noise. Finally, a dynamic saliency map M_d was obtained for each frame k (Figure 2c).

Two-pathway Fusion

The fusion of static and dynamic saliency maps is done by assigning weights to each of them using their relevant statistics, thus, combining them efficiently. For the static saliency maps, the maximum can express the power of the most salient region in a map. We observed experimentally that frames with high maximum values for static saliency maps better explain eye movements than frames with low maximum values. In case of dynamic saliency maps, we found that the maps have higher skewness when there is a small object in motion, and we wanted to enhance such maps, as they strongly predictive of eye positions. Hence, we chose the skewness to weight the dynamic saliency maps for fusion, because high skewness reflects a better predictability of eye positions. It was shown previously in [23] that our fusion method is a better predictor than a simple average fusion. Consequently, the fusion is carried out using the equation:

$$M_{sd} = \alpha M_s + \beta M_d + \alpha\beta M_s M_d$$
$$where, \begin{cases} \alpha = max(M_s) \\ \beta = skewness(M_d) \end{cases}$$

In the above equation, the fusion incorporates a term $M_s \times M_d$ that allows the salient areas in both static and dynamic pathways to be reinforced. The saliency maps M_s and M_d for the two-pathway model and their resulting master saliency map are shown in Figure 2.



(a) (b) (c) (d)

Fig. 2: From left to right the input frame with superimposed human eye positions, the static saliency map M_s, the dynamic saliency map M_d, the fusion saliency map M_{sd}.

2.2 Model with Face Feature

Face detection and face perception are active fields of research, which over the years have interested people from varying fields of research. These areas include visual cognition, neuroscience, and computer vision. Different studies claim that face detection is coded in a specific cortical area of the brain; the fusiform face area (FFA) [10, 33, 34]. Electrophysiological studies show that face processing is very fast, and human faces evoke a negative potential at 172ms (N170) [35].

A recent study [1] shows that eye movements during free viewing of static natural scenes are largely attracted by faces. Almost 80% of the participants focus on a face within the first two eye fixations. Consequently, the authors improved the predictability of their model by the inclusion of a face pathway.

In this study, we further investigated the influence of faces in dynamic stimuli (videos with various content). We used the Viola-Jones face detector [29] to extract faces in a visual scene along their confidence scores. These scores were put to work for the rejection of bad face detection, and to build a face saliency map. To sum up, the proposed model introduces a face as an important information feature which is extracted in parallel alongside other classical visual features (static and dynamic features). The improvement of performance of the spatio-temporal model using face features is critical because,

- In the recorded eye positions, we observed that faces attract gaze [36]. This behavior towards faces was anticipated, as it has already been studied in the context of free-viewing of static images [1].
- In the spatio-temporal model, faces were not emphasized enough either in static or dynamic saliency maps. In the static pathway, the textured parts of a face (eyes, nose, mouth) can be salient but their saliency could be outperformed by salient regions in the background, while for dynamic pathway, faces could be salient, unless they are moving.

Implementation of the Face Pathway

The proposed face pathway uses the Viola-Jones object detector [29]. This object detector is based on the detection of specific features that carry information about the class of object to be detected such as faces, cars, or any other object. This information can be coded by Haar-like features that are sensitive to the orientation

of contrasts among regions. In our case, a human face can be represented as a set of features exhibiting the relationship of contrast of different regions like eyes, nose, mouth, etc. The Viola and Jones Haar-like feature set defines 2-rectangle, 3-rectangle and 4-rectangle features. Each feature determines the presence or absence of certain characteristics in the image, such as edges or changes in texture. For example, a 2-rectangle feature can indicate the boundary between a dark region and a light region.

A Haar-like feature considers adjacent rectangular regions in the detection window, and computes an average pixel value for each region. Then the difference between these values is compared to an already learned threshold to separate non-objects from objects in the detection window. A large number of Haar-like features are required to detect an object robustly, as each represents a 'weak classifier' or 'low-feature detector'. Therefore, these 'weak classifiers' are organized in a cascade of classifiers, which achieves increased detection performance while reducing computation time. Here, the initial cascade starts off with very simple features rejecting the vast majority of image regions. This makes the process simpler, and the cascade becomes more meaningful as it progresses down.

Another advantage of using Haar-like features is the use of integral images also known as summed area tables. The advantage of these tables is that the sum and mean of pixel values of an area of arbitrary size can be computed in constant time. Here, each pixel is a sum of the pixels to its upper left region. It can be computed faster, and it is an effective way of calculating the sum of pixel values for the rectangular feature model. For example, the sum of rectangular areas can be computed in the image, at any position or scale, using only four lookups. Likewise, the Viola-Jones 2-rectangle features need six lookups, 3-rectangle features need eight lookups, and 4-rectangle features need nine lookups.

The working of the pathway is divided into three steps as follows:

- Pre-processing: The Viola-Jones face detection algorithm uses luminance values to extract facial features, which are prone to environmental factors such as ambient illumination. Different image enhancement methods are used to minimize the contrast of regions with over-exposure or under-exposure. In this step, we used the same retina model to extract the high spatial frequencies of the scene, since in the human visual system FFA takes its input from the parvocellular layer of the lateral geniculate nucleus (LGN) [37]. Consequently, the treatment improves the robustness and performance of the detection system in varying illumination conditions.
- Face Detector: The implementation uses the Viola-Jones face detection algorithm from OpenCV library by calling the cvHaarDetectObjects() function. The library also comes with several pre-trained cascade files to detect different types of faces. Here, we used the ones for 'frontal' and 'profile' faces. The function takes a pretreated gray scale image, and uses a search window to scan across the original image to extract facial features. The search window examines all image locations and classifies them as 'Face' or 'Not Face'. This scanning procedure is repeated on several scales to find faces of different sizes by simply resizing the classifier rather than the original image. After the completion of the search process, we obtain multiple neighboring bounding boxes in a positive face region, whereas a single bounding box is often considered as a false detection. The result of the face detector is a set of bounding boxes with

a confidence measure and the type of detection—in our case either profile or frontal face. Here, confidence is the measure of existence of an object at a location after all the information has been collected. The estimate is useful to overcome ambiguities by ranking and rejecting several detections.
We took the video database used in our experimental design to record eye positions of participants detailed in Section 3.1, and used it to evaluate the performance of the face detector. The detection was carried out on all 14155 input video frames. The resulting faces detected were hand-labeled as either a true or a false face detection (Table 1). An example of the faces detected is shown in Figure 3a.

– Post-processing: The face detector was executed twice with two different trained cascades: frontal and profile faces. Hence, the detector returns overlapping face bounding boxes for the same face regions. To judge these face detections, we first computed the overlapping regions among all the face bounding boxes using:

$$A_{i,j} = \sum_{i,j<i} \max \left\{ \frac{h' \times w'}{h_i \times w_i}, \frac{h' \times w'}{h_j \times w_j} \right\} > \tau$$

Here, h and w represent the dimensions of face bounding boxes (both frontal and profile), and h' and w' are the dimensions of their overlapping region. We used a threshold τ of 0.6 or 60% for two face bounding boxes to be considered as overlapping.
This overlapped region A is then used to reject weak detections using their confidence measures C and detected face-type T as follows:

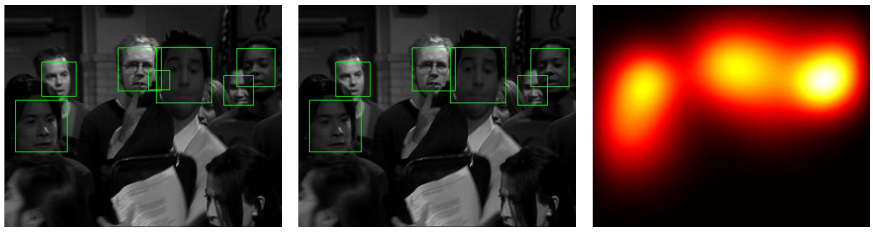
$$Face = \begin{cases} \max(Ci, Cj), & \text{if } T_i = T_j; \\ \max(Ci, Cj), & \text{if } T_i \neq T_j \ \& \ A_{i,j} > 0.8; \\ \min(h_i w_i, h_j w_j), & \text{otherwise.} \end{cases}$$

An example of post-processed faces is shown in Figure 3b.

The face saliency map M_f for the face pathway is generated by marking each detected faces bounding box by a 2D Gaussian. The dimensions of the face bounding box determine the distances from its origin in horizontal and vertical axis, while the confidence score is its width or standard deviation. The resulting face saliency map is shown in Figure 3c.

		Results of face detector	
Total detections	True positive	False positive	Percentage of true positives
7696	5424	2272	70%

Table 1: Results of the face detector for our video database.



(a) (b) (c)

Fig. 3: Raw face detections (left), post-processed face detections (middle), face saliency map M_f after post-processing (right)

Three-pathway Fusion

The fusion method modulates the static, dynamic and face saliency maps using maximum, skewness and confidence measures respectively, and fuses them together using:

$$\begin{aligned} M_{sdf} &= \alpha M_s + \beta M_d + \gamma M_f \\ &\quad + \alpha\beta M_s M_d + \beta\gamma M_d M_f + \alpha\gamma M_s M_f \\ where, &\begin{cases} \alpha = max(M_s) \\ \beta = skewness(M_d) \\ \gamma = mean(confidence) \end{cases} \end{aligned}$$

An example of this fusion is given in Figure 4. As mentioned in Section 2.1 the maximum and skewness are appropriate weightings for static and dynamic saliency maps respectively. Similarly, the weighting suitable for face saliency map is its confidence. The higher the confidence, the greater the probability of the presence of a face in an image is. Hence, this map should have a higher weight for the fusion. Furthermore, we assume that common salient regions in different saliency maps obtain the most representation in the final map. This is done by reinforcing such regions by a multiplicative fusion of different feature maps.

2.3 Model with Center Bias

The central fixation bias in visual scene viewing is selecting an optimal viewing position independent of the image features. There is a strong tendency to look more frequently around the center than in the periphery. The center might not provide an information-processing advantage, but it is an optimal position to explore the visual scene. A number of factors contribute to this effect [38–40]: high-level strategic advantage, drop in visual sensitivity in periphery, and motor bias. Moreover, people tend to direct their first saccades in the visual scene towards subjects of interest or salient locations closer to the center, as the initial saccades are a localizing response, and afterwards are the explorations of the objects [41,42]. In this study, we examine the influence of center bias on each of the three visual feature maps: static, dynamic, and face. The observations were used to propose a center model.

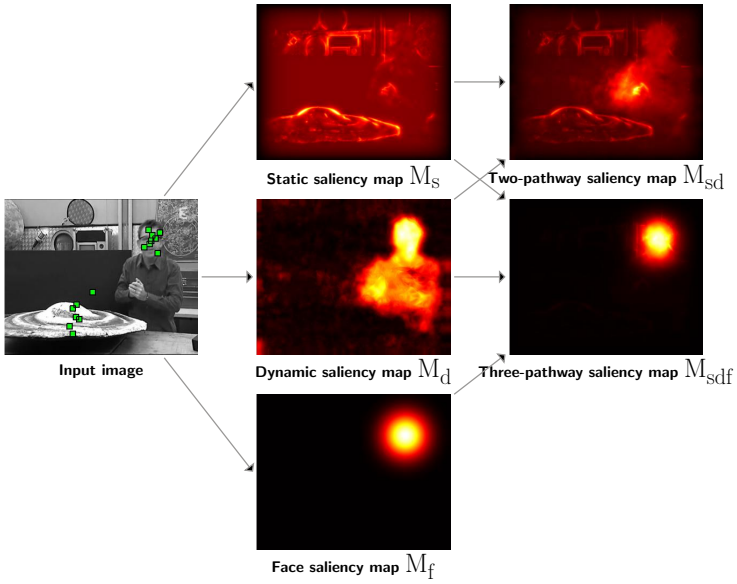


Fig. 4: Block diagram illustrates the fusion of two and three pathway visual saliency models for video database. The two-pathway saliency map M_{sd} is the result of the fusion of saliency maps M_s and M_d, whereas the three-pathway saliency map M_{sdf} also takes into account the face saliency map M_f alongside the other two saliency maps.

The Center Model

The center model is a significant predictor of eye position in arbitrary natural scenes, due to the preference of placement of focal and foreground objects in the center of the screen. This model alone outperforms models without a central bias [21, 43, 44]. Its introduction enhances the correlation between eye positions and computational model output, but it might not be useful for applications of visual attention because it is not specific to visual features. The centered model is considered as a saliency model applying the same 'central' saliency map M_c to all the frames. In our case, this map corresponds to a 2D Gaussian with sigma 10° and dimensions equal to that of the original video frame. It is applied multiplicatively to the spatial information as illustrated in Figure 5.

$$M_{m_c} = M_m \times M_c$$

Here, M_m is for the feature maps from different pathways of the visual saliency model.

Three-pathway Fusion with Center Bias

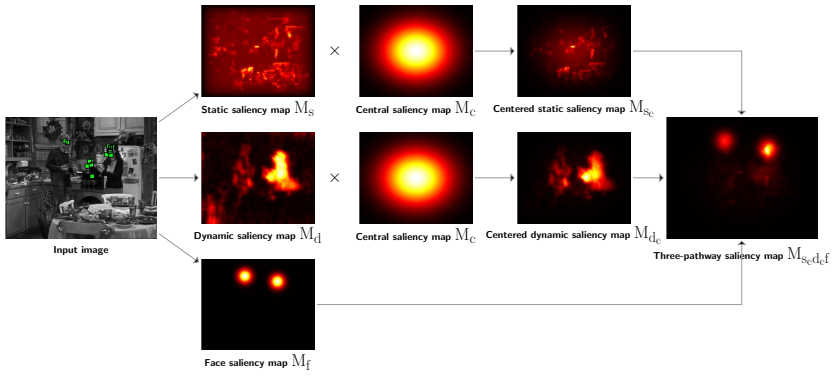
The fusion model fuses the centered static saliency map M_{s_c}, centered dynamic saliency map M_{d_c} and face saliency map M_f using maximum, skewness and

confidence scores respectively.

$$M_{s_c d_c f} = \alpha M_{s_c} + \beta M_{d_c} + \gamma M_f + \alpha\beta M_{s_c} M_{d_c} + \beta\gamma M_{d_c} M_f + \alpha\gamma M_{s_c} M_f$$

The fusion weights α, β and γ are computed for centered static saliency map M_{s_c}, centered dynamic saliency map M_{d_c} and face saliency map M_f respectively. Here, the saliency maps from the face pathway are not modulated with center bias. The reason is the attractiveness of face stimuli regardless of their location around the center of a visual scene. Also, the occasional presence of faces in the video sequences suggests that such weighting did not significantly improve the results of evaluation as shown in Table 5.

Fig. 5: Block diagram illustrates the fusion of saliency maps from the three pathways of the model. The center bias is applied to saliency maps M_s and M_d to obtain centered saliency maps M_{s_c} and M_{d_c}. These two resulting maps are fused to the face saliency map M_f to obtain final saliency map $M_{s_c d_c f}$.



3 Results

3.1 Eye Movement Experiment

The experiment aims to record eye movements of naive participants when looking at videos with various contents. The eye movement data is used to evaluate the final saliency model. Additionally, comparing this data with a saliency model helps to understand which features explain the best eye movements and fixated locations. This experiment was the same as the one described in [23].

Stimuli

To create the dynamic stimuli, we were inspired by the experiment proposed by Carmi and Itti [45]. Fifty-three videos (25 frames per second, 720 × 576 pixels per frame) were selected from heterogeneous video sources including movies, TV

shows, news, animated movies, commercials, sports, and music videos. These videos were cut every 1-3 seconds (mean $1.86s \pm 0.61$) into 305 clip snippets. The length of the clip snippets was chosen randomly with only one constraint—to avoid scene cuts inside a clip snippet. These clip snippets were strung together to obtain 20 'MTV-style' clips of 30 seconds (mean $30.20s \pm 0.81$). Furthermore, each of the clips contained at most one clip snippet from all fifty-three continuous sources. The choice of clip snippets and their duration was random to prevent observers from anticipating scene cuts. We used gray-level stimuli (14155 frames) without audio signal. Moreover, to prevent top-down effects—as the proposed model is bottom-up—we used 'MTV-style' clips rather than classical videos.

Participants

Fifteen young adults (3 women and 12 men, range 23-40 years, mean 28 years) participated in the experiment. All participants had normal or corrected to normal vision.

Data Acquisition

Participants were sitting, with their head stabilized with a chin rest, in front of a 21-inch monitor (75Hz refresh rate) at a viewing distance of 57cm ($40^\circ \times 30^\circ$ field of view). Participants were instructed to look at the videos without any particular task. All participants saw the 20 clips in a random order. Before each clip, we ensured that participants gazed at the screen center. Instantaneous eye positions were tracked by head-mounted EyeLink II cameras (SR Research) in pupil-recording mode with 500Hz temporal resolution. A 9-point calibration was carried out at the beginning of the experiment and every five clips. Drift correction was also done before each clip.

Human Eye Position Density Maps

The eye tracker recorded the two eye positions at 500Hz—20 eye positions (10 positions for each eye) per frame and per participant. The median of these positions was taken (with x-axis and y-axis median) for each frame and for each participant. Then, for each frame, the median position for fifteen participants was considered, which is the raw eye positions map M_p. A two-dimensional Gaussian was added to each position. The standard deviation of this Gaussian was chosen to obtain a diameter at mid-height equal to 0.5° of visual angle, which is close to the size of the maximum resolution of the fovea. Therefore, for each frame, we obtained a human eye position density map M_h. These maps are then used to evaluate saliency maps M_m from the proposed model.

3.2 Model Evaluation Metrics

We used two evaluation metrics to estimate the relevance between normalized predicted saliency maps from the visual saliency model and the eye-position density maps.

– NSS: Normalized Saliency Scanpath is proposed by Peters et al. [46]. It is calculated by averaging pixels that correspond to eye positions. It acts like a z-score computed by comparing a saliency map from the model to fixations of participants. The $NSS_{(i,j)}$ at positions (i, j) of a saliency map is given as:

$$NSS_{(i,j)} = \frac{M_m \cdot M_h - \bar{x}_m}{s_m}$$
$$M_m : saliency\ value\ at\ pixel\ (i,j)$$
$$M_h : human\ position\ density\ map\ at\ (i,j)$$
$$\bar{x}_m : empirical\ mean\ of\ saliency\ map\ M_m$$
$$s_m : empirical\ standard\ deviation\ of\ saliency\ map\ M_m$$

The NSS value can be: (i) zero, when there is no link between experimental eye positions and salient regions; (ii) negative, when the positions are on non-salient regions; or (iii) positive, when they are projected on the salient regions—the higher the positive values of NSS are the more the salient regions are attended.

– TC: Torralba et al. [47] propose a method to evaluate the quality of a visual saliency model. It simply estimates the ratio of the eye positions predicted by the saliency map over all the experimental eye positions. A position is predicted if it is projected on the most salient region, which is 20% of the map surface. When compared to NSS the metric needs to define a threshold, but it is simpler to calculate.

$$TC = 100 \times \frac{N_{inside}}{N_{all}}\%$$
$$N_{inside} : positions\ inside\ salient\ regions$$
$$N_{all} : total\ experimental\ eye\ positions$$

3.3 Evaluation of the Model

Table 2 presents the results for different saliency maps. Here, we calculated the sample mean $\bar{x}$ among first 70 frames for all 305 video snippets, and then took the standard deviation of this sample mean denoted as $SE_{\bar{x}}$. As shown in [23], the dynamic saliency map M_d performs better than the static one M_s for both criteria. This is due to the importance of motion in guiding attention [24, 48]. The lower results for the face saliency maps M_f is explained by the fact that faces are present only in a small percentage of the video database (35%), thus, and for the rest of the frames, the prediction score is zero. Moreover, the fusion integrating the face pathway M_{sdf} outperforms the fusion combining static and dynamic pathway M_{sd}. M_{sdf} takes into account the face information when there is a face, otherwise it is similar to M_{sd} when no face is detected. This additional information considerably improves the results because face is a powerful gaze-attractor.

In Table 2, the saliency maps for all pathways have lower prediction scores than a simple Gaussian at the center of the frame—central saliency map M_c. This demonstrates that center bias has an impact on eye positions during free viewing, and its appropriate integration will improve the model's efficiency in predicting eye movements.

Evaluation results without center model							
Criterion		M_c	M_s	M_d	M_f	M_{sd}	M_{sdf}
NSS	$\bar{x}$	1.25	0.72	0.96	0.57	0.99	1.28
	$SE_{\bar{x}}$	0.019	0.010	0.020	0.019	0.016	0.029
TC (%)	$\bar{x}$	64	49	50	10	57	58
	$SE_{\bar{x}}$	0.735	0.560	0.502	0.379	0.608	0.710

Table 2: NSS and TC results for the different pathways of the model without the center bias, except M_c that represents the center model. The sample mean $\bar{x}$ and its standard error $SE_{\bar{x}}$ were averaged over the 305 video snippets.

3.4 Interest of Separate Face Pathway

It is important to note that if we do not use a face pathway, fusion of the static and dynamic pathways is not sufficient to make faces salient. As already found in the literature, a classical visual attention model (only static and dynamic features) cannot explain gaze on face locations because they are not always emphasized in such a model. To investigate this point, we performed a simple analysis using the static saliency map and true-positive face detections obtained by comparing the detected faces to the hand-labeled ground truth. We started by calculating the mean static saliency values at face locations for all n frames with at least one face; $\{x_1, x_2, ..., x_n\}$. The mean $\bar{x}$ was used as a threshold to split the frames into two subgroups: face salient, and face non-salient frames. Subsequently, we used the two metrics to compute the scores for these subgroups of frames. The resulting scores in Table 3 show that faces have similar scores in the conditions of low or high static saliency. The NSS and TC scores are high for salient faces as expected because faces are attractive. Likewise, non-salient faces also have high scores for both metrics. We can imply based on this observation that static saliency maps do not model the saliency of faces, but are instead only activated by object contrasts. Hence, it is interesting to include a separate face pathway to improve the model's eye movement predictions.

Evaluation results for static saliency of faces		
	Face salient	Face non-salient
Samples(#)	2610	4709
NSS score	2.68	1.85
TC score	42	37

Table 3: NSS and TC results for the frames with at least one face in a condition of high or low static saliency.

3.5 Influence of Center Bias on Eye Positions

Figure 6 is an illustration of all human eye positions from the experiment superimposed as a 2D-image. We clearly observed that there is an apparent center

bias effect in play. This effect could be the result of the experimental setup where the first eye position for a clip starts from the central marker. Also, it could be a significant contribution due to the video content presented such as 'Hollywood movies' [39]. Consequently, this motivates us to incorporate a center model to enhance the relevance among experimental eye positions and visual saliency maps from the proposed model. In fact, practical applications potentially using the visual attention model might prefer to use all salient information, and hence not require this modulation.

Human data is highly center-biased as shown in Figure 6. Hence, adding a larger border will increase the overall performance of the model. The result is consistent with the fact that saliency falls off with the distance from the center. As a result, we find that the simple Gaussian technique outperformed the results of the model without center bias. Consequently, the fusion step then considered the nature of each map, and integrated a center bias when appropriate to reinforce the salient regions. The resulting master saliency map performed better than each pathway predicting independently and standalone Gaussian map from the central model.

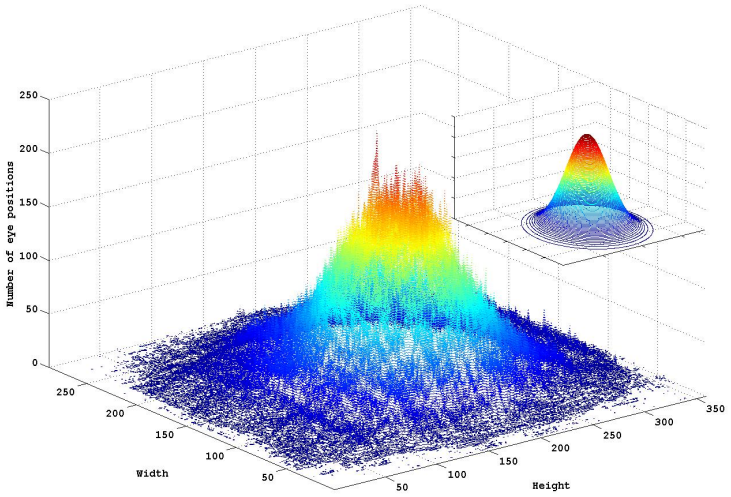


Fig. 6: 2D contour map presents the distribution of participants eye positions for video database, and the distribution after Gaussian fitting is shown in subplot.

3.6 Evaluation of the Model with Center Bias

Table 4 shows that center bias modulation of the saliency maps' results in higher scores for all maps such that M_{s_c} outperforms M_s ($F(1, 609) = 178.59, p < 0.001$)

[1] and gives similar results to M_c. The results between M_c and M_{s_c} are very close, even if M_c still outperforms M_{s_c} with respect to the NSS criteria ($F(1,609) = 4.62, p < 0.05$). M_{d_c} outperforms both M_d ($F(1,609) = 87.36, p < 0.001$) and M_c ($F(1,609) = 18.98, p < 0.001$). More than looking at the center of the frame a participant gazes at what is salient near the center of the frame, since the static and dynamic saliency maps both provide complementary information that is needed to predict the participant eye positions. In addition to the integration of center bias before its fusion into the saliency map $M_{s_c d_c}$ outperforms the simple fusion saliency map M_{sd} ($F(1,609) = 65.25, p < 0.001$) and central saliency map M_c ($F(1,609) = 4.93, p < 0.05$). Similarly, the fusion of the three pathways with center model $M_{s_c d_c f}$ gives the best results, outperforming the central saliency map M_c ($F(1,609) = 45.72, p < 0.001$), centered two-pathway saliency maps $M_{s_c d_c}$ ($F(1,609) = 19.40, p < 0.001$), and three-pathway saliency maps M_{sdf} ($F(1,609) = 34.70, p < 0.001$). Therefore, both center bias and faces are important to consider to obtain a good predictor of eye positions of participants.

		Evaluation results with center model					
Criterion		M_c	M_{s_c}	M_{d_c}	M_{f_c}	$M_{s_c d_c}$	$M_{s_c d_c f}$
NSS	$\bar{x}$	1.25	1.19	1.45	0.57	1.37	1.69
	$SE_{\bar{x}}$	0.019	0.016	0.022	0.020	0.021	0.030
TC (%)	$\bar{x}$	64	65	67	10	68	71
	$SE_{\bar{x}}$	0.735	0.675	0.623	0.375	0.666	0.673

Table 4: NSS and TC results for the different pathways of the model with the center bias, except M_c that represents the center model. The sample mean $\bar{x}$ and its standard error $SE_{\bar{x}}$ were averaged over the 305 video snippets.

3.7 Faces and Center Bias

One wonders which is the most important bias while watching videos; the center bias which tends to make a participant to gaze more at the center, or the particularity of faces which makes participants to look at a face and recognize it more rapidly than other objects. In the former case, Dorr et al. [39] discussed the impact of the center bias on different types of videos, such as professionally made 'Hollywood movies,' and amateur made 'natural movies.' They show an increased impact of center bias in 'Hollywood movies'—the kind of videos that were used to generate our clips. In the latter case, it is shown in [49] that faces are easier to detect than other objects, and the detection facilitation of such stimuli is higher even if it is presented at the periphery. They concluded that the spatial window for face detection is wider than for other objects. In our experiment, we observed similar results when considering the scores for the face saliency map M_f ($NSS = 0.51, TC = 18\%$) or the one weighted by center bias

[1] For clarity, only statistics using NSS criteria are presented since both NSS and TC generally produce the same conclusion. We took the sample mean of 70 frames from each video snippet, and then applied the significance tests.

M_{f_c} ($NSS = 0.55, TC = 19\%$) from Tables 2 and 4 respectively. This result could either be explained by a smaller impact of center bias for faces than other salient objects or by the fact that there were few faces present in the frames to make the center bias significant on face saliency maps.

To investigate further the probable cause of faces attracting human gaze independent of their location, we calculated scores for the maps M_c and M_{f_c}, as presented in Table 5. Here, the frames considered contained at least one face, yet the impact of central weighting is still marginal. Furthermore, there was no impact of the center model on the two scores (NSS and TC) for M'_f and M'_{f_c}, where we considered only the true-positive detections obtained by comparing the detected faces to the hand-labeled faces (ground truth). Therefore, our results agree with the finding of [49] that the presence of faces attracts the gaze of a participant more than their tendency to fixate on the center. This is not the case for the two other features; in fact, adding the center bias on the static and the dynamic saliency maps considerably improved the scores.

Evaluation results for face pathway					
Criterion		M_f	M_{f_c}	M'_f	M'_{f_c}
NSS	$\bar{x}$	1.75	1.75	2.34	2.32
	$SE_{\bar{x}}$	0.054	0.053	0.070	0.069
TC (%)	$\bar{x}$	30	30	38	38
	$SE_{\bar{x}}$	0.963	0.952	1.169	1.156

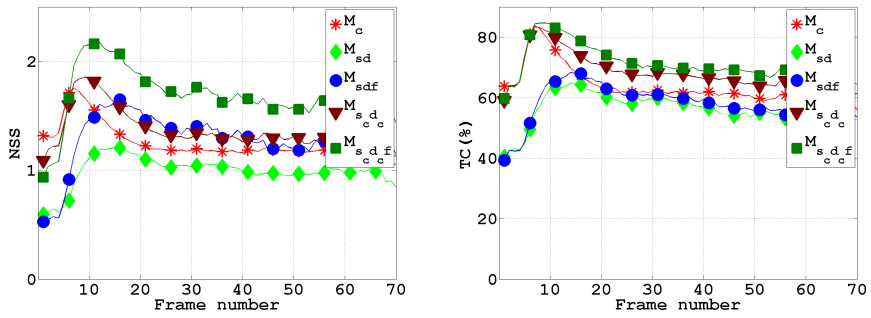
Table 5: NSS and TC results for the face saliency maps (M_f and M_{f_c}) with or without the center model. Similarly, $M_f\prime$ and $M_{f_c}\prime$ are the face saliency maps comprising of only the true-positive face detections. Here, we only consider the video frames with at least one face detected.

3.8 Temporal Evaluation of the Model

The time course of the influence of center bias has also been investigated [38, 41]. To analyze the evolution of the center bias effect on videos, different saliency map scores were plotted along frame position. For each frame position inside a snippet (independently of the snippet position on the clip), all the scores corresponding to that frame position where averaged. We averaged on the first 70 frames even if some snippets were longer.

Figure 7 illustrates that all saliency maps were more predictive at the beginning of each snippet from 5^{th} to 13^{th} frame. Afterwards, the prediction scores decreased, which is consistent with the fact that the proposed model is bottom-up. The peaks corresponding to center biased saliency maps $M_{s_c d_c}$ and $M_{s_c d_c f}$ were sharper compared to saliency maps without center bias M_{sd} and M_{sdf}. The biased saliency maps reached their maximum quickly around the 8^{th} frame, and clearly outperformed maps without center bias. However, they decreased more rapidly showing that center bias is particularly predominant at the scene onset, as also mentioned in [38]. Therefore, the influence of center bias decreases along time, letting other features to take over irrespective of their position in the visual scene.

The temporal evolution of metrics also showed that the introduction of face pathway is an improvement. In Figure 7, we find that three-pathway saliency map M_{sdf} did produce better metric scores compared to two-pathway saliency map M_{sd}. Moreover, M_{sdf} performed comparatively well against the two-pathway saliency map with center bias $M_{s_c d_c}$. This result indicated that face feature as a separate pathway certainly did increase the predictability power of the model. The saliency map was reinforced further by center bias, and the resulting saliency map $M_{s_c d_c f}$ delivered the best scores.



(a) NSS evolution (b) TC evolution

Fig. 7: Evolution of metrics (NSS and TC) for the different pathways with or without the center bias for video database.

4 Conclusion

This study presents a new bottom-up saliency model that breaks down the visual signal using three processing pathways based on different types of visual features: static, dynamic, and face. The static and dynamic pathways are inspired by the biology of the first steps of the human visual system: a retina-like filter and a cortical-like bank of filters. The static pathway extracts the texture information based on luminance. The dynamic pathway extracts information about objects' motion against background. The face pathway extracts information about the presence of faces in the frames. This model also integrates the center bias as a suitable modulation on the different saliency maps.

An eye movement experiment was used to record the gaze of participants viewing various videos freely. This experiment was used to evaluate, and also to improve the saliency model. Each pathway is effective for predicting eye movements. The face pathway is particularly effective for predicting eye movements on frames containing faces, highlighting the importance of integrating face feature in a bottom-up saliency model. The eye movement experiment enables us to study which visual features attract a participant's gaze, and how to integrate them into the saliency model, and more particularly, to the fusion step of the three pathways. The fusion of the three types of maps into a single master saliency map is optimized

by weighting the saliency maps produced by the three pathways using specific coefficients The specific coefficients correspond to particular statistics extracted from the different types of saliency maps (maximum, skewness, and confidence). These weights are then used to strengthen the most relevant feature maps.

The study concentrates on the importance of faces and center bias for the improvement of a visual saliency model. In future work, we hope to analyze the evolution of performance of the proposed model for longer videos, when bottom-up processes are no longer predominant and top-down processes might play an important role on eye movements. Thus, the bottom-up visual saliency model can be integrated with top-down weights to modulate the saliency maps as a function of the goal. The resulting saliency maps from combined stimulus-driven and goal-driven model can give better prediction of eye movements.

References

1. M. Cerf, J. Harel, W. Einhuser, C. Koch, Predicting human gaze using low-level saliency combined with face detection. in *NIPS'07* (2007)
2. T. Ro, C. Russell, N. Lavie, Changing faces: A detection advantage in the flicker paradigm. Psychol Sci **12**(1), 94 (2001)
3. P. Vuilleumier, Faces call for attention: evidence from patients with visual extinction. Neuropsychologia **38**(5), 693 (2000)
4. J. Theeuwes, S. Van Der Stigchel, Faces capture attention: Evidence from inhibition of return. Vis Cogn **13**(6), 657 (2006)
5. M. Bindemann, A.M. Burton, S.R.H. Langton, S.R. Schweinberger, M.J. Doherty, The control of attention to faces. J Vision **7**(10), 15.1 (2007)
6. H.J. Müller, J.M. Findlay, The effect of visual attention on peripheral discrimination thresholds in single and multiple element displays. Acta Psychologica **69**(2), 129 (1988)
7. H.J. Müller, P.M. Rabbitt, Reflexive and voluntary orienting of visual attention: time course of activation and resistance to interruption. J Exp Psychol Human **15**(2), 315 (1989)
8. M. Shepherd, H.J. Mller, Movement versus focusing of visual attention. Percept Psychophys **46**(2), 146 (1989)
9. M.L. Cheal, D.R. Lyon, Central and peripheral precuing of forced-choice discrimination. Q J Exp Psychol **43**(4), 859 (1991)
10. N. Kanwisher, G. Yovel, The fusiform face area: a cortical region specialized for the perception of faces. Philos Trans R Soc London, Ser B **361**(1476), 2109 (2006)
11. G. Loffler, G. Yourganov, F. Wilkinson, H.R. Wilson, fmri evidence for the neural representation of faces. Nat Neurosci **8**(10), 1386 (2005)
12. M.H. Johnson, Subcortical face processing. Nat Rev Neurosci **6**(10), 766 (2005)
13. E. Birmingham, W. Bischof, A. Kingstone, Gaze selection in complex social scenes. Vis Cogn **16**(2), 341 (2008)
14. J. Driver, G. Davis, P. Ricciardelli, P. Kidd, E. Maxwell, S. Baron-Cohen, Gaze perception triggers reflexive visuospatial orienting. Vis Cogn **6**(5), 509 (1999)
15. S. Langton, V. Bruce, Reflexive visual orienting in response to the social attention of others. Vis Cogn **6**(5), 541 (1999)
16. C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol **4**, 219 (1985)
17. J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, F. Nuflo, Modeling visual attention via selective tuning. Artif Intell **78**, 507 (1995)
18. L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis. IEEE T Pattern Anal **20**, 1254 (1998)
19. T. Ho-Phuoc, N. Guyader, A. Guérin-Dugué, A functional and statistical bottom-up saliency model to reveal the relative contributions of low-level visual guiding factors. Cogn Comput **2**(4), 344 (2010)
20. V. Yanulevskaya, J.B. Marsman, F. Cornelissen, J.M. Geusebroek, An image statisticsbased model for fixation prediction. Cogn Comput **3**(1), 94 (2011)

21. O. Le Meur, P. Le Callet, D. Barba, Predicting visual fixations on video based on low-level visual features. Vision Res **47**(19), 2483 (2007)
22. R.J. Peters, L. Itti, Applying computational tools to predict gaze direction in interactive visual environments. ACM T Appl Percept **5**(2), 1 (2008)
23. S. Marat, T.H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guérin-Dugué, Modelling spatio-temporal saliency to predict gaze direction for short videos. Int J Comput Vision **82**, 231 (2009)
24. P.K. Mital, T.J. Smith, R.L. Hill, J.M. Henderson, Clustering of gaze during dynamic scene viewing is predicted by motion. Cogn Comput **3**(1), 5 (2010)
25. M. Cerf, E.P. Frady, C. Koch, Using semantic content as cues for better scanpath prediction. in *Proceedings of the 2008 symposium on Eye tracking research & applications* (2008)
26. Y.F. Ma, X.S. Hua, L. Lu, H.J. Zhang, A generic framework of user attention model and its application in video summarization. IEEE T Multimedia **7**, 907 (2005)
27. E. Birmingham, W.F. Bischof, A. Kingstone, Saliency does not account for fixations to eyes within social scenes. Vision Res **49**(24), 2992 (2009)
28. L.Q. Chen, X. Xie, X. Fan, W.Y. Ma, H.J. Zhang, H.Q. Zhou, A visual attention model for adapting images on small displays. Multimedia Syst **9**(4), 353 (2003)
29. P. Viola, M.J. Jones, Robust real-time face detection. Int J Comput Vision **57**, 137 (2004)
30. D.H. Hubel, T.N. Wiesel, Functional architecture of macaque monkey visual cortex. Society **198**(1130), 1 (1977)
31. J.M. Odobez, P. Bouthemy, Robust multiresolution estimation of parametric motion models applied to complex scenes. J Visual Commun Image Represent **6**, 348 (1995)
32. E. Bruno, D. Pellerin, Robust motion estimation using spatial gabor-like filters. Signal Process **82**, 297 (2002)
33. A. Mechelli, C.J. Price, K.J. Friston, A. Ishai, Where bottom-up meets top-down: neuronal interactions during perception and imagery. Cereb Cortex **14**(11), 1256 (2004)
34. C. Summerfield, T. Egner, M. Greene, E. Koechlin, J. Mangels, J. Hirsch, Predictive codes for forthcoming perception in the frontal cortex. Science **314**(5803), 1311 (2006)
35. S. Bentin, T. Allison, A. Puce, E. Perez, G. McCarthy, Electrophysiological studies of face perception in humans. J Cognitive Neurosci **8**(6), 551 (1996)
36. S. Marat, N. Guyader, D. Pellerin, *Recent advances in signal processing* (In-Tech, 2009), chap. Gaze prediction improvement by adding a face feature to a saliency model, pp. 195–210. 12
37. A. Milner, M. Goodale, *The visual brain in action.* Oxford psychology series (Oxford University Press, 2006)
38. P. Tseng, R. Carmi, I.G.M. Cameron, D. Munoz, L. Itti, Quantifying center bias of observers in free viewing of dynamic natural scenes. J Vision **9**(7), 1 (2009)
39. M. Dorr, T. Martinetz, K.R. Gegenfurtner, E. Barth, Variability of eye movements when viewing dynamic natural scenes. J Vision **10**(10), 1 (2010)
40. Q. Zhao, C. Koch, Learning a saliency map using fixated locations in natural scenes. J Vision **11**(3), 1 (2011)
41. B.W. Tatler, The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. J Vision **7**(14), 4.1 (2007)
42. L.W. Renninger, P. Verghese, J. Coughlan, Where to look next? eye movements reduce local uncertainty. J Vision **7**, 1 (2007)
43. T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look. in *Computer Vision, 2009 IEEE 12th International Conference on* (2009), pp. 2106 –2113
44. L. Zhang, M.H. Tong, T.K. Marks, H. Shan, G.W. Cottrell, Sun: a bayesian framework for saliency using natural statistics. J Vision **8**(7), 1 (2008)
45. R. Carmi, L. Itti, Visual causes versus correlates of attentional selection in dynamic scenes. Vision Res **46**(26), 4333 (2006)
46. R.J. Peters, A. Iyer, L. Itti, C. Koch, Components of bottom-up gaze allocation in natural images. Vision Res **45**, 2397 (2005)
47. A. Torralba, A. Oliva, M.S. Castelhano, J.M. Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev **113**(4), 766 (2006)
48. J.M. Wolfe, T.S. Horowitz, What attributes guide the deployment of visual attention and how do they do it? Nat Rev Neurosci **5**, 1 (2004)
49. O. Hershler, T. Golan, S. Bentin, S. Hochstein, The wide window of face detection. J Vision **10**(10), 21 (2010)