Face perception: Influence of location and number in videos

Anis Rahman, Denis Pellerin, Dominique Houzet

FACE PERCEPTION: INFLUENCE OF LOCATION AND NUMBER IN VIDEOS

A. Rahman, D. Pellerin, D. Houzet

GIPSA-lab, UMR 5216, Grenoble, France

ABSTRACT

The study is about the influence of face in videos. In the experiment, the participants were instructed free viewing of various videos. The resulting eye positions are compared to the hand-labeled faces to evaluate the impact of location and number of faces in the visual field. Here, we defined three regions—Inside (I), Periphery (P), and Outside (O)—to categorize video frames with one or two faces based on the location of faces. Then we perform the evaluation of all these categories to get resulting scores. The scores indicate that the impact of face is a function of its eccentricity such that it falls as the face is far from the center of the visual scene. Similarly, the number of faces also limits face attraction.

1. INTRODUCTION

The sensitivity of primate's vision for different stimuli decreases gradually with increasing eccentricity. This eccentricity bias modulates the amount of information to determine the visibility of the given stimulus in peripheral vision. Apart from this bias, the result is also influenced by the task and the object category [1]. In the case of faces in videos, the stimulus patterns are degraded due to loss of information as the stimulus moves away from the foveal region [2].

The visual performance for face stimuli is also affected by the spatial interference of adjacent contours in the periphery [3]. A study by [4] finds that detection of faces is equally distinguishable in fovea and peripheral for one face. On the contrary, two faces presented alongside are difficult to detect and discriminate. Furthermore, object categories like faces with closely packed features tend for detailed scans to perform subtle visual tasks. On the contrary, peripheral shape information is sufficient for other object categories like places or buildings. Moreover, in the case of videos, there is crowding caused by interference between objects limiting the performance in peripheral visual field [5].

Information about the presence of face can be extracted shortly after scene onset [6]. Different studies claim that face detection is coded in a specific cortical area of the brain; the fusiform face area (FFA) [7]. Electrophysiological studies show that face processing is very fast, and human faces evoke a negative potential around 172ms (N170) [8]. In other similar studies, these early neuronal face-selective responses are found to occur around 100ms [9], and even earlier around 70ms [10]. This information can be used to control the eye movements. This explicit representation of faces suggests their importance for primates for social interaction [11].

Here, we studied the influence of faces during free viewing of videos, and the impact of location and number of faces. In the video and eye positions database used (Section 2), we observed a center bias on the scene onset followed by eye movements to attend a face, if present. We sub-divided the visual scene in three locations based on eccentricity from the fovea (Section 3). The resulting locations were used to categorize hand-labeled face maps on location and number of faces. We first presented our findings for one face in different locations (Section 4), and then for two faces (Section 5).

2. EYE MOVEMENT EXPERIMENT

We used the eye positions data from a previous experiment described in [12]. It aims to record eye movements of participants when looking freely at videos with various contents. This data is then used to understand which features explain the best eye movements and fixated locations. We recall the main aspects of the experiment:

- Stimuli: Fifty-three videos (25fps, 720×576 pixels per frame) were selected from different video sources. The videos were cut into 305 clip snippets each of 1-3s, and then strung together to obtain 20 clips of 30s.
- Data Acquisition: Participants (Fifteen adults) sitting with their heads stabilized on a chin rest, in front of a monitor at 57cm viewing distance ($40^\circ \times 30^\circ$ field of view). They were instructed to look at the videos without any particular task.
- Eye Position Density Maps: The eye tracker (SR Research EyeLink II) recorded the two eye positions at

500Hz—20 eye positions per frame and per participant. Then, for each frame, the median position for each participant was considered. A two-dimensional Gaussian was added to each position with standard deviation equal to $1.0°$. In the end, for each frame, we obtained a human eye position density map M_h.

3. METHOD

3.1. Assessment of eye positions

In this study, we consider faces in the center to have stronger impact due to the presence of center bias because of video sequences used and participants' position during experiment. This bias motivates participants to move closer to the center on visual scene onset, and then initiate scene exploration. The strategy is diminished as the scene progresses, and participants explore and fixate salient regions.

To show that there exists a center bias on the visual scene onset, we computed distance d from each frame's eye positions to center of the scene using:

$$d = \sqrt{(c_x - c_{h_x})^2 + (c_y - c_{h_y})^2}$$

where, the center points $c_{(x,y)}$ and $c_h(x, y)$ are the arithmetic mean centroids of all eye positions for the database and of eye positions for corresponding frame respectively. We then took mean distances for first 50 frames for all video snippets, and plotted them along time as shown in Figure 1. We found that the distance was minimum at the visual scene onset around the fifth frame for all video snippets.

The center for all eye positions $c_{(x,y)}$ was used instead of the actual center of the visual scene throughout our study. Moreover, similar illustrations along time were used for evaluation results.

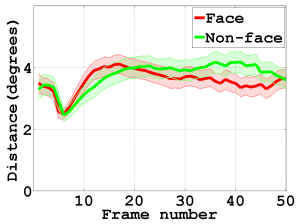


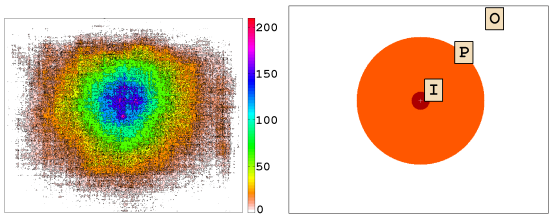
Fig. 1: Distance of eye positions from center of all eye positions.

3.2. Creation of face maps

A face map M_f was generated for each frame by hand-labeling face using bounding box, then applying 2D Gaussian upon it. The dimensions of the bounding box determine variance of the Gaussian from origin in horizontal and vertical axis, whereas amplitude of the function is identical. Consequently, we get 7599 face maps with 17580 faces in total for entire video database (14155 frames).

3.3. Definition of locations

All the hand-labeled faces for the video database were assigned to three different categories based on their location: inside (I), peripheral (P) and outside (O) faces. We started by setting center point to center of mass of all experimental eye positions illustrated in Figure 2a. Then, the visual scene was divided into three regions with eccentricities $2.0° \times 2.0°$ and $14.0° \times 14.0°$ for the inner two regions, whereas the rest was considered as the outside region. Here, inside region corresponds to the fovea, whereas outer bound of peripheral region is selected to analyze the eye movements "in near or far periphery" as defined in [1]. The resulting regions are illustrated in Figure 2b. Additionally, to reduce the influence of face size, we excluded all faces larger than radius $(r_{periphery} - r_{inside})$.



(a) (b)

Fig. 2: (a) Eye position distribution for all the database, (b) proposed regions for analysis: Inside (I), Periphery (P), and Outside (O).

3.4. Evaluation metrics

We used two evaluation metrics to estimate the relevance between the eye-position density maps M_h and the face maps M_f.

- NSS: Normalized Saliency Scanpath acts like a z-score computed by comparing a face map to participants eye positions. The $NSS_{(x,y)}$ at positions (x, y) of a face map is given as:

$$NSS_{(x,y)} = \frac{M_{f_{(x,y)}} \cdot M_{h_{(x,y)}} - \bar{x}_f}{s_f}$$

  $\bar{x}_f$: Empirical mean of face map M_f
  s_f : Empirical standard deviation of face map M_f

- TC: The metric simply estimates the ratio of predicted salient regions by face map over all experimental eye

positions.

$$TC = 100 \times \frac{N_{within}}{N_{all}}\%$$

N_{within} : Positions within salient regions
N_{all} : Total experimental eye positions

4. RESULTS: ONE FACE VERSUS ITS LOCATION

In this first case, we selected face maps with only *one* face in one of the two regions around the fovea—the periphery location or the outside location. The face maps M_f were then evaluated against the experimental eye positions.

The resulting scores for the first fixation[1] after the scene onset are detailed in Table 1. We found that a face in peripheral location attracts attention more compared to a face in outside location.

		Periphery	Outside
Frames (#)		408	227
NSS	$\bar{x}$	3.72	1.96
	$SE_{\bar{x}}$	0.160	0.187
TC (%)	$\bar{x}$	47	24
	$SE_{\bar{x}}$	1.626	1.611

Table 1: Scores for one face in peripheral or outside region. Scores for first fixation for frames 8 to 16.

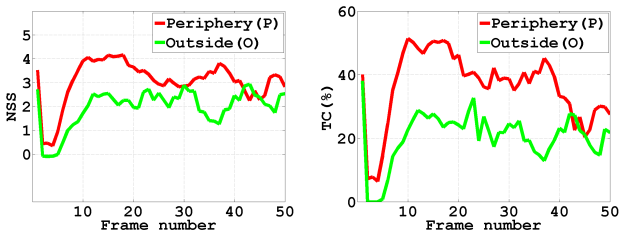
Similarly, temporal evolution illustrated in Figure 3 suggest that faces near the fovea are significantly responsive ($F(1, 203) = 22.87$, $p < 0.001$)[2] compared to faces presented in the outside region. We conclude that the influence of a face on attention is higher in peripheral region, and it gradually decreases as it appears away from the foveal region into the outside region.

5. RESULTS: TWO FACES VERSUS THEIR LOCATIONS

In this second case, we selected face maps M_f with *two* faces in *two* regions around the fovea—periphery and outside locations. We ended up with three categories: two periphery faces (PP), two outside faces (OO), and two faces in different locations (PO). The maps were then evaluated against eye position density maps.

[1]NSS and TC scores tabulated for different categories are computed for the "first fixation" with duration from 8^th to 16^th frames. This interval is mean value of the start and end of first fixation for all video snippets.

[2]For clarity, only statistics using NSS criteria are presented since both evaluation metrics NSS and TC generally produce the same conclusion. We took the mean scores for each of the 305 video snippets, and then applied analysis of repeated measures of variance (ANOVA). These significance tests were used to find the effects of face stimulus location and number.



(a) NSS evolution (b) TC evolution

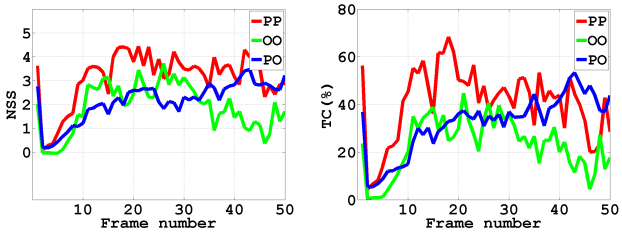
Fig. 3: One face in peripheral or outside region.

The resulting scores for the first fixation after the visual scene onset are detailed in Table 2. We found that faces in the peripheral location (PP) have stronger impact on attention compared to faces in the outside location (OO). The difference in scores was significant ($F(1, 96) = 7.02$, $p < 0.01$). Similarly, the scores for two faces in peripheral location (PP) are higher than faces in combined location (PO) with significance ($F(1, 136) = 1.49, p > 0.05$). Furthermore, the scores for two faces in location categories OO and PO are almost similar ($F(1, 137) = 2.22, p > 0.05$).

		PP	OO	PO
Frames (#)		88	79	162
NSS	$\bar{x}$	2.94	2.10	1.56
	$SE_{\bar{x}}$	0.268	0.306	0.145
TC (%)	$\bar{x}$	45	26	22
	$SE_{\bar{x}}$	4.242	3.597	2.358

Table 2: Scores for two faces in peripheral or outside region. Scores for first fixation for frames 8 to 16.

The temporal evolution in Figure 4 shows that influence of two faces depend on their eccentricities from the foveal region.



(a) NSS evolution (b) TC evolution

Fig. 4: Two faces in three categories for periphery and outside regions: PP, OO, and PO.

Our conclusion is based on results for one face that periphery region shows more sensitivity to face stimulus, and this decreases gradually as eccentricity of the face stimulus from the foveal region increases. Since the face maps consist

of two faces instead of one, visual crowding can cause a slight drop in impact of face.

5.1. Two faces in peripheral and outside regions (PO)

To investigate further results for two faces in combined periphery and outside (PO) region, we considered both faces as individual maps, and evaluated them against eye position density maps.

Table 3 summarizes evaluation scores for first fixation on the visual scene onset. The results clearly show superiority of faces in peripheral location compared to outside faces ($F(1,177) = 31.83$, $p < 0.001$). Similarly, temporal evolution in Figures 5a and 5b also indicate that contribution of peripheral faces is far greater than compared to outside faces. These results suggest that a face in the outside location cannot compete for attention with a peripheral face. Moreover, this drop is not influenced by the size of the faces since the size difference between the pairs of face is insignificant ($F(1,181) = 1.78$, $p > 0.05$), as illustrated in Figure 5c. We conclude that the face in peripheral region attract more attention than the face in outside region. This was congruent with our findings for one and two faces in peripheral region.

		P	O
NSS	$\bar{x}$	1.47	0.06
	$SE_{\bar{x}}$	0.095	0.063
TC (%)	$\bar{x}$	19	2
	$SE_{\bar{x}}$	2.389	0.419

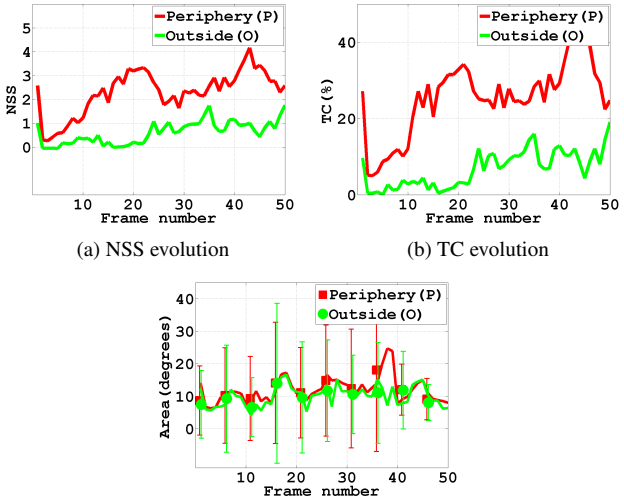
Table 3: Scores for two faces in two different regions; peripheral and outside. Scores for first fixation for frames 8 to 16.

5.2. Two faces in outside region (OO)

In this section, we tried to confirm that eccentricity also affects the influence of faces in the outside region. First, we selected face maps for two-face outside location (OO). Then, we separated the pairs of faces in two groups 'NEAR' and 'FAR' based on their distance from the center. Each face is then considered as an individual face map which is evaluated against the corresponding eye position density map. We compute distance d_f from the center of individual face bounding box $f_{(x,y)}$ to the center $c_{(x,y)}$ using:

$$d_f = \sqrt{(c_x - f_x)^2 + (c_y - f_y)^2}$$

This distance is used to separate the pair of faces (f^1, f^2) in two groups 'NEAR' and 'FAR' faces using their distances



(a) NSS evolution (b) TC evolution

(c) Size distribution

Fig. 5: Two faces in two different regions; peripheral and outside.

(d_{f^1}, d_{f^2}).

$$\begin{cases} f^1 = \text{NEAR}; f^2 = \text{FAR}, & \text{if } d_{f^1} >= d_{f^2} \\ f^2 = \text{NEAR}; f^1 = \text{FAR}, & \text{otherwise} \end{cases}$$

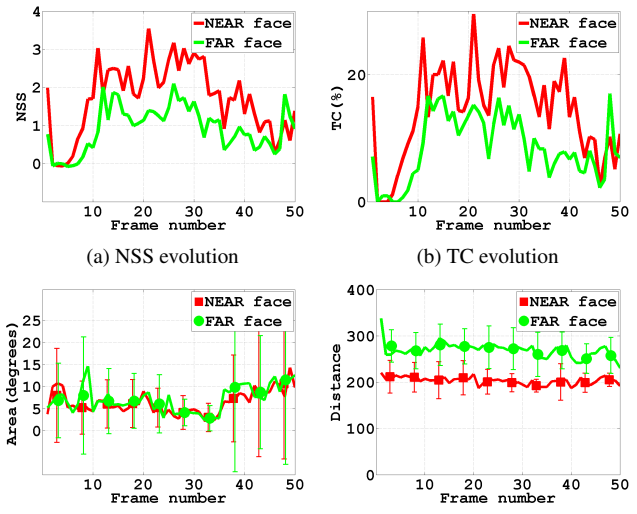
The resulting scores for the first fixation after the visual scene onset are detailed in Table 4. We found that contribution of 'NEAR' faces is greater compared to 'FAR' faces ($F(1,97) = 11.51$, $p < 0.001$). Although both faces are in the outside region, but the face with lower eccentricity from the foveal region grabs more attention than the farther one. Our finding was confirmed in Figures 6a and 6b. Moreover, Figure 6c illustrates that size of 'NEAR' and 'FAR' faces had a slight impact on evaluation scores ($F(1,99) = 0.42$, $p > 0.05$). On the other hand, Figure 6d shows the difference of distance between the two groups ($F(1,99) = 63.28$, $p < 0.001$). This difference facilitated our observation that there was an influence of eccentricity of the faces from the foveal region even in the outside region.

		Near	Far
NSS	$\bar{x}$	1.91	1.01
	$SE_{\bar{x}}$	0.232	0.231
TC (%)	$\bar{x}$	16	10
	$SE_{\bar{x}}$	1.923	2.008

Table 4: Scores for NEAR and FAR faces in outside region. Scores for first fixation for frames 8 to 16.

6. DISCUSSION

The work studies the impact of eccentricity and number of faces in videos. We use video sequences with ground-truth



(a) NSS evolution (b) TC evolution

(c) Size distribution (d) Distance distribution

Fig. 6: NEAR and FAR faces in outside region.

for faces that were sub-divided into several categories based on their eccentricity from the fovea, and the number of faces presented. These different categories were then evaluated against the human eye positions.

The results for one face confirm that when a face is in periphery region, it is more responsive as compared to a face in outside region. The same observation is repeated for two faces where faces in periphery region are more attention grabbing compared to two faces in outside region or in a combination of two regions.

Two faces in a combination of different regions present almost similar scores to two faces in the outside region. The result suggests influence faces even in the outside region. However, this is not true for two faces in a combination of regions because we find that peripheral face is the one that impact attention, but not the one in the outside region. Furthermore, in the case of two outside faces, a face with smaller eccentricity outscores its counterpart, but somewhat smaller compared to the case of two faces in different regions.

To conclude this study, we find that the influence of faces in attention is a function of their eccentricity and number of faces present. The results show that the influence decreases with increasing eccentricity by moving further into periphery, or when two faces are presented. Our study is crucial to understand the eye movements to determine the importance of complex object categories like faces, and to help their inclusion into computational models for visual attention.

Acknowledgment

The research is supported by Rhône-Alpes region (France) under the CIBLE project No. 2136.

7. REFERENCES

[1] N. Jebara, D. Pins, P. Despretz, and M. Boucart, "Face or building superiority in peripheral vision reversed by task requirements," *Advances in cognitive psychology*, vol. 5, pp. 42–53, 2009.

[2] T. Sato, "Interactions between two different visual stimuli in the receptive fields of inferior temporal neurons in macaques during matching behaviors.," *Exp. Brain Res.*, vol. 105, no. 2, pp. 209–219, 1995.

[3] A. Toet and D. M. Levi, "The two-dimensional shape of spatial interaction zones in the parafovea.," *Vision Res.*, vol. 32, no. 7, pp. 1349–1357, 1992.

[4] F. Farzin, S. M. Rivera, and D. Whitney, "Spatial resolution of conscious visual perception in infants.," *Psychol. Sci.*, vol. 21, no. 10, pp. 1502–1509, 2010.

[5] J Rovamo, P Mäkelä, R Näsänen, and D Whitaker, "Detection of geometric image distortions at various eccentricities.," *Invest. Ophth. Vis. Sci.*, vol. 38, no. 5, pp. 1029–1039, 1997.

[6] H. Liu, Y. Agam, J. R. Madsen, and G. Kreiman, "Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex.," *Neuron*, vol. 62, no. 2, pp. 281–290, 2009.

[7] N. Kanwisher and G. Yovel, "The fusiform face area: a cortical region specialized for the perception of faces," *Philos. Trans. R. Soc. London, Ser. B*, vol. 361, no. 1476, pp. 2109–2128, 2006.

[8] S. Bentin, T. Allison, A. Puce, E. Perez, and G. McCarthy, "Electrophysiological studies of face perception in humans," *J. Cognitive Neurosci.*, vol. 8, no. 6, pp. 551–565, 1996.

[9] S. M. Crouzet, H. Kirchner, and S. J. Thorpe, "Fast saccades toward faces: face detection in just 100 ms.," *J. Vision*, vol. 10, no. 4, pp. 16.1–17, 2010.

[10] S. C. A. Braeutigam, A. J. Bailey, and S. J. Swithenby, "Task-dependent early latency (30-60 ms) visual processing of human faces and other objects.," *NeuroReport*, vol. 12, no. 7, pp. 1531–1536, 2001.

[11] J. Anderson, "Social stimuli and social rewards in primate learning and cognition," *Behav. Process.*, vol. 42, no. 2-3, pp. 159–175, 1998.

[12] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *Int. J. Comput. Vision*, vol. 82, pp. 231–243, 2009.