AN EFFICIENT AUDIOVISUAL SALIENCY MODEL TO PREDICT EYE POSITIONS WHEN LOOKING AT CONVERSATIONS

Antoine Coutrot
CoMPLEX
University College London
London, United Kingdom

Nathalie Guyader
Gipsa-lab
Grenoble-Alpes University
Grenoble, France

ABSTRACT

Classic models of visual attention dramatically fail at predicting eye positions on visual scenes involving faces. While some recent models combine faces with low-level features, none of them consider sound as an input. Yet it is crucial in conversation or meeting scenes. In this paper, we describe and refine an audiovisual saliency model for conversation scenes. This model includes a speaker diarization algorithm which automatically modulates the saliency of conversation partners’ faces and bodies according to their speaking-or-not status. To merge our different features into a master saliency map, we use an efficient statistical method (Lasso) allowing a straightforward interpretation of feature relevance. To train and evaluate our model, we run an eye tracking experiment on a publicly available meeting videobase. We show that increasing the saliency of speakers’ faces (but not bodies) greatly improves the predictions of our model, compared to previous ones giving an equal and constant weight to each conversation partner.

Index Terms— saliency model, audiovisual, face, eye movements, conversations

1. INTRODUCTION

Visual attention models emphasize the regions of a visual scene most likely to attract the gaze of observers. Applications of these models are numerous, not only for cognitive sciences and neurosciences, but also for multimedia technologies, like video processing for multimedia delivery, retargeting or image quality assessment [1]. In the last decades, many different models have been proposed [2], most of them relying on the Feature Integration Theory [3]. Such models split an input visual stimulus into several feature maps like luminance, contrast, orientation, color and motion, at different scales. These feature maps are then normalized and merged into a master saliency map to emphasize the most salient regions. The efficiency of these models is tested by comparing their outputs to the eye positions of several observers recorded during eye-tracking experiments. Although reliable with many stimuli, most attention models do not consider the social nature of perception, and dramatically fail for visual scenes involving faces [4, 5]. Quite recently, visual saliency models combining faces with classic low-level features have been developed and significantly outperformed previous ones [6, 7]. All these attention models are “silent”: none of them consider sound as an input, yet ubiquitous in dynamic natural scenes. In previous studies, we showed that soundtracks significantly impacts on gaze behavior [8], and particularly when viewing conversation scenes [9]. We showed that if participants always look more at talking faces, hearing the original soundtrack makes them follow the speech turn-taking even more closely [5]. Based on these results, we proposed an audiovisual saliency model including a speaker diarization algorithm able to automatically spot “who speaks when” [10]. This algorithm allowed us to modulate the saliency weight of each conversation partner according to their speaking-or-not status.

The contribution of this paper is two-fold. Firstly, we refine our audiovisual saliency model by quantifying the relative saliency of conversation partners’ faces and bodies. Secondly, we use an efficient statistical method (Lasso) to estimate the weights of the different feature maps to be merged into the master saliency map. This method, while widespread for model selection in genetics, has never been used for attention modeling. To meet these goals, we run a new eye-tracking experiment on a publicly available meeting videobase.

2. AUDIOVISUAL SALIENCY MODEL

Our model follows the classic layout of the models inspired by the Feature Integration Theory [3]. It splits each frame in different feature maps, before merging them into a master saliency map (Figure 1). The different feature maps of the model have been fully described in [10]. In this section the latter are rapidly recalled, and a new statistical method to merge them into a master saliency map is presented.

*The first author performed the work while at Gipsa-lab, Grenoble-Alpes University, France.
2.1. Features of the Model

For each frame are computed:

- **Static and Dynamic Saliency** maps, from a classic spatiotemporal saliency model [11]. The static map emphasizes for each frame the spatial regions that differ from their context in terms of luminance, orientation and spatial frequency. The dynamic map extracted objects’ relative motion, with a preprocessing stage consisting in background motion compensation (for the videos where camera is moving).

- **Center Bias** map. As in [5,12], the center bias is modeled by a time-independent bi-dimensional Gaussian function centered at the screen center. Indeed, numerous eye-tracking studies reported that subjects tend to gaze more at the center of the image [13].

- **Face and Body** maps. The face and the body of each conversation partner are marked by a rectangle mask. A body mask contains the whole conversation partner excluding his face. The coordinates of each mask were dynamically defined for each frame using Sensarea software [14]. We visually checked the efficiency of the segmentation. While oro-facial information is obviously mandatory to understand one’s speech, body language and gestures also are crucial [15]. Here, we aim at quantifying how these two features attract observers’ gaze.

2.2. Fusion

Merging feature maps has always been a challenge, as they present different range and distribution [16]. Many different techniques have been used, from the simple average to the most complex machine learning techniques [17]. Here we propose a weighted linear combination of the feature maps. At each frame, the weight of each normalized feature is estimated from eye-tracking data with an efficient statistical method: Least Absolute Shrinkage and Selection Operator (Lasso) [18]. While widespread for model selection in genetics, this method has never been used for attention modeling. The Lasso is a regularized version of the Least Square method. Given an eye position map $Y$ obtained through an eye-tracking experiment with $N$ participants, the weights $\beta$ of the $p$ features $X$ are estimated via:

$$
\beta_{\text{Lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (Y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 \right\} \text{ with } \sum_{j=1}^{p} |\beta_j| \leq \lambda
$$

with $\lambda$ a penalization constant scaling down the number of parameters. The optimal $\lambda$ is the one leading to the model with the smallest Bayesian Information Criterion (BIC) [19]. The Matlab toolbox ”Sparse Statistical Modeling” gives an implementation of this algorithm [20]. For each frame, the best features to explain the experimental eye position map are the ones with the highest weight. The Lasso is related to the Expectation-Maximization (EM) algorithm, a popular statistical method recently applied to model eye positions on static [12,21,22] and dynamic [5] scenes. The major advantage of the Lasso is the sparsity imposed by the penalization.
as shown at the top of Figure 1. The resolution is 1232 × angles, we put side-by-side the four conversation partners, and 80 seconds. Since the meetings were shot from different meetings ((IN1008, IN1012 and IN1014) that we split into 15 videos (5 per meeting). Each video lasts between 20 different meetings. We chose 3 different meetings ((IN1008, IN1012 and IN1014) that we split into 15 videos (5 per meeting). Each video lasts between 20 and 80 seconds. Since the meetings were shot from different angles, we put side-by-side the four conversation partners, as shown at the top of Figure 1. The resolution is 1232 × 504 pixels (43.4 × 15.5 degrees), 25 fps. Dialogues are in English, sampled at 48 kHz.

Participants & Apparatus
40 participants took part in the experiment: 28 men and 12 women, from 22 to 36 years old. Participants were not aware of the purpose of the experiment and gave their informed consent to participate. This study was approved by the local ethics committee. Eye movements were recorded using an eye-tracker (Eyelink 1000, SR Research) with a sampling rate of 1000 Hz. We recorded the eye positions of the dominant eye in pupil / corneal-reflection tracking mode.

Procedure
Each video has been seen in the Visual condition (no soundtrack) and in the AudioVisual condition (original speech soundtrack) by 20 different participants. Each experiment was preceded by a calibration procedure, during which participants focused their gaze on nine separate targets in a 3 × 3 grid that occupied the entire display. A drift correction was carried out between each video, and a new calibration procedure was performed if the drift error was above 0.5 degree. To avoid any order effect, videos were randomly displayed.

3. MODEL TRAINING AND EVALUATION
To train and evaluate our model, we ran an eye-tracking experiment on a publicly available video base.

3.1. Eye-tracking Experiment
Stimuli
We used the AMI Meeting Corpus [23], comprising 100+ hours of meetings between four colleagues. We chose 3 different meetings ((IN1008, IN1012 and IN1014) that we split into 15 videos (5 per meeting). Each video lasts between 20 and 80 seconds. Since the meetings were shot from different angles, we put side-by-side the four conversation partners, as shown at the top of Figure 1. The resolution is 1232 × 504 pixels (43.4 × 15.5 degrees), 25 fps. Dialogues are in English, sampled at 48 kHz.

Procedure
Each video has been seen in the Visual condition (no soundtrack) and in the AudioVisual condition (original speech soundtrack) by 20 different participants. Each experiment was preceded by a calibration procedure, during which participants focused their gaze on nine separate targets in a 3 × 3 grid that occupied the entire display. A drift correction was carried out between each video, and a new calibration procedure was performed if the drift error was above 0.5 degree. To avoid any order effect, videos were randomly displayed.

3.2. Training
The weights of each feature map have been estimated for each frame in each experimental condition (Visual or AudioVisual) with the Lasso algorithm. To compare the attractive power of Body vs. Face maps, we ran the Lasso twice: once with all the features described section 2.1 (Static Saliency, Dynamic Saliency, Center Bias, Face and Body maps), and once without the Body maps. The results are shown on the left side of Figure 2. We see that despite their small size, Faces are by far the most important feature, more than three times more important than Bodies. Center Bias, Static and Dynamic Saliency, are barely significant. On the right side of Figure 2, we averaged the Body and Face weights of each conversation partner over their speaking and silent periods of time spotted by the speaker diarization algorithm. The weights of speaking faces are significantly greater than the weights of silent faces, particularly in the AudioVisual Condition. These results are in line with those presented in [5] with the Expectation-Maximisation method.

3.3. Evaluation
Here we compare the ability of five different master saliency maps to predict observers’ eye positions recorded in AudioVisual condition. The models differ in terms of features used and of fusion mode.

1. Static and Dynamic Saliency, Center Bias, Speakers’ Face, Addressees’ Face, weighted with the Lasso algorithm (upper part of Figure 2).
2. Static and Dynamic Saliency, Center Bias, Speakers’ Face and Body, Addressees’ Face and Body, weighted with the Lasso algorithm (lower part of Figure 2).
3. Static and Dynamic Saliency, Center Bias and Faces (equal and constant weight for every face), weighted with the Lasso algorithm (upper left part of Figure 2).
4. Simple average of Static and Dynamic Saliency, Center Bias and Faces.
5. Static and Dynamic Saliency only, combined as described in [11].

Not to evaluate the saliency maps with the same eye positions as the ones we used to estimate their feature weights, we followed a "leave-one-out" approach. More precisely, the weights used to train the model for a given video originate from the average over the weights of every video but the one
Fig. 2. Training with eye positions recorded in Visual and AudioVisual conditions. (a) Mean values of Lasso weights for Static and Dynamic Saliency, Centre Bias, Bodies and Faces masks. (b) Contributions of the speakers (S) and addressees (A) to the Bodies and Faces features in panel (a). (c and d) Idem, without the Body feature. Weights are averaged over every frame of each video, and over every video. Error bars correspond to standard errors.

We jointly used the Normalized Scanpath Saliency (NSS) [24] and the Kullback-Leibler divergence (DKL), two metrics widely used for saliency model ranking [25]. The greater the NSS and the lower the DKL, the better the model. The results shown in Figure 3 are consistent: when the NSS of a model is high, its DKL is low. We performed two ANOVAs with the different models as within-subject factors on NSS and DKL mean values. There is a main effect of the model type on the NSS (F(4,56) = 453.7, p < .001) and DKL (F(4,56) = 78.9, p < .001) values. The best model is the first one, giving different weights to speakers and addressees’ faces (Bonferroni post-hoc comparisons, all p < .001). Unexpectedly, model 2 which also considers conversation partners’ body is less efficient, with a NSS close to the one of model 3 (p = .2), and a DKL close to the one of model 4 (p = 1). Not separating speakers from addressees’ faces (models 3 and 4) also decreases model performances. As expected, not considering faces at all (model 5) leads to the worst performances. Except for the DKL values of models 3 and 5 (p = .15), all the differences presented Figure 3 are significant (all p < .001).

Fig. 3. Evaluation - Divergence of Kullback-Leibler (DKL) and Normalized Scanpath Saliency (NSS) for the different models described Section 3.3. For the models 1, 2 and 3, the feature weights have been estimated with the Lasso algorithm applied to the eye positions recorded in the AudioVisual condition of the experiment described Section 3.1.

4. CONCLUSIONS

In this paper, we used and refined an efficient audiovisual saliency model for conversation scenes. The model relies on a speaker diarization algorithm able to automatically spot “who speaks when”. It uses a statistical method (Lasso) to estimate the weights of different elementary features before merging them into a master saliency map. While many efficient but opaque machine learning techniques have been used for this purpose, the Lasso allows a straightforward interpretation of feature relevance. We ran an eye tracking experiment on a publicly available meeting videobase, the AMI Meeting Corpus. We used this new dataset to train and evaluate our model. We showed that giving a greater weight to speakers’ face significantly increases the model efficiency, but considering the whole body degrades it. This result could be imputed to the large surface of the body masks compared to its relatively low saliency. To test this hypothesis, it could be interesting to independently quantify the contribution of smaller parts of the body, like the hands or the torso.

REFERENCES


