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

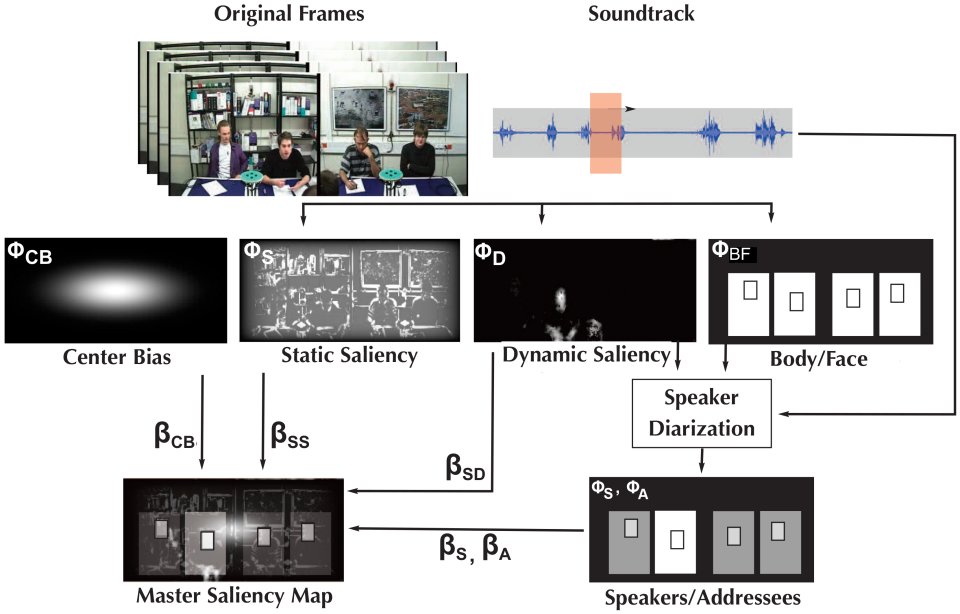

Fig. 1. Block diagram of our audiovisual saliency model. Center Bias, Static and Dynamic Saliency, Speakers and Addresses (Faces or Bodies) maps are weighted with the $\boldsymbol{\beta}^{Lasso}$ estimated weights, and merged into the audiovisual Master Saliency map.

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 $\boldsymbol{\beta}$ of the p features X are estimated *via* :

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

with λ a penalization constant scaling down the number of parameters. The optimal λ 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

constant, while the EM deals with all the parameters given as inputs, and runs the risk of overfitting.

2.3. Speaker Diarization

As shown in [5], the speakers are more likely to attract the attention than the other conversation partners. Therefore, it makes sense to give to Face and Body maps different weights according to their talking-or-not status. Speaker diarization algorithms answer to the "Who speaks when?" question. They allow to automatically distinguish a speaking from a silent conversation partner, and thus to adapt the weights of the corresponding Face and Body maps. The algorithm we use relies on both the visual and the auditory signals and comprises three stages : (1) Voice Activity Detection, (2) Speaker Temporal Clustering and (3) Speaker Spatial Labelling. More details can be found in [10].

Let's consider a conversation scene with different speakers. Once every conversation partner's speech periods have been labelled, we average the Face and Body map weights over their corresponding speaking and silent time periods, leading to "Speakers" (β_S) and "Addressees" (β_A) weights.

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 $\times$ 504 pixels (43.4 $\times$ 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 $\times$ 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 masks, 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

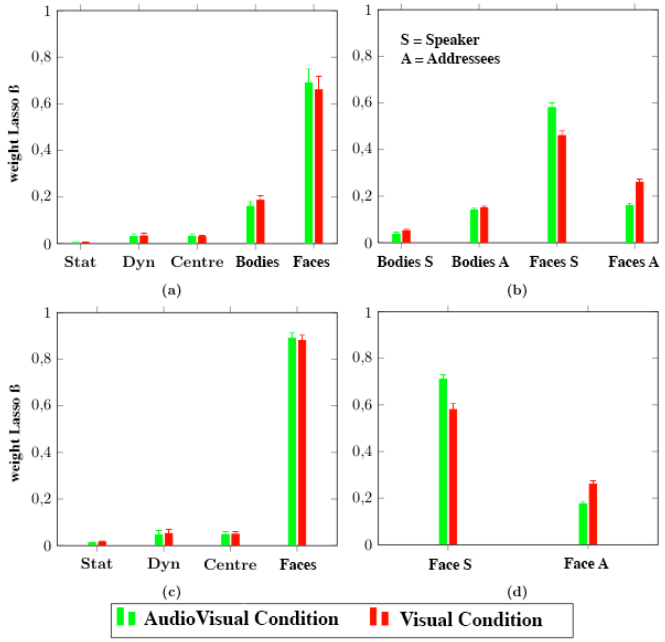

Fig. 2. **Training** with eye positions recorded in Visual and AudioVisual conditions. **(a)** Mean values of Lasso weights for Static and Dynamic Saliency, Centre Bias, Bodies and Faces masks. **(b)** Contributions of the speakers (S) and addressees (A) to the Bodies and Faces features in panel (a). **(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.

being processed.

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 addresses' faces (Bonferoni 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 (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 $p < .001$).

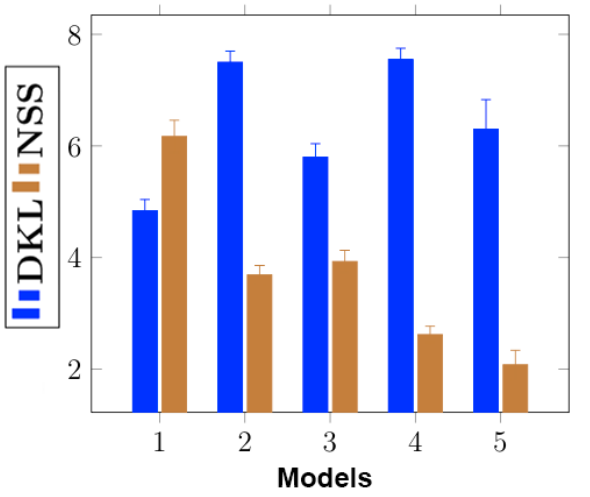

Fig. 3. **Evaluation** - Divergence of Kullback-Leibler (DKL) and Normalized Scanpath Saliency (NSS) for the different models described Section 3.3. For the models 1, 2 and 3, the feature weights have been estimated with the Lasso algorithm applied to the eye positions recorded in the AudioVisual condition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

[1] Patrick Le Callet and Ernst Niebur, "Visual Attention and Applications in Multimedia Technologies," in *IEEE Institute of Electrical and Electronics Engineers*, 2013, pp. 2058–2067.

[2] Ali Borji and Laurent Itti, "State-of-the-art in Visual Attention Modeling," *IEEE Transactions on Patterns*

Analysis and Machine Intelligence, vol. 35, no. 1, pp. 185–207, 2013.

[3] Anne M. Treisman and Garry Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, pp. 97–136, 1980.

[4] Elina Birmingham, Walter F Bischof, and Alan Kingstone, "Saliency does not account for fixations to eyes within social scenes," *Vision Research*, vol. 49, pp. 2992–3000, 2009.

[5] Antoine Coutrot and Nathalie Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," *Journal of Vision*, vol. 14, no. 8, pp. 1–17, 2014.

[6] Moran Cerf, Jonathan Harel, Wolfgang Einhäuser, and Christof Koch, "Predicting human gaze using low-level saliency combined with face detection," *Advances in Neural Information Processing Systems*, vol. 20, 2008.

[7] Sophie Marat, Anis Rahman, Denis Pellerin, Nathalie Guyader, and Dominique Houzet, "Improving Visual Saliency by Adding 'Face Feature Map' and 'Center Bias'," *Cognitive Computation*, vol. 5, no. 1, pp. 63–75, 2013.

[8] Antoine Coutrot, Nathalie Guyader, Gelu Ionescu, and Alice Caplier, "Influence of soundtrack on eye movements during video exploration," *Journal of Eye Movement Research*, vol. 5, no. 4, pp. 1–10, 2012.

[9] Antoine Coutrot and Nathalie Guyader, "Toward the Introduction of Auditory Information in Dynamic Visual Attention Models," in *14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services*, Paris, France, 2013.

[10] Antoine Coutrot and Nathalie Guyader, "An Audiovisual Attention Model for Natural Conversation Scenes," in *IEEE International Conference on Image Processing*, Paris, France, 2014.

[11] Sophie Marat, Tien Ho-Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué, "Modelling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 231–243, 2009.

[12] Josselin Gautier and Olivier Le Meur, "A Time-Dependent Saliency Model Combining Center and Depth Biases for 2D and 3D Viewing Conditions," *Cognitive Computation*, vol. 4, pp. 1–16, 2012.

[13] Po-He Tseng, Ran Carmi, Ian G M Cameron, Douglas P. Munoz, and Laurent Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *Journal of Vision*, vol. 9, no. 7, pp. 1–16, 2009.

[14] Pascal Bertolino, "Sensarea: an Authoring Tool to Create Accurate Clickable Videos," in *10th Workshop on Content-Based Multimedia Indexing*, Annecy, France, 2012.

[15] David McNeill, "So you think gestures are nonverbal?," *Psychological Review*, vol. 92, no. 3, pp. 350–371, 1985.

[16] C Chamaret, J C Chevet, and O Le Meur, "Spatio-Temporal Combination of Saliency Maps and Eye-Tracking Assessment of Different Strategies," in *IEEE International Conference on Image Processing*, Hong Kong, 2010, pp. 1077–1080.

[17] Qi Zhao and Christof Koch, "Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost," *Journal of Vision*, vol. 12, no. 6, pp. 1–15, 2012.

[18] Robert Tibshirnani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[19] Gideon E Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[20] Karl Sjöstrand, Line Harder Clemmensen, Rasmus Larsen, and Bjarne Ersbøll, "SpaSM: A Matlab Toolbox for Sparse Statistical Modeling," *Journal of Statistical Software*, pp. 1–24, 2012.

[21] Benjamin T Vincent, Roland J. Baddeley, Alessia Correani, Tom Troscianko, and Ute Leonards, "Do we look at lights? Using mixture modelling to distinguish between low - and high - level factors in natural image viewing," *Visual Cognition*, vol. 17, no. 6-7, pp. 856–879, 2009.

[22] Tien Ho-Phuoc, Nathalie Guyader, and Anne Guérin-Dugué, "A Functional and Statistical Bottom-Up Saliency Model to Reveal the Relative Contributions of Low-Level Visual Guiding Factors," *Cognitive Computation*, vol. 2, pp. 344–359, 2010.

[23] I McCowan and *al.*, "The AMI Meeting Corpus," in *International Conference on Methods and Techniques in Behavioral Research*, Wageningen, The Netherlands, 2005.

[24] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, pp. 2397–2416, 2005.

[25] Olivier Le Meur and Thierry Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior Research Methods*, vol. 45, no. 1, pp. 251–266, 2013.