IDIAP RESEARCH REPORT

Audio-Visual Speaker Tracking with Importance Particle Filters

Daniel Gatica-Perez [*] Guillaume Lathoud [*]
Iain McCowan [*] Jean-Marc Odobez [*]
Darren Moore [*]

IDIAP–RR 02-37

October 31, 2002

submitted for publication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

* IDIAP, Martigny, Switzerland

IDIAP Research Report 02-37

Audio-Visual Speaker Tracking with Importance Particle Filters

Daniel Gatica-Perez Guillaume Lathoud Iain McCowan
Jean-Marc Odobez Darren Moore

October 31, 2002

submitted for publication

Abstract. We present a probabilistic methodology for audio-visual (AV) speaker tracking, using an uncalibrated wide-angle camera and a microphone array. The algorithm fuses 2-D object shape and audio information via importance particle filters (I-PFs), allowing for the asymmetrical integration of AV information in a way that efficiently exploits the complementary features of each modality. Audio localization information is used to generate an importance sampling (IS) function, which guides the random search process of a particle filter towards regions of the configuration space likely to contain the true configuration (a speaker). The measurement process integrates contour-based and audio observations, which results in reliable head tracking in realistic scenarios. We show that imperfect single modalities can be combined into an algorithm that automatically initializes and tracks a speaker, switches between multiple speakers, tolerates visual clutter, and recovers from total AV object occlusion, in the context of a multimodal meeting room.

1 Introduction

Speaker tracking constitutes a relevant task for applications that include remote conferencing, HCI, and video indexing and retrieval. The use of audio and video as separate cues for tracking are classic problems in signal processing and computer vision. However, sound and visual information are jointly generated when people speak, and provide complementary advantages for speaker tracking if their dependencies are jointly modeled [9]. On one hand, initialization and recovery from failures - bottlenecks in visual tracking - can be robustly addressed with audio. In contrast, precise object localization is better suited to visual processing.

Probabilistic generative models, in which uncertainty is handled in a principled manner, are suitable for processing of multimodal information. For speaker tracking, several approaches have been proposed, including Bayesian networks [8], [1], and sequential Monte Carlo (SMC) [9], [10]. In particular, SMC, aka particle filters (PFs) represent an elegant methodology for data fusion [4]. For a typical state-space sequence model, a basic PF recursively approximates the filtering distribution of states given observations using a dynamical model and random sampling by (i) predicting candidate configurations, and (ii) measuring their likelihood. Tracking is therefore posed as a problem of random search in a configuration space.

Sampling from dynamics (prediction) and measuring (updating) are critical SMC stages in which data fusion can be introduced. Current formulations for AV speaker tracking fuse audio and video only at the measurement level [9], [10], thus leading to symmetrical models in which each modality accounts for the same relevance, solely depending on the dynamical model to generate candidate configurations. Additionally, AV sensors (cameras and microphones) tend to be independently calibrated for state modeling and measuring in 2-D or 3-D. Such formulations tend to overlook two important features of AV data. In the first place, audio is a strong cue to model discontinuities that clearly violate usual assumptions in dynamics (including speaker turns), speaker occlusion, and (re)initialization. Its use for sampling would therefore bring benefits to modeling realistic situations. In the second place, even though audio information might be inaccurate or biased, and visual calibration can be erroneous due to camera distortion, their joint occurrence tends to be consistent, and can be learned in a robust way from training data.

This paper presents a methodology for AV speaker tracking using PFs, and introduces novelties on data fusion and AV calibration. Given a state space defined on the image plane, audio information is used both for sampling and measuring. For sampling, 3-D audio localization computed at each frame is introduced in the PF formulation via importance sampling [5], [6], by defining an audio importance function that emphasizes the most informative regions of the configuration space. Importance PFs were introduced in [6] for visual tracking, and here we extend their use to multimodal fusion. For measuring, audio and video are jointly used to compute the likelihood of candidate configurations. We use a shape-based object representation [2], but the described approach is applicable to other visual cues. Finally, we describe a simple, yet robust AV calibration procedure that estimates a direct 3-D-to-2-D mapping for the audio localization signal onto the image plane. This procedure uses training videos of people configurations in an indoor setup, and does not require of precise geometric calibration of camera and microphones. The result is an algorithm that can robustly initialize and track a moving speaker, switch between multiple speakers, tolerate visual clutter, and recover from total AV object occlusion in a realistic video conferencing/meeting room viewed by a wide-angle camera. Other AV speaker tracking methods would find limitations in these settings.

The paper is organized as follows. Section 2 presents our algorithm. Section 3 describes the experimental setup. Section 4 presents results. Section 5 provides some final remarks.

2 Our approach

Given a discriminative object representation and a classic Markov state-space model, with hidden states $\{\mathbf{x}_t\}$ that represent an object's configuration, and observations $\{\mathbf{y}_t\}$ extracted from an AV sequence, the filtering distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ can be recursively computed by

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t) \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1}, \tag{1}$$

where $\mathbf{y}_{1:t} = \{\mathbf{y}_1, ..., \mathbf{y}_t\}$. The integral in Eq. 1 represents the prediction step, in which the dynamical model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and the previous distribution $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$ are used to compute a prediction distribution, which is used as then prior for the update step, and multiplied by the likelihood $p(\mathbf{y}_t|\mathbf{x}_t)$ to generate the current filtering distribution. A PF approximates Eq. 1 for non-linear, non-Gaussian problems, as follows. The filtering distribution is approximated by a set of weighted samples or particles $\{(\mathbf{x}_t^{(i)}, \pi_t^{(i)}), i = 1, ..., N\}$, where $\mathbf{x}_t^{(i)}$ and $\pi_t^{(i)}$ denote the i-th sample and its importance weight at the current time. The point-mass approximation is given by $\hat{p}_N(\mathbf{x}_t|\mathbf{y}_{1:t}) = \sum_{i=1}^N \pi_t^{(i)} \delta(\mathbf{x}_t - \mathbf{x}_t^{(i)})$. The prediction step propagates each particle according to the dynamics, and the updating step reweights them using their likelihood, $\pi_t^{(i)} \propto \pi_{t-1}^{(i)} p(\mathbf{y}_t|\mathbf{x}_t^{(i)})$. A resampling step using the new weights is necessary to avoid degradation of the particle set [4].

The design of a PF for AV tracking involves the definition of the object representation, the state-space, the dynamical process, the sampling strategy, the AV calibration procedure, and the probability models for sound and visual observations. Each of these issues are discussed in the following subsections.

2.1 Object modeling, state space and dynamics

Object representations and state-spaces defined either on the image plane or in 3-D space are sensible choices. However, while 3-D allows for more elaborate object modeling, it also requires precise camera calibration and computation of non-trivial features [10]. We follow an image-based standard approach in which object contours are modeled as elements of a shape-space, allowing for the description of a shape template and a set of valid geometric transformations [2]. In our case, the basic shape is a parameterized ellipse, suitable for tracking heads, and we have chosen a subspace of the Euclidean transformations comprising translation T^x, T^y and scaling s. Furthermore, a second-order auto-regressive dynamical model is defined on these parameters. With an augmented state defined by $\mathbf{x}_t = (x_t, x_{t-1})^T$, and $x_t = (T_t^x, T_t^y, s_t)$, the dynamical model is defined by $\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{w}_t$, where A, B are the parameters of the model, and $\mathbf{w}$ is a white noise process.

2.2 Importance Particle Filters

Basic PFs rely only on the dynamical model to generate candidate configurations, which -as discussed earlier- has limitations due to imperfect motion models, object occlusion, and the need for (re)initialization, (e.g. due to a speaker turn). Additional knowledge about the "true" configurations can be extracted from other modalities, and modeled via importance sampling [5], [6], by using an importance function $i_t(\mathbf{x}_t)$ that emphasizes the most informative regions of the space. The technique first draws samples from $i_t(\cdot)$ rather than from the filtering distribution, concentrating particles in better proposal regions. It then introduces a correction mechanism in order to keep the particle set as a faithful representation of the original distribution, defined by an importance ratio,

$$w_t^{(i)} = \frac{\hat{p}_N(\mathbf{x}_t^{(i)}|\mathbf{y}_{1:t-1})}{i_t(\mathbf{x}_t^{(i)})} = \frac{\sum_{k=1}^N \pi_{t-1}^{(k)} p(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(k)})}{i_t(\mathbf{x}_t^{(i)})}, \tag{2}$$

and applied to the particle weights, $\pi_t^{(i)} \propto w_t^{(i)} p(\mathbf{y}_t|\mathbf{x}_t^{(i)})$. Reinitialization is introduced by using a two-component mixture,

$$\hat{p}(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \alpha q_t(\mathbf{x}_t) + (1-\alpha)\hat{p}_N(\mathbf{x}_t|\mathbf{y}_{1:t-1}), \tag{3}$$

where $q_t(\mathbf{x}_t)$ denotes a reinitialization prior, and $\{\alpha, 1-\alpha\}$ is the prior on the mixture components. Furthermore, a mixed-state model can be introduced, in which samples are drawn from the original dynamical model, the dynamics (with IS), and the reinitialization prior [6].

We extend the previous use of I-PFs to multimodal fusion. Audio tends to be imprecise for localization, due to discontinuities during periods of silence, as well as effects of reverberation and other noise. Audio does have some important advantages however, such as the ability to provide instantaneous localization at reasonable computational expense, which is well-suited for initialization. In this paper, audio localization information is used for the IS function, the reinitialization prior, and the measurement process.

2.3 AV calibration

Current AV calibration works assume simplified configurations [9], [1], or resort to rigorous camera calibration procedures [10]. However, camera calibration models become more complex for wide-angle lenses (a usual requirement in video conferencing/meeting).

Despite the facts that audio information is usually noisy, and that visual calibration can be erroneous due to geometric distortion, their joint occurrence tends to be more consistent. We have therefore opted for a rough AV calibration procedure, which estimates a mapping from audio configurations in 3-D onto the image plane from training sequences, without requiring precise geometric calibration of audio and video. For this purpose, we collected sequences with people speaking while performing typical activities in the specific setup (walking, sitting and standing, moving on their seats). The audio localization procedure described in Section 2.5 was used to compute 3-D points X_t for each frame and a hand-initialized visual tracker was used to compute the corresponding points in the image plane. The correspondences were used to define a mapping between discrete sets $C : \mathcal{R}^3 \rightarrow \mathcal{R}^2$, such that $C(X_t) = (T^x, T^y)$. For new, unseen data, the mapping is directly implemented by nearest neighbor search.

2.4 Visual observations model

The observation model assumes that object shapes are embedded in clutter. Edge-based measurements are computed along L normal lines to a hypothesized contour, resulting in a vector of candidate positions for each line, $\mathbf{y}_t^l = \{\nu_m^l\}$ relative to the point lying on the contour ν_0^l. With some usual assumptions, the observation likelihood for L normal lines can be expressed as

$$p(\mathbf{y}_t^{vid}|\mathbf{x}_t) \propto \prod_{l=1}^{L} p(\mathbf{y}_t^l|\mathbf{x}_t) \propto \prod_{l=1}^{L} \max\left(K, \exp(-\frac{||\hat{\nu}_m^l - \nu_0^l||^2}{2\sigma^2})\right), \tag{4}$$

where $\hat{\nu}_m^l$ is the nearest edge detected on l^{th} line, and K is a constant introduced when no edges are detected.

2.5 Audio observation model

In general, audio localization methods rely on estimation of the delay between the time of arrival of a signal on a pair of microphones. We define the vector of theoretical time delays associated with a 3-D location X as $\boldsymbol{\tau}^{1:M,X} = \{\tau^{m,X}\} \triangleq \boldsymbol{\tau}^X$, where $\tau^{m,X}$ is the delay (in samples) between the microphones in pair m,

$$\tau^{m,X} = \frac{(||X - M_1^m|| - ||X - M_2^m||)\, f_s}{c} \tag{5}$$

where M_1^m and M_2^m are the locations of the microphones in pair m, and f_s is the sampling frequency. In practice, each time delay estimate $\hat{\tau}_t^m$ is calculated from the generalized cross-correlation (GCC) [7]. A phase transform (PHAT) is applied to improve the robustness to reverberation, and the GCC is interpolated to achieve sub-sample precision. Full details of this time delay estimation procedure can be found in [3]. Then, given an vector $\hat{\boldsymbol{\tau}}_t \triangleq \{\hat{\tau}_t^m\}$ of observed time delay estimates, the distribution of the observation given a speaker at location X can be modeled as $p(\hat{\boldsymbol{\tau}}_t|\boldsymbol{\tau}^X) = \mathcal{N}(\boldsymbol{\tau}^X, \Sigma^X)$, where Σ^X is the covariance matrix, which can be chosen to be independent of location. The location estimate can then be defined according to the maximum likelihood criterion as $\hat{X}_t = \arg\max_X p(\hat{\boldsymbol{\tau}}_t|\boldsymbol{\tau}^X)$. The localization estimate for each frame is found by a dynamic search over $p(\hat{\boldsymbol{\tau}}_t|\boldsymbol{\tau}^X)$ through a uniform grid of room locations. In order to eliminate low confidence values, estimates whose likelihood falls below that of a uniform distribution are labeled as silence, meaning the audio observations contain discontinuities. To synchronize the audio and video frame rates, multiple audio frames are merged by selecting only the maximum likelihood location across frames.

2.6 Defining the importance sampling function

Recall that the transformation space includes translation and scaling. Assuming independence, we define $i_t(x_t) = i_t(T_t^x, T_t^y, s_t) = \mathcal{N}(\mu_t, \Sigma_t)$. The mean $\mu_t = (\mu_t^{T^x}, \mu_t^{T^y}, \mu_t^s)$ consists of the projected 3-D audio estimate onto the image plane $C(\hat{X}_t)$ and the unit scale. The covariance matrix Σ_t is diagonal, with translation components proportional to the mean head size in the training set, and with scaling component equal to the variance in scale of head sizes in the training set. In case of silence, no importance sampling function can be generated, so the filter draws samples only from the dynamical model. The importance function is also used for the audio-based observation likelihood,

$$p(\mathbf{y}_t^{aud}|\mathbf{x}_t) \propto i_t(\mathbf{x}_t), \tag{6}$$

in case there is audio, and it is a fixed constant otherwise.

2.7 AV fusion for measurement

Observations are combined in a standard approach,

$$p(\mathbf{y}_t|\mathbf{x}_t) = p(\mathbf{y}_t^{vid}|\mathbf{x}_t)p(\mathbf{y}_t^{aud}|\mathbf{x}_t). \tag{7}$$

3 Experimental setup

Audiovisual recordings were made in a meeting room with one wide-angle camera on a wall and an 8-microphone array on the table (Fig. 1). A rate of 25 fps was used for the camera, while the audio was recorded at 16kHz, with features estimated at 62.5 fps. Images were processed in CIF format. In this setup, a human head is about 20×35 pixels (1 pixel $\approx$ 8mm).

Several parameters for the visual tracker (dynamics and observations) have been set by hand, and kept fixed for all experiments. There are of course suitable methods to estimate some of them, especially those of dynamical models [2].

4 Results

4.1 Audio localization evaluation

We projected the audio estimates of several known locations onto the vertical plane, as well as on the azimuth/elevation space (see Figure 2). A strong bias is apparent, especially on the vertical axis, and justifies the need for learning the mapping between audio estimates and position on the image plane.

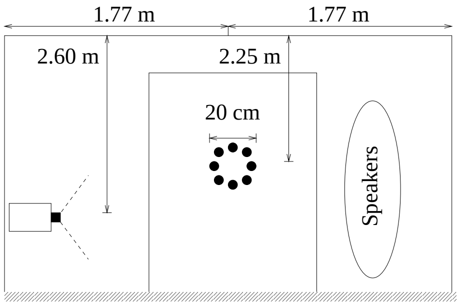

Figure 1: Meeting room

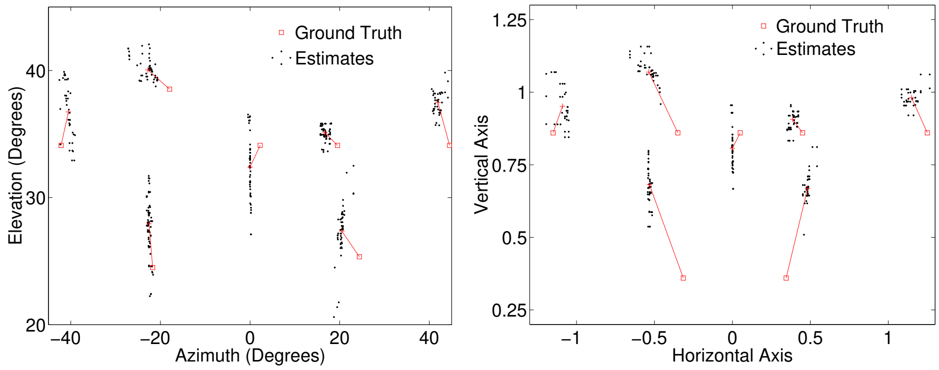

Figure 2: Audio localization of a still speaker at known locations.

4.2 AV tracking

Our research results (including a discussion on the performance of single modalities, and the effects of clutter and occlusion) should be fully appreciated by looking directly at the AV sequences, available on a website that accompanies this report[1].

Results of single-person tracking using 500 particles are shown in Fig. 3 The corresponding objective evaluation is presented in Fig. 5 (top), based on a manually generated ground-truth of the face center as the sound source. The tracker is automatically instantiated when the person enters the scene (video frame 18) and starts talking, and remains on track when speech ceases. Audio data are non-continuous (104 audio samples in 228 video frames). Mean (resp. median) errors in pixels for audio, hand-initialized video, and AV are 18.9 (11.5), 2.3 (2.0), and 2.8 (2.3), respectively. The audio localization error is the combined effect of 3-D localization and the mapping onto the image plane. The largest errors are due to the detection of footsteps as the sound source, but the tracker copes with these short-term distractions (see Fig. 3).

An example of speaker tracking in presence of visual distractions is shown in Fig. 4. A second person (who introduces visual clutter) repeatedly passes behind the speaker without distracting the tracker. Again, the system initializes automatically and tracks the speaker for almost 30 seconds. In this case, the main source of error comes from errors in the audio localization estimate and the mapping to the image plane (which momentarily drives the tracker away), and the presence of two major modes (the entire head and the face) in the shape observation likelihood.

Fig. 6 illustrates multiple speaker tracking in a meeting scenario (the original images have been cropped). Three people seated at the table speak in turn (center, right, left), starting at frames 38, 213 and 539, respectively. Evaluation on the first 700 frames is shown in Fig. 5 (bottom). The tracker requires some frames to lock to the correct speaker (mainly due to lack of audio samples from

[1] www.idiap.ch/~gatica/av − tracking.html.

which to build an IS function) but eventually succeeds. For the center and left speakers, the audio error after mapping to the image is substantially larger than for the right speaker. However, this rough initialization is good enough for head tracking. The main source of error for AV tracking is the fitting of the contour template onto the neck/shirt contour rather than onto the chin but error remains approx. below 10 pixels. A momentary sound overlap due to whispering by a second speaker (at around frame 390) confuses the AV tracker, but it rapidly relocks onto the main speaker. Once the AV tracker has automatically switched speaker, the AV error is essentially equivalent to the one obtained with hand-initialized video.

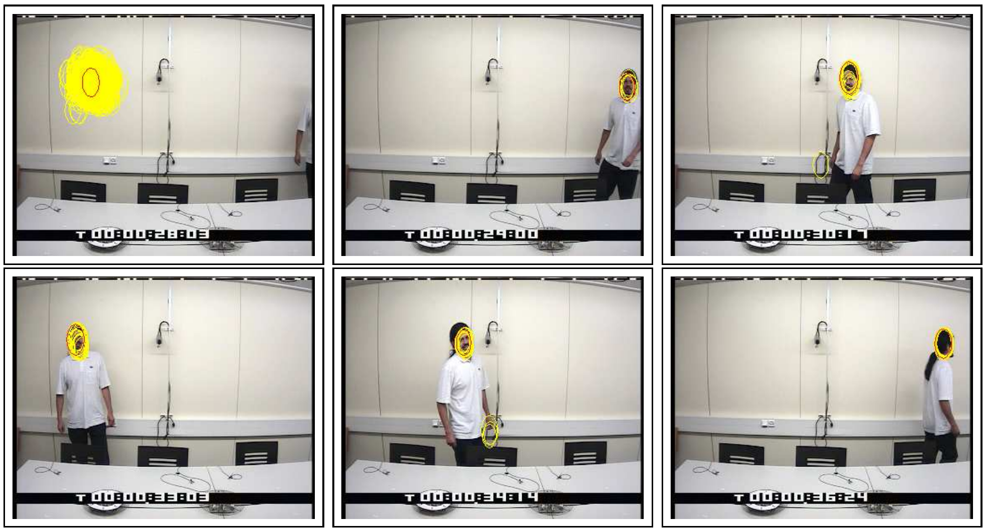

Figure 3: AV tracking. *Walking* sequence. Initialization is random, but the tracker correctly locks when a speaker enters the scene.

The effect of AV fusion in the observation likelihood is shown in Fig. 7. Video observations based only on shape are clearly limited to discriminate between two different human heads. The same would apply even for an observation likelihood based on specific face features. In presence of multiple people or visual clutter, the shape likelihood is multimodal. Each peak of the distribution corresponds to a shape configuration for which the center of most measurement lines (described in Section 2.4) happen to lie on image edges. Note that this does not necessarily translate into an "ellipse-like" shape [2]. In our case, particles with large weights are generated for both heads (Fig. 7, top row). Furthermore, the mean configuration is a bad representation of the distribution, as it lies somewhere between the peaks of the distribution without corresponding to any object (Fig. 7, middle row). In contrast, fusing audio and video in the observation process (i.e., Eq. 7) solves the above ambiguity: speaker turns are correctly tracked, and the tracker locks only onto the current speaker (Fig. 7, bottom row).

Finally, an example of speaker tracking in which there exist occlusion in both audio and video is shown in Fig. 8. The sequence is challenging: there is a physical obstacle (a second person) between the speaker and the microphone array, and due to motion, the speaker is sometimes fully occluded by another person. Using only video (again, initialized by hand) is clearly not sufficient, as the visual tracker cannot recover from a slow visual occlusion (Fig. 8, first three rows). The use of audio and video should improve performance. At the beginning of the sequence, the speaker cannot be tracked due to the lack of audio localization information (audio occlussion). However, when a direct path between the sound source and the microphone array becomes available, the algorithm locks onto the speaker. As discussed in the previous example, observations based only on video cannot discriminate between a speaker and another person, and the tracker fails in the long term (Fig. 8, three middle rows). In contrast, fusing AV in the observation process considerably improves the tracking quality. Although visual tracking by itself cannot handle occlussion, the tracker recovers in repeated cases thanks to the audio modality, using video to provide finer localization, and recovers the speaker

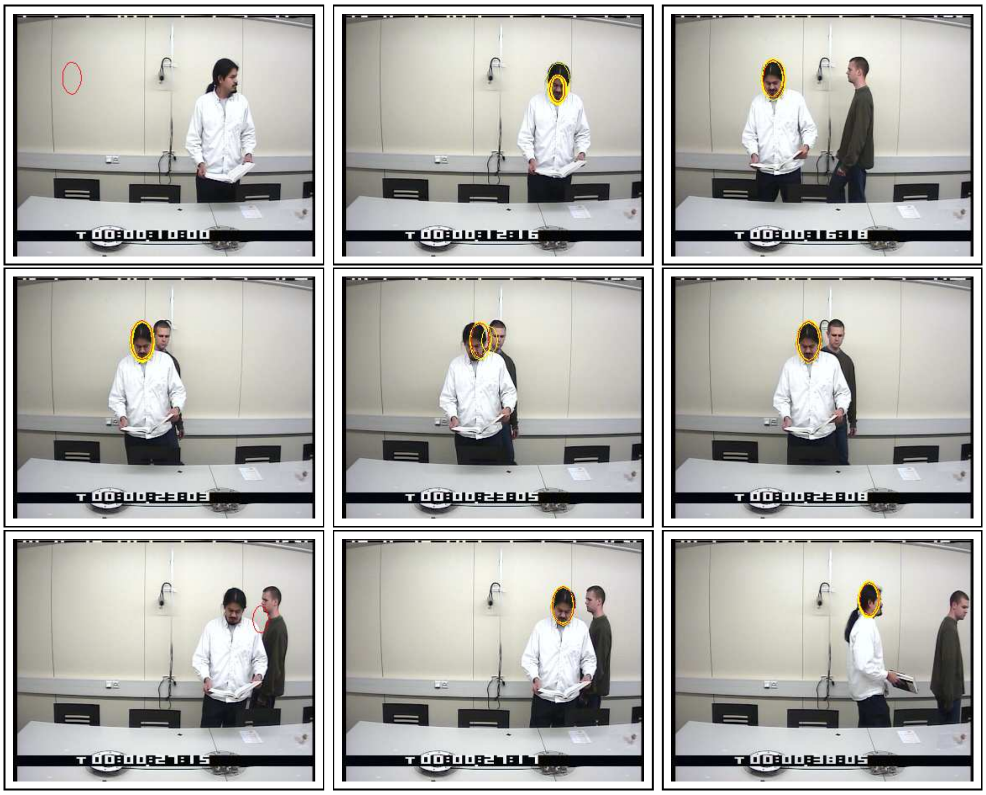

Figure 4: AV tracking. *Reading* sequence. The system automatically locks onto the speaker (top row), and stays on track despite momentary visual distractions and short-term sound-based localization errors (middle and bottom rows).

even after considerable AV occlusion (Fig. 8, last three rows). To our knowledge, previous work on multimodal speaker tracking has not fully discussed performance in cases of AV occlusion [9], [1].

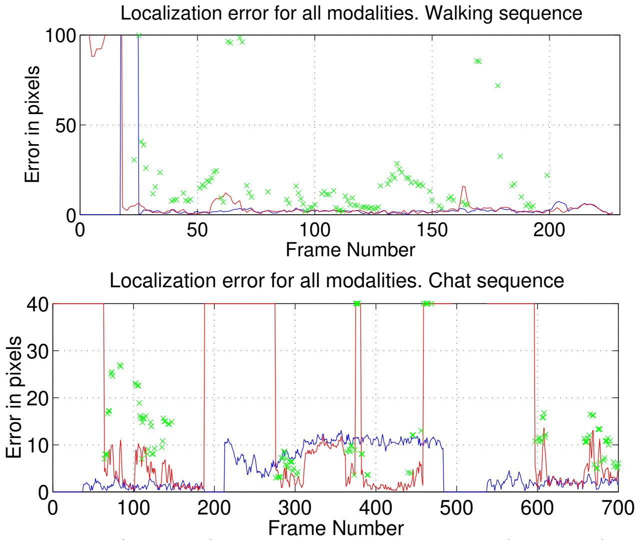

Figure 5: Tracking error for audio (green ×), video initialized by hand (blue line), and AV (red line). *Walking* and *Chat* sequences.

Figure 6: *Chat* sequence: switching between multiple speakers.

Figure 7: AV fusion in the observation likelihood (Eq. 7). The sole use of video in the computation of the likelihood is insufficient. The object model does not provide enough discrimination between different heads, so particles are locked onto two people when a speaker turn occurs (top row). The filtering distribution now has several modes, and the mean configuration therefore lies there somewhere between the modes (middle row). In contrast, fusing audio and video in the observation process solves the above ambiguity: speaker turns are correctly tracked, and the tracker locks only onto the current speaker.

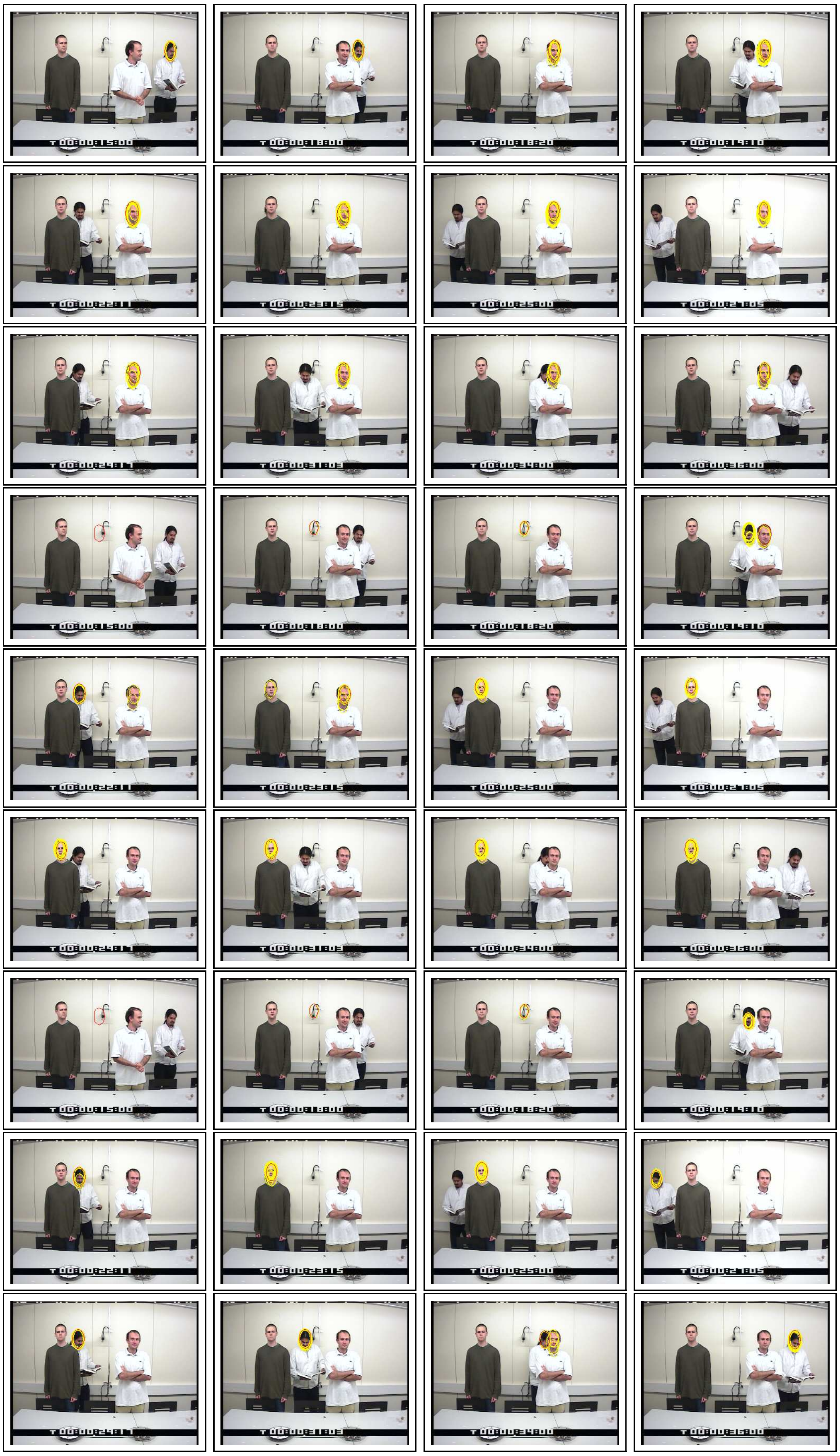

Figure 8: Speaker tracking in cases of AV occlusion. Three top rows: video-only tracker, hand-initialized. Three middle rows: AV tracker, visual observation likelihood. Three bottom rows: AV tracker, AV likelihood.

5 Conclusion and future work

We have shown that AV fusion via importance particle filters makes good use of the complementary advantages of individual modalities for speaker tracking. In addition to its principled formulation, our framework has shown to be robust in practice to many difficult situations (automatic speaker initialization anywhere in the room, and capabilities to switch between multiple speakers, tolerance to visual clutter, and repeated recovery from total AV object occlusion). Of course, there is room for improvements. Limitations for individual modalities are well-known, like the use of single visual cues, and the poor resolution of audio-only-based localization. Regarding AV tracking, audio estimates not incorporating temporal assumptions result in individual errors that could potentially drift the tracker away if they occurred often. Future work will concentrate in two areas. The first one is the integration of color distribution object models in our framework. Personalized color models usually require initialization. We are studying of generation of initial color object models, based on audio localization, and further fusion of shape/color and audio for tracking. The second area is an extension to deal with multiple simultaneous speakers.

References

[1] M. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *Proc. ECCV*, May 2002.

[2] A. Blake and M. Isard, *Active Contours*, Springer-Verlag, 1998.

[3] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proceedings of ICASSP-96*, 1996.

[4] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, 2001.

[5] A. Gelman, J. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, Chapman and Hall, 1995.

[6] I. Isard and A. Blake, "Icondensation: Unifying low-level and high-level tracking in a stochastic framework," in *Proc. ECCV*, Freiburg, June 1998.

[7] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics and Speech Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, August 2000.

[8] V. Pavlovic, A. Garg, and J. Rehg, "Multimodal speaker detection using error feedback dynamic bayesian networks," in *Proc. IEEE CVPR*, Hilton Head Island, SC, 2000.

[9] J. Vermaak, M. Gagnet, A. Blake, and P. Perez, "Sequential monte-carlo fusion of sound and vision for speaker tracking," in *Proc. IEEE ICCV*, Vancouver, July 2001.

[10] D. Zotkin, R. Duraiswami, and L. Davis, "Multimodal 3-d tracking and event detection via the particle filter," in *IEEE Workshop on Detection and Recognition of Events in Video*, Vancouver, July 2001.