MULTI-MODAL PARTICLE FILTERING TRACKING USING APPEARANCE, MOTION AND AUDIO LIKELIHOODS

*Matteo Bregonzio, Murtaza Taj, Andrea Cavallaro**

Multimedia and Vision Group – Queen Mary, University of London
Mile End Road, E1 4NS London, United Kingdom

ABSTRACT

We propose a multi-modal object tracking algorithm that combines appearance, motion and audio information in a particle filter. The proposed tracker fuses at the likelihood level the audio-visual observations captured with a video camera coupled with two microphones. Two video likelihoods are computed that are based on a 3D color histogram appearance model and on a color change detection, whereas an audio likelihood provides information about the direction of arrival of a target. The direction of arrival is computed based on a multi-band generalized cross-correlation function enhanced with a noise suppression and reverberation filtering that uses the precedence effect. We evaluate the tracker on single and multi-modality tracking and quantify the performance improvement introduced by integrating audio and visual information in the tracking process.

Index Terms— Audiovisual tracking, particle filter, multimodal processing, color histogram, change detection.

1. INTRODUCTION

The use of multiple modalities in object detection and tracking helps compensating for noisy, partial or missing observations obtained with a single modality. For example, the most appropriate camera view can be selected depending on speech activity in multi-camera video conferencing [1]. Moreover, audio can compensate for the failure of video when an object is visually occluded by vegetation or dust in surveillance scenarios [2]. Video and audio observations can be fused using Particle Filter (PF), Probabilistic Data Association (PDA), Kalman Filter (KF) [3], or Decentralized Kalman Filter [4]. The joint likelihood can be computed as the product ([5, 6, 7, 1, 8]) or as a linear combination of the single modality likelihoods [9]. An independent PF for each target is used in [10], whereas a PF tracker that uses audio information for consistent vehicles tracking during occlusions is presented in [11]. A video likelihood based on the distance from detections is fused with an audio likelihood computed on STFT coefficients in [12]. Sound source localization is performed using SPR-PHAT [7], Expectation Maximization based Maximum Likelihood, or Cross Power Spectrum based on the 2D Global Coherence Function [5].

A variety of sensor configurations have been used for audio-visual object detection and tracking. Figure 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 [13]; single or stereo cameras with multiple microphones are used in meeting rooms and teleconferencing [14, 8]. In particular, simple Stereo Audio and Cycloptic (STAC) vision sensors are suitable for wide area surveillance. STAC sensors are used to perform audio-visual tracking with a probabilistic graph model and fusion by linear mapping ([15]) or with PF ([16]). However, the position of the speaker with respect to the sensor is constrained by limiting assumptions.

In this paper, we propose a multiple object tracker for STAC sensors that is based on particle filtering and combines the audio-visual observations at the likelihood level. *Visual measurements* are derived from an appearance model based on 3D color histograms and from a motion model based on color change detection. *Audio measurements* are derived from a multi-band generalized cross correlation that is used for audio source localization. To improve the localization accuracy, we define a reverberation filtering based on onset detection. The proposed algorithm is capable of detecting and tracking an active speaker and of tracking audio-visual objects in scenes where visual occlusions occur.

The paper is organized as follows. In Section 2 we present the multi-modal detection and tracking algorithm, the reverberation reduction algorithm and data fusion using PF. Experimental results are discussed in Section 3. Finally, in Section 4 we draw the conclusions.

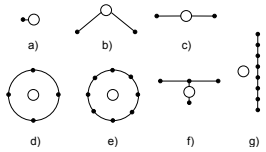


Fig. 1. *Examples of sensor configurations for audio-visual object detection and tracking (filled circles indicate microphones; empty circles indicate cameras – single or stereo): (a) single microphone-camera pair; (b-c) STAC sensors; (d-e) circular microphone array with single camera; (f) triangular microphone array with single camera; (g) linear microphone array with single camera*

2. MULTI-MODAL TRACKING

The problem of multiple audio-visual object tracking can be formalized as a continuous estimation, from audio and video observations, of the state $\mathbf{x}_t$ of each target at time t. Let us define the state as $\mathbf{x}_t = (x, y, w, h)$, where (x, y) is the position and w and h are the width and the height of the object. At any time t, one the following conditions is possible: (i) a complete audio-visual observation is available, (ii) only the sound cues is available, or (iii) only the visual cues are available. We discuss below how to improve the estimation accuracy of $\mathbf{x}_t$ by fusing audio-visual information using STAC sensors.

*The authors acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC), under grant EP/D033772/1

2.1. Particle filtering

Particle filtering solves the tracking problem based on the *state equation*

$$\mathbf{x}_t = \mathbf{f}_t(\mathbf{x}_{t-1}, \mathbf{v}_t), \tag{1}$$

and on the *measurement equation* $\mathbf{z}_t = \mathbf{h}_t(\mathbf{x}_t, \mathbf{n}_t)$, where f_t and h_t are non-linear and time-varying functions. $\{\mathbf{v}_t\}_{t=1,...}$ and $\{\mathbf{n}_t\}_{t=1,...}$ are assumed to be independent and identically distributed stochastic processes. The problem consists in calculating the *pdf* $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ at each time instant t. This *pdf* can be obtained recursively with a prediction and an update step. The *prediction step* uses $\mathbf{x}_t$ from Eq. (1) to obtain the prior *pdf* as

$$p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})d\mathbf{x}_{t-1}, \tag{2}$$

with $p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})$ known from the $t-1$ and $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ determinated by Eq. (1). Given the measurement $\mathbf{z}_t$, the *update step* is performed using the Bayes' rule

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) = \frac{p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1})}{\int p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1})d\mathbf{x}_t}. \tag{3}$$

Particle filtering approximates the densities $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ with a sum of N_s Dirac functions centered in $\left\{\mathbf{x}_k^i\right\}_{i=1,...N_s}$ as

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \approx \sum_{i=1}^{N_s} \omega_t^i \delta\left(\mathbf{x}_t - \mathbf{x}_t^i\right), \tag{4}$$

where ω_t^i are the weights associated to the particles. The weights are calculated as

$$\omega_t^i \propto \omega_{t-1}^i \frac{p(\mathbf{z}_t|\mathbf{x}_t^{i-1})p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)}{q(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i, \mathbf{z}_t)}. \tag{5}$$

$q(.)$ is the importance density function. When $q(.) = p(\mathbf{x}_t|\mathbf{x}_{t-1}^i)$, then $\omega_t^i \propto \omega_{t-1}^i p(\mathbf{z}_t|\mathbf{x}_t^{i-1})$.

Next, to avoid the degeneracy problem re-sampling is applied by setting $\omega_{t-1}^i = 1/N_s \ \forall\, i$, therefore

$$\omega_t^i \propto p(\mathbf{z}_t|\mathbf{x}_t^{i-1}). \tag{6}$$

The weights are therefore proportional to the likelihood function that will be discussed in the next sections.

2.2. Audio likelihood

A STAC sensor can estimate the horizontal position x of a target using sound source localization. Let $s_1(t)$ and $s_2(t)$ be the signals captured by the two STAC microphones. Since the microphones are spatially separated, the signal emitted by a sound source reaches the two microphones at different time instants. The signals $s_1(t)$ and $s_2(t)$ can be written as $s_1(t) = v(t) + n_1(t)$ and $s_2(t) = \lambda v(t+\tau) + n_2(t)$, where $v(t)$ is the sound wave emitted by the source, $n_1(t)$ and $n_2(t)$ are noise components, τ is the delay time of arrival of the wave to the two microphones, and λ is the attenuation component. The position x of the sound source can be estimated by computing the cross-correlation $\widehat{R}_{s_1 s_2}$ of s_1 and s_2 using the Generalized Cross Correlation function-Phase Transform (GCCF-PHAT). To reduce the effect of reverberation in the source localization process, we exploit the precedence effect and Multi-Band Frequency Analysis. The GCCF-PHAT $\widehat{R}_{s_1 s_2}$ is estimated only on ensemble

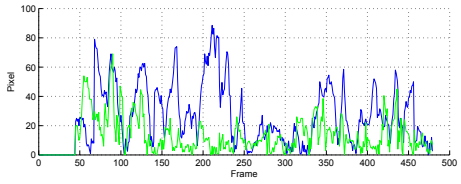


Fig. 2. *Deviation from the ground truth for source localization using the GCC-PHAT transform (blue) and the proposed method (green)*

of frames that are classified as onset $F_O(t)$ using the *precedence effect*. Onset frames $F_O(t)$ are frames containing a significant signal component and a limited or absent reverberation component caused by the signal itself. These onsets are located at the beginning of a signal audio block (the audio segment between two salient segments of the audio signal). A frame $F(t)$ is considered a signal frame if the SNR at both microphones is larger than a threshold (see Section 3). Assuming that the frame under analysis, $F(t)$, is the first frame of an onset $F_O(1)$, the subsequent T frames are processed if identified as signal frames; whereas the signal frames from $F(t+T)$ to the first *null* frame are considered reverberant frames and therefore discarded.

The *multi-band frequency analysis* is based on the observation that low frequencies are less subject to reverberation than high frequencies and that the effects of correlated noise, located in a single frequency band, can be reduced by evaluating the signal in different frequency bands [17]. The two audio signals $s_1(t)$ and $s_2(t)$ are divided into three different frequency bands. Using a normalized frequency notation, a low frequency band (B_1), a middle frequency band (B_2), and a high frequency band (B_3) are defined. The frequency band division is computed using three different 36-coefficient band-pass linear phase FIR filters, frame-by-frame, for onset frames. The cross-correlation function is then estimated for each frequency band. The final estimation of the GCC is obtained by a weighted combination of the three sub-band cross-correlations as

$$\widehat{R}_{s_1 s_2}(f) = \sum_{i=1}^{3} w_i \frac{G_{s_1 s_2}^i(f)}{\gamma|G_{s_1 s_2}^i(f)| + (1-\gamma)|N^i(f)|^2}, \tag{7}$$

where $G_{s_1 s_2}^i(f)$ is the cross power spectral density function in band B_i, $\gamma \in [0,1]$ and $N^i(f)$ is the noise spectral density in band B_i. $N^i(f)$ is estimated during the initialization assuming stationary noise. The weights w_i ($\sum_{i=1}^{3} w_i = 1$) are chosen such that higher frequency components contribute less than the low frequency ones. In the experiments they are set to $w_1 = 0.5$, $w_2 = 0.3$, $w_3 = 0.2$. A peak is retained if it is simultaneously located in the same position in the three GCCs. Peaks that appear in a single band only are reduced proportional to the weight associated. The resulting improvements compared to the plain GCCF-PHAT can be seen in Fig.2. The green line shows the distance between the ground truth and the results obtained with the proposed approach. The blue line shows the distance between the ground truth and the GCC-PHAT result. It can be seen that error for the proposed system (green line) is much smaller than that of the GCC-PHAT (blue line) result. The *audio likelihood*, $p(\mathcal{A}|\mathbf{x}_t)$, is finally computed by applying a univariate Gaussian $\mathcal{N}(\mu_\mathcal{A}, \sigma_\mathcal{A})$ to the estimated cross-correlation $\widehat{R}_{s_1 s_2}$ as

$$p(\mathcal{A}|\mathbf{x}_t) = \frac{1}{\sigma_\mathcal{A}\sqrt{2\pi}} e^{\frac{(\widehat{R}_{s_1 s_2}(f))^2}{2\sigma_\mathcal{A}^2}} \tag{8}$$

This results in a non-linear amplification of the autocorrelation function that emphasizes the major peaks.

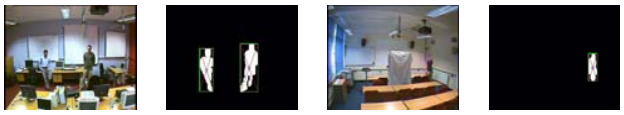


Fig. 3. *Example of change detection results used in the visual likelihood*

2.3. Visual likelihood

The visual likelihood is composed of two cues, a color measurement and a motion measurement. The *color likelihood*, $p(\mathcal{C}|\mathbf{x}_t)$, is computed in the RGB color space using 3D color histograms, uniformly quantized with 10x10x10 bins, as

$$p(\mathcal{C}|\mathbf{x}_t) = e^{-\left(\frac{d[p(\mathbf{x}),q]}{\sigma}\right)^2}, \tag{9}$$

where $d[.]$ is the distance based on the Bhattacharyya coefficient, computed as

$$d\left[p(\mathbf{x}), q\right] = \sqrt{1 - \sum_{u=1}^{m} \sqrt{p_u(\mathbf{x}) \cdot q_u}}, \tag{10}$$

with

$$p_u(\mathbf{x}) = B \sum_i K_{e,\theta}\left(\left\|\frac{\mathbf{y} - \mathbf{w}_i}{h}\right\|^2\right) \delta\left[b(\mathbf{w}_i) - u\right], \tag{11}$$

where $\mathbf{w}_i$ are the pixels of the target and $b(\mathbf{w}_i)$ associates each $\mathbf{w}_i$ to its histogram bin [18]. The elliptic kernel $K_{e,\theta}(.)$ is used to lower the weight of the pixels that are closer to the border of the target. The normalization factor B ensures that the sum of the bins is 1.

The *motion likelihood*, $p(\mathcal{D}|\mathbf{x}_t)$, is computed as distance from the results of a change detector. The change detection is computed as a thresholded absolute frame difference on each channel of the RGB color space, fused using a logical OR operator. Median filtering and morphology are then applied to reduce the resulting noise. The median filtering is applied with a $N_s = 5$ kernel size and the morphology uses a $15x5$ elliptical structuring element to perform dilation. Sample change detection results are shown in Fig. 3. The motion likelihood from the detection is finally computed by applying a multi-variate Gaussian comprising of 4 dimensions as

$$p(\mathcal{D}|\mathbf{x}_t) = \mathcal{N}_4(\mu_{\mathcal{D}}, \sigma_{\mathcal{D}}). \tag{12}$$

The visual likelihoods are then fused with the audio likelihood, as described in the next section.

2.4. Audiovisual fusion

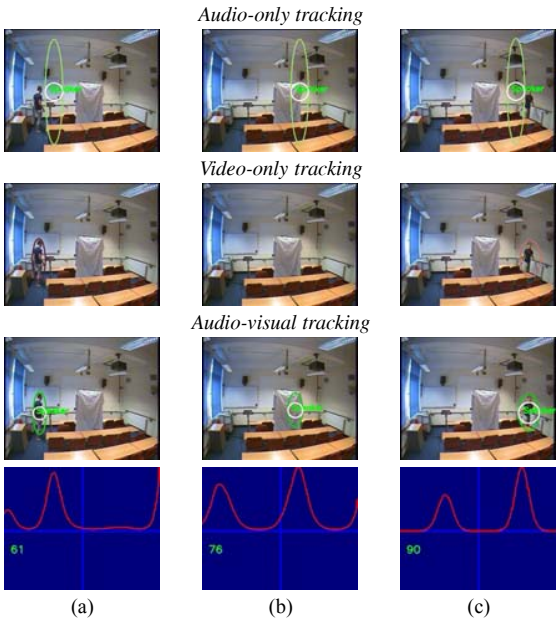
The cues are fused in the particle filter as product of the audio and visual likelihoods [12]. The overall likelihood is computed as

$$p(\mathcal{O}|\mathbf{x}_t) = p(\mathcal{D}|\mathbf{x}_t)p(\mathcal{C}|\mathbf{x}_t)p(\mathcal{A}|\mathbf{x}_t), \tag{13}$$

where $\mathcal{O}$ is the observation, $p(\mathcal{D}|\mathbf{x}_t)$ is the motion likelihood, $p(\mathcal{C}|\mathbf{x}_t)$ is the color likelihood, and $p(\mathcal{A}|\mathbf{x}_t)$ is the audio likelihood. When one modality is unavailable, its likelihood is set to 1.

Once $p(\mathcal{O}|\mathbf{x}_t)$ is computed, the weights are set proportional to the likelihood (Eq. 6). The final estimation of the state $\mathbf{x}_t$ at time t is computed based on the discrete approximation of Eq. (4) using the Monte Carlo approximation of the expectation:

$$\mathbf{E}[\mathbf{x}_t|\mathbf{z}_{1:t}] \approx \frac{1}{N_s} \sum_{i=1}^{N_s} \omega_t^i \mathbf{x}_t^i \tag{14}$$



(a) (b) (c)

Fig. 4. *Comparison of tracking results (sequence VO) using audio-only tracking (first row), video-only tracking (second row) and audio-visual tracking (third row), and the computed correlation after reverberation filtering (fourth row). Frames: (a) 804; (b) 922; (c); 996*

3. EXPERIMENTAL RESULTS

We demonstrate and evaluate the proposed multi-modal tracker on a dataset recorded with a STAC sensor composed of two Beyerdynamic MCE 530 condenser microphones and a KOBI KF-31CD analog CCD surveillance camera. The distance between the microphones is 95 cm and the video camera is located in the middle (Fig. 1(c)). The image resolution is 360x288 pixels (25 Hz) and the audio is sampled at 44.1 KHz. Sample videos are taken from the sequences *VO* (that contains a visual occlusion) and the sequence *SD* (that contains two moving speakers). The data were collected in a reverberant room with significant audio-visual background noise.

The evaluation is performed by computing the distance between the estimated track and the ground truth. The tracker is tested on scenarios without occlusions, with video occlusion, and with single and multiple targets. The normalized frequency bands are $B_1 = [0, 0.25]$, $B_2 = [0.25, 0.6]$, and $B_3 = [0.6, 1]$; with $f_{max} = 6000\,Hz$. The number of particles used by PF is $N_p = 200$. The variances $(\sigma_s, \sigma_d, \sigma_m)$ are set to 0.15 for the position parameters and to 10 for the size parameters. The onset interval is of $T = 6$ frames. *The parameters of the algorithm used were the same for all sequences.*

Figure 4 shows sample audio-visual target tracking results during a visual occlusion. The changes in the color of the ellipse correspond to the identity switches of a target. It is possible to notice that the audio-only tracker is capable of tracking the target *during* and *after* the occlusion and that there are no identity switches, although the accuracy is low. The video-only tracker fails during the visual occlusion and generates an identity switch when the target reappears. The audio-visual tracker correctly follows the target during occlusion and also improves the localization accuracy compared to the audio-only tracker. The improvement in the tracking accuracy is summarized in Table 3: an error reduction of 12-24 pixels is obtained when us-

GCC-PHAT	Plain	with RF	with RF and MB
Audio only	28.63	23.35	15.36
Audio-visual	4.47	3.93	3.47

Table 1. *Comparison of tracking accuracy results (sequence VO). Error reduction between audio-only tracking, audio-visual tracking with reverberation filtering (RF) and with RF and multi-band analysis (MB)*

ing audio-visual fusion compared to audio only. As the video-only tracker fails due to a track loss, its results are not considered in this comparison. Table 3 also shows the error reduction when using the reverberation filtering (RF vs. Plain) and the multi-band frequency analysis (RF and MB vs. RF).

Figure 5 shows an example of application of the proposed multimodal tracker to an active speaker detection and tracking scenario, with two people moving freely in a room and alternatively speaking. By comparing the sample results with the ground truth, it is possible to notice that the algorithm accurately detects the active speaker (white circle).

Although in this section the proposed method is evaluated without changing its parameter set or its configuration, the overall framework is modular and appropriate blocks can be optimized for the specific application at hand. For example, an illumination invariant change detector can be used for outdoor scenarios. Moreover, the framework is extensible and adding additional features is straightforward using the fusion at the likelihood level presented in Eq. (13).

4. CONCLUSIONS

We presented an audio-visual detection and tracking algorithm for STAC sensors. Audio observations are combined with video observations in a particle filter framework. Video observations generate two likelihoods that are based on the distance from an appearance model based on 3D color histograms and the distance from regions identified with a color change detection. Audio observations generate a source localization likelihood based on the time difference of arrival between the two microphones of the STAC sensor. The algorithm reduces localization distortions due to reverberations using the precedence effect and a multi-band frequency analysis, and can be used in indoor environments. Experimental results showed how audio information helps maintaining the track identity through visual occlusion. Moreover, the tracker can accurately select and follow an active speaker in the presence of significant noise and reverberation.

Future work includes the evaluation of the tracker in outdoor sequences and the comparison of the fusion strategy based on the product of the likelihoods with the one based on the weighted sum of likelihoods. This second solution could enable a better control of the likelihood when one of the modalities is available, but less reliable.

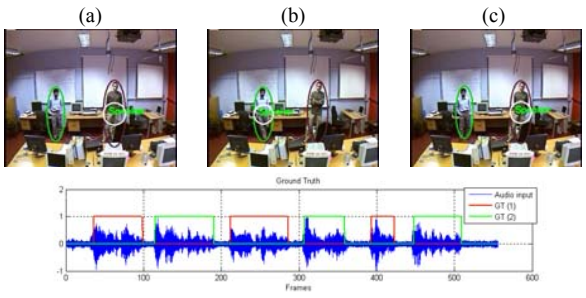


Fig. 5. *(top) Detection and tracking of alternating speakers using audio-visual cues for the sequence SD. (a) Frame 50, (b) Frame 313, (c) frame 425. (bottom) Ground truth of speaker detection: the green and the red lines represents the speaking activity of the two people*

5. REFERENCES

[1] 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*, Sept. 2001, vol. 1, pp. 354–357.

[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*, College Park, MD, USA, May 2004, vol. 3, pp. 793–796.

[3] I. Potamitis, C. Huimin, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. Speech Audio Processing*, vol. 12, pp. 520–529, Sept. 2004.

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

[5] 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, Apr. 2006.

[6] 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, pp. 25–28.

[7] I.A. McCowan D. Gatica-Perez, G. Lathoud and J.M. Odobez, "A mixed-state i-particle filter for multi-camera speaker tracking," in *IEEE Int. Conf. on Computer Vision Workshop on Multimedia Technologies for E-Learning and Collaboration*, Oct. 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, pp. 805–812.

[9] K. Nickel, T. Gehrig, H.K. Ekenel, J. McDonough, and R. Stiefelhagen, "An audio-visual particle filter for speaker tracking on the clear06 evaluation dataset," in *CLEAR*, Southampton, UK, Apr. 2006.

[10] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *IEEE Trans. Image Processing*, vol. 92, pp. 485–494, March 2004.

[11] V. Cevher, A. C. Sankaranarayanan, J. H. McClellan, and R. Chellappa, "Target tracking using a joint acoustic video system," *IEEE Trans. on Multimedia*, (to appear).

[12] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, May 2004, vol. 5, pp. 17–21.

[13] 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)*, 2000, pp. 1589–1592.

[14] 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.

[15] 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*, 2002, pp. 736–752.

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

[17] T. Sullivan and R. Stern, "Multi-microphone correlation-based processing for robust speech recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 1993, pp. 91–94.

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