12

Gaze prediction improvement by adding a face feature to a saliency model

MARAT Sophie, GUYADER Nathalie and PELLERIN Denis
Gipsa-Lab
France

1. Introduction

To understand how human observers move their eyes when viewing videos and why some locations are more often gazed at than others is of great importance for much research. Many computational models compute a *saliency map* that attributes high values to the locations that predict where humans should direct their gaze when viewing the stimulus freely. To compute saliency maps, models are generally based on low-level properties of the visual stimulus. These bottom-up models, like the human visual system, break down the visual stimulus into elementary feature maps: a colour map, spatial frequency maps, orientation maps and a motion map (motion feature is often processed separately from other static features). Hence, the visual stimulus is split into a static saliency map, a dynamic saliency map (Marat et al., 2009), and some models add a colour saliency map (Itti et al., 1998; Le Meur et al. 2006). These maps are then added together to create the master saliency map of the visual stimuli. In the research cited, models only used low-level properties of the visual stimuli, although the role of high-level (e.g. semantic information) properties is undisputed. We know, for example, that live subjects and objects play a significant role in the visual perception of static stimulus.

In this research, we focus our interest on human faces. In a recent paper, Cerf, Harel, Einhäuser and Koch (Cerf et al., 2007) showed that in images containing frontal shots of people, faces were fixated on by subjects within the first saccades and this was true independently of the task subjects were asked to complete. For this reason, they integrated into their static saliency model a special "face-detection" channel (using the Viola and Jones algorithm (Viola & Jones, 2004)). They found that this additional channel improved the performance of their static saliency model.

In this research, we show that faces are very salient even in natural dynamic stimuli. We found that when asking subjects to look at videos faces were fixated on within the first saccades as was the case for static images. We also show that to add a face saliency map to a classical saliency model, using only static and dynamic features, improves eye movement predictions. In fact, even if faces are emphasized with a static saliency map, eyes or the mouth, and with a dynamic saliency map, moving faces, adding a face saliency map considerably improves model prediction. Moreover, model predictions are improved by computing a face saliency map inversely proportional to the number of faces in the scene.

We also proposed a fusion that takes into account the special feature of each saliency map: static, dynamic and face features.

Section 2 describes the eye movement experiment. The static and dynamic pathways are presented in section 3. Section 4 tests whether faces are salient in dynamic stimuli and section 5 deals with the choice of a face detector. Section 6 describes the face pathway, and finally, the fusion of the different saliency maps and the evaluation of the model are presented in section 7.

2. Eye movement experiment

Our purpose is to analyse whether faces influence human gaze and to understand how this influence occurs. The video database was built in order to obtain videos with various contents, with and without faces, with textured backgrounds, with moving and static objects, with a moving camera etc. We were only interested in the first eye movements of subjects when viewing videos. In fact, we know that after a certain time (quite short) it is much more difficult to predict eye movements without taking into account top-down processes. In order to remove top-down effects as much as possible, we did not use classical videos. Instead, we created small concatenated clips as was done in (Carmi & Itti, 2006). We put small parts of videos together with unrelated semantic contents. In this way, we minimized potential top-down confounds without sacrificing real world relevance.

2.1.1 Participants

Fifteen human observers (3 women and 12 men, aged from 23 to 40 years old) participated in the experiment. They had normal or corrected to normal vision and were not aware of the purpose of the experiment. They were asked to look at the videos freely.

2.1.2 Apparatus

Eye tracking was performed by an Eyelink II eye tracker (SR Research[1]). During the experiment, participants were sitting, with their chin supported, in front of a 21" colour monitor (75 Hz refresh rate) at a viewing distance of 57 cm (40°x 30° usable field of view). A 9-point calibration was carried out every five trials and a corrected-drift was done before each trial.

2.1.3 Stimuli

The stimuli were inspired by an experiment proposed in (Carmi & Itti, 2006). Fifty-three videos (25 frames per seconds, 720 x 576 pixels per frame) were selected from heterogeneous sources including movies, TV shows, TV news, animated movies, commercials, sport and music clips. The fifty-three videos were cut every 1-3 seconds (1.86 ± 0.61) into *305 clip-snippets*. The length of these clip-snippets was chosen randomly with the only constraint being to obtain snippets without any shot cut. These clip-snippets were strung together to make up *twenty clips* of 30 seconds (30.20 ± 0.81). Each clip contained at most one clip-snippet from each of the fifty-three continuous sources. The choice of the clip-snippets and their duration were random to prevent subjects from anticipating shot cuts. We used grey

[1] http://www.eyelinkinfo.com/

level stimuli (14155 frames) without audio signal because the model did not consider colour and audio information. Stimuli were seen in random order.

2.1.4 Human eye position density maps

The eye tracker records eye positions at 500 Hz. We recorded twenty eye positions (10 positions for each eye) per frame and per subject. The median of these positions (X-axis median and Y-axis median) was taken for each frame and for each subject. Then, for each frame, we had fifteen positions (one per subject). Because the final aim was to compare these positions to a saliency map, a two-dimensional Gaussian was added to each position. The standard deviation at mid-height of the Gaussian was equal to 0.5° of visual angle, which is close to the size of the maximum resolution of the fovea. Therefore, for each frame k, we got a human eye position density map $M_h(x,y,k)$.

2.1.5 Metric used for model evaluation

We used the Normalized Scanpath Saliency (NSS) (Peters & Itti, 2008). This criterion was especially designed to compare eye fixations and the salient locations emphasized by a model saliency map. We computed the *NSS* metric as follows (1):

$$NSS(k) = \frac{\overline{M_h(x,y,k) \times M_m(x,y,k)} - \overline{M_m(x,y,k)}}{\sigma_{M_m(x,y,k)}} \qquad (1)$$

where $M_h(x,y,k)$ is the human eye position density map normalized to unit mean and $M_m(x,y,k)$ a model saliency map for a frame k. The *NSS* is null if there is no link between eye position and salient regions. The *NSS* is negative if eye position tends to be in non-salient regions. The *NSS* is positive if eye position tends to be in salient regions. To summarize, a saliency map is a good predictor of human eye fixations if the corresponding *NSS* value is positive and high. In the next sections, we computed the *NSS* average over several frames.

3. The static and the dynamic pathways of the saliency model

We based ourselves on the biology of the human visual system to propose a saliency model that decomposes the visual signal into a static and a dynamic saliency maps. The static and the dynamic pathways, described in detail in (Marat et al., 2008; Marat et al., 2009), were built in two common stages: a retina-like filter and a cortical-like bank of filters.

3.1 The retina and the visual cortex models

The retina model proposed split visual stimuli into different frequency bands: the high spatial frequencies simulate a "Parvocellular-like" output and the low spatial frequencies simulate a "Magnocellular-like" output. These outputs correspond to the two main outputs of the retina with a parvocellular output that conveys detailed information and a magnocellular output that responds rapidly and conveys global information about the visual scene.

V1 cortical complex cells are modelled using a bank of Gabor filters, into six different orientations and four frequency bands in the Fourier domain. The energy output of each filter corresponds to an intermediate map, m_{ij}, which is the equivalent of an elementary feature of Treisman's Theory (Treisman & Gelade, 1980).

3.2 The static pathway

The static pathway is dedicated to the extraction of the static features of the visual stimulus. This pathway corresponds to the ventral pathway of the human visual system and processes detailed visual information. It starts with the parvocellular output of the retina and is then, processed by the bank of Gabor filters. Two types of interactions between filter outputs were implemented: short interactions reinforce objects belonging to a specific orientation and long interactions allow contour facilitation.

After the interactions and after being normalized between [0,1], each map m_{ij} was multiplied by $(\max(m_{ij}) - \overline{m_{ij}})^2$ where max(m_{ij}) is the maximum value and $\overline{m_{ij}}$ is the average of the elementary feature map m_{ij} (Itti et al., 1998). Then, for each map, values smaller than 20% of the maximum value $max(m_{ij})$ were set to 0. Finally, the intermediate maps were added together to obtain a static saliency map $M_s(x,y,k)$ for each frame *k* (Fig. 1).

3.3 The dynamic pathway

The dynamic pathway, which is equivalent to the dorsal pathway of the human visual system, is fast and carries global information. Because we assumed that human gaze is attracted by motion contrast (the motion of a region against the background), we applied a background motion compensation (2D motion estimation, Odobez & Bouthemy, 1995) before the retina process. This allowed us to estimate the relative motion of regions against the background. The compensated frames were filtered by the retina model described above to form the “Magnocellular-like” output. Because this output only contains low spatial frequencies, its information would be processed by the Gabor filters with the three lowest frequency bands. For each frame, the classical optical flow constraint was applied to the Gabor filter outputs in the same frequency band. The solution of this flow constraint defined a motion vector per pixel of a frame. Then we computed for each pixel the motion vector module, corresponding to the speed, and its angle, corresponding to the motion direction. Hence, the motion saliency of a region is proportional to its speed against the background. Then, a temporal median filter was applied to remove possible noise (if a pixel had a motion in one frame but not in the previous ones). The filter was applied to five successive frames (the current frame and the four previous ones) and it was reinitialised after each shot cut. A dynamic saliency map $M_d(x,y,k)$ was obtained for each frame *k* (Fig. 1).

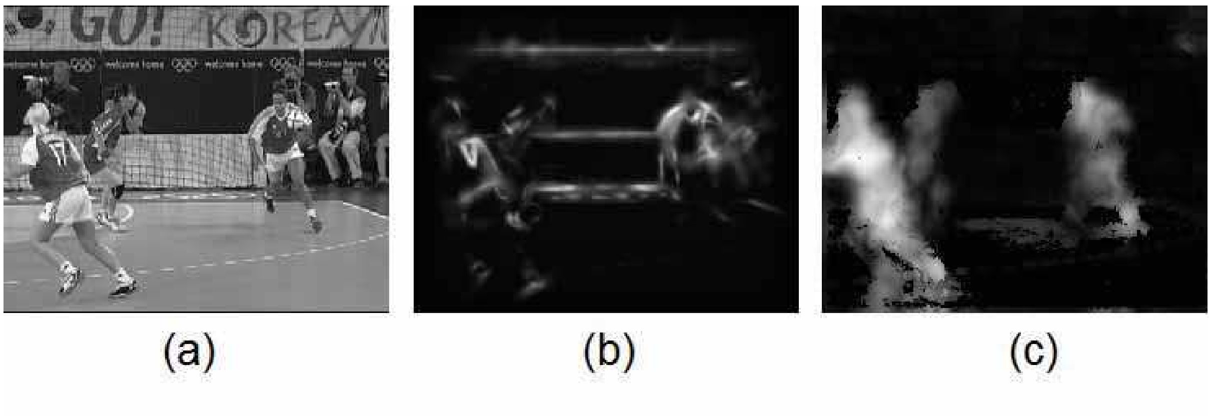


Fig. 1. Static and dynamic saliency maps: (a) Input video frame, (b) Static saliency map M_s and (c) Dynamic saliency map $M_{d.}$.

4. Face an important feature

Faces are one of the most important visual cues for communication. A lot of research has examined the complex issue of face perception (Kanwisher & Yovel, 2006; Thorpe, 2002; Palermo & Rhodes, 2007; Tsao & Livingstone, 2008; Goto & Tobimatsu, 2005), for a complete review see (Dekowska et al., 2008). In this research, we just wanted to test whether faces were gazed at during free viewing of dynamic scenes. Hence, to test if a face is an important feature in the prediction of human eye movements, we hand-labelled the frames of the videos used in the experiment described in section 2 with the position and the size of faces. We manually created a face saliency map by adding a two dimensional Gaussian to the top of each marked face: we called this saliency map the "true" face saliency map (Fig. 3). We call "face" any kind of face (frontal or profile) as long as the face is big enough for the eyes (at least one) and the mouth to be distinguished. Because it takes times to hand label all the frames and because we wanted to test the influence of faces we only used a small part of the whole database and we chose frames with at least one face (472 frames). Then, we computed the mean NSS over these 472 frames between the human eye position density maps and the different saliency model: the static saliency map, the dynamic saliency map and the "true" face saliency map (Fig. 2). As noted above a saliency map is a good predictor of human eye fixations if the corresponding *NSS* value is positive and high.

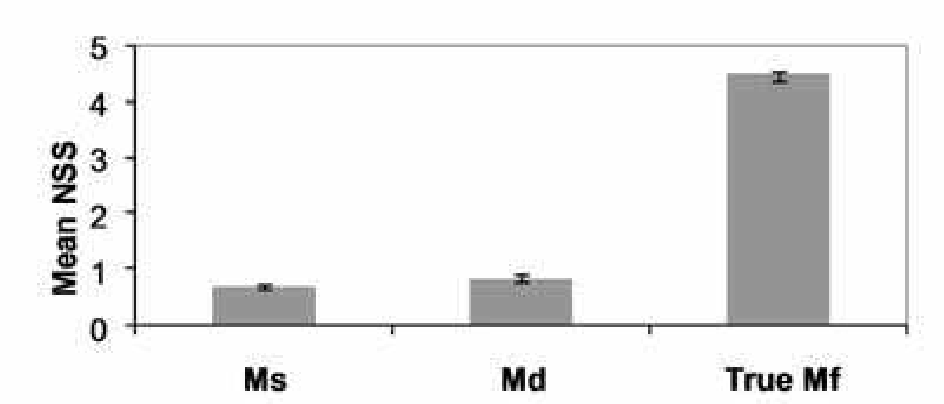


Fig. 2. Mean *NSS* values for the different saliency map: the static M_s, the dynamic M_d and the "true" face saliency map M_f.

As we can see on figure 2 the mean NSS value for the true face saliency map is higher than for the mean NSS for the static and the dynamic saliency maps (F(2,1413)=1009.81; p#0). The large difference is due to the fact that we only study frames with at least one face.

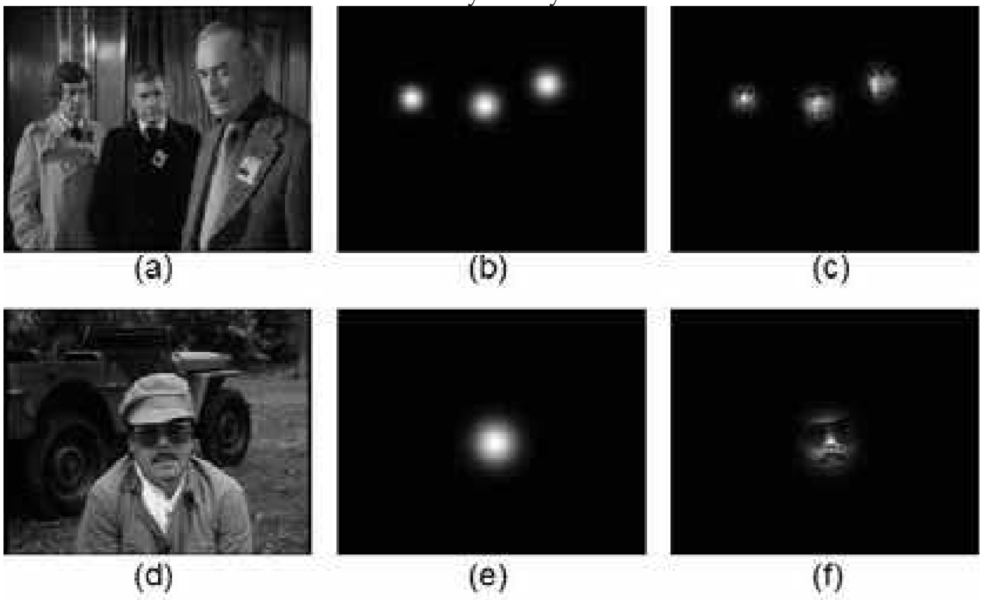


Fig. 3. Examples of the "true" face saliency maps obtained with the hand-labelled faces: (a) and (d) Input video frames, (b) and (e) Corresponding "true" face saliency maps M_f, (c) and (f) Superposition of the input frame and the "true" face saliency map.

We experimentally found that faces attract human gazes and hence computing saliency models that highlight faces improves the predictions of a more traditional saliency model considerably. We still want to answer different questions. Is a face on its own inside a scene more or less salient than a face with other faces? Is a large face more salient than a small one? To answer these questions we chose some clips according to the number of faces and according to the size of faces.

4.1 Impact of the number of faces

To see the influence of the number of faces, we split the database according to the number of faces inside the frames: three clip-snippets (121 frames) with only one face and three others (134 frames) with more than one face. We computed the *NSS* value for each frame using the "true" face saliency map and the subject's eye position density maps. Figure 4 presents the mean *NSS* value for the frames with only one face and for the frames with more than one face. A high NSS value means a good correspondence between human eye position density maps and "true" face saliency maps.

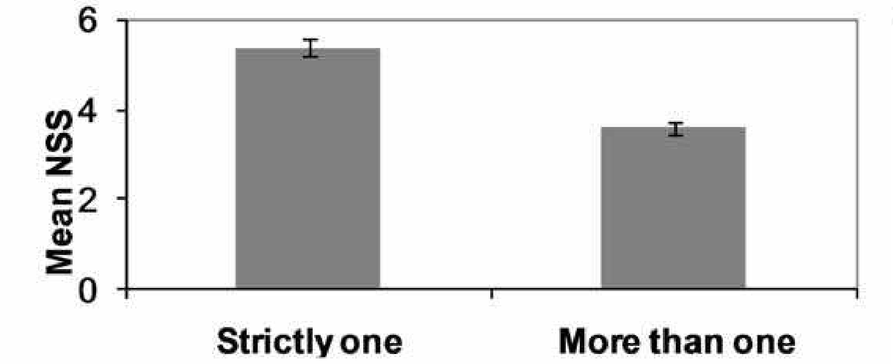


Fig. 4. Mean NSS values for the "true" face saliency maps compared with human eye positions as a function of the number of faces in frames: for frames with strictly one face (121) and for frames with more than one faces (134).

The *NSS* value is higher when there is only one face than when there are more than one face (F(1,253) =52.25; p#0). There is a better correspondence between the saliency map and eye positions. This could be predicted by the fact that if there is only one face, all the subjects would gaze at this single face whereas if there are several faces on the same frame some subjects would gaze at a particular face and other subjects would gaze at another face. Hence, a frame with only one face is more salient than a frame with more than one face, in the sense that it is easier to predict subjects' eye positions. To take this result into account, we chose to compute the face saliency map using an inversely proportional coefficient to the number of faces. That means that if there is only one face on a frame the corresponding saliency map would have higher values than the saliency map of a frame with more than one face.

An example of the eye position on a frame with three faces is presented in figure 5. Subjects' gazes are more spread out over the frame with three faces than over the frames with only one face.

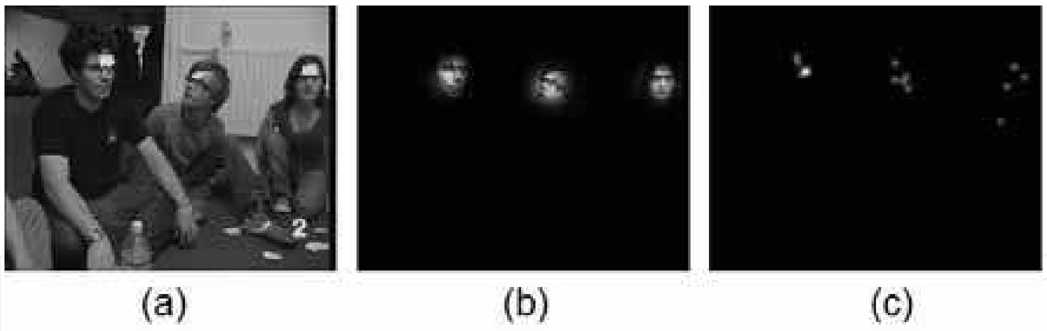


Fig. 5. Examples of eye positions on a frame with three faces: (a) Input video frame, (b) Superimposition of the input frame and the "true" face saliency map and (c) Eye positions of the fifteen subjects.

As we can see in figure 5 (c) subjects gazed at the different faces. To test how much subjects gazed at different positions in a frames we computed a criterion to measure the dispersion of eye positions between subjects using the equation (2):

$$D = \frac{1}{N^2} \sum_{i,j<i} d_{i,j}^2 \qquad (2)$$

where N is the number of subjects and $d_{i,j}$ is the distance between the eye positions of subjects i and j. Table 1 presents the mean dispersion value for frames with strictly one face and for frames with more than one face.

Number of faces	**Strictly one**	**More than one**
Mean dispersion	1 252.3	7 279.9

Table 1. Mean dispersion values of eye positions between subjects on frames as a function of the number of faces: strictly one and more than one.

As expected, the dispersion is significantly higher for frames with more than one face, than for frames with only one face (F(1,253)=269.7; p#0). This is consistent with a higher NSS for frames with only one face than more than one.

4.2 Impact of face size

The previous observations are made for faces with almost the same size (See Fig. 5). But what happen if there is one big face and two small ones? It is difficult to understand exactly how size influences eye movements as many configurations can occur: for example, if there are two faces, one may be large and the other may be small, or the two faces may be large or small, one may be in the foreground etc. Hence it is difficult to understand exactly what happens for eye movements. Let us consider clips with only one face. These clips are then split according to the size of the face: three clip snippets with only one small face (141 frames), three with a medium face (107 frames) and three with a large face (90 frames). The diameter of the small face is around 30 pixels, the diameter of the medium face is around 50 pixels and the diameter of the large face is around 80 pixels. The mean *NSS* value was computed for the frames with a small, a medium and a large face (Fig. 6).

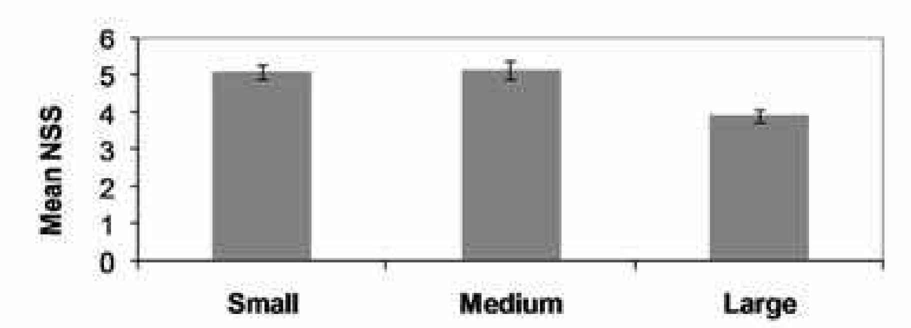


Fig. 6. Mean *NSS* value for "true" face saliency maps compared with human eye positions for frames of nine clip snippets as a function of face size.

Large faces give significantly lower results than small or medium faces (F(1,336)=18.25; p=0.00002). The difference between small and medium faces is not significant (F(1,246)=0.04; p=0.84). This could be expected in fact: when a face is small, all subjects will gaze at the same position, that is, the small face, and if the face is large, then some subjects will gaze at the eyes, other will gaze at the mouth etc. To verify this, we computed the mean dispersion of subject eye positions for the frames with small, medium or large faces in Table 2.

Face size	**Small**	**Medium**	**Large**
Mean dispersion	2 927.6	1 418.4	904.24

Table 2. Mean dispersion values of eye positions between subjects on frames as a function of face size.

The dispersion of eye positions is significantly higher for small faces (F(2,335)=28.44; p#0). The dispersion of eye positions for frames with medium faces is not significantly different from the frames with large faces (F(1,195)=2.89; p=0.09). These results are apparently in contradiction with the mean NSS values found. Hence, two main questions arise: (1) why do frames with one small face lead to a higher dispersion than frames with a larger face? And (2) why do frames that lead to more spread out eye positions give a higher NSS?

Most of the time, when a small face is on a frame it is because the character is filmed in a wide view; the frame shows the whole character and the scene behind him which may be complex. If the character moves his hand, or if there is something interesting in the foreground, some subjects will tend to gaze at the moving or the interesting thing after viewing the face of the character. On the other hand, if a large face is on a frame, this corresponds to a close-up view of the character being filmed. Hence, there is little information outside the character 's face and hence, subjects will tend to keep their focus on the only interesting area: the face, and access in more detail the different parts of the face.

A small face could lead to a high dispersion value if some subjects gaze at other areas after having gazed at the face, and a large face could lead to a low dispersion value as subject gazes tend to be spread over the face area. This is illustrated in figure 7, where eye positions were shown for a large face and for a small one. In this example a subject gazed at the device at the bottom of the frame, increasing the dispersion of eye positions. This is why we observed a high dispersion value of eye positions even for frames with a high NSS value (example of frames with a small face). A small face with few eye positions outside of the face, will lead to a high dispersion, but can thus have a higher NSS than a large face with more eye positions on the face, so lower dispersion. Hence, the NSS tends to reward fixations that are less due to chance more strongly: as the salient region for a small face is small, the eye positions that are in this region will be more strongly rewarded than the ones on a larger face.

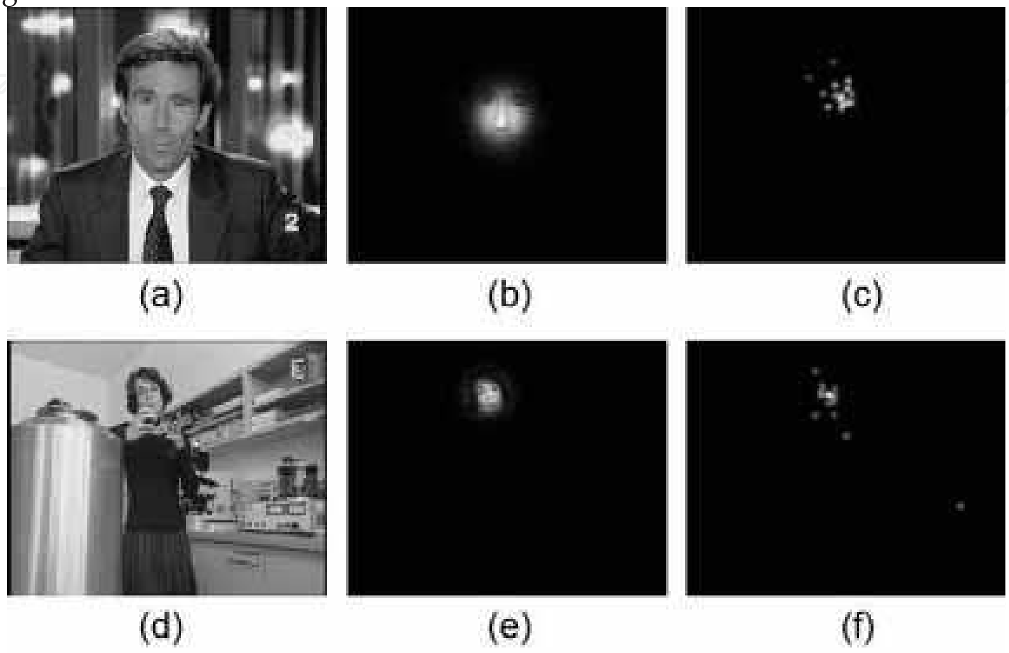


Fig. 7. Examples of eye positions on frames with a face of different sizes: (a) and (d) Input video frames, (b) and (e) Superimposition of the input frame and the face saliency map and (c) and (f) Eye positions of the fifteen subjects corresponding to the input frame.

Considering the case of only one face, face size influences eye positions. If more than one face is present, too many configurations can occur, and so, it is much more difficult to generalize the size effect. That is why for this study, the size information was not integrated to build the face saliency map from the face detector output.

5. Face detection algorithms

Various methods have been proposed to detect faces in images (Yang et al., 2002). We tested three algorithms available on the web: the one proposed by Viola[2] and Jones (Viola & Jones, 2004), the one proposed by Rowley[3] (Rowley et al., 1998) and the one proposed by Nilsson[4] (Nilsson et al., 2007) which is called the Split-up SNoW face detector. In our study, the stimuli are different from classical databases used to evaluate algorithm performance for face detection. We chose stimuli which were very different from one to another, and most faces are presented with various and textured backgrounds. The different algorithms were

[2] Viola & Jones - http://sourceforge.net/projects/openlibrary/

[3] Rowley - http://vasc.ri.cmu.edu/NNFaceDetector/

[4] Nilsson - http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=13701&objectType=FILE

compared on one of the twenty clips presented to subjects (Table 3). This clip was hand-labelled: 429 faces were marked.

Algorithms	Number of correct detections	Number of false positives
Viola & Jones, 2004	146 (34%)	77
Rowley et al., 1998	87 (20.3%)	25
Nilsson et al., 2007 Split-up SNoW[5]	97 (22.6%)	6

Table 3. Three face detection algorithms: number of correct detections (also called true positives) and false positives for one clip (745 frames with 429 faces present).

Because the videos chosen are different from traditional stimuli used to evaluate face detection algorithm, the three algorithms detected less than half the faces. During the snippets, characters are moving, can turn to profile view, can sometimes be occluded or can have tilted faces. Faces can also be blurred as the characters move fast. All these cases complicate the task of the face detection algorithms. The Viola and Jones algorithm has the highest correct detection rate but also the highest false positive rate. Most of the time, false positives are on textured regions. Because we wanted to create a face saliency map that emphasizes only areas with a face, and we wanted to prevent the highlighting of false positives, we chose to use the split-up SNoW face detector which has the lowest false positive rate.

5.1 The split-up SNoW face detector

SNoW (Sparse Network of Winnows) is a learning architecture framework designed to learn a large number of features. It can be used for a more general purpose as a multi-class classifier. SNoW has been used successfully in several applications in the natural language and visual processing domains.

If a face is detected, the algorithm returns the position and the size of a squared bounding box containing the face detected. The algorithm detects faces with frontal views, even partially occluded faces (i.e. faces with glasses) and slightly tilted faces, but it cannot retrieve faces which are too occluded or profile views. We tested the efficiency of the SNoW face detector algorithm on the whole database (14155 frames). As it takes time and it is fastidious to hand-label all the faces for all the frames, we counted the number of frames that contained at least one face and we found 6623 frames. The split-up SNoW face detector gave 1566 frames with at least a correct detection and only 147 false positives. As already said, the number of correct detections is quite low but, what is more important for our purpose is that the number of false positive is very low. Hence, using this face detection algorithm ensures that we will only emphasize areas with a very high probability of containing a face. Examples of results for the split-up SNoW face detector are given in figure 8.

[5] Results are given setting the parameter *sens* to 9 in the Matlab program.

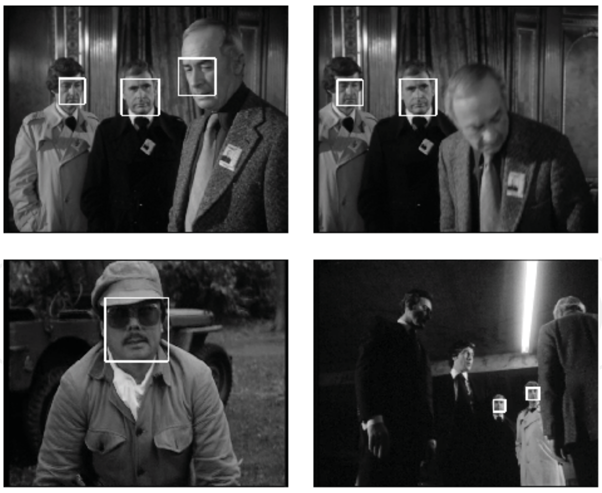


Fig. 8. Examples of correct detections (true positives) (marked with a white box) and missed detections (false negatives) for the split-up SNoW face detector.

6. Saliency model: The face pathway

The face detection algorithm output needs to be converted into a saliency map. The algorithm returns the position and the size of a squared bounding box containing the face detected. How can this information be translated into a face saliency map? The face detector gives a binary result: A pixel is equal to 1 if it is part of a face (the corresponding bounding box) and 0 otherwise. In the few papers that dealt with face saliency maps, the bounding boxes used to mark the face detected are replaced by a two-dimensional Gaussian. This induced the centre of a face to be more salient than its border. For example, in (Cerf et al., 2007) the "face conspicuity map" is normalized to a fixed range, in (Ma et al., 2005) the face saliency map values are weighted by the position of the face, enhancing faces in the centre of the frame.

As the final aim of our model is to provide a master saliency map by computing the fusion of the three saliency maps, face M_f, static M_s and dynamic M_d, the face saliency map was normalized to give values in the same range as static and dynamic saliency map values. As stated above, the face saliency map is intrinsically different from the static and the dynamic saliency maps. On one hand, the face detection algorithm returns binary information: presence or absence of face. On the other hand, static or dynamic saliency maps are weighted "by nature": more or less textured for the static saliency map and more or less rapid for moving areas of the dynamic saliency map. The face saliency map was built by replacing the bounding box of the algorithm output by a two-dimensional Gaussian. To be in the same range as the static and the dynamic saliency maps, the maximum value of the two-dimensional Gaussian was set to 5. Moreover, as stated above, a frame with only one face is more salient than a frame with more than one face. To lessen the face saliency map when more than one face is detected, the maximum of the Gaussian (after been multiplied by five) was divided by $N^{1/3}$ where N is the number of faces detected on the frame. To sum up, the Gaussian that replaced the bounding box that marked a detected face was set to $\frac{5}{N^{1/3}}$. We used the cube root of N to attenuate the effect of a high N value.

7. Evaluation

7.1. Fusions

Static, dynamic and face saliency maps do not have the same appearance. On one hand, the static saliency map exhibits a large number of salient areas, corresponding to textured areas that are spread over the whole image. On the other hand, the dynamic saliency map can exhibit only small and compact areas corresponding to moving objects. Finally, the face saliency map can be null when no face is detected.

A previous study detailed the analysis of the static and the dynamic pathways (Marat et al., 2009). This study showed that a frame with a high maximum static saliency map value is more salient than a frame with a lower maximum static saliency map value. Moreover, a frame with high skewness of the dynamic saliency map is more salient than a frame with a lower skewness value of the dynamic saliency map. A high skewness value corresponds to a frame with only one compact moving area. To add the static saliency map multiplied by its maximum to the dynamic saliency map multiplied by its skewness creating the master saliency map provides better eye movement prediction than a simple sum. The face saliency map was designed to reduce the maximum saliency value with the number of faces detected. Hence, this maximum is characteristic for the face pathway. The fusion proposed considers the particular features of each saliency map by weighting the raw saliency maps by their relevant parameters (maximum or skewness) and provides better results. The weighted saliency maps are defined as:

$$M_{s'} = \max(M_s) \times M_s \tag{3}$$

$$M_{d'} = skewness(M_d) \times M_d \tag{4}$$

$$M_{f'} = \max(M_f) \times M_f \tag{5}$$

To study the importance of the face pathway, we computed two different master saliency maps: one using only the static and the dynamic maps (6) and another using the three maps (7).

$$M_{sd} = M_{s'} + M_{d'} \tag{6}$$

$$M_{sdf} = M_{s'} + M_{d'} + M_{f'} \tag{7}$$

Note that if the face saliency map is null for a frame the master saliency map would depend only on the static and the dynamic saliency maps. Moreover, to strengthen regions that are salient in two different maps (static and dynamic, static and face or dynamic and face), a more elaborate fusion, called “reinforced” fusion (M_{Rsdf}), was proposed (8):

$$M_{Rsdf} = M_{s'} + M_{d'} + M_{f'} + M_{s'} \times M_{d'} + M_{s'} \times M_{f'} + M_{d'} \times M_{f'} \tag{8}$$

This fusion reinforces the weighted fusion M_{sdf} by adding multiplicative terms. We chose multiplicative terms with only two maps because if we chose a multiplicative term with the three maps when the face saliency map is null the multiplicative term would be null. If the face saliency map is null the “reinforced” fusion takes advantage of the static and the dynamic maps. In that case, the face saliency map does not improve the result but it does not penalize the result either. Examples of these fusions integrating the face pathway are proposed in figure 9. In figure 9 (a), the face on the right of the frame is moving, whereas the two faces on the left are not moving. In figure 9 (b) the three faces are almost equally salient, but in figure 9 (c) the multiplicative reinforcement terms increase the saliency of the moving face on the right of the frame.

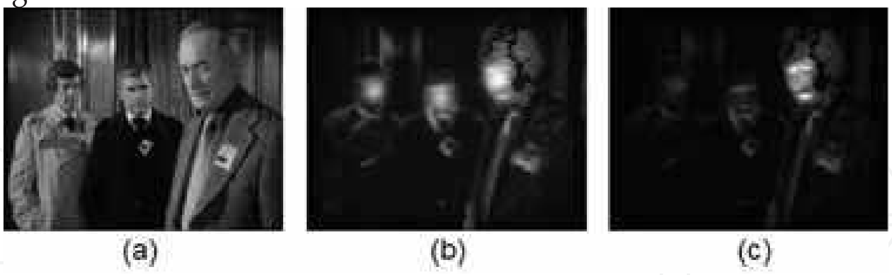


Fig. 9. Example of master saliency maps: (a) Input video frame, (b) Corresponding master saliency map computed using a weighted fusion of the three pathways M_{sdf}, (c) Corresponding master saliency map using the "reinforced" fusion of the three pathways $M_{Rsdf.}$

7.2. Evaluation of different saliency maps

The first evaluation was done on the database of "true" face saliency maps which were hand-labelled. Each saliency map was weighted as explained in section 6.1. The results are presented in Table 4.

Saliency maps	$\mathbf{M_s}$	$\mathbf{M_d}$	$\mathbf{M_f}$	$\mathbf{M_{sd}}$	$\mathbf{M_{sdf}}$	$\mathbf{M_{Rsdf}}$
Mean *NSS*	0.68	0.84	4.46	1.00	3.38	3.99
Standard deviation	0.72	1.03	2.19	0.80	1.63	2.05

Table 4. Evaluation of the different saliency map and the fusion, on the database where a "true" face saliency map was hand-labelled.

As stated above, the face saliency map gives better results than the static or the dynamic ones (F(2,1413)=1009.81; p#0). The fusion which did not take face saliency maps into account gives a lower result than the fusions with face saliency maps (F(2,1413)=472.33; p#0), and the reinforced fusion is even better than a more classical fusion (F(1,942)=25.63; $p=4.98x10^{-7}$). Subsequently, the *NSS* was computed for each frame of the whole database (14155 frames) using the different model saliency maps and the eye movement data. The face saliency map is obtained using the split-up SNoW face detector and the weighting and fusion previously explained. In order to test the contribution of face pathway, the mean *NSS* value was calculated using the saliency map given by each pathway independently and the different possible fusions. The mean NSS value is plotted for six models of saliency maps (M_s, M_d, M_f, M_{sd}, M_{sdf}, M_{Rsdf}) in comparison with human data in figure 10. The *NSS* values are given for the saliency maps (M_s, M_d and M_f) but note that the *NSS* results would be the same for the weighted saliency maps ($M_{s'}$, $M_{d'}$ and $M_{f'}$), as multiplying the saliency map by a constant did not change the NSS value.

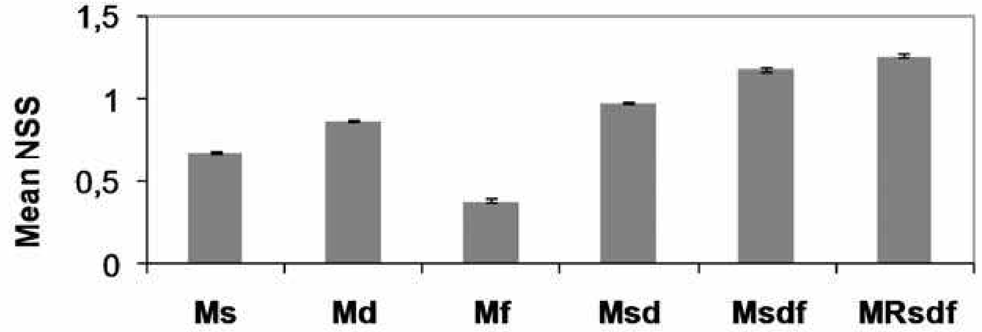


Fig. 10. Mean NSS values on the whole database (14155 frames) for six models of saliency maps (static, dynamic, face, weighted fusion of the static and dynamic pathways M_{sd}, weighted fusion of the static, the dynamic and the face pathway M_{sdf} and a "reinforced" weighted fusion M_{Rsdf}).

As presented in (Marat et al., 2009), the dynamic saliency maps are more predictive than the static ones. The fusion of the static and the dynamic saliency maps improves the prediction of the model: the static and the dynamic information needs to be considered to improve the model prediction. The results of the face pathway should not be considered; in fact, it gives the lowest results but only because a small number of frames contain at least one face detected compared to the total number of frames (12% of the whole database).The weighted fusion integrating the face pathway (M_{sdf}) is significantly better than the weighted fusion of the static and the saliency maps (M_{sd}), (F(1,28308)=255.39; p#0). Integrating the face pathway increases the model prediction; hence, as already observed, faces are crucial information to predict eye positions. The "reinforced" fusion integrating multiplicative terms (M_{Rsdf}), increasing saliency in regions that are salient in two maps, gives the best results, outperforming the previous fusion (M_{sdf}), (F(1,28308)=25.91; p=3.6x10^{-9}). The contribution of the face pathway in attracting our gaze is undeniable. The face pathway improves the results greatly, faces have to be integrated into a saliency model to make the results of the model match the experimental results more closely.

8. Conclusion

When viewing scenes, faces are almost immediately gazed on. This was shown in static images (Cerf et al., 2007). We report in this research the same phenomenon using dynamic stimuli. This means that even if there are moving objects, faces rapidly attracted gazes. To study the influence of faces on gaze, we ran an experiment to record the eye movements of subjects when looking freely at videos. We used videos with various contents, with or without faces with textured backgrounds and with or without moving objects. This experiment enabled us to check that faces are fixated on within the first milliseconds and independently of the scenes (presence or not of moving objects etc.). Moreover, we showed that a face is more salient if it is the only face on the frame. In order to take this into account, we added a "face pathway" to a bottom-up saliency model inspired by the biology. The "face pathway" uses the Split-up Snow face detector algorithm. Hence, the model splits the visual signal into static, dynamic, and face saliency maps. The static saliency map emphasizes orientation and spatial frequency contrasts. The dynamic saliency map emphasizes motion contrasts and the face saliency map emphasizes faces proportionally to the number of faces. Then, these three maps are originally fuzzed by taking into account the specificity of each saliency map. The fusion showed that the "face pathway" significantly increases the predictions of the model.

9. References

Carmi R. & Itti L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research,* Vol. 46, No. 26, pp. 4333-4345

Cerf M.; Harel J.; Einhäuser W. & Koch C. (2007). Predicting gaze using low-level saliency combined with face detection, *in Proceedings of Neural Information System NIPS 2007*

Dekowska M.; Kuniecki M. & Jaskowski P. (2008). Facing facts: neuronal mechanisms of face perception. *Acta Neurobiologiae Experimentalis*, Vol. 68, No. 2, pp. 229-252

Goto Y. & Tobimatsu S. (2005). An electrophysiological study of the initial step of face perception. *International Congress Series*, Vol. 1278, pp. 45-48

Itti L.; Koch C. & Niebur E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on PAMI*, Vol. 20, No. 11, pp. 1254-1259

Kanwisher N. & Yovel G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical transactions of the royal society Biological sciences*, Vol. 361, No. 1476, pp. 2109-2128

Le Meur O.; Le Callet P. & Barba D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Trans. on PAMI,* Vol. 28, No. 5, pp. 802-817

Marat S.; Ho Phuoc T.; Granjon L.; Guyader N.; Pellerin D. & Dugué-Guérin A. (2009). Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, Vol. 82, No. 3, pp. 231-243

Marat S.; Ho Phuoc T.; Granjon L.; Guyader N.; Pellerin D. & Dugué-Guérin A. (2008). Spatio-temporal saliency model to predict eye movements in video free viewing, *in Proceedings of Eusipco 2008,* Lausanne, Switzerland

Odobez J.-M. & Bouthemy P. (1995). Robust multiresolution estimation of parametric motion models. *Journal of visual communication and image representation*, Vol. 6, pp. 348-365

Palermo R. & Rhodes G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia*, Vol. 45, No. 1, pp. 75-92

Peters R. J. & Itti L. (2008). Applying computational tools to predict gaze direction in interactive visual environments. *ACM Trans. On Applied Perception*, Vol. 5, No. 2

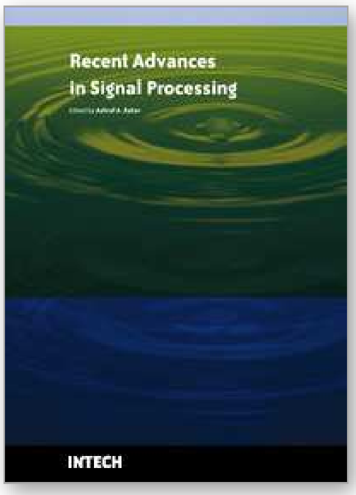
Thorpe S. J. (2002). Ultra-rapid scene categorization with a wave of spikes, *in Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, Vol. 2525, pp. 1-15

Treisman A. M. & Gelade G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, Vol. 12, No. 1, pp. 97-136

Tsao D. Y. & Livingstone M. S. (2008). Mechanisms of face perception. *Annu. Rev. Neuroscci.*, Vol. 31, pp. 411-437

Viola P. & Jones M. J. (2004). Robust real time face detection. *International Journal of Computer Vision,* Vol. 57, No. 2, pp. 137-154

Yang M.-H.; Kriegman D. J. & Ahuja N. (2002). Detecting faces in images: a survey. *IEEE Trans. on PAMI*, Vol. 24, No. 1, pp. 34-58



Recent Advances in Signal Processing

Edited by Ashraf A Zaher

ISBN 978-953-307-002-5
Hard cover, 544 pages
Publisher InTech
Published online 01, November, 2009
Published in print edition November, 2009

The signal processing task is a very critical issue in the majority of new technological inventions and challenges in a variety of applications in both science and engineering fields. Classical signal processing techniques have largely worked with mathematical models that are linear, local, stationary, and Gaussian. They have always favored closed-form tractability over real-world accuracy. These constraints were imposed by the lack of powerful computing tools. During the last few decades, signal processing theories, developments, and applications have matured rapidly and now include tools from many areas of mathematics, computer science, physics, and engineering. This book is targeted primarily toward both students and researchers who want to be exposed to a wide variety of signal processing techniques and algorithms. It includes 27 chapters that can be categorized into five different areas depending on the application at hand. These five categories are ordered to address image processing, speech processing, communication systems, time-series analysis, and educational packages respectively. The book has the advantage of providing a collection of applications that are completely independent and self-contained; thus, the interested reader can choose any chapter and skip to another without losing continuity.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821