Target detection and tracking with heterogeneous sensors

Huiyu Zhou, Murtaza Taj, Andrea Cavallaro

***Abstract*—We present a multimodal detection and tracking algorithm for sensors composed of a camera mounted between two microphones. Target localization is performed on color-based change detection in the video modality and on Time Difference of Arrival (TDOA) estimation between the two microphones in the audio modality. The TDOA is computed by multi-band Generalized Cross Correlation (GCC) analysis. The estimated directions of arrival are then post-processed using a Riccati Kalman filter. The visual and audio estimates are finally integrated, at the likelihood level, into a particle filter (PF) that uses a zero-order motion model, and a Weighted Probabilistic Data Association (WPDA) scheme. We demonstrate that the Kalman filtering (KF) improves the accuracy of the audio source localization and that the WPDA helps to enhance the tracking performance of sensor fusion in reverberant scenarios. The combination of multi-band GCC, KF and WPDA within the particle filtering framework improves the performance of the algorithm in noisy scenarios. We also show how the proposed audiovisual tracker summarizes the observed scene by generating metadata that can be transmitted to other network nodes instead of transmitting the raw images and can be used for very low bit rate communication. Moreover, the generated metadata can also be used to detect and monitor events of interest.**

***Index Terms*—Multimodal detection and tracking, Kalman filter, particle filter, heterogeneous sensors, low bit rate communication.**

I. Introduction

Localization and object tracking using audiovisual measurements is an important module in applications such as surveillance and human-computer interaction. The effectiveness of fusing video and audio features for tracking was demonstrated in [1], [2], [3]. The success of the fusion strategy is mainly because each modality may compensate for the weaknesses of the other or can provide additional information ([4], [5]). For example, a speaker identified via audio detection may trigger the camera zooming in a teleconference. The main challenges for audiovisual localization are reverberations and background noise. Therefore, the audiovisual sensor networks (with camera and microphone arrays) have been used to address these problems using a variety of sensor configurations. Fig. 1 shows a summary of these configurations, which range from a single microphone-camera pair to single or stereo cameras with stereo, circular arrays or linear arrays of microphones. Camera-microphone pairs are used for speaker detection in environments with limited reverberation under the assumption that the speakers face the microphone [6]; single or stereo cameras with multiple microphones are used in meeting rooms and teleconferencing ([7], [8]). Gatica-Perez *et al.* use cameras and eight microphones to capture interactions in meeting scenarios ([9]). A Stereo Audio and Cycloptic Vision (STAC) sensor, composed of a single camera mounted between two microphones (Fig. 1(b-c)), makes the designed system simpler, cheaper and portable and is used in this work. STAC sensors are used to perform audiovisual tracking with a probabilistic graph model and fusion by linear mapping ([10]) or with particle filtering ([11]). The cost of using such a simple sensor against an array of microphones is its sensitivity to noise and reverberations.

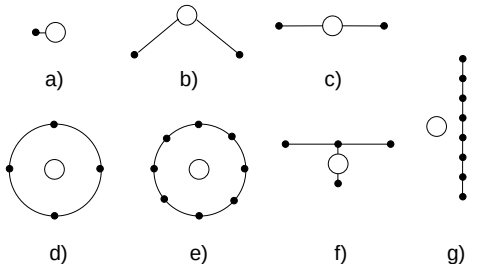


Fig. 1. Examples of sensor configurations for audiovisual object detection and tracking (filled circles indicate microphones; empty circles indicate cameras): (a) single microphone-camera pair; (b-c) Stereo Audio and Cycloptic Vision (STAC) sensors; (d-e) circular microphone array with single camera; (f) triangular microphone array with single camera; (g) linear microphone array with single camera.

In this paper, we present a target detection and tracking algorithm based on the measurements of a STAC sensor. The novelty of this approach is on the use of Kalman filter to improve the accuracy of audio source localization and on a new fusion strategy based on Weighted Probabilistic Data Association filter (WPDA), which associates the hypotheses and the measurements with a real target. WPDA takes into account the weighted probability of the detections to increase the importance of reliable audiovisual measurements in each iteration. Furthermore, we also introduce a reverberation filtering technique based on multi-band frequency analysis to reduce the erroneous peaks due to noise and reverberations in the *Generalized Cross Correlation Phase Transform* (GCC-PHAT). To further smooth the audio source localization we apply a Riccati Kalman filter. Moreover, image appearance (a 4D state space and color) of the objects in two-dimension is used to further enhance the validation of a track. These two processes are accommodated as independent observations in a particle filter that applies a WPDA in the decision stage. Par-

The authors acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC), under grant EP/D033772/1. H. Zhou, M. Taj and A. Cavallaro are with the Multimedia and Vision Group, Queen Mary University of London, e-mail: {huiyu.zhou,murtaza.taj,andrea.cavallaro}@elec.qmul.ac.uk

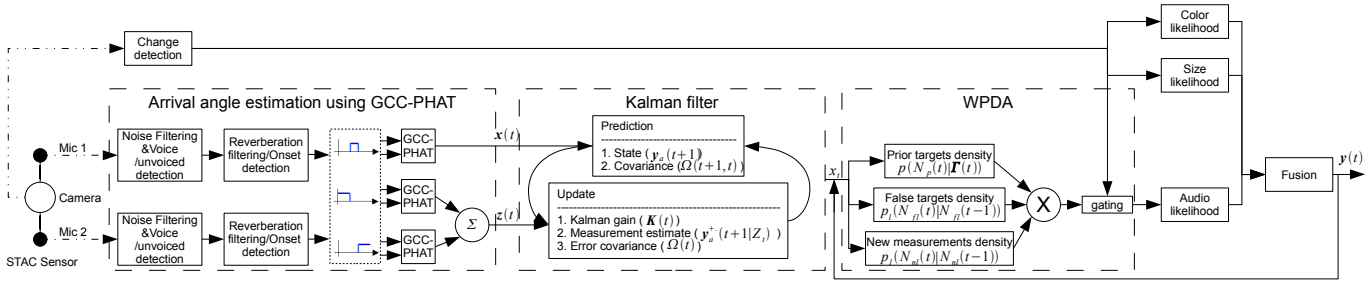


Fig. 2. Flowchart of the proposed audiovisual tracking algorithm.

ticle filtering is here applied due to its robust performance in processing multimodal information ([1]). The block diagram of the proposed audiovisual detection and tracking algorithm is shown in Fig. 2.

This paper is organized as follows: Section II introduces the related work. Section III presents the proposed audiovisual tracking algorithm. This is followed by the description of the experimental work in Section IV. Finally, conclusions and future work are drawn in Section V.

II. Related work

A significant amount of work has been reported on detecting and tracking single or multiple moving objects using Kalman filter (KF) ([12], [13]), particle filter (PF) ([3], [14], [15]) and variants of probabilistic data association (PDA) ([16], [17]). Multimodal multi-sensor configurations are used for object tracking ([14], [18], [19], [20]) to compensate for failure of each modality. Tracking can be performed using the video modality only ([21]–[25]), the audio modality only ([26]–[29]) or using audio and video simultaneously ([1], [3], [7], [9], [14], [30]–[33]).

Many approaches are addressing audiovisual tracking for smart multimodal meeting rooms ([8], [9], [30], [31], [35], [39]). Tracking of multiple non-simultaneous speakers is described in [33] whereas in [30], [35] the authors track a single speaker using variants of the classical particle filter in smart rooms. In meeting scenarios, interaction of multiple speakers is modeled using mixed-state dynamical graph models ([9], [39]). Similarly, turn taking events can be recognized by semantic analysis of the scene using trajectories generated via the audiovisual particle filter ([31]). Moving speakers can be tracked using Bayesian hidden variable sequence estimation ([8]). This approach is equivalent to extending the Bayesian network to a dynamic Bayesian network in order to account for the dynamics of the state of the sound sources ([8]). Face and upper body parts can be detected using contour extraction by performing edges and motion analysis and then combined with audio detection in particle filter framework ([1], [11], [40]). Gehrig *et al.* ([12]) apply audio detection to generate face positions that could also be observed by multiple cameras.

Unlike meeting rooms, more challenging scenarios are uncontrolled environments (e.g., indoor and outdoor surveillance) where it is not practical to use complex microphone configurations requiring sophisticated hardware for installation and synchronization. Recently, simple configurations (e.g., one camera and two microphones) were adapted using Time-Delay Neural Networks (TDNN) and Bayesian Networks (BN) ([36]). Audio features are detected by computing the spectrogram coefficients of foot-step sounds via the Short-Time Fourier Transform (STFT). TDNN are then used to fuse the audio and visual features, where the latter is generated using a modified background subtraction scheme. However, it is unclear how object detection is achieved when visual features are unavailable. Moreover, this algorithm rely on a pre-training stage that leads to intensive processing. Sensors similar to STAC sensors, with a pan, tilt and zoom (PTZ) camera are used to detect speakers in the near field with unscented particle filter for data fusion ([35]). Since STAC sensors are sensitive to reverberations, multi-band analysis and precedence effect are used in [15] and [32] to preprocess the audio signal before applying particle filter for data fusion. When the target dynamics and measurements are linear and Gaussian, a closed-form solution can be uniquely determined. Such target dynamics can be modeled using the Kalman filter to fuse the audio and video modalities ([13]). The Kalman filter cannot effectively handle nonlinear and non-Gaussian models ([13], [14], [17]), although an extended Kalman filter can linearize models with weak nonlinearities around the state estimate ([17], [41]). Particle filter is a popular choice to address nonlinearity and non-Gaussianity ([1], [11], [15], [35], [40]). Cevher *et al.* ([3]) use a particle filter to combine acoustic and video information in a single state space. The Kullback-Leibler divergence measure is adopted to decrease the probability of divergence of the individual modalities. Vermaak *et al.* ([1]) combine a particle filter based head tracking with the acoustic time difference of arrival (TDOA) measurements to track speakers in a room. Bregonzio *et al.* ([32]) use color-based change detection and TDOA for generic object tracking. In most approaches the detection mechanism uses TDOA or beamforming for audio detection. Speakers can also be detected using a recognition mechanism. In such case Mel-Frequency Cepstral Coefficients are used for speech recognition and video recognition can be done using linear subspace projection methods ([38]). A summary of multimodal tracking algorithms is presented in Table I.

III. Audiovisual tracking

The problem of audiovisual tracking involves the estimation of the arrival angle of the audio signal, video detection, filtering and smoothing of the two modalities, fusion and

TABLE I
MULTIMODAL TRACKING ALGORITHMS. (KEY: PF=PARTICLE FILTER, KF=KALMAN FILTER, DKF=DECENTRALIZED KF, LDA=LINEAR DISCRIMINANT ANALYSIS, TDNN=TIME DELAY NEURAL NETWORKS, GM= GRAPHICAL MODELS, MFA=MULTI-FEATURE ANALYSIS, HCI=HUMAN COMPUTER INTERACTION).

References	Sensor types	Algorithms	Application
[8]	Stereo camera and circular microphone array	PF	Multimodal user interface
[33]	2 cameras and 4 microphone arrays	PF	Indoor multiple person tracking
[3]	Camera and 10 element uniform circular array	PF	Outdoor surveillance
[7]	Panoramic camera and 4 omni-microphones	MFA	Face detection
[34]	Wide-angle camera and a microphone array	I-PF	Meeting rooms
[35]	PTZ camera and 2 microphones	PF	Teleconferencing
[6]	Camera and microphone	TDNN	Lip reading, HCI
[10]	Camera and 2 microphones	GM	Indoor environment
[36]		TDNN	Surveillance
[1], [11], [15], [32]		PF	Surveillance and teleconferencing
[12], [13], [37]	Multiple cameras and microphone arrays	KF, DKF	Smart rooms
[38]		LDA	Smart rooms
[9], [30], [31], [39]		PF	Meeting rooms

finally joint state estimation. Let the target state be defined as $\mathbf{y}(t) = (x, y, w, h, \mathcal{H})$, where (x, y) is the center of the ellipse approximating the object shape, (w, h) are the width and height of the bounding box and $\mathcal{H}$ is the color histogram of the object. Let the measurement equation be $\mathbf{z}(t) = \mathbf{U}\mathbf{y}(t) + \mathbf{n}(t)$, where $\mathbf{U}$ is transition matrix and $\mathbf{n}(t)$ is an independent and identically distributed stochastic process. Moreover, let $\mathbf{y}_a = (x_a)$ denote the audio only state, let $\mathbf{z}_a$ denote the audio only measurements and let $\mathbf{y}_v = (x_v, y, w, h, \mathcal{H})$ denote the video only state.

The problem of multimodal tracking can be formulated as a state estimation against time t, given the audio and visual observations. Particle filters estimate an approximate state $\mathbf{y}(t)$ on the basis of all the previous and current observations $\mathbf{z}(1:t)$, which contain visual and audio features. The tracking problem aims to estimate the posterior probability $p(\mathbf{y}(t)|\mathbf{z}(1:t))$ using a recursive prediction and update strategy ([42]). In the prediction stage, the prior probability distribution function is $p(\mathbf{y}(t+1)|\mathbf{z}(1:t)) = \int p(\mathbf{y}(t+1)|\mathbf{y}(t))p(\mathbf{y}(t)|\mathbf{z}(1:t))d\mathbf{y}(t)$, where $p(\mathbf{y}(t+1)|\mathbf{y}(t))$ is determined if an observation is available, and $p(\mathbf{y}(t)|\mathbf{z}(1:t))$ is known from the previous iteration.

The state update equation is a zero-order motion model defined as $\mathbf{y}(t+1) = \mathbf{y}(t) + \mathcal{N}(\mu, \sigma)$, where $\mathcal{N}$ is a Gaussian noise. Given the measurements $\mathbf{z}(t)$, the update step is based on the Bayes' rule:

$$p(\mathbf{y}(t)|\mathbf{z}(1:t)) = \frac{p(\mathbf{z}(t)|\mathbf{y}(t))p(\mathbf{y}(t)|\mathbf{z}(1:t-1))}{\int p(\mathbf{z}(t)|\mathbf{y}(t))p(\mathbf{y}(t)|\mathbf{z}(1:t-1))d\mathbf{y}(t)}. \quad (1)$$

The posterior probability $p(\mathbf{y}(t)|\mathbf{z}(1:t))$ can be approximated as a set of N Dirac functions: $p(\mathbf{y}(t)|\mathbf{z}(1:t)) \approx \sum_{i=1}^{N} \omega^i(t)\delta(\mathbf{y}(t) - \mathbf{y}^i(t))$, where $\omega^i(t)$ are the weights assigned to the particles, computed as

$$\omega^i(t) \propto \omega^i(t-1)\frac{p(\mathbf{z}(t)|\mathbf{y}^i(t-1))p(\mathbf{y}^i(t)|\mathbf{y}^i(t-1))}{q(\mathbf{y}^i(t)|\mathbf{y}^i(t-1), \mathbf{z}(t))}, \quad (2)$$

where $q(\cdot)$ is the proposal distribution. To avoid degeneracy whilst discarding the particles with lower weights we apply resampling ([43]). Then, we obtain the weights as $\tilde{\omega}^i(t-1) = \omega^i(t-1)/(N \cdot a^i(t-1))$, where $a^i(t-1) = \omega^i(t-1)$ because of the uniformness constraint. Thus, we obtain

$$\omega^i(t) \propto p(\mathbf{z}(t)|\mathbf{y}^i(t)), \quad (3)$$

i.e., weights are proportional to the likelihood of the observations. The likelihood $p(\mathbf{z}(t)|\mathbf{y}(\mathbf{t}))$ is the product of the likelihood estimates of the various features ($p(\mathbf{z}(t)|\mathbf{y}^\mathbf{i}(\mathbf{t}))$), as detailed in Section III-C.

A. Audio source localization using Kalman filtering

The two microphones of a STAC sensor measure the acoustic signals at different time instants (Fig. 3 and Fig. 4). The arrival angle of this signal can be estimated using Time Difference of Arrival (TDOA) or steered beamforming ([3], [26], [44]–[46]). Let us assume that one target at a time emits sound and that the sound is generated in the direction of the microphones. Then TDOA can be utilized to compute the delay τ of the wave between the reference microphone, M_1, and the second microphone, M_2 (Fig. 3). We preprocess the audio signals before the estimation of the arrival angle to reduce the effect of noise and reverberations. Similarly to background modeling in video ([47]), we assume that the first 200ms of the audio signal contains changes that are due to noise only. Next we filter unvoiced segments by analyzing the zero-crossing rate ([48]). For each windowed audio segment, zero-crossings are counted and the mean, μ, and the standard deviation, σ, are used to define high zero-crossing rate as $\varsigma > \mu + \alpha\sigma$, where ς is the zero-crossing rate for the windowed segment and α is a weight dependent on the sensor and the environment. Based on this threshold, unvoiced signal segments are filtered, as shown in Fig. 5. Using the same initial 200ms interval, we compute the signal noise level as

$$Ns_j = \frac{1}{F_n\ S_n}\sum_{n=1}^{F_n}\sum_{m=1}^{S_n} g_j(n, m), \quad with\ j = 1, 2 \quad (4)$$

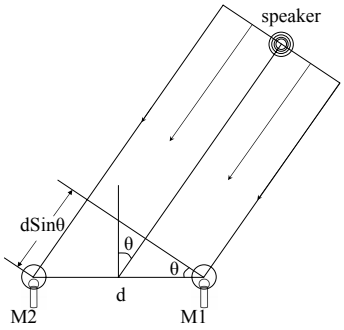


Fig. 3. Source-receiver geometry for a STAC sensor in the far field. The distance between the microphones M_1 and M_2 is denoted by d and the arrival angle by θ. The sound wave has to travel an additional distance of $d\sin\theta$ to reach microphone M_2.

where F_n is the number of audio frames, S_n is the number of samples in a frame and g_j is the audio frame from the j^{th} microphone containing 1764 (0.04 seconds at 44.1KHz) samples. The noise level is then used to detect onset frames with significant signal component without reverberation. The apparent location of a sound source largely depends on the initial onset of the sound, a phenomenon known as precedence effect or law of the first wavefront ([49]–[51]).

Let G_j be the signal levels computed as:

$$G_j(n) = \frac{1}{S_n}\sum_{i=1}^{S_n} g_j(i), \quad with\ j = 1, 2. \tag{5}$$

The frames are considered as onset frames if $G_j(n) > \beta N s_j$, where β is the noise weight. After each onset detection, the next $th_o = 6$ frames are considered as signal component while rest are assumed to be due to reverberation, and hence ignored until a null frame ($G_j(n) \leq \beta N s_j$) is detected.

We estimate the arrival angle using a *multi-band frequency analysis* to further reduce any residual reverberation effect. The angular estimates are conducted in the low ($0 - 400$Hz), middle ($400 - 960$Hz) and high-frequency ($960 - 1600$Hz) bands. We then divide these bands into two groups, namely Group 1 (middle frequency band) and Group 2 (low and high frequency bands). This division is done as some materials have higher absorptivity at high frequencies, whereas others may have higher absorptivity at lower frequencies ([52]). Let the state and the observation vectors represent Group 1 and Group 2, respectively. This means that the audio state vector (i.e. the x position of the target) is obtained from the middle frequency band whereas the observation vector is obtained from the low and high frequency estimates. To determine the audio state $\mathbf{y}_a = (x_a)$ using the estimated delay τ we apply a *Riccati Kalman filter*, as it offers better performance in noisy environments ([53]).

The Kalman filter works with a state space model consisting of a process and an observation equation:

$$\begin{cases} \mathbf{y}_a(t+1) = \mathbf{A}(t)\mathbf{y}_a(t) + \phi_1(t) \\ \mathbf{z}(t) = \mathbf{C}(t)\mathbf{y}_a(t) + \phi_2(t) \end{cases}, \tag{6}$$

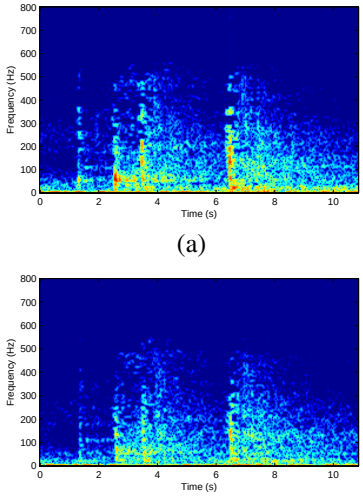


Fig. 4. Sample spectrograms from (a) microphone 1 (M_1) and (b) microphone 2 (M_2). The signal at M_2 is delayed and attenuated.

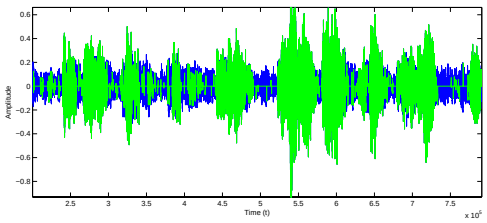


Fig. 5. Example of voiced/unvoiced segment detection for Sequence 1 (SC 1) using $\alpha = 0.8$. The original signal is shown in blue and the filtered signal is shown in green.

where

$$\mathbf{A}(t) = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}, \tag{7}$$

and $\mathbf{C} = [1, 0]$ is the transition matrix. $\phi_1(t)$ and $\phi_2(t)$ are two noise terms that are assumed to be zero mean, white Gaussian random vectors with covariance matrices defined by

$$\Psi(\phi_j(t){\phi_j}^T(k)) = \begin{cases} \mathbf{Q}_j(t) & \text{for } t = k \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

where $j = 1, 2$; $\mathbf{Q}_1$ is the covariance matrix of the process noise, and $\mathbf{Q}_2$ is the covariance matrix of the measurement noise. $\phi_1(t)$ and $\phi_2(k)$ are statistically independent and therefore $\Psi(\phi_j(t){\phi_j}^T(k)) = 0$ for all $t \neq k$.

Let $\mathbf{y}_a^+(t|Z_a(t-1))$ be the predicted state estimate of $\mathbf{y}_a(t)$ deduced from all observations $Z_a(t-1) = (\mathbf{z}_a)_{j=0}^{t-1}$ up to time $t-1$. The predicted observations is then expressed as

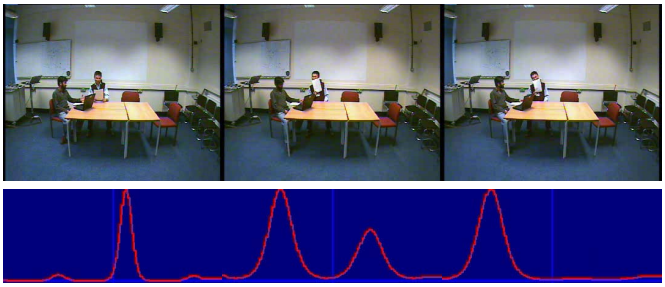
$$\mathbf{z}_a^+(t|Z_a(t-1)) = \mathbf{C}(t)\mathbf{y}_a^+(t|Z_a(t-1)). \tag{9}$$

The innovation is the difference between the actual and predicted observations:

$$\beta(t) = \mathbf{z}_a(t) - \mathbf{C}(t)\mathbf{y}_a^+(t|Z_a(t-1)). \tag{10}$$

The correlation matrix of the innovation sequence as

$$\mathbf{p}(t) = \mathbf{C}(t)\Omega(t, t-1)\mathbf{C}^T(t) + \mathbf{Q}_2(t), \tag{11}$$



(a) (b) (c)

Fig. 6. Examples of audio source mislocalization. (a): The speaker on the left is talking, but the peak indicates the person on the right. (b) and (c): The speaker on the right is talking, but the peak indicates the person on the left.

and the covariance matrix

$$\Omega(t,t-1)=\mathcal{E}[(\mathbf{y}_a(t)-\mathbf{y}_a^+(t|Z_{t-1}))(\mathbf{y}_a(t)-\mathbf{y}_a^+(t|Z_{t-1}))^T]. \quad (12)$$

The Kalman gain is defined as

$$\mathbf{K}(t)=\mathbf{A}(t+1,t)\Omega(t,t-1)\mathbf{C}^T(t)\mathbf{P}^{-1}(t). \quad (13)$$

To compute the Kalman gain, we need to estimate $\Omega(t+1,t)$, which is

$$\begin{cases} \Omega(t+1,t)=\mathbf{A}(t+1,t)\Omega(t)\mathbf{A}^T(t+1,t)+\mathbf{Q}_1(t) \\ \Omega(t)=[\mathbf{I}-\mathbf{A}(t,t+1)\mathbf{K}(t)\mathbf{C}(t)]\Omega(t,t-1) \end{cases}, \quad (14)$$

where $\mathbf{I}$ is the identity matrix. Finally, the state estimate can be updated according to the Kalman gain and innovation ([54]), that is

$$\mathbf{y}_a^+(t+1|Z_a(t))=\mathbf{A}(t+1,t)\mathbf{y}_a^+(t|Z_a(t-1))+\mathbf{K}(t)\beta(t). \quad (15)$$

B. Weighted probabilistic data association

Although the Kalman filter reduces the localization discrepancy, the estimated GCC peaks can deviate from the real audio source positions due to various noise components. As a result, the performance of the entire multimodal tracker will deteriorate, especially in the presence of adjacent objects (Fig. 6). To minimize the discrepancy between the real and the estimated positions, we propose a strategy that associates the hypotheses and the measurements with a real target, using a Weighted Probabilistic Data Association (WPDA) algorithm. Unlike PDA and Joint-PDA, WPDA takes into account a weighted probability of the detections in each iteration to increase the importance of reliable audiovisual measurements, based on the prior estimates and on validation data, and further weaken the unreliable hypotheses. The correspondence between the audio and the video modality is done using a Gaussian reliability window. Only the measurements falling within this region are considered to be valid.

Let $\mathbf{p}^t$ denote the probability of the prediction $\mathbf{\Gamma}(t)$, given the measurements $\mathbf{z}(t)$ up to time t:

$$\mathbf{p}^t \propto p(\mathbf{\Gamma}(t)|\mathbf{z}(t)). \quad (16)$$

The prediction $\mathbf{\Gamma}(t)$ can be obtained based on the prior $\mathbf{\Gamma}(t-1)$ and the association hypotheses $\Xi(t-1)$ for the current measurements. $\Xi(t-1)$ associates each measurement $\mathbf{z}(t)$ with a target. $\mathbf{p}^t$ is intractable due to the unknown association. Instead, we can estimate $p(\mathbf{\Gamma}(t),\Xi(t)|\mathbf{z}(t))$ using the Bayes' rule as

$$\begin{aligned} &p(\mathbf{\Gamma}(t),\Xi(t)|\mathbf{z}(t))= \\ &\quad = \quad c_1 p(\mathbf{z}(t)|\mathbf{\Gamma}(t),\Xi(t))p(\Xi(t)|\mathbf{\Gamma}(t))p(\mathbf{\Gamma}(t)), \end{aligned} \quad (17)$$

where c_1 is a normalizing factor and $p(\mathbf{z}(t)|\mathbf{\Gamma}(t),\Xi(t))$ can be expressed as

$$\begin{aligned} &p(\mathbf{z}(t)|\mathbf{\Gamma}(t),\Xi(t)) \propto \Pi_{n=1}^{M_t}\psi(n)= \\ &\quad = \quad cp(\mathcal{D}|\mathbf{y}(t))p(\mathcal{C}|\mathbf{y}(t))p(\mathcal{A}|\mathbf{y}(t)), \end{aligned} \quad (18)$$

where $\psi(\cdot)$ is an intermediate function, and M_t is the number of the measurements obtained by different sensors. Eq. (18) shows that the conditional probability $p(\mathbf{z}(t)|\mathbf{\Gamma}(t),\Xi(t))$ depends on the multiplication of the posterior distributions of different measurements.

The third term of the right hand side of Eq. (17), $p(\Xi(t)|\mathbf{\Gamma}(t))$, is the probability of a current data association hypotheses, given the previous prediction and estimation. Let $N_p(t)$, $N_f(t)$, and $N_n(t)$ be the measurements associated with the prior, false and new targets respectively. Considering a binomial distribution for $N_p(t)$ and the positive side of a Gaussian distribution for $N_f(t)$ and $N_n(t)$, we can express $p(N_p(t),N_f(t),N_n(t)|\mathbf{\Gamma}(t))$ as

$$\begin{aligned} &p(N_p(t),N_f(t),N_n(t)|\mathbf{\Gamma}(t))= \\ &\quad = \quad p(N_p(t)|\mathbf{\Gamma}(t))p(N_f(t),N_n(t)|N_p(t),\mathbf{\Gamma}(t)) \\ &\quad = \quad p(N_p(t)|\mathbf{\Gamma}(t))p(N_f(t)|N_p(t),N_n(t),\mathbf{\Gamma}(t))\times \\ &\qquad \times p(N_n(t)|N_p(t),N_f(t),\mathbf{\Gamma}(t)), \end{aligned} \quad (19)$$

where

$$\begin{aligned} &p(N_p(t)|\mathbf{\Gamma}(t))= \\ &\quad = \quad \begin{pmatrix} N_t(t) \\ N_d(t) \end{pmatrix} p_d(t)^{N_d(t)}(1-p_d(t))^{N_t(t)-N_d(t)}, \end{aligned} \quad (20)$$

where $N_t(t)$ and $N_d(t)$ are the numbers of previously known and currently detected targets, respectively. $p_d(t)$ can be determined using its current probability $p_{det}^{(l)}(t)$ and prior probability $p_{det}^{(l)}(t-1)$:

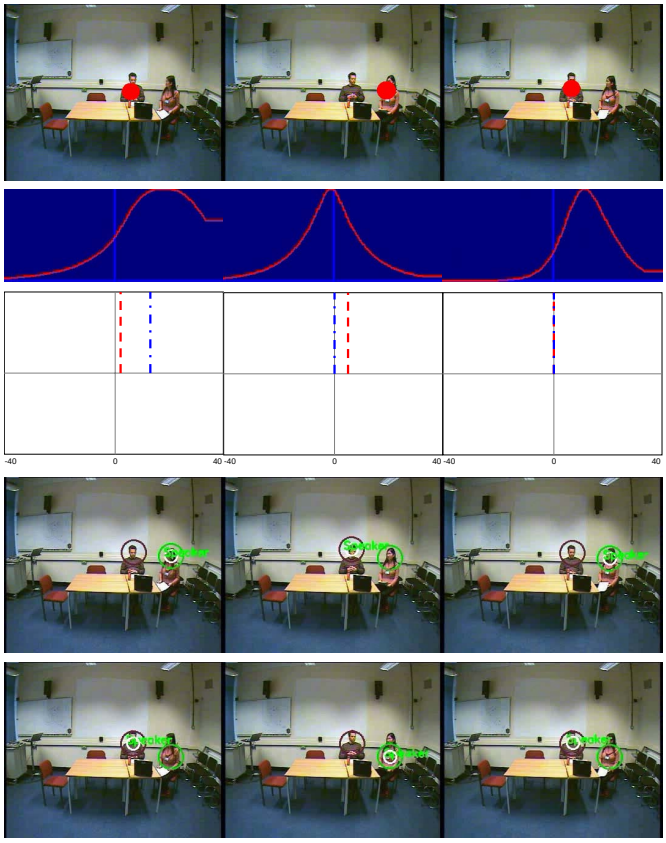
$$p_d(t)=\begin{cases} p_{det}^{(l)}(t) & \text{if } p_{det}^{(l)}(t-1)\le p_{det}^{(l)}(t) \\ p_{det}^{(l)}(t-1) & \text{otherwise} \end{cases} \quad (21)$$

$$p_{det}^{(l)}(t)=\frac{p(\mathcal{D}_l|\mathbf{y}_l(t))p(\mathcal{C}_l|\mathbf{y}_l(t))p(\mathcal{A}_l|\mathbf{y}_l(t))}{\sum_{l=1}^{M}p(\mathcal{D}_l|\mathbf{y}_l(t))p(\mathcal{C}_l|\mathbf{y}_l(t))p(\mathcal{A}_l|\mathbf{y}_l(t))}. \quad (22)$$

This strategy considers the probabilities of the previous and current measurements in addition to a normalized likelihood for the contribution of the different sensors. The target with the highest probability in a group of candidates will be the one associated to the track. Due to the contribution of previous measurements, this strategy can minimize the identity switches when the available measurements are inaccurate due to noise or errors.

The second and third terms of the right hand side of Eq. (19) can be expressed as

$$p(N_f(t)|N_p(t),N_n(t),\mathbf{\Gamma}(t)) \propto p(N_f(t)|N_f(t-1)) \quad (23a)$$



(a) (b) (c)

Fig. 7. Performance comparison of different tracking algorithms: (a) frame 157 (b) frame 435 and (c) frame 723. (Row 1): ground truth marked with filled circles; (Row 2): GCC noisy audio estimates; (Row 3): estimated arrival angles of speakers with the Kalman filter (dash lines) and without the Kalman filter (dash dots); (Row 4): Particle filter and PDA-based audiovisual tracking; (Row 5): Particle filter and WPDA-based audiovisual tracking.

and

$$p(N_n(t)|N_p(t), N_f(t), \mathbf{\Gamma}(t)) \propto p(N_n(t)|N_n(t-1)), \quad (23b)$$

where

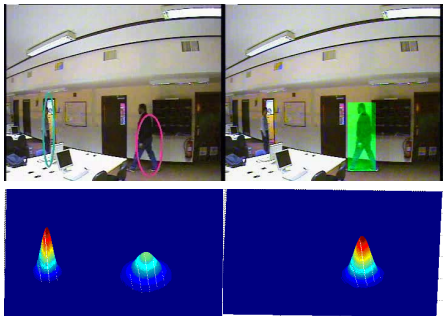
$$p(N_f(t)|N_f(t-1)) \approx \Pi_{l=1}^{M} p_l(N_{f_l}(t)|N_{f_l}(t-1)). \quad (24)$$

The right hand side of Eq. (23(a-b)) is modeled as a Gaussian distribution. The mean and variance of these distributions are computed based on N_f and N_n, respectively. The main steps of the WPDA algorithm for each frame within the particle filter framework are summarized in Algorithm 1.

Fig. 7 compares sample results from different object tracking strategies. It is possible to notice that the Kalman filter leads to smaller errors in audio source localization (Fig. 7(a-b)), and that despite the biased audio GCC estimates shown on row 2, the WPDA locates the speaker due to the correct likelihood estimation of the visual and the audio locations.

C. State and likelihood estimation

The overall likelihood is composed of visual and audio components. The visual components depend on the color histogram of the detected pixels associated to the object and



(a) (b)

Fig. 8. Sample images from the proposed audiovisual tracker. (Row 1): position estimation using (a) video and (b) audio features; (Row 2): Likelihood of the measurements: (a) visual measurements (two persons), (b) audio detection showing the speaker under the green patch associated to the change detection bounding box. The horizontal and vertical axes of the graphs in the second row correspond to the width and the height of the image, respectively. (Note that the images are tilted for improved visualization).

on a 4D state space defined by the components (x_v, y, w, h) of the bounding box associated to a detection.

Change detection is performed using a background subtraction algorithm based on absolute frame difference on each channel of the RGB color space, and the results are integrated with a logical OR. Median filtering and morphology are then applied in order to post-process the detection result. The color likelihood is then estimated using a three channel color histogram $\mathcal{H}$, uniformly quantized with $10\times10\times10$ bins

$$p(\mathcal{C}|\mathbf{y}_v(t)) = \exp\Big(-\Big(\frac{d(p(\mathbf{y}_v), \lambda)}{\sigma}\Big)^2\Big), \quad (25)$$

where λ is the reference histogram defining the target model, σ is the standard deviation and $d(\cdot)$ is the distance based on the Bhattacharyya coefficient ([55], [56]):

$$d(p(\mathbf{y}_v), \lambda) = \sqrt{1 - \sum_{u=1}^{m} \sqrt{p_u(\mathbf{y}_v)\lambda_u}}, \quad (26)$$

where m is the number of bins and $p(\mathbf{y}_v)$ is the color histogram, computed as

$$p_u(\mathbf{y}_v) = B \sum_i K_{e,\theta_r}\Big(\Big|\Big|\frac{(x,y) - \mathbf{W}_i}{\Upsilon}\Big|\Big|^2\Big)\delta(\zeta(\mathbf{W}_i) - u), \quad (27)$$

where B is the normalization factor that ensures identity of the sum of the histogram bins; $\mathbf{W}_i$ are the pixels on the target and $\zeta(\mathbf{W}_i)$ associates each $\mathbf{W}_i$ to the corresponding histogram bin; K_{e,θ_r} is the kernel profile with bandwidth Υ ([21]). To make the model robust to pose or illumination changes, the reference histogram is updated using a running average ([57]).

Since four elements have been explicitly declared in the state vector $\mathbf{y}(t)$, we can compute the 4D state space likelihood for change detection by applying a four dimensional multivariate Gaussian function:

$$p(\mathcal{D}|\mathbf{y}(t)) \sim \mathcal{N}(\mu_{\mathcal{D}}^{(4)}, \sigma_{\mathcal{D}}^{(4)}), \quad (28)$$

The audio state $\mathbf{y}_a$ is estimated as explained in Sec. III(A-B). The joint state likelihood of audio $\mathbf{y}_a(t)$ and visual $\mathbf{y}_v(t)$

Algorithm 1 *WPDA Algorithm*

1: Create samples for the target states $\mathbf{y}_l(t)$;
2: Compute the posterior distributions $p(\mathbf{z}(t)|\mathbf{\Gamma}(t), \Xi(t))$ and $p(\Xi(t)|\mathbf{\Gamma}(t))$ using Eqs. (18)-(19);
3: Compute the joint association probability $p(\mathbf{\Gamma}(t), \Xi(t)|\mathbf{z}(t))$ using Eq. (17);
4: Calculate the marginal association probability as $\gamma = \sum_{n=1}^{M_t} p(\mathbf{\Gamma}^{(n)}(t), \Xi^{(n)}(t)|\mathbf{z}^{(n)}(t))$;
5: Generate the target likelihood: $p(\mathbf{z}(t)|\mathbf{y}(t)) = \Pi_{l=1}^{M} \gamma_l p(\mathbf{z}_l(t)|\mathbf{y}_l(t))$;
6: Update the particle weights using Eq. (2);
7: Apply resampling for each target to avoid the degeneracy of the particle sets.

components is

$$p(\mathbf{y}(t)|\mathbf{y}(t-1), \mathbf{y}_a(t), \mathbf{y}_v(t)) \propto \\ \propto \quad p(\mathbf{y}_a(t), \mathbf{y}_v(t)|\mathbf{y}(t)) p(\mathbf{y}(t)|\mathbf{y}(t-1)), \qquad (29)$$

and

$$p(\mathbf{y}(t)|\mathbf{y}(t-1)) \propto c p(\mathcal{D}|\mathbf{y}_v(t)) p(\mathcal{C}|\mathbf{y}_v(t)) p(\mathcal{A}|\mathbf{y}(t)), \quad (30)$$

where c is a constant, and $p(\mathcal{A}|\mathbf{y}(t))$ accounts for the audio likelihood that is computed using a univariate Gaussian:

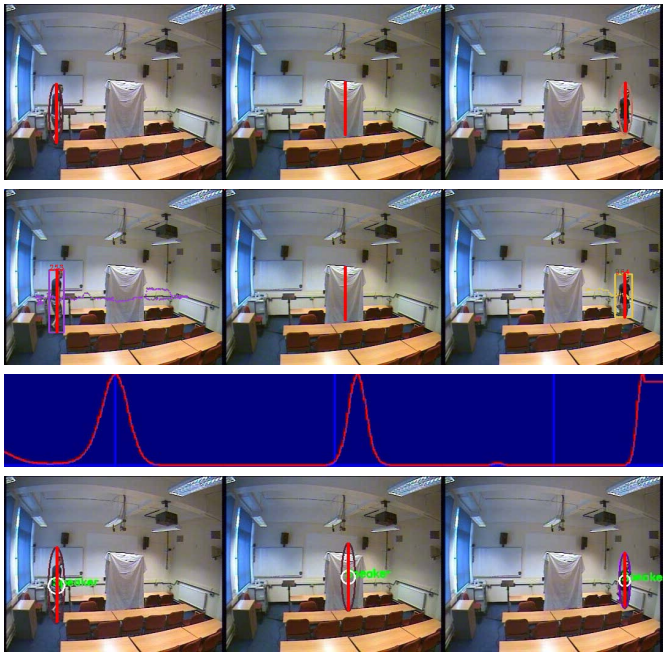
$$p(\mathcal{A}|\mathbf{y}(\mathbf{t})) \propto \frac{1}{\sigma_{\mathcal{A}}\sqrt{2\pi}} \exp\Big(-\frac{(\mathbf{y_a}(t) - x_t)^2}{2\sigma_{\mathcal{A}}^2}\Big). \qquad (31)$$

Note that if one of the measurements is unavailable (e.g., during occlusion or silence) then its corresponding probability is set to 1. Sample results from the different steps of the proposed audiovisual tracker are shown in Fig. 8.

IV. Experimental results

The proposed multimodal detection and tracking algorithm is evaluated using data collected with a STAC sensor composed of two Beyerdynamic MCE 530 condenser microphones and a KOBI KF-31CD camera. The image resolution is 360×288 pixels (25Hz) and the audio is sampled at 44.1 KHz. The audiovisual synchronization is performed using a VisioWave Discovery 300 Series recorder. We present two types of setup: in type 1, the distance between the microphones is 95 cm (experiment 1) or 124 cm (experiments 2 and 3) and the video camera is located in the middle. In type 2, the distance between the microphones is 124 cm and the camera is placed 200 cm in front of the microphones. The camera and the microphones have the same height from the floor (170 cm). The distance between the sensors and the speaker is larger than 500 cm. These datasets are available at http://www.elec.qmul.ac.uk/staffinfo/andrea/stac.html.

For a quantitative evaluation of the results, we use two scores: ϵ, the one-dimensional *Euclidean distance* between the detected x-coordinates and the ground truth, and λ, the number of lost tracks and identity switches over the entire sequence. This score is computed as $\lambda = (LT + IS)/TF$ where LT is the number of frames with lost tracks, IS is the number of frames with identity switches, and TF is the total number of frames in the sequence. Hence the lower λ, the better the performance.



(a) (b) (c)

Fig. 9. Comparison of tracking results for "1-room" (Frame numbers: (a) 814, (b) 926 and (c) 1010). (Row 1): PF; (Row 2): GM; (Row 3): GCC, (Row 4): KF-PF-P. The red bar indicates the true target position.

A. Type 1 experiments

Three indoor audiovisual datasets, namely "1-room", "2-lab-a" and "2-lab-b", are used to compare the detection and tracking performance of six different strategies: (1) vision only by PF; (2) vision only by graph matching (GM) ([58]); (3) estimation of arrival angle only before and after Kalman filtering (the former: AB-KF; the latter: AA-KF); (4) GCC-PHAT arrival angle estimation and particle filter based audiovisual tracker (GP-PF) ([32]); (5) Kalman filtering audio detection and the particle filter-based audiovisual tracker with PDA (KF-PF-P) ([15]); (6) the proposed arrival angle estimation using Kalman filtering and particle filter based audiovisual tracker with WPDA (KF-PF-WP).

In the dataset "1-room" (1077 frames), a person walks, talks and hides himself behind a barrier for about 1 second. The Kalman filter-based tracker for audio source detection generates errors in the estimation of the target's x location. These errors are due to the violation of the initial assumptions mentioned in Sec. III-A and due to the presence of a significant background noise. Fig. 9 compares sample results using the different techniques under analysis. The vision only methods

TABLE II
PERFORMANCE COMPARISON OF THE TRACKERS UNDER ANALYSIS. ABSOLUTE LOCATION ESTIMATION ERRORS (ϵ: AVERAGE IN PIXELS.)

	Seq.	PF	GM	AB-KF	AA-KF	KF-PF-P	KF-PF-WP
ϵ	1-room	6.4	6.2	18.1	14.5	5.2	4.9
	2-lab-a	5.7	5.9	23.6	21.3	5.2	4.7
	2-lab-b	7.3	6.8	25.2	23.1	5.5	5.4

TABLE III
PERFORMANCE COMPARISON OF THE TRACKERS UNDER ANALYSIS. LOST TRACKS/IDENTITY SWITCHES (λ: PERCENTAGE OVER AN ENTIRE SEQUENCE).

	Seq.	PF	GM	GP-PF	KF-PF-P	KF-PF-WP
$\lambda(\%)$	1-room	5.76	5.76	5.29	4.83	4.51
	2-lab-a	14.49	14.55	12.39	12.23	11.95
	2-lab-b	11.36	11.26	10.99	9.19	8.54

(rows 1 and 2) correctly localize the object only when it is observable. Row 3 shows the GCC estimates of the audio signals, where the estimated peaks are close to the object's position. Row 4 shows that the KF-PF-P tracker has a good accuracy in object detection and tracking (KF-PF-WP obtains comparable results) and, unlike PF and GM, does not suffer from identity switches (visualized as changes in color). Moreover, although the GCC estimation at frame 814 is not next to the person due to the violation of the initial assumptions (Sec. III-A) and because of a reverberation peak that was not filtered, the proposed tracker still estimates an accurate position of the target due to the variances of the particles along the x-axis (Eq. (31)).

The second ("2-lab-a", 1856 frames) and third ("2-lab-b", 1883 frames) experiments has a similar set-up; however the difference is in the clothing of the targets. In these experiments two persons walk from right to left and then meet in the half way generating a full occlusion (both audio and visual). The result has therefore a large number of identity switches as compared to experiment 1 (Table III). This is again due to violation of initial assumptions (Sec. III-A), however the proposed approach has minimum number or identity switches and localization error as compared to other approaches (Table II and Table III).

Table II shows that KF-PF-P and KF-PF-WP have lower average errors compared to AB-KF and AA-KF. Table III shows that KF-PF-P and KF-PF-WP have the smallest lost tracks/identity switches in the test sequences. It is worth noticing that the lost tracks are mainly due to the absence of both visual and audio signals and strong reverberations existing in the scene. To enhance the proposed tracker in audio detection, one could generate a model for the background noise.

B. Type 2 experiments

We evaluate the performance of the proposed detection and tracking algorithm (KF-PF-WP) in three sequences (SC 1, SC 2 and SC 3) of a meeting scenario and show how metadata generated automatically (object position and their sound activity) can be transferred to other sensors or multimedia receivers (e.g., mobile phones) with limited bandwidth requirements.

The first sequence (SC 1) has three subjects (sample frames and the corresponding results are shown in Fig. 10, row 1 and 2) who initially are sitting and having pair-wise conversation. Next, the person sitting in the middle stands up, moves and talks to the person on the left. This results in a difficult audio detection, as he keeps changing the direction of his face. The proposed tracking algorithm enables us to effectively detect and track the speakers: for example, the second column of row 1 and row 2 shows that the speaker sitting in the middle is correctly detected and tracked. However, in column 4 the audiovisual tracker does not detect the real speaker (the speaker on the left), due to a biased estimation of the audio GCC estimates when the person faces away from the microphones. In row 5 and 6, the proposed audiovisual tracker correctly identifies the speaker despite large measurement errors (third column of row 5 and 6). Although the audio detection deviates from the correct position, the final audiovisual result is accurate as the estimation of the speaker's position in the previous image frame is correct. In fact, this leads to a larger posterior probability of detection of the speaker on the left than that of the one on the right in the current frame, and hence the estimated position settles on the person on the left.

TABLE IV
BANDWIDTH ESTIMATES OF DIFFERENT METHODOLOGIES ON SEQUENCE 1 (SC 1), SEQUENCE 2 (SC 2) AND SEQUENCE 3 (SC 3). UNITS: KILOBYTES PER FRAME.

Seq.	MPEG-1	MPEG-2	MPEG-4	Metadata	Metadata with audio
SC 1	7.61	7.82	6.39	0.21	1.15
SC 2	6.74	7.26	5.75	0.48	1.42
SC 3	7.76	8.16	7.36	0.25	1.19

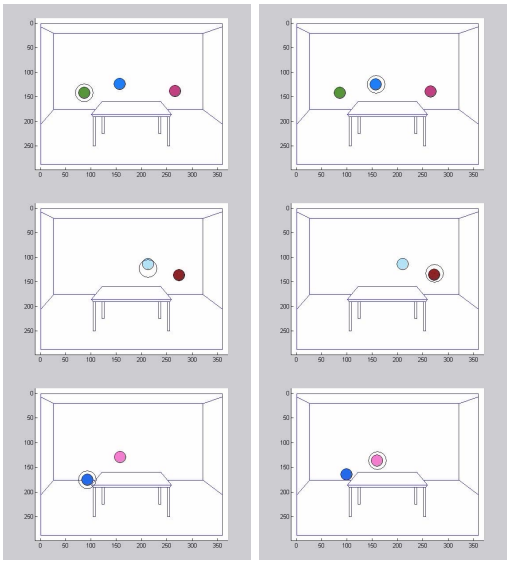


Fig. 11. Sample object animations and speaker detection created using the generated metadata. (Row 1): sequence 1, (Row 2): sequence 2, (Row 3): sequence 3. Colored circles denote the visual detection, white circles represent the audio detection, and the axes show the original image size.

Fig. 11 shows sample animations generated using the automatically extracted metadata. The comparison of the bandwidth requirement when using different coding methodologies is shown in Table IV for (1) MPEG-1, (2) MPEG-2, (3) MPEG-4 and (4) the metadata generated by the proposed multimodal tracker. The size of all the four formats also

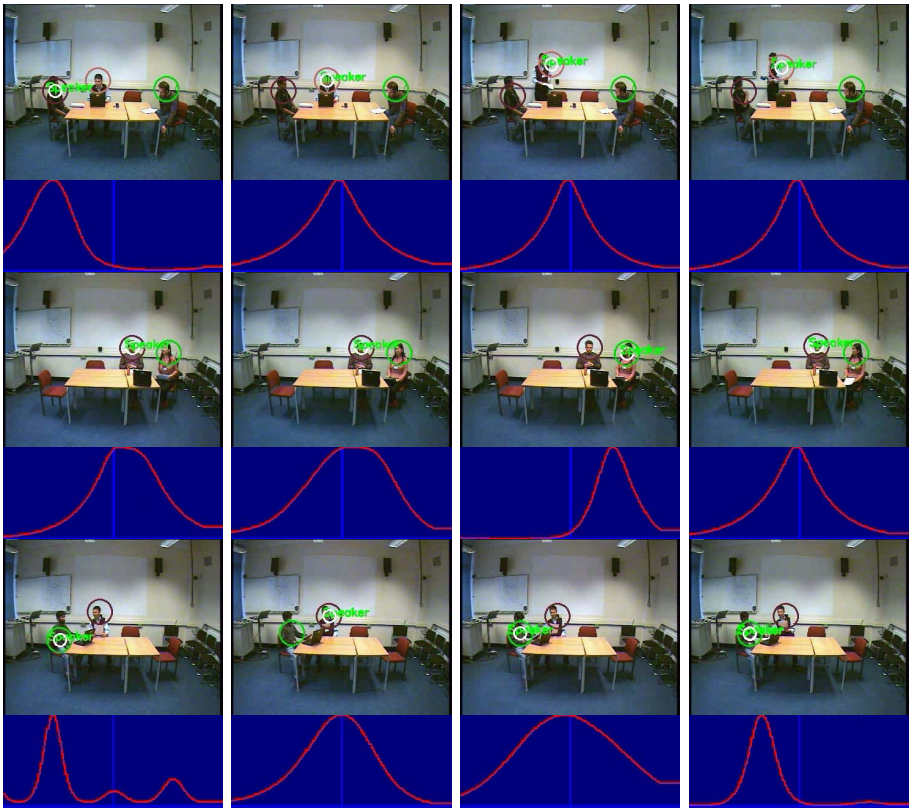


Fig. 10. Audiovisual speaker detection and tracking for sequence 1 (SC 1) (row 1 and 2), sequence 2 (SC 2) (row 3 and 4) and sequence 3 (SC 3) (row 5 and 6). Row 1,3 and 5: white circles with "speaker" indicate the detected audio source location and circles in other colors denote the visual detection and tracking. Row 2, 4 and 6 show the audio GCC estimates.

contains the audio file size. These bandwidth requirements correspond to the information to be transmitted when multiple multimodal sensors exchange the position and the activities of the observed objects.

V. Conclusions

We have presented a particle filter based tracking algorithm that integrates measurements from heterogeneous sensors and demonstrated it on audio and video signals. In order to reduce the effects of reverberations and noise, we used a Riccati Kalman filter that automatically updates the audio measurements using the historic estimates in a least squares sense as well as a WPDA scheme to associate the audio detections to the visual measurements. Another feature of the proposed framework is its modularity that allows us to replace any blocks depending on the application at hand. The experimental results demonstrated that the proposed strategy improves classical audio or video approaches in terms of tracking accuracy and performance. We have also shown the bandwidth requirements for communicating the metadata generated from the tracker to other sensors or remote devices.

Our current work addresses the use of the proposed audio-visual tracker in a sensor network and its use for distributed multimodal event detection. Moreover, we will investigate the integration of techniques like source separation and speech recognition to relax the current assumptions on the number of sound sources and the direction of the sound.

References

[1] J. Vermaak, M. Gangnet, A. Blake, and P. Pérez, "Sequential monte carlo fusion of sound and vision for speaker tracking," in *Proc. of IEEE Int. Conf. on Computer Vision*, Vancouver, Canada, July 2001, pp. 741–746.

[2] R. Chellappa, G. Qian, and Q. Zheng, "Vehicle detection and tracking using acoustic and video sensors," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Montreal, Canada, May 2004.

[3] V. Cevher, A. C. Sankaranarayanan, J. H. McClellan, and R. Chellappa, "Target tracking using a joint acoustic video system," *IEEE Trans. on Multimedia*, vol. 9, no. 4, pp. 715–727, June 2007.

[4] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *IEEE Int. Conf. on Multimedia & Expo (ICME)*, July-August 2000.

[5] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filters," in *Proc. of IEEE Int. Conf. on Image Processing*, Barcelona, Spain, September 2003.

[6] R. Cutler and L. S. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *Proc. of IEEE Int. Conf. on Multimedia and Expo (III)*, New York, NY USA, July-August 2000.

[7] B. Kapralos, M. Jenkin, and E. Milios, "Audio-visual localization of multiple speakers in a video teleconferencing setting," *Int. Journal of Imaging Systems and Technology*, vol. 13, no. 1, pp. 95–105, June 2003.

[8] H. Asoh et al., "An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion," in *Proc. of the Seventh Int. Conf. on Information Fusion*, Stockholm, Sweden, June 2004.

[9] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audio-visual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, pp. 601–616, 2007.

[10] M. J. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *Proc. of the European Conf. on Computer Vision*, Copenhagen, Denmark, June 2002.

[11] P. Perez, J. Vermaak, and A. Blake, "Data fusion for visual tracking with particles," *Proc. of IEEE*, vol. 92, pp. 495–513, March 2004.

[12] T. Gehrig, K. Nicel, H. K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York, NY, USA, October 2005.

[13] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization and tracking," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 22–31, January 2001.

[14] Y. Chen and Y. Rui, "Speaker tracking using particle filter sensor fusion," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 485–494, 2004.

[15] H. Zhou, M. Taj, and A. Cavallaro, "Audiovisual tracking using STAC sensors," in *ACM/IEEE Int. Conf. on Distributed Smart Cameras*, Vienna, Austria, September 25-28 2007.

[16] Z. Ding, H. Leung, and L. Hong, "Decoupling joint probabilistic data association algorithm for multiple target tracking," *IEE Trans. on Radar, Sonar Navigation*, vol. 146, no. 5, pp. 251–254, 1999.

[17] J. Vermaak, S. J. Godsill, and P. P. Pérez, "Monte carlo filtering for multi-target tracking and data associtation," *IEEE Trans. on Aerospace and Electronic Systems (AES)*, vol. 41, no. 1, pp. 309–332, 2005.

[18] C. Stauffer, "Automated audio-visual analysis," Tech. Rep. MIT-CSAIL-TR-2005-057, Computer Science and Artificial Intelligence Laboratory, MIT, September 20 2005.

[19] D. Smith and S. Singh, "Approaches to multisensor data fusion in target tracking: A survey," *IEEE Trans. on Knowledge and Data Engineering (TKDE)*, vol. 18, no. 12, pp. 1696–1710, 2006.

[20] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys (CSUR)*, vol. 38, no. 4, pp. 1–45, December 2006.

[21] E. Maggio and A. Cavallaro, "Hybrid particle filter and mean shift tracker with adaptive transition model," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, 19–23 March 2005.

[22] O. Lanz, "Approximate bayesian multibody tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Febuary 2006.

[23] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, May 2003.

[24] K. Shafique and M. Shah, "A noniterative greedy algorithm for multiframe point correspondence," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 51–65, January 2005.

[25] K. Zia, T. Balch, and F. Dellaert, "Mcmc-based particle filtering for tracking a variable number of interacting targets," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1805–1819, November 2005.

[26] D. B. Ward and R. C. Williamson, "Particle filter beamforming for acoustic source localisation in a reverberant environment," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, May 2002, pp. 1777–1780.

[27] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Salt Lake City, Utah, May 2001, pp. 3021–3024.

[28] E.A. Lehmann, D.B. Ward, and R.C. Williamson, "Experimental comparison of particle filtering algorithms for acoustic source localization in a reverberant room," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Hong Kong, 6–10 April 2003, pp. 177–180.

[29] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for acoustic source localization," *IEEE Trans. on Speech and Audio Processing*, pp. 826–836, Nov. 2003.

[30] R. Brunelli, A. Brutti, P. Chippendale, O. Lanz, M. Omologo, P. Svaizer, and F. Tobia, "A generative approach to audio-visual person tracking," in *CLEAR*, Southampton, UK, April 2006.

[31] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP Journal on Applied Signal Processing*, vol. 2002, pp. 1154–1164, 2001.

[32] M. Bregonzio, M. Taj, and A. Cavallaro, "Multi-modal particle filtering tracking using appearance, motion and audio likelihoods," in *Proc. of IEEE Int. Conf. on Image Processing*, San Antonio, Texas (USA), September 2007.

[33] N. Checka, K.W. Wilson, and M.R. Siracusa T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Cambridge, MA, USA, May 2004, vol. 5.

[34] D. Gatica-Perez, G. Lathoud, I. McCowan, J. Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filters," in *Proc. of IEEE Int. Conf. on Image Processing*, 2003.

[35] Y. Rui and Y. Chen, "Better proposal distributions: object tracking using unscented particle filter," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, HI, USA, December 2001, pp. 786–793.

[36] X. Zou and B. Bhanu, "Tracking humans using multi-modal fusion," in *IEEE Int. Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum (OTCBVS)*, San Diego, CA, USA, June 2005.

[37] A. Abad et al., "UPC audio, video and multimodal person tracking systems in the CLEAR evaluation campaign," in *CLEAR*, Southampton, UK, April 2006.

[38] A. Stergiou, A. Pnevmatikakis, and L. Polymenakos, "A decision fusion system across time and classifiers for audio-visual person identification," in *CLEAR*, Southampton, UK, April 2006.

[39] D. Gatica-Perez, G. Lathoud, J.M. Odobez, and I. McCowan, "Audio-visual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. Audio, Speech and Language Processing*, March 2006.

[40] A. Blake, M. Gangnet, P. Perez, and J. Vermaak, "Integrated tracking with vision and sound," in *Proc. of IEEE Int. Conf. on Image Analysis and Processing*, September 2001, vol. 1.

[41] B. D. O. Anderson and J. B. Moore, *Optimal filtering*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[42] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *Int. Journal on Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[43] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. on Signal Processing*, vol. 50, no. 2, pp. 174–188, Febuary 2002.

[44] K. Wilson, N. Checka, D. Demirdjian, and T. Darrell, "Audio-video array source separation for perceptual user interfaces," in *The 2001 workshop on Perceptive user interfaces (PUI)*, November 2001, pp. 1–7.

[45] U. Bub, M. Hunke, and A. Waibel, "Knowing who to listen to in speech recognition: visually guided beamforming," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 9-12 May 1995, vol. 1, pp. 848–851.

[46] B. Kapralos, M. Jenkin, and E. Milios, "Audio-visual localization of multiple speakers in a video teleconferencing setting," *International Journal of Imaging Systems and Technology*, pp. 95–105, 2003.

[47] A. Cavallaro and T. Ebrahimi, "Interaction between high-level and low-level image analysis for semantic video object extraction," *EURASIP Journal on Applied Signal Processing*, vol. 6, pp. 786–797, June 2004.

[48] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signal*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[49] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, pp. 1633–1654, 1999.

[50] K. Wilson and T. Darrell, "Improving audio source localization by learning the precedence effect," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, March 18-23 2005.

[51] William A. Yost and George Gourevitch, *Directional Hearing, chapter The precedence effect*, Springer-Verlag, 1987.

[52] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, John Wiley & Sons, Inc., 4^{th} edition, 2000, ch 12 pp. 341, ISBN 0-471-84789-5.

[53] C. H. Knapp and G. C. Carter, "The generalised correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech and Signal Processing (ASSP)*, vol. 24, pp. 320–327, 1976.

[54] S. M. Bozic, *Digital and Kalman Filtering*, Edward Arnold, London, UK, 1979.

[55] A. Bhattacharyya, *On a measure of divergence between two statistical populations defined by probability distributions*, vol. 35, Bulletin of the Calcutta Mathematical Society, 1943.

[56] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *IEEE Trans. on Communication Technology*, vol. 15, pp. 52–60, 1967.

[57] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "A color-based particle filter," in *Workshop on Generative-Model-Based Vision*, June 2002, pp. 53–60.

[58] M. Taj, E. Maggio, and A. Cavallaro, "Multi-feature graph-based object tracking," in *CLEAR, Springer LNCS 4122*, Southampton, UK, April 2006, pp. 190–199.