Influence of number, location and size of faces on gaze in video

Anis Rahman, Denis Pellerin, Dominique Houzet

Journal of Eye Movement Research
7(2):5, 1–11

Influence of number, location and size of faces on gaze in video

Anis Rahman
National University of Sciences and Technology (NUST), Islamabad, Pakistan

Denis Pellerin
GIPSA-lab, UMR 5216, Grenoble, France

Dominique Houzet
GIPSA-lab, UMR 5216, Grenoble, France

Many studies have reported the preference for faces and influence of faces on gaze, most of them in static images and a few in videos. In this paper, we study the influence of faces in complex free-viewing videos, with respect to the effects of number, location and size of the faces. This knowledge could be used to enrich a face pathway in a visual saliency model. We used eye fixation data from an eye movement experiment, hand-labeled all the faces in the videos watched, and compared the labeled face regions against the eye fixations. We observed that fixations made are in proximity to, or inside the face regions. We found that 50% of the fixations landed directly on face regions that occupy less than 10% of the entire visual scene. Moreover, the fixation duration on videos with face is longer than without face, and longer than fixation duration on static images with faces. Finally, we analyzed the three influencing factors (Eccentricity, Area, Closeness) with linear regression models. For one face, the $E+A$ combined model is slightly better than the E model and better than the A model. For two faces, the three variables (E,A,C) are tightly coupled and the $E+A+C$ model had the highest score.

Keywords: faces, attention, eccentricity, eye movements, video

Introduction

Gaze is highly influenced by faces in visual scenes compared to other object stimuli. Over the years, a number of studies have been conducted regarding the influence of faces on gaze, mostly using static stimuli (Hasson, Levy, Behrmann, Hendler, & Malach, 2002; Rousselet, Macé, & Fabre-Thorpe, 2003; Jebara, Pins, Despretz, & Boucart, 2009; Jacques & Rossion, 2006; Bindemann, Scheepers, & Burton, 2009; Heisz & Shore, 2010) and some on videos (Rice, Moriuchi, Jones, & Klin, 2012; Riby & Hancock, 2009; Buchan, Paré, & Munhall, 2007; Võ, Smith, Mital, & Henderson, 2012). There is enough evidence that faces can be processed at the earliest after stimulus presentation (Ro, Russell, & Lavie, 2001; Vuilleumier, 2000), and they are preferentially processed by the visual system compared to other object categories (Rossion et al., 2000). The preference is thought to be influenced by several different factors like number of faces, eccentricity of face from fixation, face surface area, and closeness to other faces. However, these influences have rarely been reported for dynamic stimuli.

The number of faces limits the preference of faces, as they compete for limited attentional resources. Studies regarding event-related response to face stimulus (the N170), decrease considerably when more stimuli are presented in the visual field (Miller, Gochin, & Gross, 1993; Rolls & Tovee, 1995). This suppression of neural representation for stimuli is referred to as competition (Kastner & Ungerleider, 2001; Jacques & Rossion, 2004, 2006). A recent study (Jacques & Rossion, 2004) showed that the response to foveal faces is reduced when another face is presented parafoveally. The suppression remained even when a scrambled face was presented as competing stimuli. This suggests that there is certainly some sensory competition rather than simply an effect of reduced spatial attention.

In videos, object stimulus patterns degrade due to the loss of information as it moves away from the foveal region. This degradation of information seems likely to influence the preference for faces in videos (Sato, 1995). In still images, the influence of peripheral vision on face has been thoroughly

This research was supported by Rhone-Alpes region (France) under the CIBLE project No. 2136. Thanks to Silvain Gerber, Lionel Granjon, Sophie Marat and Nathalie Guyader for helpful suggestions and eye movement experiment.

studied, with controlled presentation of visual stimuli at predefined locations on rings of different eccentricities (Paras, Yamashita, Simas, & Webster, 2003; Reddy, Reddy, & Koch, 2006; Jebara et al., 2009; Hershler, Golan, Bentin, & Hochstein, 2010; Rigoulot et al., 2011). Most found a drop in performance of object stimuli; in the case of faces in periphery, a steep drop in face-selective responses was observed. The main question in this study is to evaluate how face eccentricity-dependent sensitivity loss which occurs when viewing dynamic stimuli, could be used in an improved saliency model.

Both these limiting factors (number of faces and eccentricity) can be alleviated by the size of faces, which causes stimulus magnification to maintain foveal performance of faces and to diminish the effects of competition. Evidence shows that the accuracy and quality of visual performance in the periphery is identical to that in foveal vision (Still, Thibos, & Bradley, 1989; Banks, Sekuler, & Anderson, 1991), but the drop is due to progressive undersampling of information presented away from fixation. Consequently, it limits the capacity of visual systems to extract information, which is important to attend to objects in a natural scene. This direct effect of eccentricity on stimuli can be compensated by using some linear eccentricity-dependent magnification, or size scaling (Virsu & Rovamo, 1979; Dow, Snyder, Vautin, & Bauer, 1981; Levi, Klein, & Aitsebaomo, 1985; Johnson & Gurnsey, 2010).

The purpose of the current research is to study the influence of faces on gaze during free-viewing of videos, and analyze the effects of number, location and size of faces. We know that faces attract gaze in videos. We put forward the hypothesis that preference for faces depends not only on eccentricity but also on their area and number. The study reported here evaluates different eye fixation attributes: distance, proportion and duration. We also analyze the different influencing factors—number, location and size of faces—with statistical models, and test the hypothesis by analyzing different combinations of these influencing factors using a comparison criterion. The results obtained from this work could support the improvment of a separate face pathway to a visual saliency model—to more accurately predict eye movements.

Eye movement experiment

We used the eye position data from a previous experiment described in (Marat et al., 2009). The experiment aimed to record eye movements of participants when looking freely at videos with various contents. We used this data to understand the features that best explain eye movements and fixated locations. Here, we recall some of the main aspects of this experiment.

- **Stimuli:** Fifty-three videos (25 fps, 720×576 pixels per frame) were selected from different video sources, for example: indoor scenes, outdoor, scenes of day and night (Figure 1). The videos are converted to grayscale before presenting them to the participants.

Figure 1. : Some sample frames from different video source.

The videos were cut into 305 *clip snippets* each of 1-3s. This was done in order to obtain snippets with minimum change in plane. The aim here is to minimize potential top-down influence on eye movement. Finally, these *clip snippets* were strung together to obtain 20 *clips* of 30s. The duration of the *clip* was random, to eliminate any anticipation of transition made by the participants during viewing.

- **Participants:** Fifteen young adults (3 women and 12 men, aged 23-40 years) participated in the experiment. All participants had normal, or corrected to normal vision. Each participant, sitting with his/her head stabilized on a chin rest, in front of a monitor at 57cm viewing distance ($40^\circ \times 30^\circ$ field of view), was instructed to look at the videos without any task.
- **Apparatus:** An eye tracker (SR Research EyeLink II) was used to record eye movements. It is composed of three miniature cameras mounted on a helmet, two in front of each eye to provide binocular tracking, with the third on a head-band for head tracking. The recordings from the first two cameras, when compensated for head movements, give the gaze direction of the participant.

Method

In this study, we test a video database comprising faces to analyze their interest during free-viewing. We also evaluate the influence of different factors on the

interest of faces, such as number of faces, face eccentricity, face surface area, and closeness between two faces. In this section, we first present the data used for the evaluation that includes the hand-labeled faces of the entire video dataset, and the eye fixations recorded during the eye movement experiment. We use at most five fixations overtime that roughly equal 250ms after the scene onset. Second, we define these influencing factors. Third, we detail several evaluation metrics to analyze the influencing factors. Last, we summarize the methods used for statistical analysis of the data.

Database

The video database comprises a variety of face content, such as scenes with cases of one or more faces at different locations. Moreover, the faces are of different sizes. We labeled 14,600 frontal and upright faces in total for the entire video database (14,155 frames), to create a face ground-truth for this study. We also labeled turned faces when the facial features such as eyes and mouth regions were distinguishable. Moreover, background faces with blurred features were ignored in favor of foreground faces.

During the experiment, the eye tracker recorded participants' eye movements at 500 Hz—20 recordings for two eyes per frame and per participant. The recordings are then used to calculate corresponding fixations and saccades. In this study, we used these eye fixations to study different factors influencing the interest of faces in a dynamic scene. The distribution of eye fixations is shown in Figure 2.

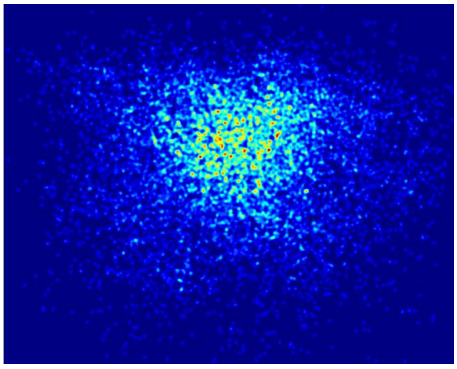


Figure 2. : Representation of the positions of all fixations for all participants in a scene. These surface maps were created by placing a Gaussian with a diameter of 2° of visual angle (equivalent to the fovea) centered at each fixation point, adding the Gaussians, and normalizing the height of the resulting sums.

In total, 23,797 fixations were recorded for 15 participants, with 11,155 fixations on scenes with at least one face. The number of fixations for one, two or more faces in a scene are summarized in Table 1.

Table 1
: Total number of face frames and fixations for one, two, or more faces in a scene. In the study, we consider frames, with corresponding fixations, for scenes with one or two faces.

	One face	Two faces	More than two faces
No. of sample frames	3,335	2,317	1,151
Total Fixations	5,425	3,937	1,793

Influencing factors

The study was designed to provide an insight into the extent to which different factors of faces affect their perceived interest. We annotate each frame in a scene with the number of faces present, and each face in the frame with its eccentricity and surface area. We also computed closeness of a face to another face in the frame. All these mentioned factors are measured as follows:

Number: is a simple count of faces present. It determines the complexity of the scene. For clarity, we only consider cases of frames with one face and two faces.

Eccentricity: is the relative distance from a participant's fixation on screen to the edge of nearest face ellipse in degrees. In Figure 3, $(d - r(\alpha))$ is eccentricity E of face ellipse with origin (O_x, O_y) from fixation position (C_x, C_y).

Area: is the two-dimensional surface of face ellipse in squared degrees. It is calculated as $\pi r_a r_b$, where r_1 and r_2 are the face ellipse's major and minor radii respectively.

Closeness: between the faces f^1 and f^2 in the case of two faces is the euclidean distance between the two face regions. In Figure 4, $(d - (r(\alpha) + r(\beta))$ is the closeness C between the face ellipses with origins O^{f^1} and O^{f^2}.

Metrics

We used several evaluation metrics (minimum fixation distance, fixation proportion, fixation duration) to investigate the effects of faces on eye fixations during free-viewing of a visual scene. We analyzed the inflencing factors number, eccentricity, area, and closeness of

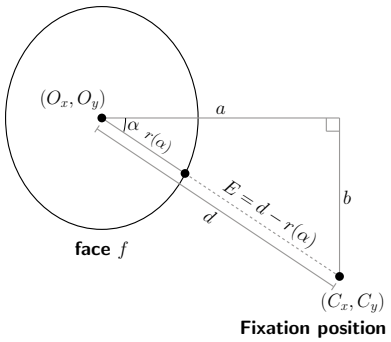


Figure 3. : Eccentricity of face E presented on screen from fixation. Consider a face ellipse f with major and minor radii, r_a and r_b, corresponding to face dimensions. $r(\alpha)$ is the radius to the position on the ellipse at angle α of a right angle triangle with sides of length a and b. The angle is measured from the axis of the face ellipse f to the fixation position (C_x, C_y). Finally, the radius $r(\alpha)$ is subtracted from the euclidean distance d between the origin of the face ellipse (O_x, O_y) and fixation position to obtain the eccentricity of the face.

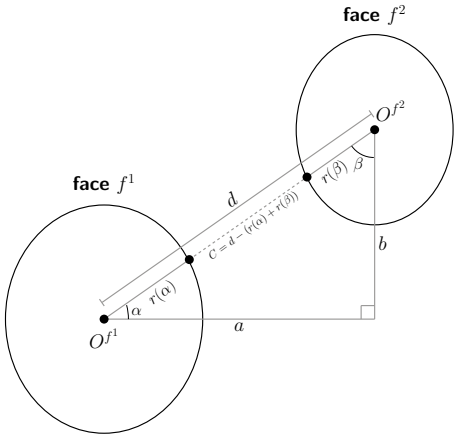


Figure 4. : Closeness C between two faces presented on screen. Consider two face ellipses f^1 and f^2 with major and minor axis equal to the respective dimensions of the faces. $r(\alpha)$ and $r(\beta)$ are the radii to positions on the ellipses at angles α and β of a right angle triangle with sides of length a and b. The angles are measured from the major axis of one face ellipse to the origin of the counterpart face ellipse. Finally, the radii $r(\alpha)$ and $r(\beta)$ are subtracted from the euclidean distance d between the origins O^{f^1} and O^{f^2} to obtain closeness between faces.

faces with linear regression models based on comparisons with face maps.

Minimum fixation distance: or *Shortest Euclidean distance* (Wang & Pomplun, 2012) from faces in a scene to fixation. The distance is computed from the fixation position to the face region of interest—the edge of face ellipse. Essentially, it is equal to the eccentricity E of the face closest to the fixation.

$$d_{min} = arg\,min_E$$

Fixation proportion We categorized the fixations on scenes with faces into two types: fixations landing inside a face, called *'on-face' fixations (oF)*, and fixations landing outside a face, called *'not-on-face' fixations (nF)*. This was done by comparing fixation coordinates to a face, represented by an elliptical mask equal to the face dimensions plus 1° of margin. In the study, we used the proportion of the two types of fixations, normalized by the total surface area of the faces. Here, we did not consider fixations for scenes with no faces to fixate upon.

Fixation duration. Cognitive systems interact with the scene to determine where, and how long to fixate. The position of fixation points toward the region of interest, while its duration amounts to the attentional processing directed to that location (Just & Carpenter, 1976; Rayner, 1998; Henderson, 2007).

Comparison with face maps. Different criteria are used to predict the likelihood of different regions attracting attention in a scene. It is often done by comparing such regions of interest to participant eye movements (Itti, Koch, & Niebur, 1998; Parkhurst, Law, & Niebur, 2002; Tatler, Baddeley, & Gilchrist, 2005; Peters, Iyer, Itti, & C., 2005; Torralba, Oliva, Castelhano, & Henderson, 2006; Le Meur, Le Callet, Barba, & Thoreau, 2006).

In this study, we are interested in analyzing the influence of faces in a scene. We used a comparison criterion to measure the correspondence between regions predicted to be fixated and regions fixated by participants, represented as face maps and eye fixation maps respectively.

- **Face maps:** We computed face map M^f (Figure 7b) for each frame by hand-labeling the position of the face using a bounding box, and then applying a 2D Gaussian to it. The dimensions of the bounding box determine the variance of the 2D Gaussian from its origin in horizontal and vertical axis, whereas the amplitude of the function was kept constant for all faces. All values outside the Gaussianed elliptical face were set to zero.
- **Eye fixation maps:** The eye fixation maps were defined for each fixation made by a participant. It is simply the fixation position Gaussianed for one participant equivalent to 0.5° of the visual field—the

size of the fovea with highest resolution. These maps, denoted as M^h, were used to evaluate faces using the comparison criterion. A sample M^h map is illustrated in Figure 5c.

(a) (b) (c)

Figure 5. : From left to right (a) the input frame with faces, (b) the face map M^f for the input frame with 2D-Gaussianed faces from the ground truth, and (c) the frame's corresponding fixation Gaussianed for one participant ($\sigma = 0.5°$), referred to as M^h.

• **Comparison criterion:** To compute the comparison criterion, for instance for the first fixation, we compare M^h for each participant to all M^f maps for the entire duration of the fixation. The values are then averaged to get a score for the participant. Likewise, this process is repeated for all participants. Finally, all individual scores from all participants are averaged to get the score for this first fixation. We do the same for the five fixations and process a mean score of the five scores. We looked at each fixation separately but obtained no difference compared to the mean of five fixations. Note that in the case of face maps M^f, the dimensions of the face define the standard deviations of the applied Gaussian where all values lying outside the resulting face ellipse are set to zero. In this study, we perform *ROC* (*Receiver Operating Characteristics*) analysis between two maps: a face map M^f and a eye fixation map M^h. The maps are processed as a binary classifier applied to every pixel; classified as fixated (or salient) or as not fixated (or not salient). A simple threshold is systematically moved between minimum and maximum values of the map. For each pair of thresholds, we get four numbers: the true positives (TP), the false positives (FP), the false negatives (FN) and the true negatives (TN). A *ROC* curve plots the false positive rate as a function of the true positive rate. The *ROC* area or the *AUC* (*Area Under Curve*) obtained by a trapezoid approximation, measures the classification performance. The trapezoidal rule used:

$$A = \frac{1}{2}\sum_{i=2}^{N}(x_i - x_{i-1}) \cdot (y_i + y_{i-1})$$

It is usually taken as a scalar value, such that a value $A = 0.5$ reflects random forecasts, whereas $A = 1.0$ implies perfect forecasts.

Statistical analysis

To measure the influence of *'number of faces'* in a scene, we compute a linear regression with levels (one face, two faces and no faces) for all video snippets—166 with one face, 120 with two faces and 191 with no faces. Likewise, this is done for all the evaluation metrics. To determine the influence of *'eccentricity of faces'*, *'area of faces'* and *'closeness between two faces'*, the *AUC* comparison criterion was averaged across subjects for 166 video snippets in the three cases of video snippets. The correlation of the interest of faces with these influencing factors (independent variables) may be checked by multiple linear regression analysis. The statistical significance of the correlation (interaction effects) and of each independent variable (main effects) is given as an F statistic from which a p value can be determined. Note that regression was used because the dependent or predictor variables were continuous.

Results

In this study, we evaluate the interest in faces using different metrics: fixation duration, minimum fixation distance from face, 'on-face' and 'not-on-face' fixation proportion and comparison with face maps.

Minimum fixation distance

Distance of fixation from the closest face averaged across subjects for video snippets (166 with one face and 120 with two faces) with faces is shown as a function of fixation number in Figure 6. We observe that distance from the closest face decreased on scene onset, reached its minimum at the second fixation of about 2.5°, and afterward remained steady in the following fixations. In the case of two faces, the distance to the closest face was smaller compared to one face, due to the presence of multiple regions of interest. This observation for fixation distance from face indicates one face attended yo in the first couple of fixations followed by exploratory fixations on the rest of the scene. This is not necessarily true in the case of two faces.

To get a baseline distance, we consider the fixation distribution of the entire eye fixation data. We take the median of this distribution as a baseline position, and compute the baseline using the distance of this position to the closest face, represented by a red line in Figure 6. The result shows that there is a significant difference between distance from a face and distance to the baseline point; $t_{(570)} = -13.534$, $p < 0.001$ [1]. In the

[1] 572 samples in total = 2 distance values per {166 (one face) +120 (two faces) video snippets}

case of two faces, participants remain in proximity of faces over time. However, they comprise a significant proportion of the total fixations for scenes with faces.

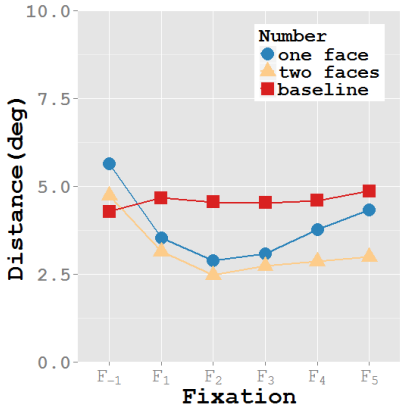


Figure 6. : Minimum fixation distance from face for *number of faces*—one and two faces. We took the first five fixations $\{F_1, F_2, F_3, F_4, F_5\}$ after current scene onset and fixation F_{-1} from the previous clip snippet (fixation just before onset of current scene).

We conclude that faces are considered potential regions of interest when present in a scene, and participants remained in proximity to them over time.

Fixation proportion

Initially, the proportion of fixations made on face regions in a scene is smaller, which is attributed to a center-looking strategy on scene onset. As the scene progresses, at the second fixation F_2, the proportion of fixations landing within the face regions becomes more than 50% of the total fixations. The preference of fixating faces becomes more prominent with the observation that the surface area occupied by face regions is much smaller compared to the rest of the visual stimuli. Here we normalized the proportion with area and analyzed it over time and found this large proportion remains true in the following fixations for both one and two faces (Figure 7). Moreover, statistical test on proportions of first five fixations shows significant difference between *'on-face' (oF)* and *'not-on-face' (nF)* fixations; $t_{(8)} = 53.080$, $p < 0.001$ in case of one face and $t_{(8)} = 30.283$, $p < 0.001$ in case of two faces.

We conclude that faces, when present in a complex scene, are considered potential regions of interest. We found that participants remained in proximity to face over time. In addition to the inherent interest of faces, their eccentricity and area are also important, since these two factors could lead to less distorted information to form a recognizable object, resulting in fixations made much closer to a face in a scene.

Fixation duration

Fixation durations averaged across subjects for video snippets with faces (286 with one or two faces, and 192 with no faces) show that snippets with no faces have a shorter average duration of fixation (7.45 frames, 298ms) than snippets with faces (10.42 frames, 417ms), and the difference is significant $t_{(476)} = 6.288$, $p < 0.001$. Since we use dynamic stimuli without sound, there is no impact of auditory cues on visual information extracted. We can imply that the presentation of visual stimuli alone leads to the centralization of gaze, resulting in longer fixations to extract maximum information from faces.

Comparison with face maps

Faces in a scene are certainly the regions of interest in dynamic stimuli, and they influence eye movements of the participants. In this section, we used a comparison criterion, the *AUC* score, to evaluate the face maps M^f against the eye fixation maps M^h. The aim was to evaluate the spatial importance of face regions.

Case of one face

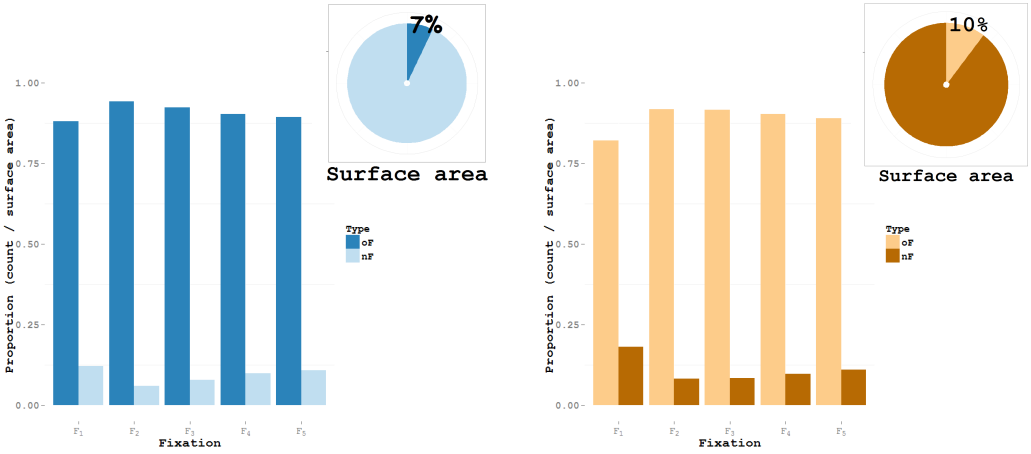
AUC scores averaged across subjects for video snippets with one face (166 in total) are shown as a function of eccentricity and area in Figure 8. We observe that performance of faces decreased with increasing eccentricity; that is, faces presented in the foveal region of the current fixation tend to attract gaze compared to faces presented in the peripheral region in a quite linear way. In contrast, the performance of faces increased with increasing surface area. Faces with larger surface areas attract more gaze compared to ones with smaller surface areas.

Different models with eccentricity and area have been tested in the case of one face (Table 2). The results of linear regression for these models are presented in Table 3. For all cases, p-value < 0.05, then all proposed models (E, A, and $E + A$) are significantly better than the constant model.

Table 2
: List of statistical models in case of one face.

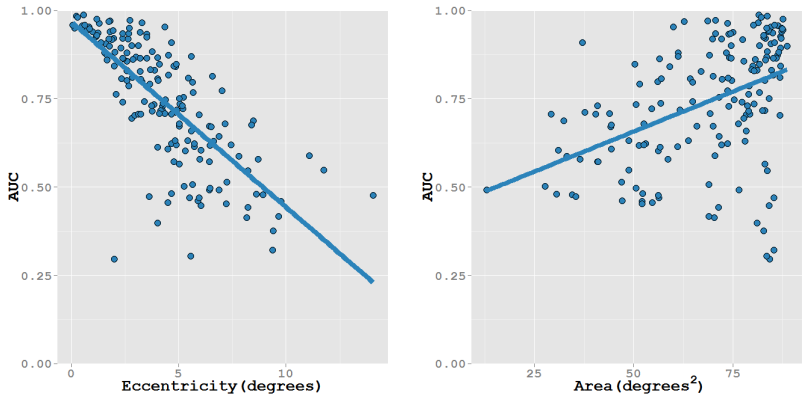
Short name	Models
Constant	$Y = b_0 + \varepsilon$
E	$Y = b_0 + b_1 E + \varepsilon$
A	$Y = b_0 + b_2 A + \varepsilon$
$E + A$	$Y = b_0 + b_1 E + b_2 A + \varepsilon$

In order to compare the different models, we consider f^2 criterion of Cohen (Cohen, 1988) whose defini-



(a) One face (b) Two faces

Figure 7. : Area-normalized proportion of fixations made on or off face regions, denoted as *'on-face' (oF)* and *'not-on-face' (nF)* fixations respectively. The pie chart represents the surface area of the face ellipses f_i computed as $\sum_i \pi r_a r_b$ and $(40^\circ \times 30^\circ) - \sum_i \pi r_a r_b$ for oF and nF regions respectively. The latter region uses the $40^\circ \times 30^\circ$ field of view. Here, We took the first five fixations $\{F_1, F_2, F_3, F_4, F_5\}$ after current scene onset.



(a) *'Eccentricity'* of face (b) *'Area'* of face

Figure 8. : Scatterplots with regression lines for AUC comparison criterion for one face.

tion is $f^2 = R^2/(1-R^2)$ where R^2 is the square of the coefficient of multiple correlation. Cohen suggested that f^2 values of 0.02, 0.15 and 0.35 represent respectively small, medium and large effect size. We can use it to compare models with the same number of variables. We also consider the Akaike information criterion (*AIC*) which is a measure of the relative quality of a statistical model for a given set of data. It takes into account both how well the model fits and its complexity. It provides a mean for model selection (The best model corresponds to the largest negative value of *AIC*).

For one face (Table 3), f^2 criterion is very large with E model and medium with A model, thus E model is better than A model to explain the data. f^2 is large for $E+A$ model too. *AIC* criterion shows that the best model is the $E+A$ model, which confirms the interest of associating area variable to the classical eccentricity variable to explain the attraction of faces during eye movement. Let us note that we verified the applied condition of the models (independence, normality and homoscedasticity of residuals) by an analysis of studentized residuals.

Table 3
: Results of linear regression for models tested in case of one face.

Model	F	p	f^2	AIC
E	$F(1,164) = 204.8$	< 0.001	1.248	-239.3
A	$F(1,164) = 38.67$	< 0.001	0.235	-139.9
$E+A$	$F(2,163) = 104.8$	< 0.001	1.280	-240.0

Case of two faces

AUC scores averaged across subjects for video snippets with two faces (120 in total) are shown as a function of eccentricity of face, area of face and closeness between faces in Figure 9. Similar to one face, we observe that performance of faces decreases with increasing eccentricity to the nearest face and decreasing area of face. In addition, closer face regions result in higher scores compared to faces which are farther apart.

Different models with eccentricity, area and closeness were tested in case of two faces (Table 4). The results of linear regression for these models are presented in Table 5. For all cases (E, A, C, $E+A$, $E+C$, $A+C$, and $E+A+C$) p-value is < 0.05, and hence significantly better than the constant model.

Table 4
: List of statistical models in case of two faces.

Short name	Models
Constant	$Y = b_0 + \varepsilon$
E	$Y = b_0 + b_1 E + \varepsilon$
A	$Y = b_0 + b_2 A + \varepsilon$
C	$Y = b_0 + b_3 C + \varepsilon$
E+A	$Y = b_0 + b_1 E + b_2 A + \varepsilon$
E+C	$Y = b_0 + b_1 E + b_3 C + \varepsilon$
A+C	$Y = b_0 + b_2 A + b_3 C + \varepsilon$
E+A+C	$Y = b_0 + b_1 E + b_2 A + b_3 C + \varepsilon$

For two faces (Table 5), f^2 criterion is medium for all models. We can use it to compare models with the same number of variables. So, E model is better than A model and C model. Also, $A+C$ model is better than $E+A$ model and $E+C$ model. AIC criterion shows that the best model is the $E+A+C$ model, which confirms the interest of associating area and closeness variables to the eccentricity variable to explain the attraction of faces during eye movement. Let us note that we verified the applied condition of the models (independence, normality and homoscedasticity of residuals) by an analysis of studentized residuals.

Table 5
: Results of linear regression for models tested in case of two faces.

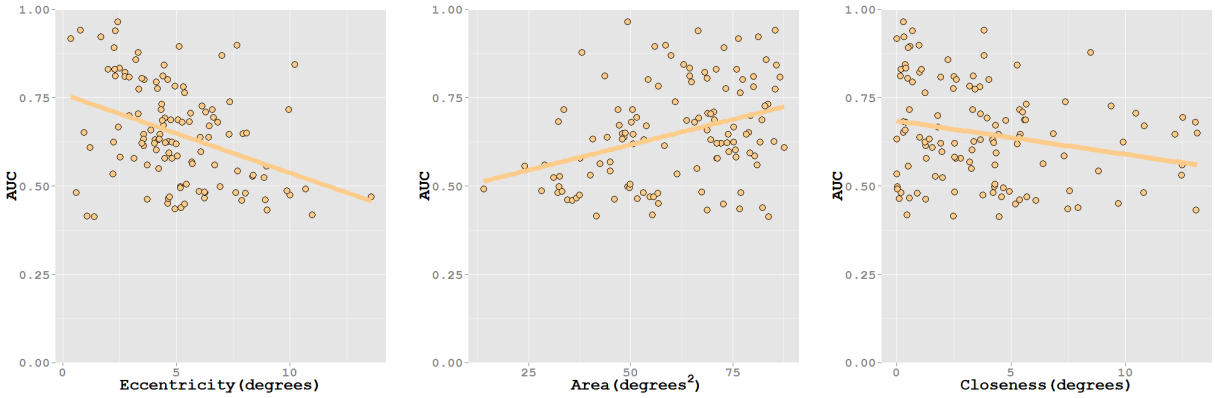
Model	F	p	f^2	AIC
E	$F(1,118) = 20.11$	< 0.001	0.170	-141.7
A	$F(1,118) = 16.23$	< 0.001	0.137	-138.3
C	$F(1,118) = 6.634$	0.011	0.056	-129.4
$E+A$	$F(2,117) = 12.68$	< 0.001	0.216	-144.4
$E+C$	$F(2,117) = 11.1$	< 0.001	0.189	-141.7
$A+C$	$F(2,117) = 13.29$	< 0.001	0.227	-145.4
$E+A+C$	$F(3,116) = 10.15$	< 0.001	0.262	-146.8

Discussion

In summary, we measured and found significant effects of influencing factors on the interest of faces with dynamic stimuli. A number of studies have been conducted evaluating the influence of faces on gaze with static stimuli, very few with dynamic stimuli. They showed that participants tend to fixate faces based on saliency (Birmingham, Bischof, & Kingstone, 2009), or due to their social importance (Birmingham, Bischof, & Kingstone, 2008). Dynamic stimuli can offer more information compared to static stimuli, as it contains information about social status, identity, and emotions (Buchan et al., 2007; Foulsham, Cheng, Tracy, Henrich, & Kingstone, 2010).

Several works have been published using dynamic social scene-viewing. Some compared visual fixation patterns of two groups, autistic and normal children (Rice et al., 2012; Riby & Hancock, 2009). They measured fixation time proportion on scene parts (face, body, background) and response times showing differences between the two groups. Other works used calibrated stimuli from interviews. They evaluated how gaze is dynamically directed to the eyes, nose, or mouth (Buchan et al., 2007; Võ et al., 2012) on single large face with or without sound. They measured fixation proportions on face parts showing a higher proportion on mouth when sound is present. Recent work (Song, Pellerin, & Granjon, 2013) using complex scenes with free-viewing, showed that only sounds produced by humans on-screen influence participants gaze. In this case, human gaze was attracted by a talking or singing face.

Our contribution here is to evaluate dynamic stimuli on the impact of faces on gaze with several fixation attributes and influencing factors, for one face and for two faces. The database is based on dynamic complex scenes from films without calibration and without sound. The varying face content makes the video database useful to study the interest of face in real videos. Since little of the face content used comes from famous movies, less than 5%, the impact of familiar faces was negligible. Our study involved an experiment with free-viewing participants. (Dorr, Martinetz, Gegenfurtner, & Barth, 2010) using complex videos to evaluate the variability of gaze positions from free-viewing participants showing the interest of such an experiment. First, we quantified eye fixation attributes: distance of fixations to faces, proportion of fixations on faces, and duration of fixations. Then we evaluated three influencing factors (eccentricity,



(a) *'Eccentricity'* of face (b) *'Area'* of face (c) *'Closeness'* between faces

Figure 9. : Scatterplots with regression lines for AUC comparison criterion for two faces.

area, closeness), with AUC criterion applied to fixation and face maps, and modeled the interactions of the influencing factors with statistical models.

With the distance of fixations to face criterion, we conclude that participants remained in proximity to faces over time (mean of 2.5°, similar to the literature (Buchan et al., 2007)). The short distance is significantly lower than the baseline which is constant over time. We quantified the proportion of fixations on faces. We obtained a ratio of 50% of fixations on faces corresponding to only about 10% of screen. This is coherent with previous studies obtaining from 50% to 80% of fixations on face (Rice et al., 2012; Riby & Hancock, 2009; Smith & Mital, 2013). Fixation durations in a dynamic scene with faces are longer than with no face (around 400ms for one and two faces versus around 300ms with no face) because they offer more visual information for perception. We found that they were significantly longer in presence of faces, hinting that faces in a scene trigger fixations preceded by small saccades, essentially motivated to perform detailed analysis of the facial features. The durations were longer than usually reported in the literature for static images, reporting 200-250ms with no face and around 300ms with faces (Pannasch, Helmert, Herbold, Roth, & Henrik, 2008; Smith & Mital, 2013; Guo, Mahmoodi, Robertson, & Young, 2006), which supports the idea of long explorations of few regions of interest. These results on fixation attributes are coherent with the literature and add new results on fixation duration on dynamic scenes with faces.

We also evaluated three influencing factors (eccentricity, area, closeness) with AUC criteria applied to comparison of fixation maps with ground truth face maps. First we consider the case of one face evaluating two variables (eccentricity and area). We observed that AUC drops as faces are presented farther away in the periphery or when their area is reduced. These observations are coherent with findings of different studies on static stimuli (with calibrated stimuli): that the performance of faces drops with increasing eccentricities with limited spatial information (Thorpe, Gegenfurtner, Fabre-Thorpe, & Bülthoff, 2001; Tomalski, Johnson, & Csibra, 2009; Guo, Liu, & Roebuck, 2011). Area of face can limit and compensate for the effects of eccentricity already reported for static stimuli (Johnson & Gurnsey, 2010), and it can also reduce the crowding of facial information, increasing the influence of faces: that is, face becomes more recognizable when larger. We evaluated the effect of Eccentricity and Area with linear regression models (E, A and $E+A$ models). f^2 criterion is high for E model showing that E model fits the data well. f^2 criterion is only medium for A model. AIC criterion shows that the $E+A$ combined model is slightly better than the E model and better than the A model.

In the case of two faces we evaluated three variables (eccentricity, area and closeness) with AUC criteria. This case is particularly difficult because it involves complex scenes with various faces of different size and position, and evolving in front of complex backgrounds. We observed for one face that AUC drops for increasing eccentricity and decreasing area. Moreover, the competing faces influence each other based on their relative location (closeness). The closer the two faces are, the greater their spatial strength to attract gaze is, hence limiting the effects of competition between the two faces. Thus, the AUC score increases when faces are closer. The results obtained on competition of faces corroborate previous findings on static images with calibrated stimuli showing that competition reduces performance of faces (Jacques & Rossion, 2004; Nagy, Greenlee, & Kovács, 2011). The performance with increasing eccentricity drops further compared to one

face. In a related study on static stimuli (Guo et al., 2011) when a face is presented alongside similar faces (with same characteristics), the performance to fixate a face drops, possibly a consequence of the limited information-processing capacity of peripheral vision. We evaluated the effect of the three variables: Eccentricity, Area and Closeness, with seven linear regression models (using E, A and C). In fact the three variables are not independent, they are tightly coupled. *AIC* scores are comparable for E and A models due to face competition (C model is a bit lower). The three variables provide information, then the combined models ($E+A$, $A+C$, $E+C$) obtain higher scores than single variable models. For the two faces case, the adding of a closeness variable to the traditional eccentricity and area variables improves the quality of the model (highest score for $E+A+C$ model).

In conclusion, we have evaluated with statistical models the influence of one or two faces that decreases with increasing eccentricity, with decreasing area and with increasing closeness. This study is important to understand eye movements for free viewing of complex object categories like faces in videos, in particular to build computational models for visual attention. The results could be helpful to support the adding of a modulation to a face pathway as for recently proposed visual saliency models (Cerf, Harel, Einhäuser, & Koch, 2007; Marat, Rahman, Pellerin, Guyader, & Houzet, 2013), to better predict eye movements. The models already take into account area of faces by modifying the representative Gaussians in face saliency maps. These could be improved by modulating the amplitude of Gaussians based on closeness. Furthermore, a coefficient based on area and closeness could be introduced to weight the face saliency map during the fusion phase with other saliency maps of the model. However, the use of eccentricity for modulation is not straightforward as it depends on human eye fixation unlike area and closeness. Nevertheless, factors like moving faces, emotion/expression could be investigated to improve the face pathway, and then the modulation.

References

Banks, S. M., Sekuler, A. B., & Anderson, S. J. (1991, Nov). Peripheral spatial vision: limits imposed by optics, photoreceptors, and receptor pooling. *Journal of the Optical Society of America A*, *8*(11), 1775–1787.

Bindemann, M., Scheepers, C., & Burton, A. M. (2009). Viewpoint and center of gravity affect eye movements to human faces. *Journal of Vision*, *9*(2), 7.1–16.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008). Gaze selection in complex social scenes. *Visual Cognition*, *16*(2), 341–355.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research*, *49*(24), 2992–3000.

Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience*, *2*(1), 1–13.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection. In *Nips'07*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates.

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, *10*(10), 1–17.

Dow, B. M., Snyder, A. Z., Vautin, R. G., & Bauer, R. (1981). Magnification factor and receptive field size in foveal striate cortex of the monkey. *Experimental Brain Research*, *44*, 213–228.

Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., & Kingstone, A. (2010). Gaze allocation in a dynamic situation: Speaking. *Cognition*, *117*, 319–331.

Guo, K., Liu, C. H., & Roebuck, H. (2011). I know you are beautiful even without looking at you: discrimination of facial beauty in peripheral vision. *Perception*, *40*(2), 191–195.

Guo, K., Mahmoodi, S., Robertson, R. G., & Young, M. P. (2006). Longer fixation duration while viewing face images. *Experimental Brain Research*, *171*(1), 91-98.

Hasson, U., Levy, I., Behrmann, M., Hendler, T., & Malach, R. (2002). Eccentricity bias as an organizing principle for human high-order object areas. *Neuron*, *34*(3), 479–490.

Heisz, J. J., & Shore, D. I. (2010). More efficient scanning for familiar faces. *Journal of Vision*, *8*(1), 1–10.

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science*, *16*(4), 219–222.

Hershler, O., Golan, T., Bentin, S., & Hochstein, S. (2010). The wide window of face detection. *Journal of Vision*, *10*(10), 21.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 1254–1259.

Jacques, C., & Rossion, B. (2004). Concurrent processing reveals competition between visual representations of faces. *Neuroreport*, *15*(15), 2417–2421.

Jacques, C., & Rossion, B. (2006). The time course of visual competition to the presentation of centrally fixated faces. *Journal of Vision*, *6*(2), 154–162.

Jebara, N., Pins, D., Despretz, P., & Boucart, M. (2009). Face or building superiority in peripheral vision reversed by task requirements. *Advances in Cognitive Psychology*, *5*, 42–53.

Johnson, A., & Gurnsey, R. (2010). Size scaling compensates for sensitivity loss produced by a simulated central scotoma in a shape-from-texture task. *Journal of Vision*, *10*(12), 1–16.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480.

Kastner, S., & Ungerleider, L. G. (2001). The neural basis of biased competition in human visual cortex. *Neuropsychologia*, *39*(12), 1263–1276.

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 802-817.

Levi, D. M., Klein, S. A., & Aitsebaomo, A. P. (1985). Vernier acuity, crowding and cortical magnification. *Vision Research*, *25*(7), 963–977.

Marat, S., Phuoc, T. H., Granjon, L., Guyader, N., Pellerin, D., & Guérin-Dugué, A. (2009). Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, *82*, 231–243.

Marat, S., Rahman, A., Pellerin, D., Guyader, N., & Houzet, D. (2013). Improving visual saliency by adding 'face feature map' and 'center bias'. *Cognitive Computation*, *5*(1), 63–75.

Miller, E. K., Gochin, P. M., & Gross, C. G. (1993). Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus. *Brain Research*, *616*(1-2), 25–29.

Nagy, K., Greenlee, M. W., & Kovács, G. (2011). Sensory competition in the face processing areas of the human brain. *PLoS ONE*, *6*(9), e24450.

Pannasch, S., Helmert, J. R., Herbold, A.-K., Roth, K., & Henrik, W. (2008). Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research*, *2*(2), 1–19.

Paras, C. L., Yamashita, J. A., Simas, M. L., & Webster, M. A. (2003). Face perception and configural uncertainty in peripheral vision. *Journal of Vision*, *3*(9), 822.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visaul attention. *Vision Research*, *42*, 107-123.

Peters, R. J., Iyer, A., Itti, L., & C., K. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*, 2397-2416.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Reddy, L., Reddy, L., & Koch, C. (2006). Face identification in the near-absence of focal attention. *Vision Research*, *46*(15), 2336–2343.

Riby, D., & Hancock, P. J. (2009). Looking at movies and cartoons: eye-tracking evidence from williams syndrome and autism. *Journal of Intellectual Disability Research*, *53*(2), 169–181.

Rice, K., Moriuchi, J. M., Jones, W., & Klin, A. (2012). Parsing heterogeneity in autism spectrum disorders: Visual scanning of dynamic social scenes in school-aged children. *Journal of the American Academy of Child and Adolescent Psychiatry*, *51*(3), 238–248.

Rigoulot, S., D'Hondt, F., Defoort-Dhellemmes, S., Despretz, P., Honoré, J., & Sequeira, H. (2011). Fearful faces impact in peripheral vision: behavioral and neural evidence. *Neuropsychologia*, *49*(7), 2013–2021.

Ro, T., Russell, C., & Lavie, N. (2001). Changing faces: A detection advantage in the flicker paradigm. *Psychological Science*, *12*(1), 94–99.

Rolls, E. T., & Tovee, M. J. (1995). The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. *Experimental Brain Research*, *103*(3), 409–420.

Rossion, B., Gauthier, I., Tarr, M. J., Despland, P., Bruyer, R., Linotte, S., & Crommelinck, M. (2000). The n170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain. *Neuroreport*, *11*(1), 69–74.

Rousselet, G. A., Macé, M. J. M., & Fabre-Thorpe, M. (2003). Is it an animal? is it a human face? fast processing in upright and inverted natural scenes. *Journal of Vision*, *3*(6), 440–55.

Sato, T. (1995). Interactions between two different visual stimuli in the receptive fields of inferior temporal neurons in macaques during matching behaviors. *Experimental Brain Research*, *105*(2), 209–219.

Smith, T. J., & Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision*, *13*(8), 1–24.

Song, G., Pellerin, D., & Granjon, L. (2013). Different types of sounds influence gaze differently in videos. *Journal of Eye Movement Research*, *6*(4), 1–13.

Still, D. L., Thibos, L. N., & Bradley, A. (1989). Peripheral image quality is almost as good as central image quality. *Investigative Ophthalmology and Visual Science*, *30*, 52.

Tatler, B., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, *45*, 643-659.

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bülthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience*, *14*(5), 869–876.

Tomalski, P., Johnson, M. H., & Csibra, G. (2009). Temporal-nasal asymmetry of rapid orienting to face-like stimuli. *NeuroReport*, *20*(15), 1309–1312.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006, October). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, *113*(4), 766–786.

Virsu, V., & Rovamo, J. (1979). Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research*, *37*(3), 475-494.

Võ, M. L.-H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? dynamic allocation of attention when viewing moving faces. *Journal of Vision*, *12*(13), 1–14.

Vuilleumier, P. (2000). Faces call for attention: evidence from patients with visual extinction. *Neuropsychologia*, *38*(5), 693 - 700.

Wang, H.-C., & Pomplun, M. (2012). The attraction of visual attention to texts in real-world scenes. *Journal of Vision*, *12*(6), 1–17.