LOOK WHO'S TALKING: DETECTING THE DOMINANT SPEAKER IN A CLUTTERED SCENARIO

Eleonora D'Arca, Neil M. Robertson and James R. Hopgood

Joint Research Institute for Signal and Image Processing,
Heriot-Watt University & University of Edinburgh, UK
visionlab.eps.hw.ac.uk

ABSTRACT

In this work we propose a novel method to automatically detect and localise the dominant speaker in an enclosed scenario by means of audio and video cues. The underpinning idea is that gesturing means speaking, so observing motions means observing an audio signal. To the best of our knowledge state-of-the-art algorithms are focussed on stationary motion scenarios and close-up scenes where only one audio source exists, whereas we enlarge the extent of the method to larger field of views and cluttered scenarios including multiple non-stationary moving speakers. In such contexts, moving objects which are not correlated to the dominant audio may exist and their motion may incorrectly drive the audio-video (AV) correlation estimation. This suggests extra localisation data may be fused at decision level to avoid detecting false positives. In this work, we learn Mel-frequency cepstral coefficients (MFCC) coefficients and correlate them to the optical flow. We also exploit the audio and video signals to estimate the position of the actual speaker, narrowing down the visual space of search, hence reducing the probability of incurring in a wrong voice-to-pixel region association. We compare our work with a state-of-the-art existing algorithm and show on real datasets a 36% precision improvement in localising a moving dominant speaker through occlusions and speech interferences.

Index Terms— Audio-Video Correlation, Speaker Tracking, Speaker Recognition, Multimodal tracking, AV Tracking

1. INTRODUCTION AND RELATED WORK

Tracking a speaker in an enclosed scenario has become an increasingly interesting topic over the last twenty years. The establishment of the digital era has created a number of applications which combine or not the usage of voices and videos to different aims, i.e. detecting a threatening behaviour in a public place or understanding how people interact during a meeting. Such applications mostly involve analysing large cluttered areas where no constraints on people movements exist. In such scenarios, designing novel joint audio-video (AV) systems may require a lower complexity than using state-of-the-art stand alone systems, as audio waves reverberations and visual clutter are very difficult to predict and this severely compromises the estimation. Existing AV speaker tracking systems [1–6] treat the two signals as they were independent processes. Conversely, little attention has been given to the exploitation of underlied relations between audio and video to detect and localise a moving speaker over time in large indoor environments, whereas instead event anomalies detection literature is widely based on inferring AV signal correlation [7–10]. This work proposes to borrow the consolidated anomaly detection techniques [9] to be novelly applied together with AV tracking techniques [11] in far distance cluttered scenes (e.g. *cocktail party* scenarios) where the dominant speaker must be detected and tracked. Nevertheless, the cited event detection techniques are normally applied in close-up scenes in which speaking sources are mostly stationary. In light of this, the presented work contributions are: *a*) AV correlation techniques are extended to a larger range of data i.e. complex scenarios; *b*) in complex scenarios we enhance on a state-of-the-art algorithm to make it fully automatic, by adding in the audio localisation information, so that distracting correlated AV moving objects do not compromise the dominant speaker estimation.

2. ALGORITHM DESCRIPTION

The assumptions behind this work comes from the observation that very often locating gesturing in a conversational scene gives an indication of where speaking activity spatially originates. In fact, it has been proved in several ways that the "gesturer is the speaker" [12], meaning gesturing is almost always (80%-90% of times [13, 14]) associated to speaking activity [13–17] and that gesticulation and speaking activity are the fundamental cues to define dominance [18–20]. For this reason, we attempt to recognise and exploit a somewhat inherent correlation between audio and video signal as done previously in [9]. In particular, we compute canonical correlation analysis (CCA) between audio features i.e. Mel-

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/J015180/1 and the MOD University Defence Research Collaboration in Signal Processing.

(a) (b) (c) (d)

Fig. 1: **Baseline Method ([9]) application example.** In (a) one of the analysed video frames from the baseline method dataset is shown. (b) presents the QuickShift spatial segmentation which is obtained by overlapping colour, motion velocity and acceleration features of frame (a), (c) shows its K-means spatio-temporal segmentation. At last, (d) represents the final AV correlation result as an overlaid heatmap, which represents the probability of the ball to be the moving object which is the most correlated to the the audio signal, hence which is generating the sound. This results has no audio localisation information. We seek to improve on the state-of-the-art by including tracking information at data decision level.

frequency cepstral coefficients (MFCC) and Δ -MFCC and video features (optical flow velocity and acceleration) associated with the scene. Hence, we infer a speaker likelihood in every video frame, by selecting the frame segment where most likely the sound has originated from, to eventually conclude the actual/dominant speaker is within that pixel region. In [9], they improve on this procedure by manually selecting AV foreground and background points in the frame. They show this step increase results accuracy in complex scenarios where distracting, occluding and correlated motion may appear. We further enhance this approach to make it fully automatic, by substituting the user with the audio localisation information i.e. the position calculated by evaluating the time delays of arrival of the audio signals at the microphones.

Video Features Extraction. To extract video features we first compute the forward and backward dense optical flow of each image frame. Then, we calculate velocity and acceleration of two adjacent frames motion. If $U^+(\mathbf{p},t)$ represents the optical flow (u,v) at pixel position $\mathbf{p}=(i,j)$, at time t, calculated between frames F_t and F_{t+1} and analogously $U^-(\mathbf{p},t)$ the flow vector computed over time between F_t and F_{t-1}, then the velocity and acceleration vectors are defined as:

$$vel = U^+(\mathbf{p},t), \quad acl = U^+(\mathbf{p},t) - (-U^-(\mathbf{p},t)). \qquad (1)$$

Hence, we combine the RGB colour, velocity and acceleration of each pixel in a frame $\mathbf{p}$ into a single feature vector: $v_{ij} = (\mathbf{p}, col, vel, acl)$. Thus, we spatially segment every frame using the QuickShift algorithm. Furthermore, we compute across frames a K-means spatio-temporal segmentation. In consequence of that, when the processing ends, every pixel in a frame can be ascribed to the spatio-temporal centre of mass of the k-th segment found by K-means. The k final segments $S_k(k=1,...,K)$ are described by the averaged normalised velocity and acceleration of the pixels they enclose, in addition to their mean RGB colour : $v_{ij} = (\mu_{\mathbf{p}}, \mu_{col}, \mu_{vel}, \mu_{acl})$. Finally, the m_1 top segments for velocities and the m_2 top for acceleration are chosen to compose the final video features vector $\mathbf{v}$. In practice, $\mathbf{v}$ is a $m \times t$ matrix whose columns correspond to frames.

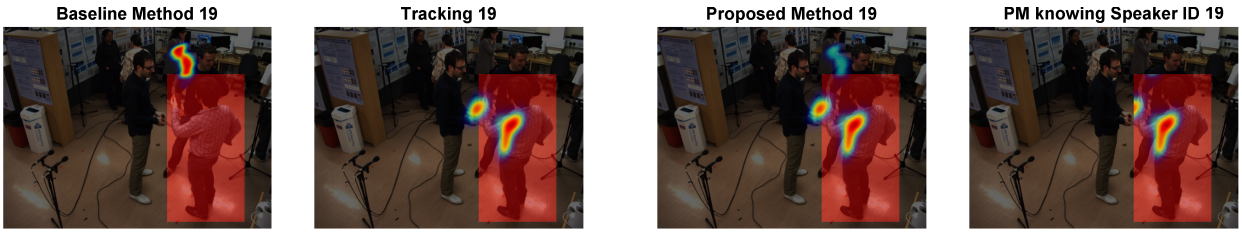
Audio Features Extraction. Audio feature vectors are represented by the first $n/2$ MFCC coefficients [21] (audio signal velocity) and their $n/2$ derivatives (audio signal acceleration). The audio feature vector $\mathbf{a}$ is a $n \times t$ matrix whose columns correspond to frames. Note that the audio signal must be windowed and processed accordingly to the video frame rate in order for the CCA to be based on the same number of observations.

2.1. Audio Video Correlation and Tracking Data Fusion

Audio Video Correlation. We use canonical correlation analysis (CCA) [22] to seek audio and video feature vectors correlation, hypothesising a hidden correspondence between the image motion velocity together with the audio MFCC, and the motion acceleration with the MFCC derivatives (Δ -MFCC) exists. Canonical correlation analysis allows to find a common coordinate system where $\mathbf{a}$ and $\mathbf{v}$ can be projected, and also to immediately know their maximised correlation. This ensure the retrieved video segment to be the one that maximise the correlation between audio and video data, hence to be associated with the dominant audio source. Specifically, the CCA problem between two random variables has the closed form solution:

$$\begin{cases} C_{vv}^{-1} C_{va} C_{aa}^{-1} C_{av} \mathbf{w}_v = \lambda^2 \mathbf{w}_v \\ C_{aa}^{-1} C_{av} C_{vv}^{-1} C_{va} \mathbf{w}_a = \lambda^2 \mathbf{w}_a, \end{cases}$$

where $\hat{C} = \begin{pmatrix} C_{vv} & C_{va} \\ C_{av} & C_{aa} \end{pmatrix}$ represents the total covariance matrix and $\mathbf{w}_v$ and $\mathbf{w}_a$ are the canonical basis of $\mathbf{v}$ and $\mathbf{a}$. The largest CCA eigenvectors w_{v^1} and w_{a^1}, which correspond to the largest eigenvalue λ_1^2 are the ones which give the larger contribution to the maximum audio and video correlation, hence they maximises the canonical variates $v_1^{'} = w_{v^1}^T \mathbf{v}$ and $a_1^{'} = w_{a^1}^T \mathbf{a}$. If we assume that only a single dominant audio source exists, the first of these eigenvectors w_{v^1} is chosen and the corresponding frame segments $\overline{\mathbf{S}}$ are said to be the ones where the sound is originating. Only the normalised elements of w_{v^1} largest then a predefined threshold are selected,



(a) (b) (c) (d)

Fig. 2: ***'Cocktail Party'* Frame 19 Results.** This figure shows the *Second Speaker* talking while the other two people are listening without moving. In (a) the results of the baseline method [9] are given whereas (b) shows the result for the speaker localisation (SL) algorithm. (c) presents the output of the proposed method. Finally, (d) shows what is the results when the information about the speaker identity is given. Ground truth is shown in red.

thus those segments are identified by a binary confidence map smoothed over space and time by a Gaussian kernel. Figure 1 shows the reimplementation of the baseline method [9] as a walk-through example.

Correlation and Sound Localisation Fusion. Details of the triangulation, by mean of an extended Kalman filter (EKF) can be found in [11, 23]. Briefly, we feed the time difference of arrival (TDOA) computed for each microphones pair to an EKF over time to iteratively calculate the dominant speaker position $\mathbf{x}_{SL} = (x, y)$ on the ground plane.

The integration between audio speaker localisation (SL) data (i.e. speaker trajectory) and the CCA result is carried out at confidence map level. In other words, we project the audio source trajectory $\mathbf{x}_{SL}(t)$ onto the pixel domain. Thus, at every time step we associate the trajectory points to the k-th segmented region to which they belong i.e. $(x,y)_{SL} \mapsto (i,j)_{SL} \in S_{k_{SL}}$. Therefore, we set $S_{k_{SL}}$ as a further confidence map (other than the ones already given by the first base eigenvector coefficients) and define a smoothing Gaussian kernel as said above to finally obtain a heatmap to be overlaid on the image. Ultimately, we treat the resulting kernel as an extra first base eigenvector coefficient adding up its contribution to the CCA result, according to an averaged sum decision rule.

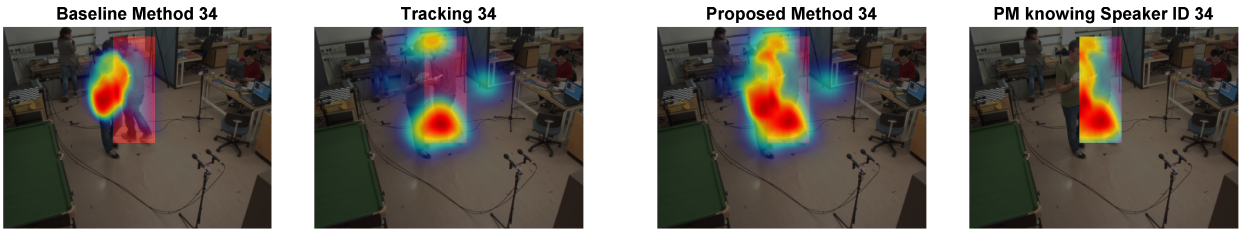
3. EXPERIMENTATION AND RESULTS

We now present comprehensive results on real data. No comparison is made for them since, as far as we are aware of, no other works exist with same experiment setup. Neither the authors of [9] used more microphones for localisation purposes. We analyse a real indoor room where people can freely move. In particular, audio and video data are gathered in a typical open office room, whose size is 111.44 m^2, where the area considered of interest is 12 m^2. Also we make no attempt to reduce normal background noise (desk fans, footsteps, talking etc.). A significant reverberation time ($T_{60} \approx 0.5$ s) is measured. Ground-truth data is hand labelled considering feet position to 10 cm of accuracy on a ground plane common to the cameras and the microphones. Synchrony of data is obtained by processing audio and video signals accordingly to the cameras frame rate ≈ 7.5 Hz. Only 4 pairs of directional microphones are used. The EKF filter is initialised using a video detected position of the targets and static matrices Q and R [23], whose values is chosen on the basis of an optimisation step. Audio is sampled at 44.1 KHz, and framed with 50 % overlap. 10 MFCC coefficients are computed, as well as their first 10 derivatives (Δ - MFCC). The QuickShift algorithm parameters used are $\gamma = 0.25$,$\sigma = 1$ and $\tau = 15$ i.e. the same as in [9]. The number of clusters in the K-means algorithm is set to be 30. And the smoothing Gaussian kernel has a variance of $\sigma = 5$.

Experiment *'Cocktail Party'* (Fig. 2) is a recording of several people having a conversation in groups and some passer-by. Speakers are at least 50 cm far from the microphones. They stand still and move around. The ground-truth consists of the speaker on the left foreground and the speaker on the right foreground whereas a third person in the foreground (blue jumper) is just paying attention to the conversation while producing some distracting fine motion by slightly moving his body on a side. Note that another group of speaking people is in the background. This results in challenging speech interferences and occlusions.

Experiment *'Occlusion'* (Fig.3) shows two people who look alike walking while having a conversation. They meet along a diagonal where they keep on walking past each other causing an occlusion in the resulting image. Also two moving people external to the main scene are in the room.

Results. In first instance we give a qualitative description of the preliminary results we have obtained referring to their corresponding figures. In particular, Figure 2a shows frame 19 baseline method results. The actual speaker is about to raise his hand while the listener has been moving his body resulting in false positive detections. This can be only mitigated by the SL corresponding segment (2b), so that the fusion results, despite pointing out the correct speaker, still presents false detection trails corresponding to the other people movements (2c). When we can also recognise the speaker identity, as we did previously in [6], we can actually further filter out these trails as shown in Fig.2d. Figure 3a shows results of the baseline method applied to the second dataset

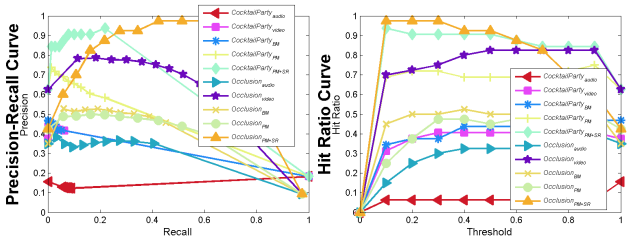


(a) (b) (c) (d)

Fig. 3: ***'Occlusion'* Frame 34 Results.** In (a) the results of the baseline method [9] are given whereas (b) shows the result for the SL results projected onto the image plane. In (c) the output of the proposed method is given, whereas (d) presents results when the information about the speaker identity is given. Ground truth is shown in red.

at the moment of occlusion. The segments corresponding to the tracked position of the actual speaker are given in Fig. 3b whereas Fig. 3c shows the results of the proposed method. Again, knowing the information about the speaker identity the results are ascribed at the current speaker (Fig. 3d).

We evaluate our method performance against the baseline method ([9]) by using a precision-recall measure. In particular, we first manually define the moving pixel ground truth by selecting those region of the video which correlated with the dominant speaker's voice. In practice, as this method is meant to be used for tracking purposes this is always represented by a bounding box including the speaker's body pixel. This region is denoted as R_c ,whereas R_d is the pixel region detected by the method. Hence, the two curves are defined as: $Pr = {}^{R_c \cap R_d}/_{R_d}$ and $Rec = {}^{R_c \cap R_d}/_{R_c}$. The precision-recall curve is given by letting vary a threshold between zero and one for every frame, thus we present the average curve for all the video frames. Figure (4a) shows that the proposed method (PM) precision is higher than the one of audio and video only and than the baseline method (BM) over the entire range of recall, although when the recall value increases both curves drops dramatically. However, this is largely expected as the ground truth size is larger if compared to the recovered segments size, which decreases the accuracy of the methods by definition. Nevertheless, the size of the segments cannot be increased, as clutter will take over the segmentation phase and foreground region would be blended in to the background. On the other hand, for detection for tracking purposes we cannot restrict the size of the ground truth regions to just the joint of a person. Note that also the results of the proposed method using the information about the speaker identity (PM+SR) is shown. This is because for speaker diarisation purposes, the pixel based precision-recall metric does not make much sense. Hence, by defining R_d as the detected pixel which actually belong to the current speaker, we can evaluate the goodness of the method in recognising the actual speaker among other potential speakers. At last, to capture the temporal aspect of the methods performances we show in Fig.4b their hit-ratio curves. Note that a hit occurs in a frame if $Pr > 0.5$.



(a) (b)

Fig. 4: **Comparison to [9]**. Precision-recall and Hit Ratio curves for the testing videos averaged over the total number of frames. On average, PM+SR improves on speaker ID recognition by 23% and 59% over audio only and video only systems and by 36% over BM [9].

4. CONCLUSION AND FUTURE WORK

This paper has presented a new approach to audio-video (AV) speaker detection and localisation in a large unconstrained environment. We have shown that we improve a state-of-the-art AV correlation technique by adding speaking localisation data. In particular, we have reported preliminary results of the baseline method failing when distracting and interfering/occluding AV sources exist in the scene and we have provided for an alternative solution, showing that the speaker detection and localisation precision improves. [9] results deeply rely on the chosen segmentation techniques. Hence, it may be worthy investigating new segmentation methods to work in more visually challenging scenarios.

5. RELATION TO PRIOR WORK

This paper stems from the work of [9] which shows how audio and video signals correlation at feature level allows to detect the dominant source of audio in sanitized scenarios where stationary moving objects emit some sound. Hence, it is is not at all suited to moving targets in the prototypical *cocktail party* or video diarisation (even indoor surveillance) scenarios, however we show with a small extra cost in sensing overheads the method may be adapted for wider use.

6. REFERENCES

[1] N. Checka, K.W. Wilson, M.R. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," vol. 5, pp. V–881–4 vol.5, May 2004.

[2] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 601–616, Feb. 2007.

[3] Huiyu Zhou, M. Taj, and A. Cavallaro, "Target detection and tracking with heterogeneous sensors," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 2, no. 4, pp. 503–513, Aug. 2008.

[4] Yeongseon Lee and R. Mersereau, "Data association for people tracking using multiple cameras," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 312008-april4 2008, pp. 2585 –2588.

[5] Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *J. Sel. Topics Signal Processing*, vol. 4, no. 5, pp. 882–894, 2010.

[6] E D'Arca, M.N. Robertson, and J. Hopgood, "Using the voice spectrum for improved tracking of people in a joint audio-video scheme," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013.

[7] M. Andersson, S. Ntalampiras, T. Ganchev, J. Rydell, J. Ahlberg, and N. Fakotakis, "Fusion of acoustic and optical sensor data for automatic fight detection in urban environments," in *Information Fusion (FUSION), 2010 13th Conference on*, 2010, pp. 1–8.

[8] Marco Cristani, Manuele Bicego, and Vittorio Murino, "Audio-visual event recognition in surveillance video sequences," *Multimedia, IEEE Transactions on*, vol. 9, no. 2, pp. 257–267, Feb. 2007.

[9] H. Izadinia, I. Saleemi, and M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," *Multimedia, IEEE Transactions on*, vol. 15, no. 2, pp. 378–390, 2013.

[10] Z. Barzelay and Y.Y. Schechner, "Harmony in motion," pp. 1–8, June 2007.

[11] Eleonora D'Arca, Neil Robertson, and James Hopgood, "Audio-video tracking of active speakers through occlusion," in *In Proc. of the 9th IET Data Fusion and Target Tracking Conference*, 2012.

[12] B.G. Gebre, P. Wittenburg, and T. Heskes, "The gesturer is the speaker," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 3751–3755.

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

[14] N. Campbell and N. Suzuki, "Working with very sparse data to detect speaker and listener participation in a meeting corpus," 2006.

[15] P. Feyereisen and J.D. de Lannoy, *Gestures and Speech: Psychological Investigations*, Cambridge University Press, 1991.

[16] J.M. Iverson and S. Goldin-Meadow, "What's communication got to do with it? gesture in children blind from birth.," *Developmental Psychology*, vol. 33, no. 3, pp. 453–67, 1997.

[17] R. I. Mayberry and J. Jaques, "Gesture production during stuttered speech: Insights into the nature of gesture-speech integration," pp. 199–213, 2000.

[18] Alex Pentland, "Socially aware computation and communication," *IEEE Computer*, vol. 38, no. 3, pp. 33–40, 2005.

[19] D.B. Jayagopi, H. Hung, Chuohao Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 3, pp. 501–513, March 2009.

[20] H. Hung, Yan Huang, G. Friedland, and D. Gatica-Perez, "Estimating dominance in multi-party meetings using speaker diarization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 847–860, May 2011.

[21] M. Grimm and K. Kroschel, *Robust Speech Recognition and Understanding*, I-Tech Education and Publishing, 2007.

[22] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, Dec. 1936.

[23] T. Gehrig, K. Nickel, H.K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," pp. 118–121, Oct. 2005.