An audiovisual attention model for natural conversation scenes

Antoine Coutrot, Nathalie Guyader

IEEE International Conference on Image Processing (ICIP) 2014 - Oct 27-30, Paris, France

AN AUDIOVISUAL ATTENTION MODEL FOR NATURAL CONVERSATION SCENES

Antoine Coutrot & Nathalie Guyader

Gipsa-lab, CNRS & Grenoble-Alpes University, France
firstname.lastname@gipsa-lab.fr

ABSTRACT

Classical visual attention models neither consider social cues, such as faces, nor auditory cues, such as speech. However, faces are known to capture visual attention more than any other visual features, and recent studies showed that speech turn-taking affects the gaze of non-involved viewers. In this paper, we propose an audiovisual saliency model able to predict the eye movements of observers viewing other people having a conversation. Thanks to a speaker diarization algorithm, our audiovisual saliency model increases the saliency of the speakers compared to the addressees. We evaluated our model with eye-tracking data, and found that it significantly outperforms visual attention models using an equal and constant saliency value for all faces.

Index Terms— audiovisual saliency model, eye movements, speaker diarization, speech, social gaze

1. INTRODUCTION

Imagine yourself at a poster session in a noisy conference room, witnessing a discussion between a poster presenter and a couple of colleagues. You will probably look at the poster presenter to focus your auditory attention, but also at your colleagues to check if someone is about to take the floor, and show that you are listening. This example illustrates that in complex natural scenes, your attention is driven by many features, including dynamic and social ones [1]. Yet, to predict the most attractive areas in a scene, models of visual attention mostly rely on low-level visual features like luminance, contrast, orientation and motion, and do not take account of these crucial features (see [2] for a taxonomy of 65 models). Therefore, they are unlikely to generalize to social contexts [3, 4]. To address this issue and consider faces as particular visual features, few visual saliency models combining "faces" with classical low-level features have been developed. These models significantly outperform classical visual saliency models to predict observers' eye movements [5, 6]. Despite these significant efforts in attention modeling, an important factor has been left aside: auditory information. So far, visual saliency models do not consider sound, even when dealing with dynamic scenes. When running eye tracking experiments with videos, authors never mention soundtracks or explicitly remove them, making participants look at mute movies, which is of course not a natural situation. Indeed, we live in a multimodal world and our attention is constantly guided by the fusion between auditory and visual information. When looking at natural dynamic scenes, sound has been shown to significantly impact on eye movements [7]. In particular, few very recent papers investigated how speech turn-taking affects the gaze of non-involved viewer of natural conversations [8, 9]. These eye-tracking studies presented conversations to participants with speech soundtracks, or without any sound. They both showed that sound changes the timing of looks. With sound, speakers are fixated more often and more quickly after they start speaking, leading to a greater attentional synchrony. In a recent study, we quantified the relative contributions of faces and of other visual features to explain the eye movements recorded when viewing conversations [10]. We found that non-involved observers looked more at faces than at any other visual features, and that talking faces were more gazed at than mute faces (talking and mute faces were manually labelled). These experimental results stress the need to take into account speech turn-taking in saliency models.

In this paper, we present an audiovisual saliency model able to predict the eye movements of observers viewing natural conversations. Contrary to previous dynamic saliency models giving an equal and constant saliency value to every detected faces [5, 6], we propose to modulate faces' saliency by increasing the saliency of speakers. To this end, the audiovisual saliency model we present here includes a quite simple audiovisual speaker diarization algorithm, spotting which conversation partner is speaking and which is not. To evaluate our model, we compared the predicted saliency values to the eye positions of naive observers viewing conversations embedded in complex natural scenes.

2. AUDIOVISUAL SALIENCY MODEL

In this section, we describe the general framework of the proposed audiovisual saliency model (Figure 1).

2.1. Model Layout

Each video frame is first decomposed into three classical visual maps that have been shown to play a role in eye move-

ment guidance [10].

- Low-Level Saliency The low-level saliency map of each video frame was computed using a biologically-inspired saliency model [11]. This model, based on luminance information, decomposes video frames into static and dynamic features. Static features rely on high spatial frequencies to emphasize areas with high contrast. Dynamic feature extraction rely first on a camera motion compensation to only detect motion relative to the background. Second, low spatial frequencies from two consecutive frames are used to extract motion, under the assumption of luminance constancy. Finally, static and dynamic features are merged into a low-level saliency map Φ_{LS} (see Figure 1).

- Faces The face of each conversation partner was marked by an oval mask, giving a Face map Φ_F. Since faces were moving, the coordinates of each mask were dynamically defined for each video. We used Sensarea, an in-house authoring tool that automatically or semi-automatically performs spatio-temporal segmentation of video objects [12].

- Center Bias Eye-tracking studies reported that subjects tend to gaze more at the center of the image. Several hypotheses have been proposed to explain this bias. Some are stimuli-related, like the photographer bias (one often places regions of interest at the center of the picture), others are inherent to the oculomotor system (motor bias) or to the observers' viewing strategy [13]. The center bias is modeled by a time-independent bi-dimensional Gaussian function centered at the screen center, Φ_{CB}.

These maps Φ_k were computed for each frame f and linearly combined into a master audiovisual saliency map M:

$$M(f) = \sum_{k \in \{LS,F,CB\}} \alpha_k(f)\Phi_k(x,y,f), \quad \text{with} \sum_k \alpha_k(f) = 1.$$

The weights α_k were adjusted with the Expectation - Maximization (EM) algorithm, a statistical method using observations to estimate the relative importance of each feature in order to maximize the global likelihood of the mixture model [14]. The observations were eye positions recorded during an eye-tracking experiment further described in [10]. The EM algorithm returned a value for each feature and each frame. For each feature map k, we averaged the $\alpha_k(f)$ over all frames of all videos, and used these constants to weight the feature maps. The results were $\alpha_{CB} = 0.05$, $\alpha_{LS} = 0.2$ and α_F=0.75. To separate the contribution of talking and mute faces, we averaged $\alpha_F(f)$ over the manually labelled talking and mute time periods. We found that α_F can be broke into $\alpha_{MF} = 0.25$ for mute faces and $\alpha_{TF} = 0.50$ for talking faces. This illustrates that, if all faces attract attention, speakers are more looked at than mute conversation partners. Let's consider a scene where N persons are present, one talking and $N-1$ mute. In the following, we give to the speaker face a α_{TF} weight and to each mute face a $\alpha_{MF}/(N-1)$ weight. If all faces are in the same state (*i.e.* speaking or mute), we give to each face the same α_F/N weight. The novelty of the proposed audiovisual saliency model is the use of these coefficients to differently weight faces across frames, depending on whether they are currently speaking or not. To automatically distinguish talking from mute faces, we propose a speaker diarization algorithm.

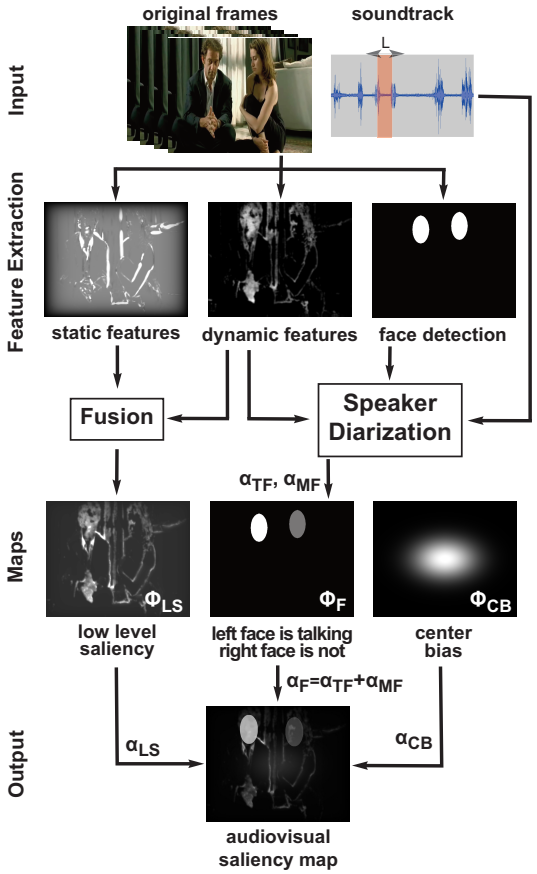

Fig. 1: Block diagram of the proposed audiovisual saliency model. The low-level saliency map (Φ_{LS}), the talking face map (Φ_{TF}), the mute face map (Φ_{MF}) and the center bias map (Φ_{CB}) are respectively weighted with $\alpha_{LS}, \alpha_{TF}, \alpha_{MF}$ and α_{CB} (adjusted to fit experimental data), and merged into the final audiovisual saliency map.

2.2. Speaker Diarization Algorithm

Speaker diarization has emerged as an increasingly important field of speech recognition and relates to the problem of determining who spoke when [15]. In this paper, we propose a simple algorithm that does not require training. Our algorithm is based on two assumptions: each speech turn-taking is separated by a silence, and speakers move more than other conversation partners [16, 17]. The speaker diarization algorithm relies on different stages described below: voice activity detection (VAD), audio speaker clustering (BIC framework) and motion detection to attribute each audio cluster to the right speaker.

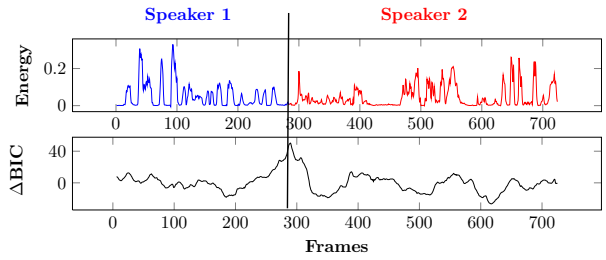

Fig. 2: Energy of a soundtrack (top) and its corresponding ΔBIC (bottom). The ΔBIC maximum matches with the turn-taking.

2.2.1. VAD & Speaker Clustering

We first extracted and appended the speech segments from the soundtrack, discarding silence or noise. To do so, we used a energy-based algorithm available online [18]. To decide whether two successive speech segments are delivered by the same speaker or if a turn-taking occurred, we used audio and visual features.

Audio - To describe the speech segments, we extracted the first 26 Mel Frequency Cepstral Coefficients (MFCCs) on 10 ms intervals. Denote $\mathbf{z} = \{z_i \in \mathbf{R}^d, i = 1, ..., N\}$ as the sequence of MFCCs describing an audio segment of length N (duration N×10 ms) . Then, we used a method based on the Bayesian Information Criterion (BIC), that has proven effective for audio classification [19, 20]. On each sample s of the speech signal was centered a symmetrical and fixed-size analysis window (L = 200 ms). We tested the hypothesis that a change occurred at sample s. We compared a model assuming that the samples contained in the window $\mathbf{w} = \{z_i \in \mathbf{R}^d, i = s - \frac{L}{2}, ..., s + \frac{L}{2}\}$ were drawn from an independent multivariate Gaussian process: $\mathbf{w} \sim N(\mu_w, \Sigma_w)$ with μ_w and Σ_w the mean and standard deviation of $\mathbf{w}$; versus a model with two Gaussians: one for the first half of the window $\mathbf{x} = \{z_{s-L/2}, ..., z_s\} \sim N(\mu_x, \Sigma_x)$, and one for the other half $\mathbf{y} = \{z_s, ..., z_{s+L/2}\} \sim N(\mu_y, \Sigma_y)$. To decide which model is the best fit, we used the Bayesian Information Criterion (BIC), a likelihood criterion penalized by the model complexity. The analysis window was slid through the successive speech segments (orange rectangle at the top of Figure 1), and the difference between the two BIC was computed at each sample:

$$\Delta BIC(\mathbf{x}, \mathbf{y}) = L \log|\Sigma_w| - \frac{L}{2} \log|\Sigma_x| - \frac{L}{2} \log|\Sigma_y| - \lambda P$$

with the penalty $P = \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log L$, the dimension of the space d (here d=26) and the penalty weight $\lambda = 1$. A ΔBIC value was computed for each sample, and a local maximum was extracted from each speech segment. The higher it was, the more likely a speaker transition occurred (Figure 2).

Visual - We used the hypothesis that speakers move more than other conversation partners. For each frame, we took the dynamic feature map defined in [11] (see Figure 1). For each speaker, we summed the pixels of the dynamic map contained in their corresponding face masks. Thus, we had the frame-by-frame evolution of the "activity" of each conversation partner. Then, we standardized these values and compared their mean over each speech sequence. For each conversation partner, the higher the modulus of the difference between two successive speech segments was , the more likely this person began or stopped moving.

Finally, we standardized and added the audio and visual "transition probabilities" for each speech segment. If this combination was higher than an empirical threshold T (we took T = 1), the speech segments were said to be delivered by different speakers. Else, the speech segments were merged.

Table 1: Length, speech time proportion, number of turn switches and number of faces for the 14 videos used to evaluate the audiovisual saliency model.

	Length (s)	Speech (%)	Switches	Faces
vid1	11.4	92	3	2
vid2	19.3	96	2	2
vid3	24.2	87	0	2
vid4	20.5	47	2	2
vid5	11.6	72	0	2
vid6	19.3	38	7	2
vid7	22.8	42	2	2
vid8	20.8	89	6	2
vid9	29.3	72	6	2
vid10	15.1	74	0	2
vid11	12.3	95	0	2
vid12	22.9	69	7	4
vid13	18.8	95	2	3
vid14	13.4	65	2	2

2.2.2. Cluster Labelling

To attribute each speech cluster to the right speaker, we used the same dynamic low-level saliency maps as described above. We summed the pixel values contained in each mask to get the activity of the corresponding conversation partner. We then averaged these activities over each speech cluster. The corresponding speech sequence was attributed to the most "active" conversation partner.

3. RESULTS

3.1. Eye-Tracking Experiment

Comparing the predicted saliency values to the eye positions of naive participants is a classical way to evaluate visual attention model. 18 participants took part in the experiment (all French native speakers). The stimuli consisted of 14 one-shot conversation scenes extracted from French

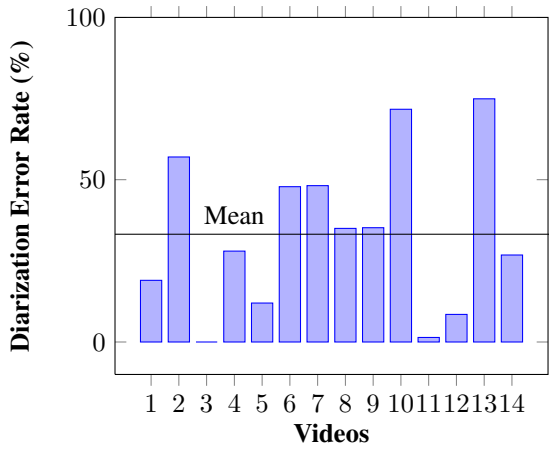

Fig. 3: Diarization error rate for the 14 videos used to evaluate the audiovisual saliency model. The black horizontal line corresponds to the mean (33.2%).

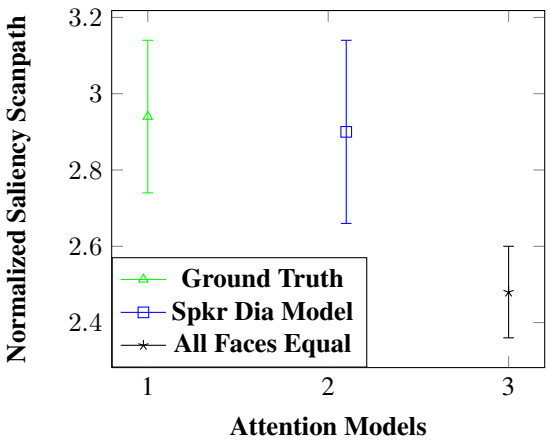

Fig. 4: Normalized Scanpath Saliency values for (1 & 2) the audiovisual saliency model presented in this paper (in green, the distinction Talking Face / Mute Face is made on the ground truth basis, in blue according to the speaker diarization algorithm presented above); (3) the same model, without differentiating talking and mute faces.

"Hollywood-like" movies, with monophonic soundtracks. The experiment is further described in [10, 21]. The stimuli and the eye-tracking data are available at *http://www.gipsa-lab.fr/∼antoine.coutrot/DataSet.html*.

3.2. Evaluation of the Speaker Diarization Algorithm

We evaluated the proposed speaker diarization algorithm on the 14 videos from the experiment (7300 frames), see Table 1 for details. The Voice Activity Detection stage gave 500 misclassified frames (silence instead of speech or speech instead of silence), *i.e.* 6.9% of the total amount. The diarization performance was measured on the speech segments using the diarization error rate (DER). To compute the DER, we performed a frame-by-frame comparison between the ground truth and the predicted speech clusters (model). We obtained an average DER of 33.2%, see Figure 3 for details. This result is comparable to other state-of-the-art algorithms [17, 22].

3.3. Evaluation of the Attention Models

To evaluate our model, we used the Normalized Scanpath Saliency (NSS) [23]. It acts like a z-score computed by comparing a saliency map M_m from an attention model to the eye position density maps M_p of participants. An eye position density map was computed for each frame, by adding a 15 pixels wide patch to each of the 18 eye positions.

$$NSS = \frac{M_m.M_p - mean(M_m)}{std(M_m)}$$

The higher the values of NSS are the more the salient regions are attended. We compared the NSS of our model to (1) the NSS of the same model, but with a ground truth manual speaker diarization and (2) the NSS of the same model, but without differentiating talking and mute faces, *i.e.* giving the same weight α_F/N to all faces (see Figure 4). The NSS of the models making the distinction between talking and mute faces are significantly higher than the NSS of the model giving the same and constant weight to all faces (paired t-test, $p < 0.001$).

4. CONCLUSIONS

Previous visual attention models neither consider social cues, such as faces, nor auditory cues, such as speech. Thus, they dramatically fail in many experimental contexts. The few existing saliency models featuring face detectors do not use the auditory information, which has yet proven to significantly influence gaze. In particular, when viewing natural conversation scenes, it has been shown that observers look more at the speakers than at the addressees. In this paper, we presented an audiovisual saliency model featuring a speaker diarization algorithm able to temporally distinguish talking conversation partners from silent ones. We proposed a very simple algorithm that did not require training. We obtained an average DER of 33.2%. A direct comparison of our results with previous ones is hazardous given the high sensitivity of the DER to video features [22]. However, this performance is sufficient to significantly improve our saliency model, compared to a model giving equal and constant weights to all faces. In future work, we hope to evaluate our model with standard audiovisual corpora (*e.g.* The AMI Meeting Corpus), so it could be compared with other algorithms in an unbiased way.

5. REFERENCES

[1] Tom Foulsham, Joey T Cheng, Jessica L Tracy, Joseph Henrich, and Alan Kingstone, "Gaze allocation in a dy-

namic situation: Effects of social status and speaking," *Cognition*, vol. 117, no. 3, pp. 319–331, 2010.

[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, 2012.

[3] 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.

[4] Benjamin W. Tatler, Mary M Hayhoe, Michael F Land, and Dana H Ballard, "Eye guidance in natural vision: Reinterpreting salience," *Journal of Vision*, vol. 11, no. 5, pp. 1–23, 2011.

[5] 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.

[6] 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.

[7] 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.

[8] Lotta Hirvenkari, Johanna Ruusuvori, Veli-Matti Saarinen, Maari Kivioja, Anssi Peräkylä, and Riitta Hari, "Influence of Turn-Taking in a Two-Person Conversation on the Gaze of a Viewer," *PLoS ONE*, vol. 8, no. 8, pp. 1–6, 2013.

[9] Tom Foulsham and Lucy Anne Sanderson, "Look who's talking? Sound changes gaze behaviour in a dynamic social scene," *Visual Cognition*, vol. 21, no. 7, pp. 922–944, 2013.

[10] Antoine Coutrot and Nathalie Guyader, "How Saliency, Faces and Sound influence gaze in Dynamic Social Scenes," *Journal of Vision*, in press.

[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] Pascal Bertolino, "Sensarea: an Authoring Tool to Create Accurate Clickable Videos," in *10th Workshop on Content-Based Multimedia Indexing*, Annecy, France, 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] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[15] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

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

[17] Binyam Gebrekidan Gebre, Peter Wittenburg, and Tom Heskes, "The Gesturer Is the Speaker," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, Vancouver, Canada, 2013, pp. 1–5.

[18] Theodoros Giannakopoulos, "Silence removal in speech signals," http://www.mathworks.com, 2010.

[19] Himanshu Vajaria, Sudeep Sarkar, and Rangachar Kasturi, "Exploring Co-occurence between Speech and Body Movement for Audio-guided Video Localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1608–1617, 2008.

[20] Shih-Sian Cheng, Hsin-Min Wang, and Hsin-Chia Fu, "BIC-based Speaker Segmentation Using Divide-and-Conquer Strategies with Application to Speaker Diarization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 1, pp. 141–157, 2010.

[21] 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 (WIAMIS 2013)*, Paris, France, 2013, pp. 1–4.

[22] Nikki Mirghafori and Chuck Wooters, "Nuts and Flakes: a Study of Data Characteristics in Speaker Diarization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, Toulouse, France, 2006, pp. 1017–1020.

[23] 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.