Signal Processing 129 (2016) 137–149

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Robust indoor speaker recognition in a network of audio and video sensors ☆

Eleonora D'Arca [a,*], Neil M. Robertson [a], James R. Hopgood [b]

[a] Visionlab, ISSS, Heriot Watt University, Edinburgh EH14 4AS, UK
[b] University of Edinburgh, Edinburgh EH9 3JG, UK

ARTICLE INFO

Article history:
Received 4 October 2015
Received in revised form
26 April 2016
Accepted 28 April 2016
Available online 4 June 2016

Keywords:
Surveillance
Speaker diarisation
Security biometric
Audio–video speaker tracking
Multimodal fusion

ABSTRACT

Situational awareness is achieved naturally by the human senses of sight and hearing in combination. Automatic scene understanding aims at replicating this human ability using microphones and cameras in cooperation. In this paper, audio and video signals are fused and integrated at different levels of semantic abstractions. We detect and track a speaker who is relatively unconstrained, i.e., free to move indoors within an area larger than the comparable reported work, which is usually limited to round table meetings. The system is relatively simple: consisting of just 4 microphone pairs and a single camera. Results show that the overall multimodal tracker is more reliable than single modality systems, tolerating large occlusions and cross-talk. System evaluation is performed on both single and multi-modality tracking. The performance improvement given by the audio–video integration and fusion is quantified in terms of tracking precision and accuracy as well as speaker diarisation error rate and precision–recall (recognition). Improvements vs. the closest works are evaluated: 56% sound source localisation computational cost over an audio only system, 8% speaker diarisation error rate over an audio only speaker recognition unit and 36% on the precision–recall metric over an audio–video dominant speaker recognition method.

1. Introduction

The establishment of the digital era has created applications which combine audio and video to automate human activity analysis and understanding. We highlight the main areas of interest. First, for surveillance applications, i.e., detecting a person's biometric features to ensure that there are no intruders in a restricted area [1]. Second, understanding people social behaviour and interaction to determine their "role" and their intentions [2]. Third, detecting a possible threat in a public place [3] and, consequently, beam-forming and segmenting a dialogue [4]. Typical surveillance scenarios are characterised by the use of many wide area, distributed sensors covering unconstrained scenarios. Scene monitoring is often required to be real-time, thus computationally inexpensive algorithms are fundamental to the development of an effective system, but this is not always evident in the literature at present, and this is the challenge we address in this work.

1.1. Related work

The first step to full audio–video (AV) human activity analysis and understanding systems is detecting and tracking speakers through significant occlusions. State-of-the-art sound source localisation algorithms [5,6] are still computationally expensive, hence they are not suitable for "real-time" (or frame-rate) applications. Solving large video occlusions is still an inherently challenging research problem: many existing papers solve the problem by using advanced multi-camera 3-dimensional (3D) systems [7] which are prone to error when the camera views do not overlap. They are computationally expensive, requiring GPU/FPGA implementations to function at frame-rate when parallelisation is possible. Complementary use of audio and video is able to compensate for noisy, missing and erroneous data, reducing the number of sensors and the computational resources required at the expense of minimal effort in integrating or fusing signals [8–11,2,12–15,3,16–22].

Audio and video fusion can be achieved in several ways chiefly using variations of sampling techniques [8,14,15,19,21]. Existing AV person tracking system architectures work well only in highly sanitised, i.e., constrained and predictable scenarios: principally meeting rooms and diarisation [13,16,18,20] in which the person motion is either stationary, e.g., when people are talking seated

☆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.

* Corresponding author.
E-mail addresses: eleonoradarca@hotmail.it (E. D'Arca), N.Robertson@qub.ac.uk (N.M. Robertson), James.Hopgood@ed.ac.uk (J.R. Hopgood).

http://dx.doi.org/10.1016/j.sigpro.2016.04.014
0165-1684/

around a fixed table. Existing systems use large sensor networks in which microphones and cameras are often very close or even *attached* to people [13–16]. A hierarchical system is more likely to achieve robust situational awareness. These are more robust, as accurate and require lower algorithmic and hardware complexity [2,16]. The weakness is that such systems often treat the two signals as if they were derived from truly independent processes: assuming one source of noise can affect only one kind of signal. None of the previous work explores whether an underlying relation between audio and video exists and seeks to exploit it fully.

AV event or anomaly detection literature is generally based on inferring AV signal correlations to recognise whether a relevant event has happened in some scenario of interest [3,11,12,17]. The same correlation approach may be used to pick out the dominant speaker from a group of speaking people, without audio beamforming, filtering, blind source separation and data association [23,24]. The definition of dominant speaker is clearly useful: a high degree of gesticulation and speaking activity are the fundamental cues to define dominance [25–28]. In fact, gesturing is 80–90% of the time associated to speaking activity [29]. Focussing on gesticulation detection is particularly suitable for low resolution video, where fine lip motion detection is not applicable and where close microphones may not be available.

To aid the reader, a schematic is shown in Fig. 1 and links to the different sections of the papers are explicitly made in the caption. Section 2 presents the integration of audio and video data at the signal level and their fusion at decision level for speaker detection and tracking (see [30]). A speaker voice recognition unit is implemented to make the multimodal tracking robust to occlusions (see also [31]). In Section 3, the experiments and the results related to the first part of the system are described. Here, the benefits of fusing multimodal data are highlighted remarking that standalone trackers have worse performances than the AV solution. Then, a possible solution to the problem of tracking the current speaker identity through occlusions by recognising speakers voices is demonstrated. Section 4 presents how to visualise in large indoor surveillance-like scenarios the dominant speaker identity when multiple people speak contemporaneously without resorting to sophisticated algorithms (see also [32]). Finally, in Section 5 the conclusions of this research study are highlighted and future avenues of research enumerated.

The exact contributions of this work relative to the published literature are: (a) definition of a new, high accuracy, fast audio source localisation algorithm augmented by video (stochastic region contraction with height estimation (SRC-HE)) which outperforms the baseline method stochastic region contraction (SRC) of Do et al. [6]; (b) extension of AV techniques for speaker tracking and event detection where people dynamically move and interact which outperforms the baseline method of Izadinia et al. [17]; (c) exploitation of a small sensor network, deploying only a single

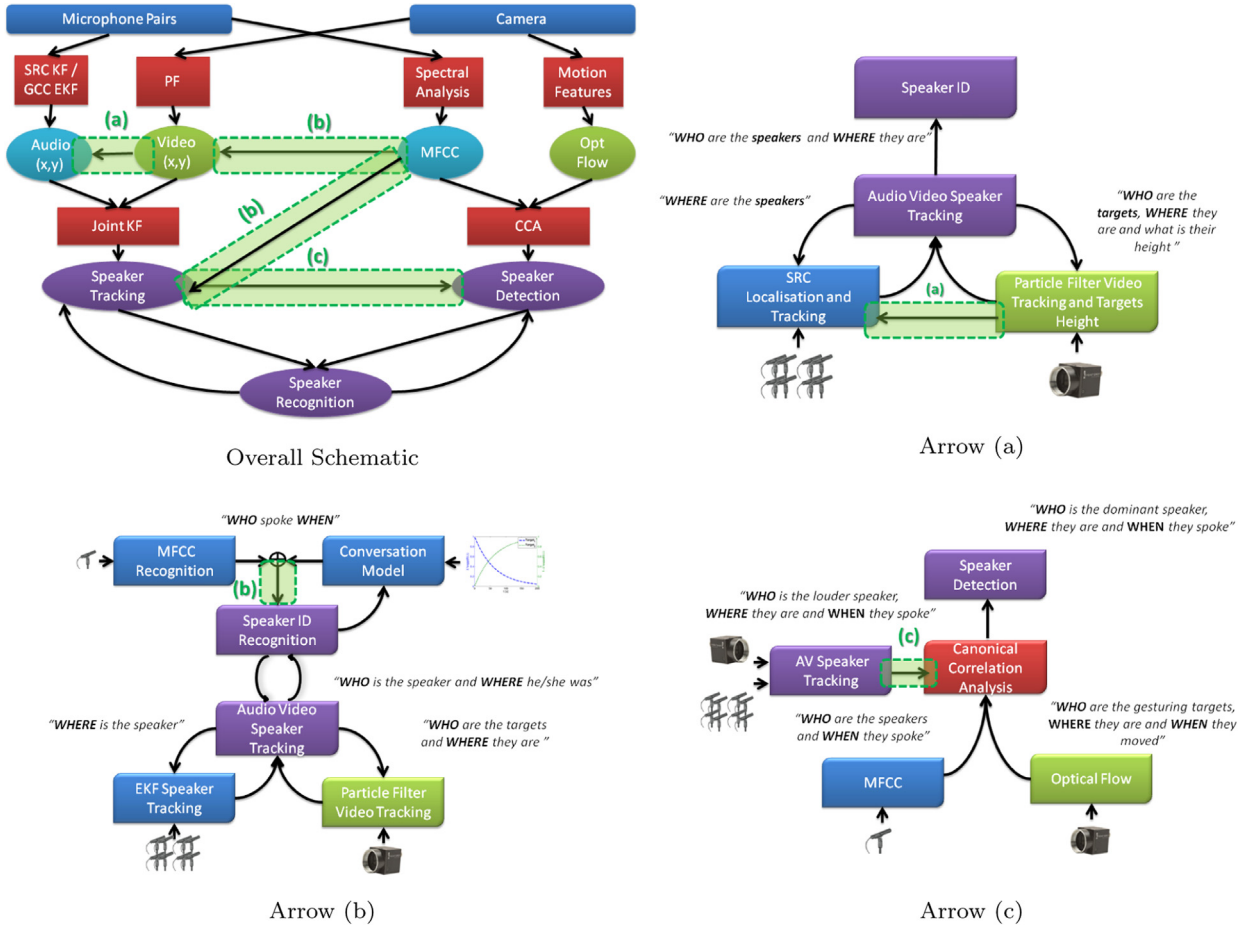


Fig. 1. A detailed schematic diagram of the overall system presented in this paper. The schematic in (a) shows how the audio and video features cooperate at different levels of semantic abstraction. Block cooperations are represented by highlighted arrows which coincide with the novelties of this work. In (b) an audio localisation algorithm is cued by video data which becomes faster and not less accurate (see Section 2.1). In (c) it is shown how Mel-frequency cepstral coefficients (MFCC) voice signature recognition helps video ID tracking to be consistent through occlusions and ID swaps (see Section 2.7). In (d) the system describes how the correlation between optical flow associated with gesturing and sound signature of the scene helps the speaker ID recognition through speech interferences (see Section 4.3). Fundamentally, this system represents the combination of the detections of three "weak" classifiers into one robust process.

Table 1
Test bed comparison to closest works. Abbreviation mic denotes microphones. Abbreviation cam denotes cameras.

Room type	Reference	Room size (m)	Analysed area (m)	Sensor equipment
Meeting	[13]	8.2×3.6	4.8×1.2	6 cam, 2×8 mic
	[13]	8.2×3.6	3.75×1.2	6 cam, 2×8 mic
	[16]	8×3.6	4.8×1.2	4 cam, 6×4 mic
Open	[8]	–	2×4	4 cam, 4×4 mic
	[58]	8×3.6	4.8×1.2	4 cam, 5×4 mic
	Us	$\mathbf{11 \times 10.1}$	$\mathbf{3 \times 4}$	**1 cam, 2×4 mic**

Table 2
How the presented system compares to the literature.

Room type	Reference	Tracking algorithm	Occlusion	Overlaps
Meeting	[13]	MCM-PF	YES(partial)	YES
	[16]	MID+HMM	YES	NO
Open	[8]	PF	NO	NO
	[58]	MID	YES	NO
	Us	**EKF**	**YES**	**YES**

camera and 8 microphones, which operates in open rooms vs. "standard" meeting rooms with constrained participants; (d) detection and tracking of speaker identity through occlusion and speech overlaps in a joint audio–video algorithm outperforming the state-of-the-art (Tables 1 and 2).

Early elements of this work were already presented in [30–32] and this paper makes two additional contributions relative to these papers. First, we give a unified presentation of the earlier work in a broader and fuller context; second, in this paper we present additional, new material, specifically graphs (Figs. 1a and 4), Tables 3 and 4 and original results (Figs. 3 and 6).

2. Audio–video speaker tracking

Bayesian inference is the foundation of most of the existing joint AV tracking schemes. The Kalman Filter and its Extended version [10,15], the Particle Filter [8,19,21] as well as hybrid approaches using Monte Carlo Markov chains [13] have been all used to tackle the problem. However, these filters work in meeting room scenarios and use close-field sensors deployed in large array configurations [13–16]. In this section, an AV speaker identity (ID) localisation and tracking algorithm which works in more unconstrained situation using a small sensor network is presented. Participants are not forced to wear sensors or to orient themselves towards the sensor. Audio source position estimates are computed

Table 3
A tabular summary of the following experiments and their rationale.

Experiment name	Figure	Algorithms	Rationale
'Formal conversation'	7(a)	AVT + SRC	Audio source localisation improves using video localisation.
'Informal conversation'	7(b)(c)		
'Single speaker'	8	AVT+ SR	Speaker ID recognition improves using audio–video tracking. Conversely, video ID tracking improves using speaker ID recognition
'Abandoning'	9		
'Crossing'	10		
'Crossing'	11	CCA + AVT + SR	Detecting the dominant speaker in babble noise using classical correlation algorithms can be aided by audio–video ID tracking systems
'Surveillance'	12		

Table 4
Simplified synoptic table of symbols used to defined the MOTP and MOTA indexes first defined for the CHIL meetings [2] as a benchmark for the CLEAR2007 datasets and others.

o_i	Audio–video object position
h_j	Tracking hypothesis
(o_i, h_j)	One possible mapping between an object and a track
d_t^i	Distance between object i and the mapped hypothesis
c_t	# Current matching pairs at time t
mme_t	# Mismatches errors made at time t
fp_t	# False positives at time t
m_t	# Missed objects at time t
g_t	# Objects present at time t

by the SRC sound source localisation algorithm [6]. The novelty here is that SRC is aided by available video information which estimates head height over the whole scene and gives a speed improvement of the 56% over the original SRC algorithm [6]. We call this approach SRC-HE. Finally, audio and video data are combined in a Kalman filter (KF) which fuses person-position likelihoods and tracks speaker positions and identities through occlusions demonstrating that the global audio–video tracker (AVT) outperforms single modality trackers.

2.1. Feature extraction

Height detection and video tracking: The appearance-based video tracker extracting person height is based on a GPU-accelerated particle filter with ellipsoid models [33]. Implementation is first described by Limprasert [34] and we direct the reader there for more details. In our work the video data coming from a single camera is exploited. Height measurements $\mathbf{z}_h$ for each detected target $i = 1, 2, \ldots, N_v$ (N_v number of detected targets) are extracted to cue the audio localisation algorithm (Fig. 2), since they directly correspond to a good estimate of the speaker's head position. Each detected position $\mathbf{z}_v$ can be described at each time step t as a $N_v \times 2$ vector, i.e., $\mathbf{z}_v = \{(\mathbf{x}_{vid}(t))\}$.

Audio source localisation cued by video height and audio tracking: A popular method of audio source tracking is extracting maximal time-difference of arrival (time difference of arrival (TDOA)) values from the generalised cross correlation phase transform (GCC-PHAT) of the signals from a pair of microphones, in the frequency domain (see Knapp for full details [35]).

A method more robust to reverberation, the Steered Response Power (steered response power (SRP)), makes use of the GCC-

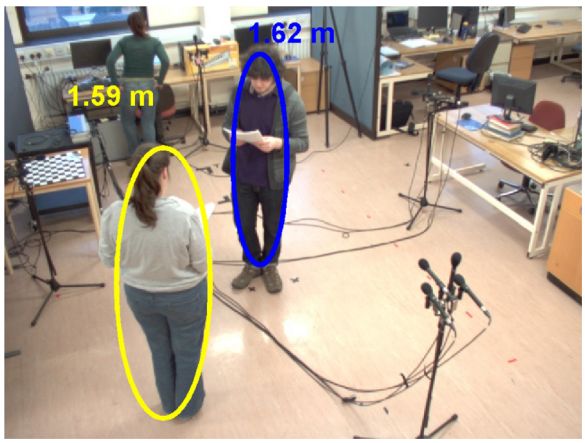


Fig. 2. Video detected height data is used to reduce the search space for the audio sound source localisation (SSL) algorithm SRC [6]. Note that the third person is not picked up from the height estimation algorithm as she was part of the background right from the start of the video signal processing. In fact, the last updates on the basis of a background subtraction algorithm.

PHAT to build an energy map in a system with multiple microphones [6]. This is the sum over all pairs $(m,\ n)$ of microphones of the corresponding value of the GCC-PHAT for the TDOA. Evaluating the SRP across an entire room is computationally costly. In this work, an enhanced version of the SRC algorithm to localise quicker and better an audio source is used. SRC works by sampling the SRP randomly and choosing a subset of the largest samples to form a new region to sample within. This is repeated until the process has discovered a maximum. In order to further improve upon the SRC, instead of sampling uniformly over height, a different sampling distribution is used, centred around a head height.

Around each person, a tracking algorithm can be relatively confident of their height. Further away from them, the decreasing confidence is modelled by increasing the variance of the sampling probability density function (PDF). Hence, the variance at a distance l metres from a speaker is chosen to be modelled by a sigmoid function q, such as (1), which is a scaled error function:

$$\mathrm{q}(l) = \alpha_1\ \mathrm{erf}\big(\alpha_2 l\big) \tag{1}$$

This function is zero at the origin and asymptotically approaches a constant as its argument tends towards infinity. All the variances around each detected speaker height are combined to form a global variance in the following equation:

$$\mathbb{L}_{\mathbf{p},\mathbb{T}} = \{l\colon (\exists\ \mathbf{q} \in \mathbb{T})(l = \overline{\mathbf{pq}})\},$$
$$\hat{q}_{\mathbf{p},\mathbb{T}} = \min_{l \in \mathbb{L}_{\mathbf{p},\mathbb{T}}} q(l) \tag{2}$$

At any point $\mathbf{p}$ in space, the appropriate variance $\hat{q}$ to use will be the sigmoid function q of the minimum of the set of all 2-dimensional (2D) Euclidean distances $\overline{\mathbf{pq}}$ to known sources, where the set of known source locations is denoted as $\mathbb{T}$ and an element from the set of known sources is denoted as $\mathbf{q}$. The minimum is chosen to ensure that the change in variance remains smooth even for overlapping sigmoids from multiple sources.

From a sparse set of people, the head height at every x–y coordinate in the SRP map needs to be defined. This is achieved using interpolation and extrapolation. When doing the interpolation, there is a trade-off between the smoothness of the curve produced and the size of ripples produced. The interpolation should not contain severe ripples as they would lead to large errors in the head height estimation across the room. Ideally, it should be monotonic and one way to achieve this is to use Delaunay triangulation [36] on the set of speakers, which creates a surface which can be evaluated at any 2D point.

To choose head height, existing knowledge of the current positions and heights of people in a room which is obtained from a camera (Fig. 2) is used (SRC-HE). In particular, the height data is updated on each iteration to the height of the last SRP peak found. Finally, the height h_{sub} to use at each time step for every 2D point $\mathbf{p} = (x_{p_2},\ y_{p_2})$ is drawn from (3) where, as said, $\mathbb{T}$ is the set of known speaker locations and $\mathbf{H}$ is the set of interpolated heights:

$$\varphi(z_h|\mathbf{p}_2) = \alpha_0 \mathcal{N}\Big(\mu_h,\ \sigma_h^2\Big) + (1-\alpha_0)\mathcal{U}\Big(0,\ h_r\Big)$$
$$\mu_h = \mathbf{H}[\mathbf{p}_2]$$
$$\sigma_h^2 = \hat{q}(\mathbf{p}_2,\ \mathbb{T}) \tag{3}$$

This mixes a Gaussian with a Uniform distribution across h_r, the entire height of the room. The resulting SRP value for the point $\mathbf{p}_2$ then is given by $SRP_{\mathbf{p}_2} = \max_{z_h}[S(x_{p_2}, y_{p_2})]$ (see Algorithm 1 and Fig. 1b). The SRC-HE algorithm allows for direct speaker position calculation. Nevertheless, speaker position estimations are characterised by missing and false detections. This is mostly due to speech pauses and room reverberation respectively.

Algorithm 1. Finding the global maximum using video height.

Input: video detected heights $\mathbf{z}_h$
Output: speaker position $SRP_{\mathbf{p}_2}$

Initial search for speech source
while *running* **do**
  $\hat{\mathbb{T}} = \mathbb{T}$
  for all room corners **do** ▷ add corners to $\hat{\mathbb{T}}$
    $\mathbf{n} \leftarrow (x_{\text{corner}}, y_{\text{corner}}, z_{\text{nearest member of } \mathbb{T}})$
    $\hat{\mathbb{T}} \leftarrow \hat{\mathbb{T}} \cup \{\mathbf{n}\}$
  end for
  $\hat{\mathbb{H}} \leftarrow \mathrm{DT}(\hat{\mathbb{T}})$ ▷ Delaunay triangulation
  for all $\mathbf{p}_2 = (x_{p_2}, y_{p_2}) \in \mathbb{A}$ **do** ▷ whole area
    $\hat{\mathbb{H}}_0 \leftarrow h_{sub} \sim \varphi(z_h|\mathbf{p}_2)$ ▷ video cueing
  end for
  Perform SRC-HE
  $\hat{\mathbb{T}} \leftarrow \hat{\mathbb{T}} \cup$ ▷ new speaker position
end while
return $SRP_{\mathbf{p}_2} = \max_{z_h}[S(x_{p_2}, x_{p_2})]$

Thus, SRC estimated positions are filtered by a KF. The signal vector obtained $\mathbf{z}_a$ can be written as $\mathbf{z}_a = \left\{ (x_{aud}(t), y_{aud}(t))^T \right\}$, to which the speaker ID at any given time t, $S_A(t)$, is assigned.

2.2. Fusion of audio and video decisions

As previously stated, to speed up SRC search time, the speaker's height (computed by the video particle filter (PF)) is input into the audio unit to drive height sampling (SRC-HE). Then, after the audio and video data have been aligned, the posteriors of the KF audio tracker and of the video PF, $\mathbf{x}_a$ and $\mathbf{x}_v$ respectively, are fused in a common KF node. As data are gathered simultaneously and used all at once in a centralised fashion, audio and video *pdfs* are assumed to be independent of one another. On the basis of the a priori local estimates for the state predicted by the single-modality trackers at each time step, we evaluate the joint state estimate (Algorithm 2).

The final, joint AV output is fed back into the individual audio and video trackers as the best estimate of the previous time step to improve the single modality estimation. It is important to note that, since the assumption that people speak alternatively (which is a strong assumption for a normal conversation) has been made, a single audio signal corresponds to several video measurements at a time, one for each of the detected targets. By basing the audio-to-video data association step on spatial proximity, i.e., nearest neighbour (NN) (more than one speaker cannot exist at the same point in space) speaker segmentation and ID recognition can also be obtained as long as people are resolved by the AV tracker. Its measurements can be considered robust with respect to the speaker motion model (see Algorithm 2).

In particular, the speaker ID inferred by the joint AVT is equal to the one of the i-th target if $S_{AV} = \mathrm{argmax}_i \left\{ p\left(\mathbf{z}_a, \mathbf{z}_{v_{(i)}} \middle| \mathbf{x}\right) \right\},\ i = 1, \ldots, N_v$.

Once a visual ID i has been assigned to every target in an image, the speaker change detection output by the audio unit is used to solve video occlusion. In particular, when a pair of video detections fall within a certain region D which depends on the video tracker accuracy $(\| \mathbf{z}_{v_{(i)}} - \mathbf{z}_{v_{(j)}} \| \le D)$ for each pair i, j of video detected targets), audio only contributes to KF filter innovation. If audio and

video do not both fall within a certain region A, based on both audio and video tracker accuracy, ($\| \mathbf{z}_a - \mathbf{z}_{v_{(i)}} \| \geq A$), then a new speaker is conservatively considered to be detected according to the audio ID guess (S_A), successfully resolving occlusions (Algorithm 2). However, in a large reverberant room audio false positives do exist and compromise the speaker ID recognition based on positional data only. The integration of a speaker recognition (SR) module is proposed to make the multimodal AVT more robust to video occlusions in reverberant rooms where people move around.

Algorithm 2. Audio video tracking algorithm.

Input: Audio $\mathbf{z}_a$ and video $\mathbf{z}_v$ measures
Output: Position $\mathbf{x}_{av}$ and identity S_{AV} of actual speaker

for every time step t **do**
  $\mathbf{x}_{av} = \mathbf{P}_{av}\left\{ \mathbf{P}_a^{-1}\mathbf{x}_a + \mathbf{P}_v^{-1}\mathbf{x}_v \right\}$ ▷ $\mathbf{P}_{av} = \mathbf{P}_a^{-1} + \mathbf{P}_v^{-1}$
  $S_{AV} = \text{argmax}_i \left\{ p\left(\mathbf{z}_a, \mathbf{z}_{v_{(i)}} \middle| \mathbf{x}\right) \right\}$
  if $\| \mathbf{z}_{v_{(i)}} - \mathbf{z}_{v_{(j)}} \| \leq D$ **then** ▷ occlusion
    $\mathbf{x}_{av} = \mathbf{x}_a$
    if $\| \mathbf{z}_a - \mathbf{z}_{v_{(i)}} \| \geq A$ **then**
      $S_{AV} = S_A$ ▷ speaker change
    end if
  end if
end for
return $\mathbf{x}_{av}$ and S_{AV}

2.3. Dealing with occlusion

In Section 2.2 it is pointed out that when people occlude each other, as in normal social interactions behaviours, Bayesian multimodal speaker tracking based on audio and video position detections in certain situations cannot distinguish the actual speaker ID in a conversation. This mainly occurs when the video tracks merge or cross over and the signal to reverberation ratio (SRR) is too low. As long as the video target ID recognition is based on general properties such as characteristic clothing, the natural dynamic and ambiguous behaviour of such a feature may lead to situations like occlusions in which they are completely useless, e.g., two people who wear clothes of the same colours will have an associated histogram of colours very similar (see Fig. 9 for an example of such a situation). In the literature, this is normally solved either by using proximity models or placing physical constraints on people. However, if target ID is decided on the basis of a more specific feature such as voice, the fact that it is seldom observable could reduce the number of cases in which visual ID determination is compromised, representing a more elegant and less invasive solution. Voice *spectral features* are now calculated for each speaker and such information is incorporated into the AVT, so as to simplify the video-to-audio data nearest neighbour association step. By doing so, it is demonstrated that the AVT ID tracking performance improves. In turn, when speakers are distant from the microphones, recognising a speaker by their voice can be very complicated [37,38]. Thus, exploiting audio–video positional cues also benefits the speaker voice recognition task at a distance (see Section 2.5).

2.4. SRC-HE vs. GCC-PHAT audio tracking

Despite the fact that SRC-HE reduces the number of FEs, audio measurements extraction based on SRC would still be not suitable for real-time applications [39]. The previous SRC-HE module is then replaced by the generalised cross correlation phase transform (GCC-PHAT) introduced in Section 2.1, as this does not involve cumbersome point function estimations. The drawback is that the basic GCC algorithm can only detect one source at a time and it is known to be sensitive to room reverberations [5], however it is still effective under moderate reverberant environments ($T_{60} \approx 0.3$ s) [40]. For these reasons, at first experiments where only a speaker is active at any given time are carried out, as it often happens in a polite conversation between two or more people. Speech segments using a voice activity detector (VAD) [41] are further extracted and processed using a GCC-PHAT step, for the signal to be more robust to reverberations. Thus, the measure vector obtained $\mathbf{z}_a$ (see Section 2.1) can now be rewritten as $\mathbf{z}_a = \left\{ \tau_m(t) \right\}$, where each component τ_m is the TDOA collected at the m-th microphone pair at each time step t. Since TDOAs are not linear in the speaker position, they must be input into an extended Kalman filter (EKF), as in [10] to get an audio position estimation.

2.5. Text-independent speaker recognition

We propose that, since the microphones already gather audio information for tracking purposes, the temporal spectral content of the signal can be used to extract speakers voice features and recognise their ID. Specifically, a SR module is chosen which performs text-independent speaker identification based on Gaussian mixture model (GMM) [42], under the assumptions that there exist N_v possible speaker identities (as many as the detected video targets), whose "voiceprints" models $p(s_i)$ are learned beforehand.[1] In particular, speaker voice models are calculated on the base of 60s training signal for each speaker. From every voice sequence 12 sets of Mel-frequency cepstral coefficients (MFCC) [43] are extracted. Each model is represented by a 32-mixture GMM, whose parameters are estimated on the base of the extracted MFCC vectors by expectation maximisation (EM) [44]. The test conversation sequences, not recorded in matching conditions, as it would be in a surveillance scenario characterised from different noises day by day, are framed in small speech-only subsegments which are considered to be long enough to detect a speaker change. For each speech subsegment its MFCCs are extracted and compared to the available database of speaker models to determine the likelihood $p(S|s_i)$ of a particular speaker ID to be the one who uttered the actual speech subsegment S. Finally, the speaker ID opinion is output as: $S_{SR} = \text{argmax}_i \left\{ p(S|s_i) \right\}$, $i = 1, \ldots, N_v$ and its GMM's likelihood is used as a confidence measure. Our experiments here are characterised by just one speaker change detection point. The performance of the SR unit is better evaluated in terms of speaker verification [45]. Hence, Fig. 3 shows the comparison of each voice in our database against each other voice model. Fig. 3a illustrate performances for 1 close microphone pair recording (2 channels) and Fig. 3b for 3 far microphone pairs recording (6 channels) to highlight the difficulties of detecting speaker at a distance despite the increased number of channels. In particular, the equalisation error rate (EER) [46] is 0 in the first case whereas it raises to 5.12% in the second. Finally, Fig. 3c shows results from all 8 channels; note that despite adding in the 2 more close recordings used for the first measure (Fig. 3a), the far distance microphones detrimentally affect the global performance whose EER is still as high as 4.96%.

2.6. Speaker conversation model

A new speaker switching probability is now introduced to

[1] 13 close microphone English recordings in our database from native and non-native speakers in different background conditions.

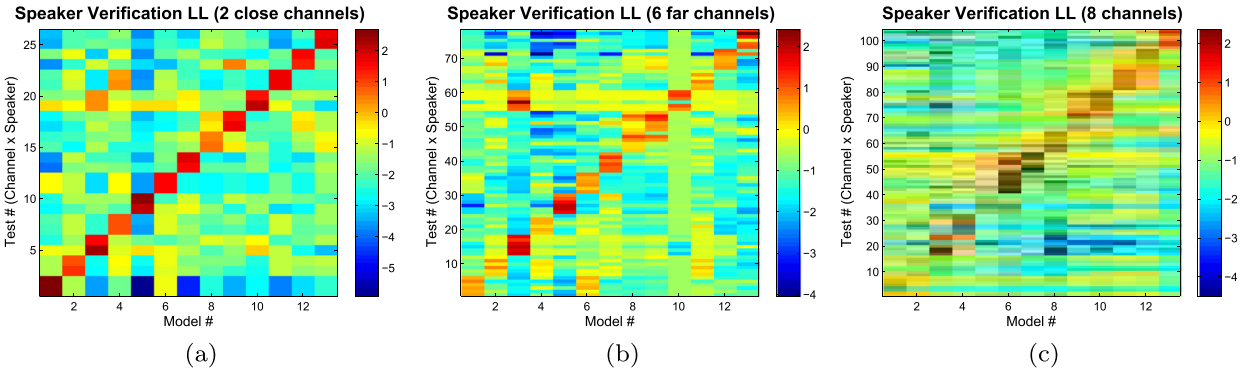


Fig. 3. The speaker verification confusion matrix for close and far microphone setups. The speaker verification confusion matrix for close and far microphone setups. This figure shows the ability of the implemented SR unit to verify ID of people in the recorded pool of voices for (a) 2 close microphones (1 close field microphone pair); (b) 6 far microphones (3 far field microphone pairs); and (c) the total 8 channels (4 microphone pairs: 1 in the close field plus 3 in the far field). Results show that the best performance is obtained for close distance recording, i.e., (a), whereas in (b) and (c) speaker recognition is severely compromised due to the 3 pairs of far distance microphones.

model the amount of time that has to be elapsed before a person stop talking and hand over to the next speaker. This acts as a smoothing prior on person ID recognition. In particular, this is defined by an exponential probability density function, i.e., $p\left(s_i(\delta t);\ \lambda\right) = \lambda e^{-\lambda \delta t} H(\delta t)$, where $H(\delta t)$ represents the Heaviside step function. The remaining $N_v - 1$ potential speakers are characterised by a probability of starting the conversation which is the complementary speaking probability scaled by $N_v - 1$. We call this a conversation model (CM). The CM is initially triggered by the i-th speaker ID detection obtained as a weighted speaker score fusion of the AVT and the SR modules [38]. The actual speaker ID in this case is given by: $S_{CM} = \arg\max_i \left\{ p\left(S_{CM} | s_i\right) \right\},\ i = 1, \ldots, N_v$.

Algorithm 3. Audio video tracking aided by speaker recognition algorithm.

Input: Audio $\mathbf{z}_a$ and video $\mathbf{z}_{v_{(i)}}$ measurements, S_{SR} and S_{CM} $(i = 1, 2, \ldots, N_v)$
Output: Position $\mathbf{x}_{av}$ and identity S of actual speaker

for every time step t **do**
  if $t=1$ **then** ▷ initialisation
    $S_{CM} \leftarrow w_{AV} S_{AV} + w_{CM} S_{CM}$
  end if
  $\mathbf{x}_{av} = \mathbf{P}_a^{-1}\mathbf{x}_a + \mathbf{P}_v^{-1}\mathbf{x}_v$
  if $\| \mathbf{z}_{v_{(i)}} - \mathbf{z}_{v_{(j)}} \| \leq D$ **then**
    $S = w_{SR} S_{SR} + w_{CM} S_{CM}$
    $S \leftrightarrow id \Longrightarrow z_{v_{(i)t}} \leftarrow z_{v_{(id)t}}$ ▷ $id = 1, .., N_v$
  end if
  $S_{AV} = \text{argmax}_i \left\{ p\left(\mathbf{z}_a, \mathbf{z}_{v_{(i)}} | \mathbf{x}\right) \right\}$
  $S \leftarrow w_{AV} S_{AV} + w_{SR} S_{SR} + w_{CM} S_{CM}$
end for
return $\mathbf{x}_{av}$ and S

2.7. Fusion of audio–video tracking and speaker recognition scores

Once a video ID i has been assigned to every target in each frame, the person recognition score derived from a SR+CM combination may be used in order to recover tracking ID data when occlusions occur. In such a case, competitive association hypotheses exist for the AVT, i.e., the AVT confidence drops below a certain threshold; thus, the SR and the CM opinions ratify the actual speaker ID, according to a weighted sum fusion rule [47] where weights are decided on the base of their estimate confidences [38]. Hence, the speaker ID is first fed back into the AVT to aid resolving the nearest neighbour AV association and successively to correct the wrongly inferred speaker ID. Secondly, it
is sent to the video tracker to indirectly re-assign the correct appearance models to the targets thus resolving the occlusion (see Algorithm 3 and Fig. 1c).

3. Experimental evaluation

Since the presented overall system (Fig. 4) is composed of several modules, in order to show the validity of the proposed approach, results are now presented separately to aid readability. Nevertheless, a summary of the overall conducted experiments is already presented in Table 3 to clarify the evolution of their rationale.

It is worth nothing that, given the wide range of human activity analysis applications, scenarios of interest and sensors configurations are varied, hence no standard data set has yet been collected for general purpose benchmarking. Systematic evaluation and comparison of the different fusion techniques for the specific AV speaker localisation and tracking is not possible and for this work it has been decided to develop a custom setup with less constraint on people, in contrast to classic meeting room applications.

1 camera and 4 directional microphones pairs are used to record AV data in a typical open office room, whose size is

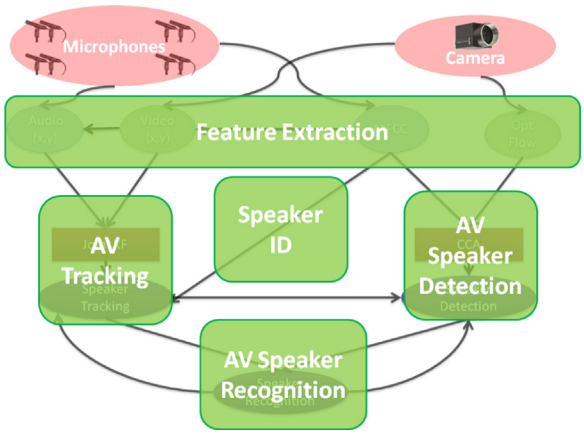


Fig. 4. A high level schematic diagram of the overall system presented in this paper. The presented high level diagram depicts the combination of the detections of three AV "weak" classifiers into one robust AV speaker recognition process.

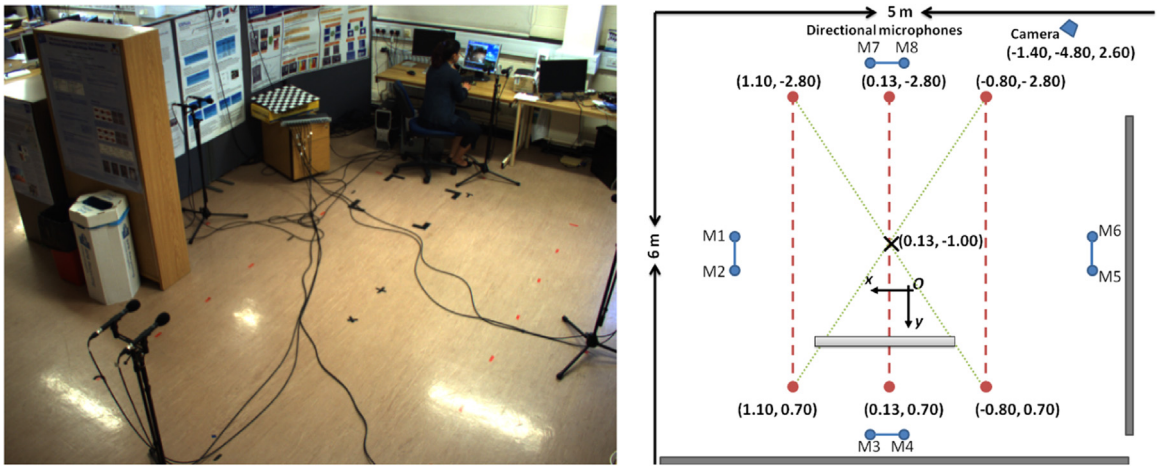
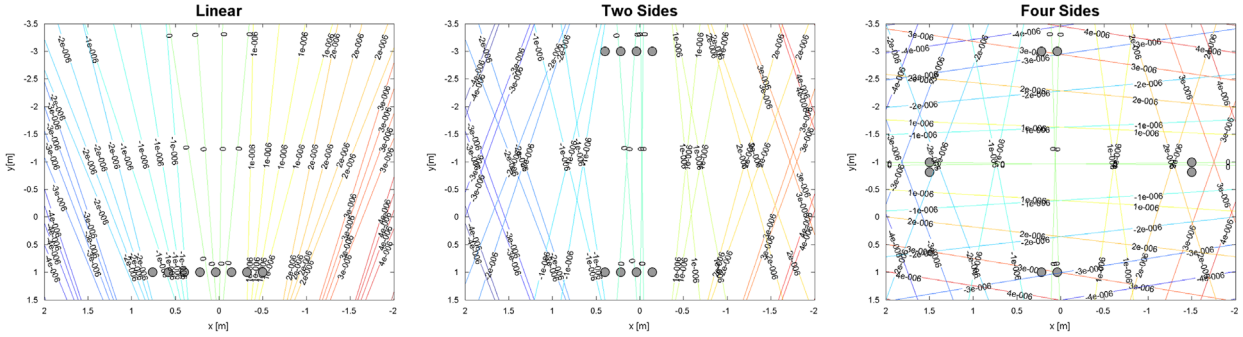


Fig. 5. In (a) a picture of the room used for our set of experiments is shown. (b) illustrates its layout and sensor setups.



(a) One linear array (8 microphones).

(b) 2 linear arrays (4 microphones each).

(c) 4 microphone pairs.

Fig. 6. Isolines, i.e., hyperbolas loci. An isoline indicates all points of a constant difference in distance from two points. 2D contour maps in the *xy* plane when $z = 1.7$ m for the (a) linear array of microphones, (b) 2 linear arrays of 4 microphones on two sides of the room, (c) 4 pairs of microphones of the 4 sides of the room. The last, has the more dense number of intersecting isolines, i.e., a higher number of possible solutions.

11 m × 10.1 m, where the area considered of interest is 3 m × 4 m and where people can freely move. A picture of the room and the sensor layout is presented in Fig. 5 and a graphical explanation of why the microphones were placed is presented in Fig. 6. Such a positioning for the microphones pairs was chosen to maximise the performance of the TDOA estimation in the analysed room. In particular, three configurations of microphones were compared, i.e., (a) 1 linear array of 8 microphones (1×8), (b) 2 arrays composed of 4 microphones (2×4) at two sides of the area, and (c) 4 pairs of microphones (4×2) at the four sides of the analysed area, among which the winning solution is the one proposed (4 pairs of microphones). Fig. 6 shows the contour maps of the room in the *xy* plane. As it can be seen, the contours for the 4 microphone pairs are the most dense and distinct all around the room unlike the other configurations.

Ground-truth data were hand labelled on a ground plane common to camera and microphones. Audio signals were sampled by the audio interface with a 24-bit precision resolution at 44.1 kHz, whereas the camera recorded the 640×480 RGB video frames at a rate of ≈ 7.5 Hz. No attempt to reduce normal background noise (desk fans, footsteps, talking, etc.) was made and a reverberation time $T_{60} \approx 0.5$ s was measured [48]. Synchronisation of the data was achieved by processing audio and video streams according to the camera frame rate, i.e., each 133 ms. Filters were initialised using the video detected position of their correspondent targets and static matrices Q and R [10], whose values were chosen on the basis of an optimisation step.

Results are described in terms of multiple object tracking precision (MOTP) and multiple object tracking accuracy (MOTA) [49]:

$$MOTP = \frac{\Sigma_t^i d_t^i}{\Sigma_t c_t}, \tag{4}$$

$$MOTA = 1 - \frac{\Sigma_t (mme_t + fp_t + m_t)}{\Sigma_t g_t}, \tag{5}$$

The tracker is considered to have correctly hit the target if the distance between its output and the ground truth is within 0.5 m. Furthermore, the ability of the system to detect speaker ID by localising their voice is measured in terms of Diarisation Error Rate (DER) [46], expressing the speaker error only parameter, i.e., percentage of speech assigned to the wrong speaker.[2]

[2] Note that all, but the last following experiments, are characterised by only one speaker change throughout the entire recorded spoken signal, therefore no clustering technique, typical of SR systems, is used to detect speaker changing points [2]. Hence, a small diarisation error in proximity of the only speaker change point is expected. Nonetheless, it will be taken into account into the final system analysis.

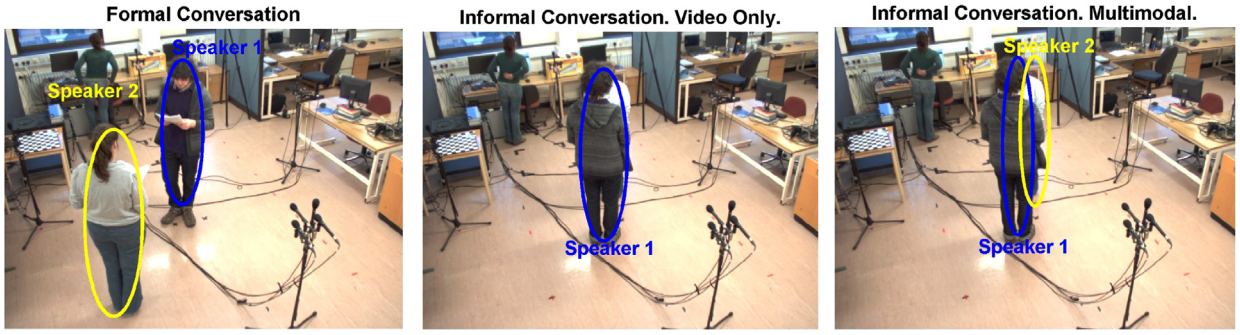


Fig. 7. 'Formal Conversation' and 'Informal Conversation' localisation results. In (a) a '*Formal Conversation*' is shown, the video tracker as well as the multimodal AVT can detect and recognise that there are two targets speaking alternatively and their output is the same. (b) shows an '*Informal Conversation*', targets are so close that the video tracker cannot distinguish them. In (c) the AVT instead correctly localises the actual speaker despite the occlusion. Note that (c) showing two speakers talking contemporary is only for an illustration purpose to highlight AVT that can discriminate identities. In reality, as said, speakers talk in turns.

Table 5
Performance comparison for '*Formal Conversation*' and '*Informal Conversation*' experiments. In (a) we enumerate SRC vs SRC-HE raw speaker position detections. Results are shown in terms of sound source localisation (SSL) accuracy and number of functional evaluations (FEs) calculations. In (b) we present MOTP, MOTA and DER of the joint AVT against single modality trackers.

(a)

Experiment	System	SSL accuracy (%)	FEs
'*Formal*'	SRC-HE	69.07	**23,601**
	SRC [6]	62.50	56,742
'*Informal*'	SRC-HE	51.22	**25,992**
	SRC [6]	47.30	55,821

(b)

Experiment	System	MOTP (m)	MOTA (%)	DER (%)
'*Formal*'	AVT	0.34	90	**7**
	Audio Tracker	0.46	72	15
	Video Tracker	**0.06**	**100**	–
'*Informal*'	AVT	**0.10**	**99**	18
	Audio Tracker	0.30	80	**16**
	Video Tracker	0.53	46	–

3.1. Experiments and results

The first set of experiments is designed to simulate a 60 s long personal and intimate conversation between two people, according to Hall's classification of the social interpersonal distance in relation to physical interpersonal distance [50]. Specifically:

Experiment '*Formal Conversation*' (Fig. 7a) considers two people whom throughout the experiment are separated by a distance of approximately 1 m.

Experiment '*Informal Conversation*' (Fig. 7b, c) considers two people whom throughout the experiment are at a distance of approximately 0.4 m, resulting in an occlusion for the video tracker.

3.1.1. Results

Results are shown in Table 5a for 2 off-line cycles of the SRC-HE detection algorithm against the original SRC. Sound source localisation (SSL) accuracy changes by 4% when adding up extracted video height info. More interesting is the number of functional evaluations (FEs) which on average is reduced by 56% (FEs 56,281 vs 24,797) for the SRC-HE implementation, meaning that narrowing down the space of search our algorithm effectively speeds up the localisation task. In Table 5b performances of the multimodal AVT against single modality trackers are introduced. Results averaged over both the experiments and 100 Monte Carlo runs performance comparison show fusion of audio and video data that improves on single modality trackers when an occlusion occurs (see *'Informal Conversation'* results). In particular, by fusing audio and video the AVT results in a 53% higher MOTA, which is reflected also in a far higher MOTP (80%) with respect to the video-only solution which just half the time of experiments tracks the correct person ID. In fact, the video tracker on its own cannot resolve occlusions. At last, note that the DER is 8% better as expected for the multimodal AVT solution with respect to the audio only system in the *'Formal'* experiment, whereas in the *'Informal'* one is slightly worst. This is due to the fact that the appearance based video ID estimations are completely wrong and they corrupt the multimodal decision. This motivated the further integration of the speaker voice features in the system. Hence, next experiments demonstrate how this algorithm can more robustly maintain and recover tracking ID through occlusions by recognising people voice signatures.

In the following experiments, every dataset is normally 2–5 min long and features people speaking in turns in a non-meeting scenario. The focus is on ID recognition results, rather than on the precision ones, which are obviously not high in such a challenging scenario if any further signal processing is used, as stated also in [51]. Note that we have deliberately recorded our unique set of audiovisual data. This choice was made as classical AV datasets [52,51] are not suitable for our purposes: none provide speaker's voices recording for recognition purposes, as their principal aim is tracking people ID by means of video cues only.

Experiment '*Single Speaker*' (Fig. 8) considers a person speaking along a rectangular trajectory for two times its perimeter, appearing and disappearing from behind an occlusion. His trajectory is shown in the lower left corner of Fig. 8 as detected by the video tracker.

Experiment '*Abandoning*' (Fig. 9) shows a person walking and talking along a rectangular trajectory, as in the previous experiment, disappearing behind an occlusion. Then a second person, who looks like the first one and who is speaking as well, reappears from behind the occlusion and walks along the same trajectory till the point he disappears again.

Experiment '*Crossing*' (Fig. 10) shows two people with very similar appearance 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. Again, trajectory is shown in the lower right corner of Fig. 10.

Results are now presented in Table 6.[3] Here, it is worth noting that there is no real improvement between the AVT, AVT+SR and

[3] For illustrative video results refer to http://visionlab.eps.hw.ac.uk/AV.

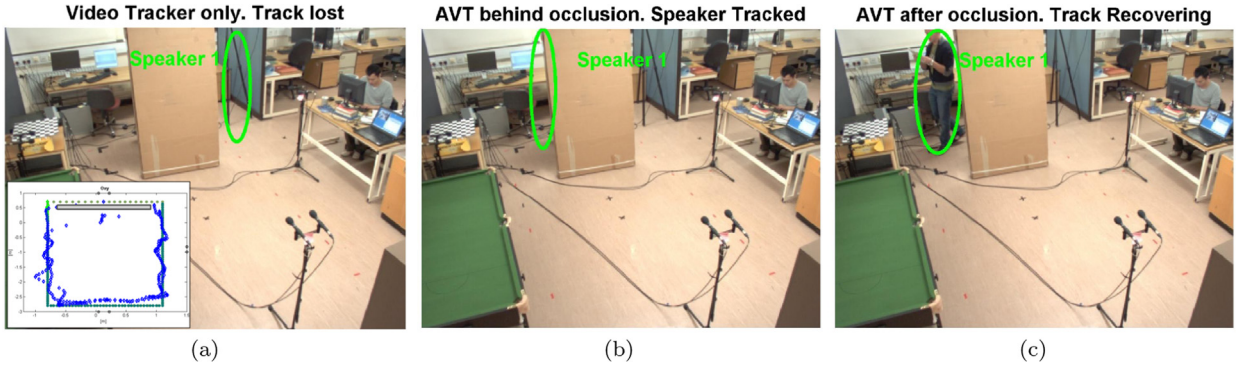


Fig. 8. '*Single Speaker*' *tracking results*. In (a) the video tracker only loses the speaker track when a long occlusion occurs. In turn, (b) shows the AVT correctly locating the speaker through the occlusion at the same time instant of (a). Finally (c) shows speaker track recovering (the video tracker alone is not capable of achieving this result).

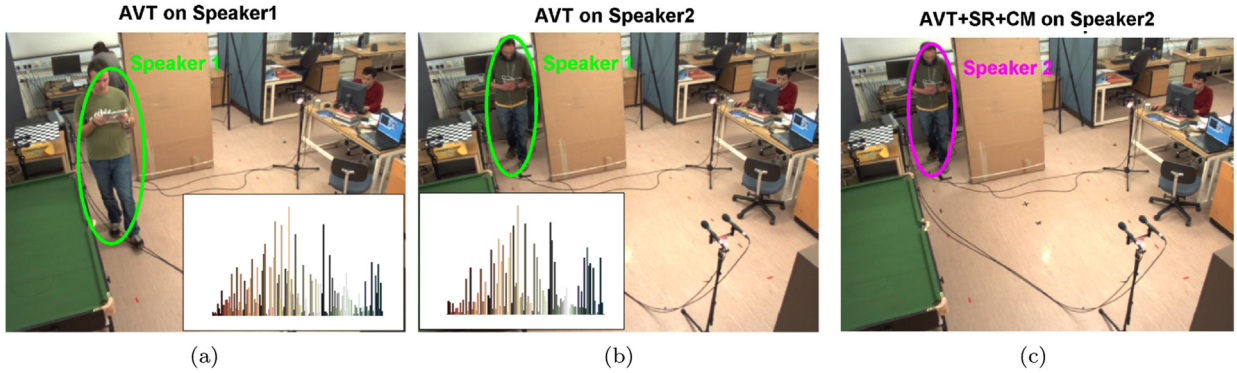


Fig. 9. '*Abandoning*' *tracking results*. (a) Shows the AVT locked onto *Speaker1*. In (b) the other person appears while *Speaker1* has left the scene. The ID assessed by the AVT is still *Speaker1* meaning the AVT cannot make a distinction between IDs. In fact, the video tracker features for the two people, i.e., the histogram of colours at the bottom of (a, b), are very similar. In (c) instead, the AVT+SR+CM solution correctly the person ID is *Speaker2*. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

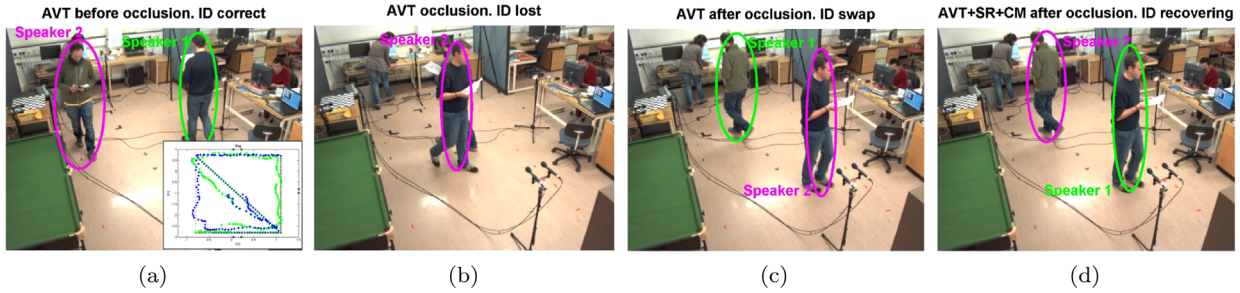


Fig. 10. '*Crossing*' *tracking results*. In (a) the AVT correctly identifies both people ID. In (b) a short term occlusion leads track to merge. This results in (c) in an ID swap as the ellipses colors have exchanged. On the other hand (d) presents the AVT+SR+CM result for the same situation, i.e., correct ID recovering after the occlusion. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

AVT+SR+CM tracking results as the final decision on the speaker voiceprint influences the AVT performance only when their confidences ratio is large. In this case, as the audio tracker is quite confident in its estimation, this ratio is close to 1 for almost the whole trajectory, therefore the DER is the index which shows the benefits of adding speaker voice to the system. Results averaged over all the experiments and 100 Monte Carlo runs show that AVT+SR is better by 13% than AVT, whereas AVT+SR+ CM outperforms the AVT by 27% because of the conversation smoothing prior. Furthermore, a very good result when detecting speaker ID from far-field microphones is achieved, i.e., almost 5% in the worst case. In fact, a 2 m distant microphone normally shows a $\approx 20\%$ DER [38]. In the presented experiments instead, an average 8% DER improvement of the AVT+SR+CM multimodal system over the SR only results (very first column of Table 6) is measured.

4. An indoor surveillance scenario

In Section 2 it is highlighted that the generalised cross correlation audio localisation function is not useful for cross-talking situations such as general security and surveillance (e.g., see Section 2.4). Automatic speaker recognition fails when two people are talking at once [23]. In fact, as the CLEAR 2007 evaluation proved [51], temporal overlaps accounted for more than 70% of error for the speaker ID recognition task. However, the grand aim of automatic surveillance applications is to correctly detect the "dominant" speaker in a large scenario where speech overlaps are highly probable. A speaker is dominant in that their speaking energy is higher with respect to the other people who are talking whom can be instead considered as babble noise. This for example, may be the case of a bank where isolating individual sources is useful for safety reasons. Audio wise, such a task would normally require filtering techniques, beam-forming, non-trivial data association

Table 6
Joint audio–video tracking aided by speaker recognition experiment results. Note that, as it can be inferred in [51] which report the results of the CLEAR 2007 evaluation in real-word interactive seminar scenarios, perfect tracking of multiple people in such challenging situations is still unrealistic. Moreover, this statement refers to meeting room scenarios equipped with large sensor networks, thus more constrained and densely covered with sensors than scenes such as ours. Most significant here are the MOTA the DER indices which express the ability of the system to maintain the correct speaker ID.

Experiment	System	MOTP (m)	MOTA (%)	DER (%)
'*Single Speaker*'	AVT + SR	0.25	94	**4.7**
SR_{DER} = 17.68%	AVT	0.25	94	4.7
'*Abandoning*'	AVT + SR + CM	0.25	91	**2.2**
SR_{DER} = 23.5%	AVT + SR	0.30	84	23.5
	AVT	0.30	84	43.7
'*Crossing*'	AVT + SR + CM	0.47	72	**2.1**
SR_{DER} = 20.6%	AVT + SR	0.55	55	9.6
	AVT	0.56	55	14.8

and blind source separation [23,24]. The concept of dominance in the literature has multiple definitions often used as equivalent [53]. Nevertheless, many studies do agree that speaker loudness or energy and speaking time and rate, as well as gesture based cues are the fundamental features to define dominance [26,54]. In this section, a novel method to automatically detect and localise the actual (dominant) speaker in an enclosed and cluttered scenario is introduced. Specifically, one more video feature is added on top of the system presented in Section 2.7, i.e., optical flow velocity and acceleration and Δ-MFCC, and audio and video are finally combined across semantic data levels. The motivating insight is that *gesturing means speaking*. This implies that observing strong motion implies an audio signal may be causally linked to such a video signal. We seek the correlation between the optical flow in a scene and its associated audio MFCC coefficients (see Section 2.5). Furthermore, audio and video position estimates of the actual speaker given by the AVT+SR+CM (see Section 2.2) are used and combined with correlation cues at the feature level to narrow down the visual space of search of the correlation algorithm, hence reducing the probability of inferring a wrong sound-to-pixel region association. Using this solution we further improve on ID recognition-at-a-distance in a surveillance scenario.

4.1. Feature extraction

4.1.1. Optical flow video features

The video features for AV correlation computation are at first computed as the forward and backward dense optical flow of each image. Then, velocity and acceleration of two adjacent frames motion is calculated. 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:

$$vel = U^+(\mathbf{p}, t), \quad acl = U^+(\mathbf{p}, t) - (-U^-(\mathbf{p}, t)). \tag{6}$$

Hence, we combine the RGB colour, velocity and acceleration of each pixel $\mathbf{p}$ in a frame into a single feature vector: $v_{ij} = (\mathbf{p}, col, vel, acl)$. Thus, we spatially segment every frame using the QuickShift [55] algorithm with $\gamma = 0.25$, $\sigma = 1$ and $\tau = 15$, i.e., the same as in [17]. Furthermore, we compute across frames a K-means [56] spatio-temporal segmentation where K=30. 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 $(k = 1, \ldots, K)$ 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.

Organisation of MFCC audio features: To compute AV correlation, audio features are now represented by the first $n/2$ MFCC (see Section 2.5) coefficients, i.e., signal velocity and their $n/2$ derivatives, i.e., signal acceleration. The audio feature vector $\mathbf{a}$ is a $n \times t$ matrix whose columns correspond to frames. Note that for the following experiments also the audio signal MFCC derivatives (Δ-MFCC) have been computed.

4.2. Audio video correlation

Canonical correlation analysis (CCA) is used to seek audio and video feature vectors correlation, hypothesising that a hidden correspondence between the image motion velocity and the audio MFCC exists, as well as between the image motion acceleration and the MFCC derivatives (Δ-MFCC) [57]. CCA computes a common coordinate system where $\mathbf{a}$ and $\mathbf{v}$ can be projected, and where their maximised correlation is immediately known. This ensures the retrieved video segment to be the one which maximises 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 $\mathbf{a}$ respectively. 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 maximise 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

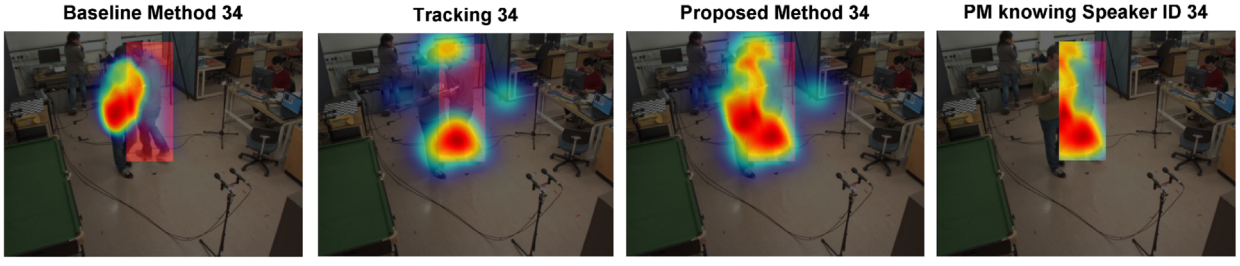


Fig. 11. '*Crossing*' *results*. In (a) the results of the baseline method [17] are presented; (b) shows the result for the tracking results projected onto the image plane. In (c) the output of the proposed method is displayed, whereas (d) presents results when the information about the speaker ID is given. Ground truth is shown in red. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

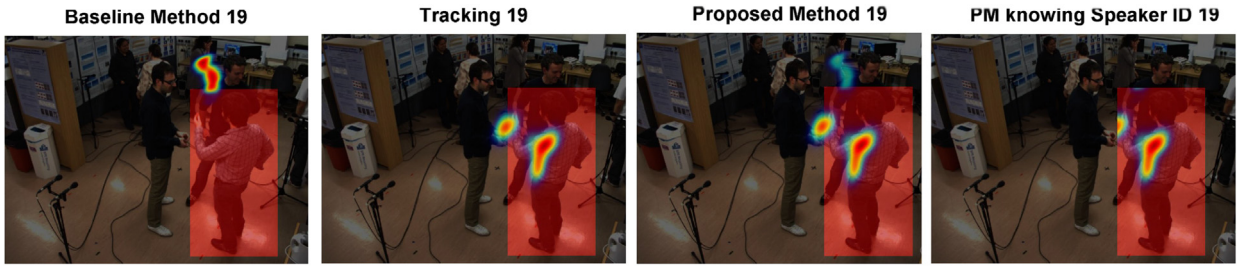


Fig. 12. *'Surveillance' results*. This figure shows the *Second Speaker* talking while the other two people are listening without moving. In (a) the results of the baseline method [17] are given whereas (b) shows the result for the tracking algorithm. (c) presents the output of the proposed method. Finally, (d) shows what is the results when the information about the speaker ID is given. Ground truth is shown in red. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

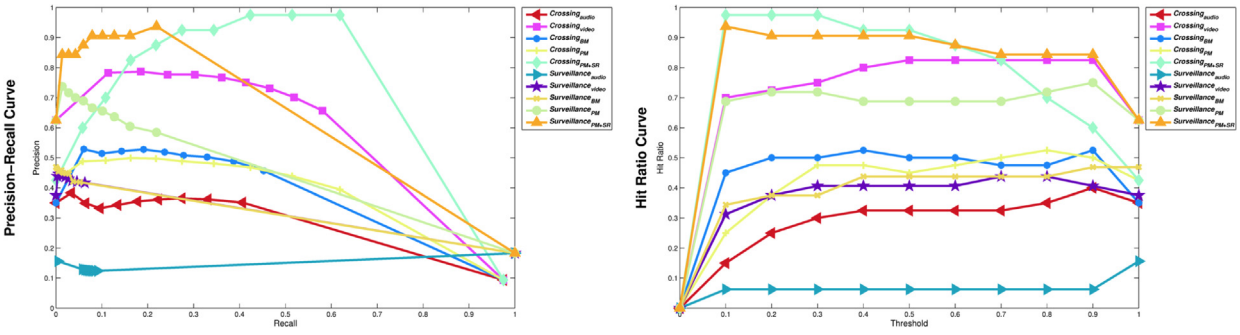


Fig. 13. Comparison to baseline method (BM) [17]. Precision–recall and hit ratio curves for the testing videos averaged over the total number of frames. Results of the proposed method (PM) are compared first against audio and video only results, then against the baseline method (BM) of [17]. Besides, the proposed method aided by the information about the speaker ID, i.e., PM + SR is given, showing that the method is actually suitable for diarisation purposes in cross-talking scenarios whenever a record of the speaker ID over time is kept.

$S_{k_{v^1}}(i_{v^1}, j_{v^1})$ are said to be the ones where the sound is originating. Only the normalised elements of w_{v^1} largest then a predefined threshold is selected, thus those segments are identified by a binary confidence map $Q(S_{k_{v^1}})$ smoothed over space and time by a Gaussian kernel $G(i_{v^1}, j_{v^1})$, whose variance is $\sigma = 5$. Finally a heatmap $\mathcal{H}_{CCA}(i_{v^1}, j_{v^1})$ is overlaid on the selected segments to visually locate sound in an image frame.

Algorithm 4. Dominant speaker detection algorithm.

Input: Audio **a** and video **v** features, AVT positions $\mathbf{x}_{av}$ and identities S_{AV} ▷ see Section 2.2
Output: A visual heatmap $\mathcal{H}_F$ showing the dominant speaker in video

Calculate CCA between **a** and **v** ▷ see Section 4.2
for each frame F_t **do**
  Find w_{v^1} which maximise correlation
  $w_{v^1} \mapsto S_{k_{v^1}}(i_{v^1}, j_{v^1})$
  Define binary map $Q(S_{k_{v^1}})$
  $\mathcal{H}_{CCA}(i_{v^1}, j_{v^1}) \leftarrow Q(S_{k_{v^1}}) * G(i_{v^1}, j_{v^1})$
  $\mathbf{x}_{av}(t) \mapsto \boldsymbol{p}(t)$ ▷ image plane projection
  $\boldsymbol{p}(t) \mapsto S_{k_{\boldsymbol{p}}}(i_{\boldsymbol{p}}, j_{\boldsymbol{p}})$
  Define binary map $Q(S_{k_{\boldsymbol{p}}})$
  $\mathcal{H}_{AVT}(i_{\boldsymbol{p}}, j_{\boldsymbol{p}}) \leftarrow Q(S_{k_{\boldsymbol{p}}}) * G(i_{\boldsymbol{p}}, j_{\boldsymbol{p}})$
  $\mathcal{H}_{F_t} \leftarrow \mathcal{H}_{CCA}(i_{v^1}, j_{v^1}) + \mathcal{H}_{AVT}(i_{\boldsymbol{p}}, j_{\boldsymbol{p}})$
end for
return video $\mathcal{H}_F$

4.3. Fusion of audio–video correlation and audio–video tracking decisions

The integration of the speaker trajectory and the CCA result is carried out at confidence map level. For every frame F_t we project the actual audio source trajectory calculated by the AVT $\mathbf{x}_{av}$ (Section 2) onto the pixel domain $\boldsymbol{p}(t)$. Then, said trajectory points are associated to the k-th segmented region to which they belong $S_{k_{\boldsymbol{p}}}(i_{\boldsymbol{p}}, j_{\boldsymbol{p}})$. Successively, a second confidence map $Q(S_{k_{\boldsymbol{p}}})$ is set for $S_{k_{\boldsymbol{p}}}(i_{\boldsymbol{p}}, j_{\boldsymbol{p}})$, other than the ones already given by the first base eigenvector coefficients as described in Section 4.2. Furthermore, a smoothing Gaussian kernel on the segment $S_{k_{\boldsymbol{p}}}$ is defined, which we denote with $G(i_{\boldsymbol{p}}, j_{\boldsymbol{p}})$. By doing this, the heatmap $\mathcal{H}_{AVT}(i_{\boldsymbol{p}}, j_{\boldsymbol{p}})$ is finally obtained; this has to be overlaid on the image according to the AVT estimation. That done, such a map $(\mathcal{H}_{AVT}(i_{\boldsymbol{p}}, j_{\boldsymbol{p}}))$ is obtained as if it was resulting from an extra first base eigenvector coefficient adding up its contribution to the CCA result, i.e., $\mathcal{H}_{CCA}(i_{v^1}, j_{v^1})$, according to a sum decision rule (see Algorithm 4 and Fig. 1d).

4.4. Experiments and results

Results of dominant speaker detection on real data are now presented and evaluated against audio-only and video-only methods as well as against the baseline method presented by Izadinia et al. [17]. An indoor room where people can freely move and talk together is our experimental region.

Experiment 'Crossing (Fig. 11) is used again (see Section 3.1) to demonstrate in this case that extending correlation techniques to scenarios where distracting motion and occlusion exist can be

done if more cues, as speaker positions, are used.

Experiment 'Surveillance' (Fig. 12) is a recording of several people having a conversation in groups and some passer-by. Speakers are at least 0.5 m far from the microphones. They stand still and move around. The ground-truth consists of the left and right people in the foreground who are having a conversation. Meanwhile a third person, frontal facing in the foreground, is just listening to the conversation and producing some distracting fine motion slightly moving his body on a side. Note that another group of speaking people is in the background. In total, at every moment, 4–5 people are speaking contemporary, resulting in challenging speech interferences. This experiment is designed to demonstrate the power of the method to detect the loudest (dominant) source among a group of speaking people in a cluttered scenario.

At first a qualitative evaluation of results performance against the baseline [17] is given. Fig. 11a shows results of the baseline method applied to the first dataset at the moment of occlusion. The segments corresponding to the AVT tracked position of the actual speaker are given in Fig. 11b, whereas Fig. 11c shows the results of the proposed method. In Fig. 11d knowing the information about the speaker ID, i.e., using the AVT+SR+CM output, the results are ascribed at the current speaker. Fig. 12a shows one frame of '*Surveillance*' for the 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 AVT speaker position $\mathbf{x}_{av}$ (see Section 2.2) corresponding segment (12b), so that the fusion results, despite pointing out the correct speaker, still present false detection trails corresponding to the other people movements (12c). When it is possible to recognise the speaker ID from the AVT+SR+CM, these trails can be actually further filtered out as shown in Fig. 12d.

To measure quantitatively performances of the presented method against [17], the precision–recall measure given in their paper is calculated. Specifically, the moving pixel ground truth is manually defined by selecting those regions of the video which correlated with the dominant speaker's voice. In practice, as this method is ultimately meant to be used for recognition and 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 $Pr = R_c \cap R_d/R_d$, $Rec = R_c \cap R_d/R_c$. Note that, for detection of tracking purposes the size of the ground truth regions cannot be restricted to just the physical (anatomical) joint of a person. Hence, by defining R_d as the detected pixels which actually belong to the current speaker, the goodness of the method in recognising the dominant speaker among other potential speakers can now be evaluated using this metric rather than the DER, as the last is more specific to diarisation systems. 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. At last, to capture the temporal aspect of the methods performances we calculate their hit-ratio curves; we assume a hit that occurs in a frame if $Pr > 0.5$.

Precision–recall curve and hit-ratio curves are shown in Fig. 13. The proposed method+speaker recognition (PM+SR), i.e., the CCA+ AVT+SR+CM precision is higher than the one of both the proposed method (PM), i.e., the CCA+AVT, and the baseline method (BM) over the entire range of recall, although when the recall value increases all curves drop 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. On the other hand, the size of the segments cannot be increased, as clutter will take over the segmentation phase and foreground region would be blended into the background. Nevertheless, the PM+SR solution improves on average on speaker ID recognition through occlusions and interferences by 23%, 59% over audio-only and video-only systems and by 36% over the baseline method [17].

5. Conclusion

In this paper a hierarchical AV tracking and recognition system based on novel audio and video feature integration and fusion is introduced. Specifically, the system carries out a finer independence-based AV localisation and a coarser AV correlation-based scene analysis to robustly track the dominant speaker through general (babble) noise in an open room scenario using a small sensor network. This can be useful in a number of general contexts which range from surveillance applications to the prototypical "cocktail party". Results show that we can rely on low complexity techniques even in unconstrained scenarios, without resorting to more cumbersome audio-only or video-only methods.

5.1. Future work

We highlight that the problem of detecting a speaker in a non-obtrusive fashion and in a natural environment is extremely challenging. And so a number of assumptions had to be made in order to make the problem tractable. We therefore suggest that the following future work could be undertaken: (a) defining a sounding calibration procedure independent from sensors movement, (b) learning gestures associated with specific person roles in a conversation, (c) speeding up optical flow computation using variational methods or sparse techniques, (d) developing an improved speech overlap recognition system to further decrease the diarisation error rate and, consequently, on the dominant speaker detection error, (e) developing a fully probabilistic scheme, i.e., a dynamic Bayes network (DBN) where Fig. 1a would represent one-time slice of the system. The hidden variables would be the speaker position and identity whilst the observables would be the audio and video detected speaker locations, spatio-temporal features and optical flow. With such a fully probabilistic model, any further features may be integrated.

References

[1] S. Ben-Yacoub, Y. Abdeljaoued, E. Mayoraz, Fusion of face and speech data for person identity verification, IEEE Trans. Neural Netw. 10 (5) (1999) 1065–1074, http://dx.doi.org/10.1109/72.788647.

[2] K. Bernardin, R. Stiefelhagen, Audio–visual multi-person tracking and identification for smart environments, in: Proceedings of the 15th International Conference on Multimedia, MULTIMEDIA '07, ACM, New York, NY, USA, 2007, pp. 661–670.

[3] M. Andersson, S. Ntalampiras, T. Ganchev, J.A. Rydell, N. Fakotakis, Fusion of acoustic and optical sensor data for automatic fight detection in urban environments, in: FUSION 2010, 2010.

[4] A. O'Donovan, R. Duraiswami, N. Gumerov, Real time capture of audio images and their use with video, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2007, pp. 10–13. http://dx.doi.org/10.1109/ASPAA.2007.4393037.

[5] J.H. Di Biase, H.F. Silverman, M.S. Brandstein, Robust localization in reverberant rooms, in: M. Brandstein, D. Ward (Eds.), Microphone Arrays, Digital Signal Processing, Springer, Berlin, Heidelberg, 2001, pp. 157–180.

[6] H. Do, H. Silverman, Y. Yu, A real-time srp-phat source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2007, ICASSP 2007, vol. 1, 2007, pp. I-121–I-124. http://dx.doi.org/10.1109/ICASSP.2007.366631.

[7] W. Limprasert, A.M. Wallace, G. Michaelson, Real-time people tracking in a camera network, IEEE J. Emerg. Sel. Top. Circuits Syst. 3 (2) (2013) 263–271.

[8] N. Checka, K. Wilson, M. Siracusa, T. Darrell, Multiple Person and Speaker Activity Tracking with a Particle Filter, vol. 5, 2004, pp. 881–884. http://dx.doi.org/10.1109/ICASSP.2004.1327252.

[9] P. Perez, J. Vermaak, A. Blake, Data fusion for visual tracking with particles, Proc. IEEE 92 (3) (2004) 495–513, http://dx.doi.org/10.1109/

JPROC.2003.823147.
[10] T. Gehrig, K. Nickel, H. Ekenel, U. Klee, J. McDonough, Kalman filters for audio–video source localization, 2005, pp. 118–121. http://dx.doi.org/10.1109/ASPAA.2005.1540183.
[11] Z. Barzelay, Y. Schechner, Harmony in Motion, 2007, pp. 1–8. http://dx.doi.org/10.1109/CVPR.2007.383344.
[12] M. Cristani, M. Bicego, V. Murino, Audio–visual event recognition in surveillance video sequences, IEEE Trans. Multimed. 9 (2) (2007) 257–267, http://dx.doi.org/10.1109/TMM.2006.886263.
[13] D. Gatica-Perez, G. Lathoud, J.M. Odobez, I. McCowan, Audiovisual probabilistic tracking of multiple speakers in meetings, IEEE Trans. Audio Speech Lang. Process. 15 (2) (2007) 601–616.
[14] Y. Lee, R. Mersereau, Data association for people tracking using multiple cameras, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008, 2008, pp. 2585–2588. http://dx.doi.org/10.1109/ICASSP.2008.4518177.
[15] H. Zhou, M. Taj, A. Cavallaro, Target detection and tracking with heterogeneous sensors, IEEE J. Sel. Top. Signal Process. 2 (4) (2008) 503–513, http://dx.doi.org/10.1109/JSTSP.2008.2001429.
[16] S.T. Shivappa, B.D. Rao, M.M. Trivedi, Audio–visual fusion and tracking with multilevel iterative decoding: framework and experimental evaluation, J. Sel. Top. Signal Process. 4 (5) (2010) 882–894.
[17] H. Izadinia, I. Saleemi, M. Shah, Multimodal analysis for identification and segmentation of moving-sounding objects, IEEE Trans. Multimed. 15 (2) (2013) 378–390, http://dx.doi.org/10.1109/TMM.2012.2228476.
[18] M. Barnard, P. Koniusz, W. Wang, J. Kittler, S. Naqvi, J. Chambers, Robust multi-speaker tracking via dictionary learning and identity modelling, IEEE Trans. Multimed. 16 (3) (2014) 864–880.
[19] V. Kiliç, M. Barnard, W. Wang, J. Kittler, Audio assisted robust visual tracking with adaptive particle filtering, IEEE Trans. Multimed. 17 (2) (2015) 186–200, http://dx.doi.org/10.1109/TMM.2014.2377515.
[20] I.D. Gebru, S. Ba, G. Evangelidis, R. Horaud, Audio–visual speech-turn detection and tracking, in: Latent Variable Analysis and Signal Separation: Proceedings of 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25–28, 2015, pp. 143–151. Springer International Publishing, Cham, 2015. http://dx.doi.org/10.1007/978-3-319-22482-4_17.
[21] M. Ohkita, Y. Bando, Y. Ikemiya, K. Itoyama, K. Yoshii, Audio–visual beat tracking based on a state-space model for a music robot dancing with humans, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015, pp. 5555–5560. http://dx.doi.org/10.1109/IROS.2015.7354164.
[22] A. Deleforge, R. Horaud, Y.Y. Schechner, L. Girin, Co-localization of audio sources in images using binaural features and locally-linear regression, IEEE/ACM Trans. Audio Speech Lang. Process. 23 (4) (2015) 718–731, http://dx.doi.org/10.1109/TASLP.2015.2405475.
[23] S. Makino, T. Lee, H. Sawada, Blind Speech Separation, Springer, Netherlands, 2007.
[24] C.Y. Chong, Tracking and data fusion: a handbook of algorithms (Bar-Shalom, Y. et al. 2011) [bookshelf], IEEE Control Syst. 32 (5) (2012) 114–116, http://dx.doi.org/10.1109/MCS.2012.2204808.
[25] A. Pentland, Socially aware computation and communication, IEEE Comput. 38 (3) (2005) 33–40.
[26] H. Hung, Y. Huang, G. Friedland, D. Gatica-Perez, Estimating dominance in multi-party meetings using speaker diarization, IEEE Trans. Audio Speech Lang. Process. 19 (4) (2011) 847–860, http://dx.doi.org/10.1109/TASL.2010.2066267.
[27] D. McNeill, So you think gestures are nonverbal? Psychol. Rev. 92 (3) (1985) 350–371.
[28] N. Campbell, N. Suzuki, Working with Very Sparse Data to Detect Speaker and Listener Participation in a Meeting Corpus, 2006.
[29] B. Gebre, P. Wittenburg, T. Heskes, The gesturer is the speaker, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 3751–3755. http://dx.doi.org/10.1109/ICASSP.2013.6638359.
[30] E. D'Arca, A. Hughes, N. Robertson, J. Hopgood, Video tracking trough occlusions by fast audio source localisation, in: 2013 19th IEEE International Conference on Image Processing (ICIP), 2013.
[31] E. D'Arca, N. Robertson, J. Hopgood, Using the voice spectrum for improved tracking of people in a joint audio–video scheme, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.
[32] E. D'Arca, N. Robertson, J. Hopgood, Look who's talking: detecting the dominant speaker in a cluttered scenario, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.
[33] D. Eberly, Perspective Projection of an Ellipsoid, 1998. URL ⟨http://www.geometrictools.com⟩.
[34] W. Limprasert, A.M. Wallace, G. Michaelson, Accelerated people tracking using texture in a camera network, in: VISAPP (2)'12, 2012, pp. 225–234.
[35] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, IEEE Trans. Acoust. Speech Signal Process. 24 (4) (2003) 320-327, URL ⟨http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1162830⟩.
[36] D. Lee, B. Schachter, Two algorithms for constructing a Delaunay triangulation, Int. J. Comput. Inf. Sci. 9 (3) (1980) 219–242.
[37] T. Gustafsson, B. Rao, M. Trivedi, Source localization in reverberant environments: modeling and statistical analysis, IEEE Trans. Speech Audio Process. 11 (6) (2003) 791–803.
[38] Q. Jin, Y. Pan, T. Schultz, Far-field speaker recognition, in: International Conference on Acoustic, Speech, and Signal Processing (ICASSP), 2006.
[39] J. Dmochowski, J. Benesty, S. Affes, A generalized steered response power method for computationally viable source localization, IEEE Trans. Audio Speech Lang. Process. 15 (8) (2007) 2510–2526, http://dx.doi.org/10.1109/TASL.2007.906694.
[40] M. Fallon, Acoustic Source Tracking Using Sequential Monte Carlo (Ph.D. thesis), Darwin College, University of Cambridge, September 2008.
[41] J. Sohn, N.S. Kim, W. Sung, A statistical model-based voice activity detection, IEEE Signal Process. Lett. 6 (1) (1999) 1–3, http://dx.doi.org/10.1109/97.736233.
[42] D. Reynolds, R. Rose, Robust text-independent speaker identification using gaussian mixture speaker models, IEEE Trans. Speech Audio Process. 3 (1) (1995) 72–83, http://dx.doi.org/10.1109/89.365379.
[43] M. Grimm, K. Kroschel, Robust Speech Recognition and Understanding, I-Tech Education and Publishing, Vienna, 2007.
[44] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. Ser. B (Methodol.) 39 (1) (1977) 1–38.
[45] S. Sadjadi, M. Slaney, L. Heck, Msr Identity Toolbox v1.0: A Matlab Toolbox for Speaker-Recognition Research. ⟨http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox⟩. Doi http://research.microsoft.com/apps/pubs/default.aspx?id=205119.
[46] J.G. Fiscus, J. Ajot, M. Michel, J.S. Garofolo, The Rich Transcription 2006 Spring Meeting Recognition Evaluation, NIST, 2006.
[47] F. Roli, J. Kittler, G. Fumera, D. Muntoni, An experimental comparison of classifier fusion rules for multimodal personal identity verification systems, in: F. Roli, J. Kittler (Eds.), Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 2364, Springer, Berlin, Heidelberg, 2002, pp. 325–335.
[48] M. Horvat, K. Jambrosic, H. Domitrovic, Methods of measuring the reverberation time, in: Proceedings of the Third AAAA Congress, Graz, Austria, 2007.
[49] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: the clear mot metrics, J. Image Video Process. 2008 (2008) 1:1–1:10, http://dx.doi.org/10.1155/2008/246309.
[50] J.M. Matthias Wolfel, Distant Speech Recognition, Wiley, 2009.
[51] R. Stiefelhagen, J. Garofolo, Multimodal technologies for perception of humans: international evaluation workshops clear 2007 and rt 2007. Lecture Notes in Computer Science, no. 4625, Springer, Baltimore, MD, USA, 2008.
[52] G. Lathoud, J.M. Odobez, D. Gatica-Perez, Av16.3: an audio–visual corpus for speaker localization and tracking, in: MLMI, 2004, pp. 182–195.
[53] Q.M. Rojas, D. Masip, J. Vitria, Predicting dominance judgements automatically: a machine learning approach, in: 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), 2011, pp. 939–944. http://dx.doi.org/10.1109/FG.2011.5771377.
[54] N. Dunbar, J. Burgoon, Perceptions of power and interactional dominance in interpersonal relationships, J. Soc. Pers. Relationsh. 22 (2) (2005) 207–233.
[55] A. Vedaldi, S. Soatto, Quick shift and kernel methods for mode seeking, in: D. Forsyth, P. Torr, A. Zisserman (Eds.), Computer Vision – ECCV 2008, Lecture Notes in Computer Science, vol. 5305, Springer, Berlin Heidelberg, 2008, pp. 705–718.
[56] J. MacQueen, Some Methods for Classification and Analysis of Multivariate Observations. URL ⟨http://projecteuclid.org/euclid.bsmsp/1200512992⟩, 1967.
[57] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936) 321–377, http://dx.doi.org/10.2307/2333955.
[58] S. Shivappa, M. Trivedi, B. Rao, Hierarchical audio–visual cue integration framework for activity analysis in intelligent meeting rooms, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009, 2009, pp. 107–114.