Look Who's Talking

Eleonora D'Arca, Neil M. Robertson*, James Hopgood†

*Heriot Watt University, Edinburgh, UK, {ed88,n.m.robertson}@hw.ac.uk
†University of Edinburgh, Edinburgh, UK, James.Hopgood@ed.ac.uk

Abstract. This paper proposes a method to automatically detect and localise the dominant speaker in a conversation by using audio and video information. The idea is that gesturing means speaking, so we look for people hands or heads movements to infer a person is talking. In a normal conversational context with two or more people, we learn Mel-frequency cepstral coefficients (MFCC) and find how they correlate with the optical flow associated with moving pixel regions by canonical correlation analysis (CCA). In complex scenarios, this operation could be resulting in associating pixel regions to sounds which actually are not really correlated. Therefore, we also triangulate the information coming from the microphones to estimate the position of the actual audio source, narrowing down the visual space of search, hence reducing the probabilities 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 data the improvement in dominant speaker localization.

1. Introduction

Tracking a speaker in an unconstrained environment has become an increasingly studied problem over the last few years only. In fact, speaker localisation and person tracking in general, finds a number of applications such as video surveillance, security, home automation, hospital care and so on. Such applications mostly involve analysing large uncontrolled areas where no constraints on people movements exist. In such scenarios, speaker tracking by means of microphones is subject to reverberation and background noise which dramatically decrease standalone system performances. In addition to clutter, obstructions and changing light conditions which on the other hand affects video person tracking. Thus, designing a multi-modal system which integrates and/or fuses audio and video data may lead to a better speaker detection and localisation result.

State-of-the-art speaker tracking systems [1–4] which aim at detecting the number of speakers in a room and localising them over time, normally treat the two cues as they did not relate to each other so that the data, different by nature, are integrated/fused as if they were independent variables, as e.g. in [5]). Conversely, little attention has been directed towards the exploitation of audio and video signals underlie relation to track the actual speaker. However, audio-video (AV) correlation has been widely used to recognise AV event anomalies [6, 7] in large unconstrained spaces whereas on the contrary AV signal correlation has been used for speaker detection only in small controlled environments [8, 9]. In particular, experiments have been carried out for scenes where people speak next to microphones and cameras while some distracting sources are playing in the background. This work proposes to extend those techniques to track the dominant speaker in a large uncontrolled scenario, where people are having a conversation (e.g. *cocktail party* scenario). However it is important to highlight that the aforementioned AV correlation techniques are mostly proved to be effective [8,9] in very stable scenarios, where sound sources are stationary and no distracting motion no occlusions exist. On the contrary, large unconstrained scenarios such as surveillance ones, are normally low resolution and large field of view, meaning the subjects of interest are far apart from the sensors and normally described by a few number of pixel within a frame. Furthermore, surveillance scenes often focus on crowds fluxes or, to a smaller extent, *cocktail party* scenarios, where more than 2 sources of motion and speech are recorded and where people often occlude each other. Thus, to effectively applying the said algorithms to such complex scenarios, it is necessary to integrate or fuse a further cue to decrease the roughness of the detections. To such aim, we triangulate also the audio information gathered by the microphones so as to localise the dominant audio source. Hence, this paper contribution is twofold i.e: *a*) it extends the use of AV correlation analysis to large uncontrolled environments i.e. for low resolution scenes and for non stationary sound sources; *b*) it resolves the problem of extra video motion correlated with the dominant audio signal for AV canonical correlation analysis (CCA).

2. Algorithm Description

Assuming that observing gesturing in a video stream often means locating speaking activity, we attempt to recognise and exploit a somewhat inherent correlation between audio and video signal as done similarly in [8]. In particular, they learn speaker Mel-frequency cepstral coefficients (MFCC) and find how MFCC correlate with the optical flow associated to moving visual objects by canonical correlation analysis (CCA). Hence, they define a smoothed speaking likelihood in the video segments, which sound is supposed to be the most correlated with, to eventually infer the actual/dominant speaker is within that pixel region. In their approach, they further improve on this segmentation by manually selecting points in the image which indicates AV foreground and background. In particular,

they show this step improve results in complex scenarios where distracting, occluding and correlated motion may appear. We enhance this approach to make it fully automatic, by substituting the user by the audio localisation information i.e. the position calculated by evaluating the time delays of arrival of the audio signals at the microphones.

2.1 *Video Features*

The video features extraction procedure consists of several steps. First, we compute the forward and backward dense optical flow of each image frame. Such information are combined to 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, the RGB colour, velocity and acceleration of each pixel $\mathbf{p}$ in a frame is combined into a single feature vector $v_{ij} = (\mathbf{p}, col, vel, acl)$. Then a spatial segmentation, based on the QuickShift algorithm, is performed in a per-frame fashion. This first segmentation is followed by a second one which is carried out across frames in order to further reduce the data dimension. In particular we calculate a K-means spatio-temporal segmentation so that, at the end of the processing, every pixel in a frame can be ascribed to the spatio temporal centre of mass of the k-th segment found by K-means i.e. final segments $S_k(k=1,...,K)$ are identified by the normalised velocity and acceleration of their centre of mass in addition to their mean RGB colour : $v_{ij} = (\mu_{\mathbf{p}}, \mu_{col}, \mu_{col}, \mu_{col})$. Segments which in time corresponds to image patches with constant motion are set to zero whereas the m_1 top segments for velocities and the m_2 top for acceleration are selected in order to compose the final video feature vector. The feature vector $\mathbf{v}$ for a video, is basically represented by an $m \times t$ matrix whose columns correspond to frames.

2.2 *Audio Features*

Audio feature vectors are represented by the first $\frac{n}{2}$ MFCC coefficients [10] (audio signal velocity) and their $\frac{n}{2}$ derivatives (audio signal acceleration). The feature vector $\mathbf{a}$ for a video, $n \times t$ matrix whose columns correspond to frames. Note that this means 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.

2.3 *Audio Video Correlation*

Audio and video feature vectors correlation is sought under the hypotheses that there exist some kind of hidden correspondence between the two signals i.e. motion velocity of the image is related to the audio MFCC, whereas motion acceleration is related to the MFCC derivatives (Δ -MFCC). To this aim canonical correlation analysis (CCA) [11] is used. In fact, it allows to not only find a common coordinate system where $\mathbf{a}$ and $\mathbf{v}$ can be projected, but also to immediately know as a by product the maximised correlation. This is very important in order to be sure the video segment retrieved is associated with the dominant audio source i.e. the one that maximise the correlation between audio and video data. Specifically, the CCA problem between two random variables has the closed form solution:

$$\begin{cases} C^1_{vv} C_{va} C^1_{aa} C_{av} w_v = \lambda^2 w_v \\ C^1_{aa} C_{av} C^1_{vv} C_{va} w_a = \lambda^2 w_a, \end{cases}$$

where C represents the correlation matrix and w_v and w_a the canonical basis of $\mathbf{v}$ and $\mathbf{a}$ respectively. This means the largest CCA eigenvector w_{v^1} corresponding to the largest eigenvalue λ_1^2 is the one which give the larger contribution to the maximum correlation between audio and video i.e. which maximise the canonical variates, $v_1^{'} = w_{v^1}^T \mathbf{v}$ and $a_1^{'} = w_{a^1}^T \mathbf{a}$. If it is assumed that only a single dominant audio source exists, the first of these eigenvectors w_{v^1} is chosen and the frame segments that it identifies $\overline{\mathbf{S}}$ are said to be the one where the sound is originating. A binary confidence map is set for the selected segments and, finally, the confidence maps are convolved with a 2D spatial Gaussian kernel function and a 1D temporal Gaussian mask, to smooth the results over the spatio-temporal volumes associated with the segments. In practice only the normalised elements of w_{v^1} largest then a predefined threshold are selected.

3. Correlation and Localisation data Fusion

Details of the triangulation, by mean of an extended Kalman filter (EKF) can be found in [5, 12]. 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 CCA result is intuitively done 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 which they belong to i.e. $(x,y)_{SL} \mapsto (i,j)_{SL} \in S_k$. Therefore, we set S_k 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. Ultimately, we treat this kernel as if it was another first base eigenvector coefficient adding up his contribution to the CCA result.

4. Experimentation

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 [8] 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

Figure 1. In (a) the layout of the experiments setup is given. (b) shows the result of the baseline method [8] applied on the Basketball data provided for one of the video frame. This output comes from the final (QuickShift) spatial segmentation shown in (c) which is obtained by overlapping the colour, motion velocity and acceleration of the frame. In (d) the final (K-means) spatio-temporal segmentation is presented.

area considered of interest is 12 m^2 (as seen in Figure 1a). 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 their correspondent targets and static matrices Q and R [12], 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 [8]. The number of clusters in the K-means algorithm is set to be 30. And the smoothing gaussian kernel has a variance of $\sigma = 5$. For each experiment, we analyse approximately 4 *s* of recorded audio.

Experiment *'Cocktail Party'* (Figures 2, 3, 4, 5) shows several people having a conversation in groups and some passer-by. There is a dominant group in the foreground while another group is in the background. This results in challenging speech overlaps and occlusions. Speakers are at least 50 *cm* far from the microphones. They stand still while some passer-by walks in the background.

Experiment *'Crossing'* (Figures 6, 7) shows two people who look alike walking while having a conversation and other people in the background. They meet along a diagonal where they keep on walking past each other causing an occlusion in the resulting image. Also two people external to the main scene are in the room.

4.1 *Dominant Speaker Detection Results*

In the following we report a qualitative description of the preliminary results we have obtained referring to their corresponding figures. As said, the AV correlation method was implemented on the base of [8]. Figure 1b shows the output of such a method on the "Basketball" video sequence provided by the authors. Figures 1c and 1d illustrate respectively the results of the Quickshift spatial segmentation and the K-means spatiotemporal segmentation for the analysed frame.

Experiment *'Cocktail Party'* ground-truth consists of the speaker on the left (*First Speaker*) talking for the first part of the video (≈ 1 *s*) whereas the one with the check shirt (*Second Speaker*) speaks for the remaining (≈ 3 *s*). The third person (*Listener*) in the video foreground (blue jumper) is paying attention to the conversation while producing some distracting fine motion by slightly moving his body on a side. Other people are having a conversation in the background. Figure 2a shows frame 2 of the video. It is clear that the AV correlation results are good as only the *First Speaker* is gesturing while the other two are politely listening (2b). Also the speaker localisation (SL) data shows good results (2c), only they are more oriented towards the floor. The fusion output in Figure 2d shows how the localisation and correlation data combine: the final kernel is more oriented to the segment pointed by the SL data as its corresponding kernel has a larger domain, hence it is assumed to be more reliable. In the successive frame in Figure 3a, *Second Speaker* has started to move his hands while the *Listener* has been moving his body resulting in false positive detections of the CCA approach (Figure 3b). This can be only mitigated by the SL corresponding segment (3c), so that the fusion results, despite pointing out the correct speaker, still presents false detection trails corresponding to the other people movements (3d). In figure 4a the conversation has just been handed over to the *Second Speaker*, the *First Speaker* is still gesturing while the *Listener* has shifted his body to the right to meet the *Second Speaker*'s gaze. This uncorrelated motion reflects on a AV correlation total failure. Adding the speaker position corresponding segment (Figure 4b) improves the CCA approach as can be seen in Figure 4d. It is important to highlight that, the nature of the experiments scenarios themselves (characterised by reverberation, speech far from the sensors, small amount of

data processing, speech overlaps and background noise) make TDOA-based SL a rough estimator of speaker position, even if it may still represent a decent result for speaker detection. On the other hand, SL also may benefit from using AV correlation data for systems in which speaker position must be determined to the 10 *cm* resolution. Frame 5a is very interesting in this sense. In this case in fact, the SL is pretty rough as the speaker position falls into a segment which corresponds to the speaker's arm (*Second Speaker*) as can be seen in Figure 5c. The AV correlation output which is itself wrong (Figure 5b), if combined with the SL, results in the segment corresponding to the actual speaker (Figure 5d). Note that in a speaker tracking system, its centroid may be re-projected back onto the ground plane to better the recursive estimation of the speaker trajectory.

Experiment *'Crossing'* only speaker is the person moving from the right side of the image to the left side (speaker). A second person (distracting person) walks in the opposite fashion causing an occlusion and gross distracting motion. In Figure 6a the AV CCA output points originally out the current AV source in the image (Figure 6b). This is actually better once the "localisation" segment (Figure 6c) has been fused into the final estimation (Figure 6d). Nevertheless, after the occlusion has occurred in Figure 7a the AV correlation algorithm detect an area which does not correspond to the speaker (7b) nor to the distracting person. This is mainly due to the fact that this particular output segment has a larger contribution to the AV correlation over the entire video sequence. In fact, that area shows high motion velocity and acceleration values in all the analysed frames. Figure 7d shows the AV foreground detection recovering obtained after adding in the SL data (Figure 7c).

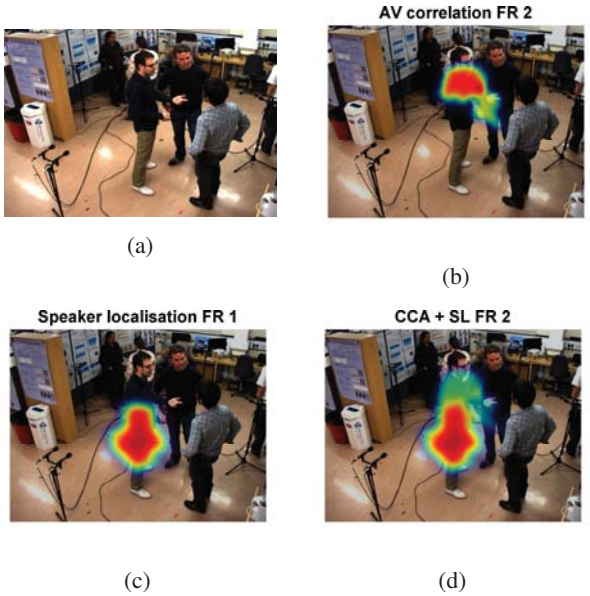


Figure 2. *'Cocktail Party'* Frame 2 Results. (a) shows the *First Speaker* talking while the other two people are listening without moving . In (b) the results of the AV CCA analysis are given whereas (c) shows the result for the speaker localisation (SL) algorithm. Finally, (d) presents the fusion of CCA and SL data.

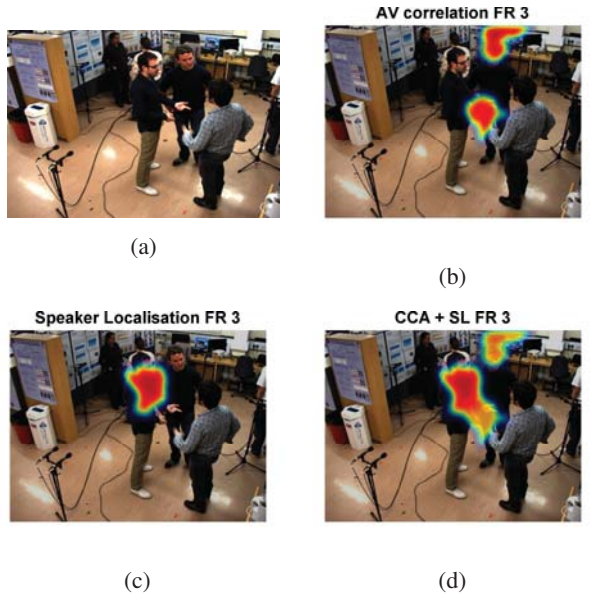


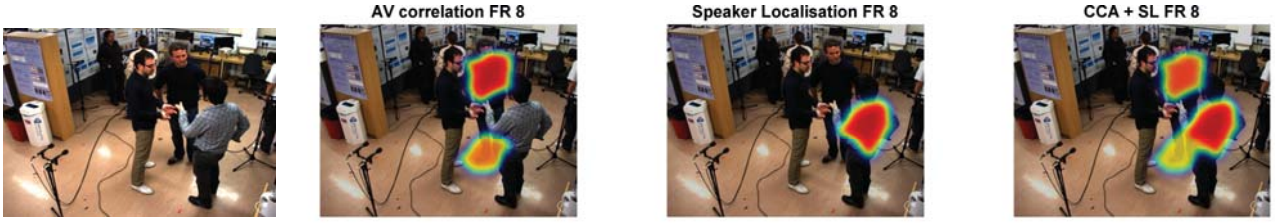
Figure 3. *'Cocktail Party'* Frame 3 Results. (a) shows the *First Speaker* still talking while the *Second Speaker* has started to do some gesture and the *Listener* has moved slightly his head towards the person who is going to speak. In (b) the results of the AV CCA are given whereas (c) shows the result for the SL algorithm. Finally, (d) presents the fusion of CCA and SL data. These do not totally resolve the problem, but reduce the probability for the *Listener* to be the speaker and add on a high probability for the *First Speaker* to be yet the actual one.

5. 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 occluding/overlapping AV sources exist in the scene and we have provided for an alternative solution, showing that the speaker detection and localisation qualitatively improves.
Further investigations shall be done on the fusion of the localisation data at different abstraction levels. For example, it may give better results combining straight into the video feature vector the velocity and acceleration which characterised the "speaker trajectory segments"; in fact, they may have been replaced during the *m* dimension selection by highest motion velocity and acceleration segments which actually belong to distracting motion non corresponding to the dominant audio source. Furthermore, we want to implement a composite AV dynamic Bayes network (DBN), including the presented work, which encompasses several audio-video (AV) weak classifiers fusing audio and video on different level of abstraction. Such a network shall infer the role of the detected speaking person e.g. speaker, listener, person talking on a phone, passer-by.

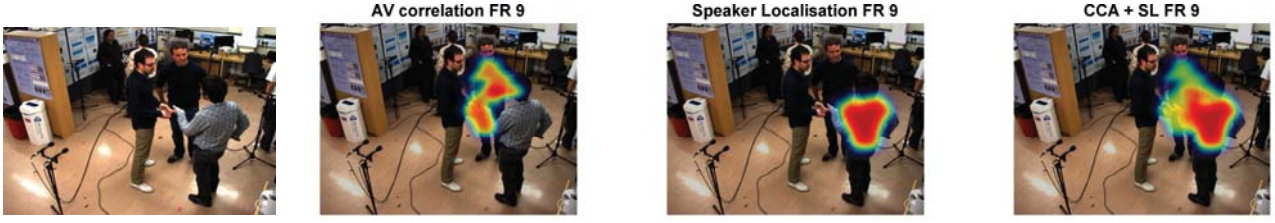
Acknowledgements

This research is supported by the EPSRC research grant number EP/K014277/1.



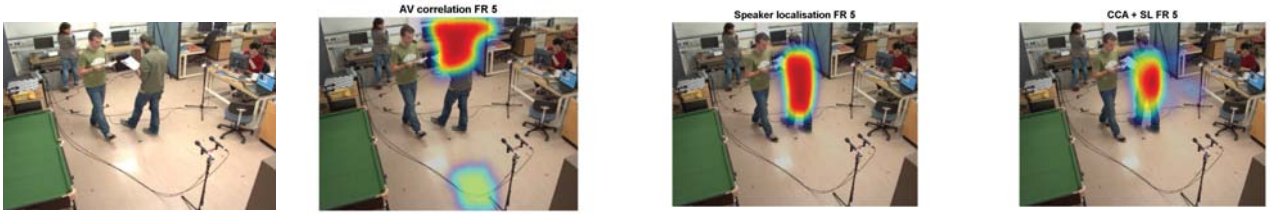
(a) (b) (c) (d)

Figure 4. *'Cocktail Party'* Frame 8 Results. (a) shows the *First Speaker* now silent together with the *Listener* after the conversation has been handed over to the *Second Speaker*. Note that the *Listener* has been shifting his body in a way such to turn better in the direction of the speaker's (*Second Speaker*) gaze direction. In (b) the results of the AV CCA are given whereas (c) shows the result for the SL algorithm. Finally, (d) presents the fusion of CCA and SL data. These do not totally resolve the problem, but reduce the probability for the *Listener* to be the speaker and add on a high probability for the *Second Speaker* to be the actual one.



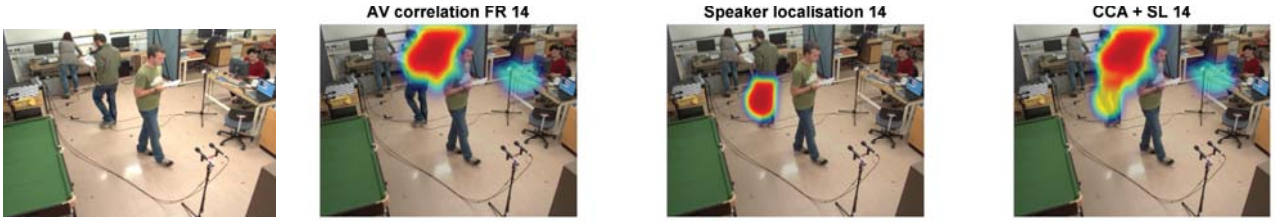
(a) (b) (c) (d)

Figure 5. *'Cocktail Party'* Frame 9 Results. (a) shows the *Second Speaker* talking while the other two people are listening. The *Listener* is moving slightly on his right. In (b) the results of the AV CCA are given whereas (c) shows the result for the SL algorithm. Finally, (d) presents the fusion of CCA and SL data.



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

Figure 6. *'Crossing'* Frame 5 Results. (a) shows the speaker moving while reading a book towards the bottom of the room while the distracting person is moving to reach the top. (b), (c) and (d) respectively presents the AV CCA algorithm, the SL and the final fusion results.



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

Figure 7. *'Crossing'* Frame 14 Results. (a) shows the speaker almost reaching the bottom of the room while the distracting person has almost reached the top. (b), (c) and (d) respectively presents the AV CCA algorithm, the SL and the final fusion results. Note that now in (d) the area occupied by the speaker has been assigned a higher probability

References

[1] N. Checka, K. W. Wilson., M.R.Siracusa, T.Darrell. "Multiple person and speaker activity tracking with a particle filter", *Acoustics, Speech, and Signal Processing Proceedings (ICASSP). IEEE International Conference on*, **vol.5**, pp. 881–884, (2004).

[2] H. Zhou, M.Taj, A .Cavallaro. "Target Detection and tracking with heterogeneous sensors", *Delected Topics in Signal Processing, IEEE Journal of*, **volume 2, no.4**, pp. 503–513, (2008).

[3] Y. Lee, R. Merserau. "Data Association for people tracking using multiple cameras",*Acoustics, Speech, and Signal Processing Proceedings (ICASSP). IEEE International Conference on*, pp. 2585-2588, (2008).

[4] S. T.Shivappa, B. D. Rao, M. M. Trivedi. "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation", *J. Sel. Topics Signal Processing*, **volume4, no.5**, pp. 882–894, (2010).

[5] E. D'Arca, N. M. RObertson, J. Hopgood. "Audio-video tracking of active speakers trough occlusion", *In Proc. of the 9th IET Data Fusion and Target Tracking Conference (DF TT): Algorithms Applications*, pp. 1–6, (2012).

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

[7] M. Cristani, M. Bicego, V. Murino. "Audio-visual event recognition in surveillance video sequences", *Multimedia, IEEE Transactions on*, **volume 9, no. 2**, pp. 257–267, (2007).

[8] H. Izadinia, I. Saleemi, M.Shah. "Multimodal analysis for identification and segmentation of moving-sounding objects", *Multimedia, IEEE Transactions on*, **volume 15, no. 2**, pp. 257–267, (2013).

[9] Z. Barzelay, Y.Y. Schechner. "Harmony in motion", *Computer Vision and Pattern Recognition (CVPR). IEEE Conference on*, pp. 1–8, (2007).

[10] M. Grimm, K. Kroschel. "Robust Speech Recognition and Understanding", *InTech Education and Publishing*, (2007).

[11] H. Hotelling. "Relations between two sets of variates", *Biometrika*,**volume 28, no.3/4**, pp. 321–377, (1936).

[12] T. Gehrig, K. Nickel, H.K. Ekenel, U.Klee, J. McDonough. "Kalman Filters for audio-video source localisation", *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pp. 118–121, (2005).