IDIAP RESEARCH REPORT

AUDIO-VISUAL PROBABILISTIC TRACKING OF MULTIPLE SPEAKERS IN MEETINGS

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

IDIAP–RR 05-27

JUNE 2005

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 Research Report 05-27

AUDIO-VISUAL PROBABILISTIC TRACKING OF MULTIPLE SPEAKERS IN MEETINGS

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

JUNE 2005

SUBMITTED FOR PUBLICATION

Abstract. Tracking speakers in multiparty conversations constitutes a fundamental task for automatic meeting analysis. In this paper, we present a probabilistic approach to jointly track the location and speaking activity of multiple speakers in a multisensor meeting room, equipped with a small microphone array and multiple uncalibrated cameras. Our framework is based on a mixed-state dynamic graphical model defined on a multiperson state-space, which includes the explicit definition of a proximity-based interaction model. The model integrates audio-visual (AV) data through a novel observation model. Audio observations are derived from a source localization algorithm. Visual observations are based on models of the shape and spatial structure of human heads. Approximate inference in our model, needed given its complexity, is performed with a Markov Chain Monte Carlo particle filter (MCMC-PF), which results in high sampling efficiency. We present results -based on an objective evaluation procedure- that show that our framework (1) is capable of locating and tracking the position and speaking activity of multiple meeting participants engaged in real conversations with good accuracy; (2) can deal with cases of visual clutter and partial occlusion; and (3) significantly outperforms a traditional sampling-based approach.

1 Introduction

The automatic analysis of meetings recorded in multi-sensor rooms is an emerging research field in various domains, including audio and speech processing, computer vision, human-computer interaction, and information retrieval [23, 46, 32, 41, 8, 30, 47, 10]. Analyzing meetings poses a diversity of technical challenges, and opens doors to a number of relevant applications, including automatic structuring and indexing of meeting collections, and facilitation of remote meetings.

In the context of meetings, localizing and tracking people and their speaking activity play fundamental roles in two areas. The first one is media processing: speaker location is useful to select or steer a camera as part of a visualization or production model, to enhance the audio stream via microphone-array beamforming for speech recognition, to provide accumulated information for person identification, and to recognize location-based events (e.g. a presentation). The second one is human interaction analysis: social psychology has highlighted the role of non-verbal behavior (e.g. gaze and facial expressions) in interactions, and the correlation between speaker turn patterns and aspects of the behavior of a group [31]. Extracting cues to identify such multimodal behaviors requires reliable speaker tracking.

Although the tracking task in meetings is facilitated by the constraints of the physical space and the expected type of human activities, the multimodal multispeaker tracking problem poses various challenges. These include algorithms for AV data fusion, that make use of the modalities' complementarity, and for tractable joint multiperson models (which represent each individual state, while accounting for the constraints introduced by their interaction). In meetings, interaction in its simplest form relates to proximity in the visual modality (occlusion being the fundamental case), and to overlapping speech in the audio modality (commonly found in spontaneous conversations [38]). Approaches addressing some of these issues have begun to appear [5, 6].

In this paper, we address the problem as one of approximate inference in a dynamical graphical model, using particle filtering (PF) techniques [18, 11], building on recent advances in the field [21]. For a state-space model, a PF recursively approximates the filtering distribution of states given observations, using a dynamical model, an observation model, and sampling techniques, by predicting candidate configurations and measuring their likelihood. Our model uses a mixed-state, multi-object state-space, which in addition to being mathematically rigorous, allows for the integration of a pairwise person occlusion model, through the addition of a Markov Random Field (MRF) prior in the multi-object dynamic model. To address the problems of traditional PFs in handling high-dimensional spaces (defined by the joint multi-object configurations), we combine Markov Chain Monte Carlo (MCMC) techniques with the PF framework, allowing for efficient sampling [28, 21]. Our work integrates data captured by a small microphone array and multiple cameras with mostly non-overlapping fields-of-view through the definition of a novel observation model of AV features. Based on an objective evaluation of the quality of estimated head location and speaking activity, and using small-group discussion data, we show that our framework is capable of jointly tracking the location and speaking activity of multiple meeting participants with good accuracy, dealing with realistic conditions, and outperforming a traditional PF model.

The paper is organized as follows. Section 2 discusses related work. Section 3 summarizes our framework. Section 4 describes the multi-person dynamical model. Section 5 presents the multi-person AV observation model. Section 6 describes the sampling technique. Section 7 discusses the case of varying numbers of people. Section 8 describes and discusses experiments and results. Finally, section 9 provides some concluding remarks.

2 Related work

Localizing and tracking speakers in enclosed spaces using AV information has increasingly attracted attention in signal processing and computer vision [36, 17, 7, 34, 13, 43, 48, 1, 3, 6, 5], given the complementary characteristics of each modality. Broadly speaking, the differences among existing works arise from the overall goal (tracking single vs. multiple speakers), the specific detection/tracking framework, and the AV sensor configuration. Much work has concentrated on the single-speaker case, assuming either single-person scenes [7, 34, 1], or multiperson scenes where only the location of the current speaker needs to be tracked [36, 17, 13, 43, 48, 3]. Many of these works used simple sensor configurations (e.g. one camera and a microphone pair) [7, 34, 43, 3]. Among the existing techniques, probabilistic generative models based on exact [34] or approximate inference methods (both variational [3] and sampling-based [43, 48]) appear to be the most promising, given their principled formulation and demonstrated performance.

None of the above works can handle the problem addressed here: continuously inferring, from audio and video data, the location and speaking status for several people in a realistic conversational setting. In fact, although audio-based multispeaker tracking and vision-based multiobject tracking have been studied for a few years as separate problems in signal processing [42, 37, 45, 25] and computer vision [19, 35], respectively, the AV multispeaker tracking problem has been studied only relatively recently, making use of more complex sensor configurations [8, 20, 39, 6, 5]. While single cameras are useful for remote conferencing applications, multiperson conversational settings like meetings often call for the use of multiple cameras and microphones to cover an entire workspace (table, whiteboards, etc.) [8, 6, 5]. More specifically, the work in [8] described a system based on a device that integrates a small circular microphone array and several calibrated cameras, whose views are composed into a panorama, The system, in which each person is tracked independently, consists of three modules: AV auto-initialization, (using either a standard acoustic source localization algorithm or visual cues), visual tracking using a Hidden Markov Model (HMM), and tracking verification. The work in [20] described a non-probabilistic multispeaker detection algorithm using an omnidirectional camera (which has limitations of resolution) and a microphone array, calibrated with respect to each other. At each video frame, the method extracts skin-color blobs by traditional techniques, and then detects a sound source using standard beamforming on the small set of directions indicated by the skin-blob locations. The work in [39] described an AV multispeaker system, based on a stereo camera and a linear microphone array, consisting of three separate modules: stereo-based, visual tracking of 3-D head location and pose for each person independently, estimation of the audio signal's direction of arrival with the microphone array, and estimation of audio-visual synchronous activity. Two hypothesis tests are used to make independent decisions about the speaking activity and visual focus of the speakers, based on simple statistical models defined on the observations derived from each module.

To the best of our knowledge, the closest works to ours are [5, 6], both based on PF techniques. The work in [5] used two calibrated cameras and four linear sub-microphone arrays on a wall, and was based on the model proposed in [19], defining a multi-person state-space in which the number of people can vary over time. An full-body multi-person observation model was defined by two terms: one for video, derived from a pixelwise background substraction model, and one for audio, derived from a set of short-time Fourier transforms computed on each microphone's signal. The PF relied on importance sampling (IS), and is thus likely to rapidly become inefficient as the number of objects increases. The work in [6] used the same calibrated sensor setup as [8], and tracked multiple speakers with a set of independent PFs, one for each person. Each PF uses a mixture proposal distribution, in which the mixture components are derived from the output of single-cue trackers (based on audio, color, or shape information). This proposal increases robustness in case of tracking failure in single modalities.

As we describe in the remainder of the paper, our work substantially differs from previous work in AV multispeaker tracking with respect to multi-object dynamic and AV observation modeling, and to the sampling mechanism. Building on the model in [21], our model has two advantages over [5, 6]. First, unlike [5, 6], we use a multi-person dynamical model that explicitly incorporates a pairwise person interaction prior term. This model is especially useful to handle person occlusion. Second, unlike [5], we use efficient MCMC sampling, that allows to track several objects in a tractable manner (effectively close to the case of independent PFs), while preserving the rigorous joint state-space formulation. Finally, we objectively evaluate the performance of our algorithm, in more detail than that of [5, 6].

3 Model formulation

We use a generative approach to model the tracking problem [18, 11]. Given a Markov state-space model, with hidden states $\mathbf{X}_t$ representing the joint multi-object configuration (e.g. position, scale, etc.), and AV observations $\mathbf{Y}_t$, the filtering distribution $p(\mathbf{X}_t|\mathbf{Y}_{1:t})$ of $\mathbf{X}_t$ given the observations $\mathbf{Y}_{1:t} = (\mathbf{Y}_1, ..., \mathbf{Y}_t)$ up to time-step t can be recursively computed using Bayes' rule by

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t}) \propto p(\mathbf{Y}_t|\mathbf{X}_t) \cdot \int_{\mathbf{X}_{t-1}} p(\mathbf{X}_t|\mathbf{X}_{t-1})p(\mathbf{X}_{t-1}|\mathbf{Y}_{1:t-1})d\mathbf{X}_{t-1}, \quad (1)$$

where $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ is the dynamical model of the temporal evolution of the multi-object state-space, and $p(\mathbf{Y}_t|\mathbf{X}_t)$ denotes the observation likelihood, which measures how well the observations fit the multi-object predictions.

PFs approximate Eq. 1 for non-linear, non-Gaussian problems [11]. From the various available formulations, we follow the one described in [18]. The basic PF represents the filtering distribution using a weighted set of samples $\{(\mathbf{X}_t^{(n)}, w_t^{(n)}), n = 1, ..., N\}$, where $\mathbf{X}_t^{(n)}$ and $w_t^{(n)}$ denote the n-th sample and its associated weight at each time-step, and updates this representation as new data arrive. With this representation, Eq. 1 can be approximated by a mixture model,

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t}) \approx \mathcal{Z}^{-1} p(\mathbf{Y}_t|\mathbf{X}_t) \sum_n w_{t-1}^{(n)} p(\mathbf{X}_t|\mathbf{X}_{t-1}^{(n)}), \quad (2)$$

using IS ($\mathcal{Z}$ is a normalization constant). Given the particle set at the previous time-step, $\{(\mathbf{X}_{t-1}^{(n)}, w_{t-1}^{(n)})\}$, a set of new configurations at the current time-step are drawn from a proposal distribution $q(\mathbf{X}_t) = \sum_r w_{t-1}^{(r)} p(\mathbf{X}_t|\mathbf{X}_{t-1}^{(r)})$. The weights are then computed as $w_t^{(n)} \propto p(\mathbf{Y}_t|\mathbf{X}_t^{(n)})$.

A state at time-step t is defined by $\mathbf{X}_t = (\mathbf{X}_{i,t}), i \in \mathcal{I}_t$, where $\mathcal{I}_t$ is the set of object identifiers in the configuration, $m_t = |\mathcal{I}_t|$ denotes the number of objects, and $|\cdot|$ indicates set cardinality. Each object has a unique identifier, given by the position occupied by their configuration in the state vector. In what follows, we assume $\mathcal{I}_t$ to be fixed (the case when $\mathcal{I}_t$ varies over time is discussed in Section 7). A mixed state-space is defined for single-object configurations $\mathbf{X}_{i,t}$, where both the geometric transformations of a person's model template in the image plane and the speaking activity are tracked. In the specific implementation described in this work, a person is represented by the elliptical silhouette of the head in the image plane (Fig. 1(left)). Furthermore, a single-object state $\mathbf{X}_{i,t} = (\mathbf{x}_{i,t}, k_{i,t})$ is composed of a 3-D continuous vector $\mathbf{x}_{i,t} = (u_{i,t}, v_{i,t}, s_{i,t})$, defined over a subspace of affine transformations comprising 2-D translation and scaling, and a discrete binary variable $k_{i,t}$, which models each participant's speaking status (0: silent, 1: speaking).

The generative model in Eq. 1 and its approximation in Eq. 2 require the specification of the dynamical and observation models. Additionally, the dimension of the multi-object state-space grows linearly with the number of objects, so that even for a small group discussion (4-5 participants) and the compact single-object state-space described above, the dimension of the joint state-space is prohibitively high for IS, which calls for a more efficient sampling scheme. These issues are discussed in the following sections.

4 Multi-object dynamical model

The dynamical model includes two factors: one that describes interaction-free, single-object dynamics, and another one that explicitly models interactions (e.g. occlusion), constraining the dynamics of each object based on the state of the others, via a pairwise MRF prior [26, 21]. The field is defined on an undirected graph, where the graph vertices are the objects, and the links are defined between object pairs at each time-step. The dynamical model is expressed as

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) \propto \left(\prod_{i \in \mathcal{I}_t} p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1}) \right) \left(\prod_{(i,j) \in \mathcal{C}} \phi(\mathbf{X}_{i,t}, \mathbf{X}_{j,t}) \right), \quad (3)$$

where $p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1})$ denotes the dynamics of the i-th object, and the prior is a product of pairwise potentials, denoted by $\phi(\mathbf{X}_{i,t}, \mathbf{X}_{j,t})$ over the set of cliques $\mathcal{C}$ (i.e., pairs of connected nodes) in the graph. The approximation in Eq. 2 is now given by

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t}) \approx \mathcal{Z}^{-1} p(\mathbf{Y}_t|\mathbf{X}_t) \left(\prod_{(i,j)\in\mathcal{C}} \phi(\mathbf{X}_{i,t},\mathbf{X}_{j,t}) \right) \left(\sum_n w_{t-1}^{(n)} \prod_{i\in\mathcal{I}_t} p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1}^{(n)}) \right), \tag{4}$$

where the interaction term can be moved out of the sum over all particles, as it does not depend on past information [21]. Furthermore, assuming that the motion and the speaking activity are independent, each single-object dynamical model is factorized as

$$p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1}) = p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})p(k_{i,t}|k_{i,t-1}),$$

where the continuous distribution $p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})$ is classically modeled as a second-order auto-regressive model [18], and $p(k_{i,t}|k_{i,t-1}) = \left[\begin{smallmatrix}\beta_{00} & \beta_{01}\\ \beta_{10} & \beta_{11}\end{smallmatrix}\right]$ is a 2×2 transition probability matrix (TPM) with parameters $\{\beta_{00},\beta_{01},\beta_{10},\beta_{11}\}$ ($\beta_{01} = 1-\beta_{00}$ and $\beta_{10} = 1-\beta_{11}$).

The interaction model we adopt takes into account visual information, and penalizes large visual overlaps between objects [21], which reduces the possibility of associating two configurations to one single object when people occlude each other momentarily. Let the spatial supports of $\mathbf{x}_{i,t}$ and $\mathbf{x}_{j,t}$, i.e., the application of the continuous transformation to the object template on the image plane, be denoted by $\mathcal{S}_{i,t}$ and $\mathcal{S}_{j,t}$, respectively. The overlap measures are the well-known precision (ν) and recall (ρ) measures from information retrieval [2]. Assuming that $\mathcal{S}_{i,t}$ is the reference, the measures are given by

$$\nu(\mathcal{S}_{i,t},\mathcal{S}_{j,t}) = \frac{|\mathcal{S}_{i,t}\cap\mathcal{S}_{j,t}|}{|\mathcal{S}_{j,t}|},\ \rho(\mathcal{S}_{i,t},\mathcal{S}_{j,t}) = \frac{|\mathcal{S}_{i,t}\cap\mathcal{S}_{j,t}|}{|\mathcal{S}_{i,t}|}. \tag{5}$$

As $\nu(\mathcal{S}_{i,t},\mathcal{S}_{j,t}) = \rho(\mathcal{S}_{j,t},\mathcal{S}_{i,t})$, the (symmetric) pairwise potentials in the MRF can be defined as

$$\phi(\mathbf{X}_{i,t},\mathbf{X}_{j,t}) \propto \exp\left(-\lambda_\phi(\nu(\mathcal{S}_{i,t},\mathcal{S}_{j,t}) + \rho(\mathcal{S}_{i,t},\mathcal{S}_{j,t})\right), \tag{6}$$

where λ_ϕ is a model parameter. Precision and recall take their maximum value (unity) when the spatial support of two objects perfectly match, which corresponds to the minimum value of $\phi(\cdot,\cdot)$, effectively penalizing such object overlap. In contrast, both precision and recall reach their minimum (zero) when the objects have no overlap, which corresponds to the maximum value of $\phi(\mathbf{X}_{i,t},\mathbf{X}_{j,t}) \propto 1$.

5 Audio-visual observation model

Observation models are derived from audio and video. Both shape and spatial structure of human heads are used as visual cues, so the three types of observations are defined as $\mathbf{Y}_t = (\mathbf{Y}_t^a, \mathbf{Y}_t^{sh}, \mathbf{Y}_t^{st})$, where the superindices stand for $a$udio, shape, and spatial structure, respectively. As other works [35], we further assume that observations are extracted for each object, and that the different observations are conditionally independent given the single-object states, producing the following factorized representation,

$$p(\mathbf{Y}_t|\mathbf{X}_t) = \prod_{i\in\mathcal{I}_t} p(\mathbf{Y}_{i,t}^a|\mathbf{X}_{i,t})p(\mathbf{Y}_{i,t}^{sh}|\mathbf{X}_{i,t})p(\mathbf{Y}_{i,t}^{st}|\mathbf{X}_{i,t}). \tag{7}$$

All terms in Eq. 7 are defined in the following subsections.

5.1 Audio observations

Audio observations are derived from the microphone array signals to produce 2-D location estimates in the corresponding image plane. The audio observation likelihood is then defined using such estimates. The procedure consists of three steps: audio source localization, speech/non speech classification, and mapping of speaker location estimates onto the image plane [14]. Each of these steps are described in the following subsections.

5.1.1 Source localization

A simple single source localization technique based on Time Delay of Arrival (TDOA) is used to generate candidate 3-D speaker locations. In particular, we use the Steered Response Power - Phase Transform (SRP-PHAT) measure [9], due to its low computational requirements and suitability for reverberant environments.

We define a vector of theoretical time-delays associated with a 3-D location $Z \in \mathbb{R}^3$ as $\boldsymbol{\tau}^Z \triangleq (\tau^{1,Z}, \ldots, \tau^{\psi,Z}, \ldots \tau^{N_\psi,Z})$, where N_ψ is the number of pairs and $\tau^{\psi,Z}$ is the delay (in samples) between the microphones in pair ψ, defined as $\tau^{\psi,Z} = \frac{f_s}{c_s}\left(||Z - M_1^\psi|| - ||Z - M_2^\psi||\right)$, where $M_1^\psi, M_2^\psi \in \mathbb{R}^3$ are the locations of the microphones in pair ψ, $||\cdot||$ is the Euclidean norm, f_s the sampling frequency, and c_s the speed of sound. Note that for a given time-delay τ_0 and pair ψ, there exists a hyperboloid of locations Z satisfying $\tau^{\psi,Z} = \tau_0$.

From two signals $\mathbf{s}_1^\psi(t)$ and $\mathbf{s}_2^\psi(t)$ of a given microphone pair ψ, the frequency-domain GCC-PHAT [22] is defined as

$$G_{PHAT}^{\psi}(f) \triangleq \frac{\mathbf{S}_1^\psi(f) \cdot \left[\mathbf{S}_2^\psi(f)\right]^*}{\left|\mathbf{S}_1^\psi(f) \cdot \left[\mathbf{S}_2^\psi(f)\right]^*\right|}, \tag{8}$$

where $\mathbf{S}_1^\psi(f)$ and $\mathbf{S}_2^\psi(f)$ are Fourier transforms of the two signals and $[\cdot]^*$ denotes the complex conjugate. Typically the two Fourier transforms are estimated on Hamming-windowed segments of 20-30 ms. By performing an Inverse Fourier Transform, and summing the time-domain GCC-PHAT $R_{PHAT}^\psi(\tau)$ across pairs, we obtain the SRP-PHAT measure,

$$P_{SRP-PHAT}(Z) \triangleq \sum_{\psi=1}^{N_\psi} R_{PHAT}^\psi(\tau^{\psi,Z}). \tag{9}$$

From this, the source location is estimated as

$$\hat{Z} = \arg\max_{Z \in \mathbb{R}^3} [P_{SRP-PHAT}(Z)]. \tag{10}$$

From geometrical considerations, at least three microphone pairs ($N_\psi \geq 3$) are required to obtain a unique peak.

The maximization is implemented through exhaustive search over a fixed grid of points, $H \subset \mathbb{R}^3$ such that $\forall Z \in \mathbb{R}^3$, $\exists Z_H \in H$ such that $\Gamma(Z, Z_H) \leq \gamma_0$, where $\Gamma(Z_1, Z_2)$ is the distance in time-delay space,

$$\Gamma(Z_1, Z_2) \triangleq \sqrt{\frac{1}{N_\psi}\sum_{\psi=1}^{N_\psi}\left(\tau^{\psi,Z_1} - \tau^{\psi,Z_2}\right)^2}, \tag{11}$$

and γ_0 is the desired precision in samples. Since we typically upsample $R_{PHAT}^\psi(\tau)$ with a factor α_{up}, the desired precision is set accordingly to $\gamma_0 = 1/\alpha_{up}$. The grid H is built by picking points heuristically on a few concentric spheres centered on the microphone array. The spheres' radii were also determined by γ_0. Conceptually this approach relates to [16]. Note that in practice, the estimated range is imprecise, and only azimuth and elevation are significant. Finally, for each time frame, our implementation approximates Eq. 10 by

$$\hat{Z} \approx \arg\max_{Z \in H} [P_{SRP-PHAT}(Z)]. \tag{12}$$

5.1.2 Speech/non-speech classification

In the second step, a speech/non-speech classification algorithm based on short-term clustering of the localization results is used to filter out noisy speaker location estimates. Conventional single-channel speech/non-speech segmentation approaches are based upon energy, SNR estimation [9] or more complex estimators such as zero-crossing rate [29]. While relatively robust, techniques based on energy thresholding often miss low-energy beginnings of words and short speaker turns. Furthermore, they can provide a significant amount of erroneous audio estimates to the observation model. Unlike traditional approaches, we pose the problem of speech/non-speech classification in the framework of localization, making decisions purely based on the location information. We first run single source localization on each time frame, then classify each frame as speech or non-speech, relying on short-term clustering of location estimates. Our motivation for short-term clustering is that noisy location estimates feature high variations over time, while location estimates are consistent during speech periods. The algorithm has three steps: (1) build short-term clusters of frames whose location estimates are close to each other; (2) retain only significant clusters by applying a duration constraint; and (3) label those frames belonging to any significant cluster as speech, others as non-speech.

In step 1, two frames t_1 and t_2 belong to the same cluster if $d(\hat{Z}_{t_1}, \hat{Z}_{t_2}) < d_0$ and $|t_2 - t_1| \leq T_0$, where d_0 and T_0 are thresholds in space and time respectively. $d(\hat{Z}_{t_1}, \hat{Z}_{t_2})$ is a distance defined according to the setup. With a single, planar microphone array it is reasonable to use the difference in azimuth between $\hat{Z}_{t_1}$ and $\hat{Z}_{t_2}$. T_0 should be close to the length of a phoneme.

For step 2, we find the longest segment of contiguous frames within each cluster. If that segment lasts more than a threshold T_S, the cluster is kept as "significant", otherwise it is dropped. Simpler criteria such as minimum cluster duration or the minimum number of frames within the cluster did not prove adequate. Additionally, to eliminate diffuse or minor far-field noise sources (e.g. PC), we also discard clusters whose mean square SRP-PHAT value is below a threshold $T_{SRP-PHAT}$.

In step 3, frames belonging to any significant cluster are labeled as speech, others as non-speech. In the usual case where the audio frame rate is higher than the video frame rate, we downsample the audio by grouping audio 3-D estimates between consecutive video frames. For example, with audio frame rate 62.5 fps and video frame rate 25 fps, there can be zero (non-speech), one, two or three (speech) audio 3-D estimates $\{\hat{Z}_t\}$ per video frame.

5.1.3 AV calibration

Mapping the 3-D audio location estimates onto the camera image planes requires a form of sensor calibration. As discussed in Section 2, most previous works have either assumed simplified configurations [43, 3] or resorted to rigorous camera calibration procedures [48, 5, 6]. Note that, in general, cameras and microphones might not necessarily remain in the same location for long term, so practical calibration procedures are needed to cope with sensor changes. Unlike such previous approaches, we use a nearest-neighbor approach to project 3-D audio estimates on the corresponding 2-D image planes, exploting the fact that, although audio localization estimates are usually noisy, and visual calibration is affected by geometric distortion, their joint occurrence tends to be more consistent [14]. The procedure requires an off-line rough AV calibration procedure between the sensors, without requiring precise geometric calibration of audio and video. The procedure uses training data collected by having a person talking while performing activities in the meeting room in typical locations (walking, sitting, moving while seated, standing at the whiteboard and projector screen areas). The correspondences between 3-D and 2-D+camera-index points are obtained from the audio estimates, as described earlier in this subsection, and from the output of a single-person PF visual tracker, respectively. For non-overlapping fields-of-view (FOVs), the set of correspondences obtained for the training set defines a mapping between discrete sets $\Gamma : \mathbb{R}^3 \rightarrow \mathbb{R}^2 \times \{0, .., N_{CAM-1}\}$, where N_{CAM} is the number of cameras, such that 3-D positions are mapped into vectors containing image position (u_t, v_t) and camera index cam_t, $\Gamma(Z_t) = (u_t, v_t, cam_t)$. Note that when several image views are concatenated into a single image (as we do in Section 8), the camera index simply results in an fixed 2-D translation term added to (u_t, v_t). Finally, the mapping for unseen data is computed via nearest neighbor search.

5.1.4 Observation model

The audio observation likelihood is finally defined on the image domain, relating the Euclidean distance between the 2-D audio location estimates and the candidate particles. Let $x^a_{i,t} = (u^a_{i,t}, v^a_{i,t})$ denote the audio estimate closest to the translation components $x_{i,t} = (u_{i,t}, v_{i,t})$ of the i-th object. We define a distribution for each value of the speaking status variable,

$$p(\mathbf{Y}^a_{i,t}|\mathbf{x}_{i,t}, k_{i,t}=1) \propto \begin{cases} K^a_1, & ||x_{i,t} - x^a_{i,t}|| \leq \mathcal{R}^a, \\ K^a_2, & \text{otherwise}, \end{cases} \tag{13}$$

$$p(\mathbf{Y}^a_{i,t}|\mathbf{x}_{i,t}, k_{i,t}=0) \propto \begin{cases} K^a_1, & ||x_{i,t} - x^a_{i,t}|| \geq \mathcal{R}^a, \\ K^a_2, & \text{otherwise}, \end{cases} \tag{14}$$

where $\mathcal{R}^a$ defines a radius around the translation components of $\mathbf{X}_{i,t}$, and $K^a_1 > K^a_2$ are constant terms introduced to refect the desired situation: the likelihood of a person actively speaking must be large when there exists a nearby audio estimate, and small if such condition does not hold (e.g., $\frac{K^a_1}{K^a_2} = 10$). In case no audio location estimates exist, $u^a_{i,t}$ and $v^a_{i,t}$ are set to an arbitrarily large number.

5.2 Shape observations

Assuming that shapes are embedded in clutter, edge-based observations are computed, based on a classic model, along L normal lines to a hypothesized contour [18]. This results in a vector of candidate positions for each line l, $\{z^l_{i,c}, l \in \{1, ..., L\}, c \in \{1, ..., C_l\}\}$, relative to $z^l_{i,0}$, the point lying on the contour. With some typical assumptions, the shape-based likelihood for each object is given by

$$p(\mathbf{Y}^{sh}_{i,t}|\mathbf{X}_{i,t}) \propto \prod_{l=1}^{L} \max\left(K^{sh}, \exp(-\frac{\|z^l_{i,\hat{c}} - z^l_{i,0}\|^2}{2(\sigma^{sh})^2})\right), \tag{15}$$

where $z^l_{i,\hat{c}}$ is the nearest edge detected on the l^{th} line, σ^{sh} is a standard deviation parameter, and K^{sh} is a constant that limits the influence of cases when no edges are detected.

5.3 Spatial structure observations

We propose an observation model of spatial structure of human heads, based on a parametric representation of the overlap between skin-blobs and head configurations. The model is based on the fact that the presence of skin pixels in a typical head blob is usually limited to specific regions inside and outside a head elliptical configuration. Skin pixels are mainly distributed in the central and lower regions of a head, but also outside the head ellipse, e.g. on the neck region. Additionally, the blobs corresponding to a head often appear connected to other skin-color blobs due to body postures of people in conversations (e.g. resting the head on a hand).

Skin-color blobs are extracted at each frame according to a standard procedure. A 20-component Gaussian Mixture Model (GMM) of skin color in RGB space was estimated from a training set of people participating in real meetings in the room, including Caucasian, Indian, and Latin-American individuals, and collected over several days [30]. Skin pixels were classified based on thresholding of the skin likelihood, followed by morphological postprocessing to extract blobs (Fig. 1).

Given a set of skin-color blobs and a single-object configuration $\mathbf{X}_{i,t}$, and assuming that the candidate configuration is the reference, the recall between the spatial support $\mathcal{S}_{i,t}$ and each blob is computed. Let $\mathcal{S}^B_{i,t}$ denote the spatial support of the blob with the largest recall. A head template is further decomposed into three non-overlapping parts with spatial support $\mathcal{S}^l_{i,t}$, $l \in \{1,2,3\}$, $\mathcal{S}_{i,t} = \cup_l \mathcal{S}^l_{i,t}$ (Fig. 1). With this representation, precision and recall are computed for each of the head parts ($\nu(\mathcal{S}^l_{i,t}, \mathcal{S}^B_{i,t})$ and $\rho(\mathcal{S}^l_{i,t}, \mathcal{S}^B_{i,t})$), and for the whole head, ($\nu(\mathcal{S}_{i,t}, \mathcal{S}^B_{i,t})$ and $\rho(\mathcal{S}_{i,t}, \mathcal{S}^B_{i,t})$). Although the proposed features are obviously not as discriminant as the ones used in dedicated face processing algorithms [44], they are reasonable in realistic conditions, including out of plane rotation, and partial occlusion. The features define an eight-component observation space $\mathbf{Y}^{st}_{i,t}$, that is modeled by a mixture model composed of GMM with diagonal covariance matrices, and a uniform distribution $U(\cdot)$ used to limit the effect of very low likelihood values,

$$p(\mathbf{Y}^{st}_{i,t}|\mathbf{X}_{i,t}) \propto \omega_0 U(\mathbf{Y}^{st}_{i,t}) + \sum_{l=1}^{N_{st}} \omega_l \mathcal{N}(\mathbf{Y}^{st}_{i,t}, \mu_l, \Sigma_l), \tag{16}$$

where $\{\omega_0, \omega_l, \mu_l, \Sigma_l\}$ are model parameters. When no blobs are detected, the likelihood is set to a constant value.

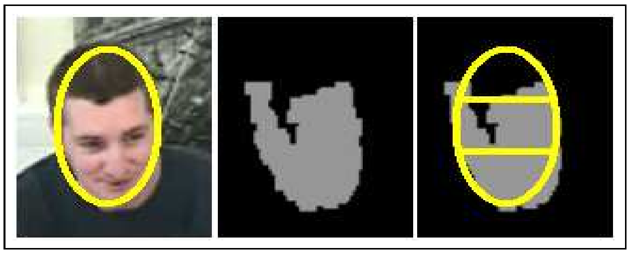

Figure 1: Spatial structure observations. Given a configuration (in yellow), and a skin blob (in gray), part-based precision/recall features are computed from the spatial support of configuration parts and blob (see text).

6 MCMC sampling

Inference with a traditional particle filter (based on IS) on the high-dimensional space defined by several objects being tracked is computationally infeasible, given that the number of particles required for a given performance grows roughly exponentially with the dimensionality of the space [28, 11]. In order to efficiently place samples as close as possible to regions of high likelihood, we build on recent work and propose to approximate Eq. 4 with MCMC techniques [28], using a Metropolis-Hastings (MH) sampler at each time-step [21]. MCMC methods produce a sequence of samples from a Markov chain whose stationary distribution corresponds to the target distribution (the filtering distribution in the tracking case), after running the sampler long enough, and discarding the initial part of the run, called burn-in period [28]. The MH algorithm consists of two iterative steps. First, given a current configuration $\mathbf{X}$, a new sample $\mathbf{X}^*$ is drawn from a proposal distribution $q(\mathbf{X}^*|\mathbf{X})$. Then, the proposed sample is accepted as the new configuration in the Markov chain with probability (also called acceptance ratio),

$$\alpha = \min\left(1, \frac{p(\mathbf{X}^*)q(\mathbf{X}|\mathbf{X}^*)}{p(\mathbf{X})q(\mathbf{X}^*|\mathbf{X})}\right), \tag{17}$$

where $p(\mathbf{X})$ denotes the target distribution (in our case $p(\mathbf{X}) = p(\mathbf{X}_t|\mathbf{Y}_{1:t})$). If the move is not accepted, the chain remains in the same configuration. The sample set obtained by the MH sampler is a fair sample from the true filtering distribution, and so all particle weights are equal to $\frac{1}{N}$ [28].

In the tracking case, we run a MH sampler at each time-step. However, for computational tractability, a proposal distribution that simplifies the evaluation of the acceptance ratio is needed, as Eq. 17 involves the evaluation of Eq. 4 (a sum over all particles). We define a mixture model over all objects, where one object is chosen at each step in the chain to attempt a move,

$$q(\mathbf{X}_t^*|\mathbf{X}_t) = \sum_i q(i)q(\mathbf{X}_t^*|\mathbf{X}_t, i), \tag{18}$$

where $q(i)$ is the prior over object indices, and $q(\mathbf{X}_t^*|\mathbf{X}_t, i)$ are the mixture components. To construct a candidate configuration $\mathbf{X}_t^*$ from the current configuration $\mathbf{X}_t$, an object index i^* is first chosen with probability $q(i = i^*)$. A move will be attempted on i^*, while the rest of the multi-object configuration is left unchanged. The mixture components are defined so that

$$q(\mathbf{X}_t^*|\mathbf{X}_t, i) = \begin{cases} \frac{1}{N}\sum_n p(\mathbf{X}_t^*|\mathbf{X}_{t-1}^{(n)}) & i = i^*, \\ \frac{1}{N}\sum_n p(\mathbf{X}_t^*|\mathbf{X}_{t-1}^{(n)}) & i \neq i^*, \mathbf{X}_t^* = \mathbf{X}_t, \\ 0 & i \neq i^*, \mathbf{X}_t^* \neq \mathbf{X}_t, \end{cases} \tag{19}$$

which implies that given i^*, the new configuration for object i^* is sampled from $p(\mathbf{X}_{i^*,t}^*|\mathbf{X}_{i^*,t-1}^{(n^*)})$, using a randomly chosen particle n^* from the previous time, while keeping all the other object configurations fixed. Using the Dirac delta function, the specific expression fulfilling Eq. 19 is

$$q(\mathbf{X}_t^*|\mathbf{X}_t, i) = \frac{1}{N}\sum_n p(\mathbf{X}_{i,t}^*|\mathbf{X}_{i,t-1}^{(n)}) \cdot \prod_{l \in \mathcal{I}_t - \{i\}} p(\mathbf{X}_{l,t}|\mathbf{X}_{l,t-1}^{(n)})\delta(\mathbf{X}_{l,t}^* - \mathbf{X}_{l,t}).$$

This distribution fulfills the desired property of cancelling all the factors that involve summations over particles in the acceptance ratio. It is not difficult to show that the acceptance probability is simplified to

$$\alpha = \min\left(1, \frac{p(\mathbf{Y}_t|\mathbf{X}_t^*)\prod_{(i^*,j)\in\mathcal{C}_{i^*}}\phi(\mathbf{X}_{i^*,t}^*,\mathbf{X}_{j,t}^*)}{p(\mathbf{Y}_t|\mathbf{X}_t)\prod_{(i^*,j)\in\mathcal{C}_{i^*}}\phi(\mathbf{X}_{i^*,t},\mathbf{X}_{j,t})}\right), \tag{20}$$

where $\mathcal{C}_{i^*}$ denotes the set of pairwise cliques that involve object i^*. For the factorized form for the multi-object likelihood (Eq. 7), the expression can be further simplified to

$$\alpha = \min\left(1, \frac{p(\mathbf{Y}_t|\mathbf{X}_{i^*,t}^*)\prod_{(i^*,j)\in\mathcal{C}_{i^*}}\phi(\mathbf{X}_{i^*,t}^*,\mathbf{X}_{j,t}^*)}{p(\mathbf{Y}_t|\mathbf{X}_{i^*,t})\prod_{(i^*,j)\in\mathcal{C}_{i^*}}\phi(\mathbf{X}_{i^*,t},\mathbf{X}_{j,t})}\right), \tag{21}$$

which only involves the evaluation of single-object AV likelihood terms. The MH sampler improves the predictions of multi-object configurations by accepting, at each step, single-object candidates closer to a region of high likelihood, without discarding good candidates already accepted for other objects. Note that other formulations combining MCMC iterations with PF techniques exist in the statistics literature [4, 27, 15, 12], but they all differ from the algorithm presented here.

Finally, the mean estimate is approximated by the marginal mean estimates for each object, $\hat{\mathbf{X}}_t = (\hat{\mathbf{X}}_{i,t}), i \in \mathcal{I}_t$. Each $\hat{\mathbf{X}}_{i,t}$ is computed as usual in mixed-state models, first computing the maximum a posteriori (MAP) estimate for the discrete variable $k_{i,t}$, and then the weighted mean of the continuous component $\mathbf{x}_{i,t}$ given the MAP discrete estimate [18],

$$\hat{k}_{i,t} = \arg\max_m \sum_{n\in\mathcal{J}_{i,m}} w_t^{(n)};\ \hat{\mathbf{x}}_{i,t} = \frac{\sum_{n\in\mathcal{J}_{i,\hat{k}_{i,t}}} w_t^{(n)}\mathbf{x}_{i,t}^{(n)}}{\sum_{n\in\mathcal{J}_{i,\hat{k}_{i,t}}} w_t^{(n)}}, \tag{22}$$

where $\mathcal{J}_{i,m} = \{n|k_{i,t}^{(n)} = m\}$. The full MCMC-PF algorithm is summarized in Fig. 2.

`Generate` N samples $\{\mathbf{X}_t^{(n)}, w_t^{(n)}\}$ from $\{\mathbf{X}_{t-1}^{(n)}, w_{t-1}^{(n)}\}$.

- `Initialize` the MH sampler, by sampling $\mathbf{X}$ from the purely predictive distribution $\sum_n w_{t-1}^{(n)} p(\mathbf{X}_t|\mathbf{X}_{t-1}^{(n)})$. This implies randomly choosing a particle n^* from $\{\mathbf{X}_{t-1}^{(n)}, w_{t-1}^{(n)}\}$, and then sampling from $p(\mathbf{X}_t|\mathbf{X}_{t-1}^{(n^*)})$.
- `MH sampling`. Draw $B + N$ samples, where B and N denote the number of particles in the burn-in and fair sample sets, respectively. For each sample,
  - `Sample` $\mathbf{X}^*$ from $q(\mathbf{X}^*|\mathbf{X})$ (Eq. 18).
  - `Compute acceptance ratio` α (Eq. 21).
  - `Accept` $\mathbf{X}^*$ ($\mathbf{X} \leftarrow \mathbf{X}^*$) with probability α.
  - `Add` $\mathbf{X}$ to the set $\{\mathbf{X}_t^{(n)}, w_t^{(n)}\}$, with $w_t^{(n)} = 1/N$.
- `Compute mean estimate` $\hat{\mathbf{X}}_t$.

Figure 2: MCMC-PF algorithm.

7 Varying number of people

Although the MCMC-PF could formally integrate birth-death processes as part of the filtering recursion (e.g., via reversible-jump MCMC [28]), this would require a multi-person observation model that allowed for the comparison between configurations containing varying number of people [19, 5]. The factorized observation model in Eq. 7 is not suitable for such a case. For this reason, we opted for a simple process in which, at each time frame, the set of people of the scene $\mathcal{I}_t$ is first established, and then the MCMC-PF is applied on the detected $\mathcal{I}_t$. It has been argued that, unless a clear ambiguity in the number of scene objects exists, the detection-then-tracking mechanism can be more efficient than the case in which particles with varying number of objects coexist in the same time-step, as particles corresponding to the "wrong" number of objects are effectively wasted [35].

New objects are handled as follows. All skin-color blobs inside a set of birth-likely scene regions, and not overlapping with existing objects, are probed as candidates. Given a standard ellipse template, and a new object ID i^* (chosen as the next available object ID in a list), a number of single-object samples $\{\mathbf{X}_{i^*,t}^{(r)}\}$ is constructed by drawing samples $\{\mathbf{x}_{i^*,t}^{(r)}\}$ from a Gaussian distribution (with mean translation equal to the blob centroid, mean scaling set to unity, and diagonal covariance matrix set to explore a relatively short space around the mean), while setting $\{k_{i^*,t}^{(r)}\}$ to zero. The set of samples is ranked based on their visual likelihood, $p(\mathbf{Y}_{i^*,t}^{sh}|\mathbf{X}_{i^*,t})p(\mathbf{Y}_{i^*,t}^{st}|\mathbf{X}_{i^*,t})$, and the presence of a new object is decided by thresholding the likelihood of the best configuration. The best configuration is used in the prior in the MCMC-PF for the new object. Needless to say, more robust people detectors could be integrated in our approach [44]. Object disappearance is declared whenever a configuration leaves the image limits, or when a configuration has too low visual likelihood. Finally, continuing people are handled as in the case of fixed number of objects.

8 Experiments and Results

8.1 Data collection

Data are recorded in a 8.2m×3.6m×2.4m meeting room containing a 4.8m×1.2m rectangular meeting table, and equipped with fully synchronized video and audio capture devices. The video equipment includes three identical uncalibrated CCTV cameras [33]. Two cameras on opposite walls record frontal views of participants, including the table and workspace area, and have non-overlapping fields-of-view (FOVs). A third wide-view camera looks over the top of the participants towards the white-board and projector screen. The audio equipment consists of an eight-element circular equi-spaced microphone array centered on the table, with diameter 20cm, and composed of high quality miniature electret microphones. Video was captured at 25 fps (288×360 pixels), while audio was recorded at 16kHz, with features estimated at 62.5 fps. Training data to estimate the GMM parameters for skin-color models and spatial structure features, and for the AV calibration procedure were additionally recorded in the meeting room. In Section 8.4, we present results on two two-camera sequences, (*meeting1* and *meeting2*, 1715 and 1200 video frames, respectively), and one three-camera sequence (*meeting3*, 1200 video frames). The sequences are composed by concatenating the different views into one larger merged image. The first two sequences have non-overlapping FOVs, while in the third there is some overlap.

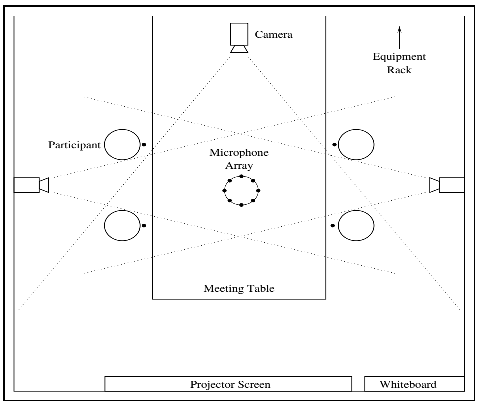

Figure 3: Meeting recording configuration.

8.2 Parameter setting

For the audio source localization algorithm, we used $\alpha_{up} = 20$. The thresholds in the speech/non-speech classification algorithm were set to $d_0 = 5^o$, T_0 = 200 ms, T_S = 100 ms, and $T_{SRP-PHAT} = 0.03$. The first two thresholds were intuitively chosen, based on experiments conducted on a separate test data set [14], and on the typical phoneme length, respectively. The third threshold was chosen on a single, separate test case [14]. Regarding the video observations, the GMM parameters for skin-color pixels and spatial structure features were estimated by standard Expectation Maximization (EM). Model selection was automatically done using the minimum description length principle. All other parameters were set by hand to sensible values, and kept fixed for all experiments. Regarding the dynamical model for the single-object continuous dynamics $p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})$, we use an augmented continuous state, $\tilde{\mathbf{x}}_{i,t} = (\mathbf{x}_{i,t}, \mathbf{x}_{i,t-1})$, and express the dynamics as $\tilde{\mathbf{x}}_{i,t} = A\tilde{\mathbf{x}}_{i,t-1} + B(\omega_t, 0)^T$, with $A = \left[\begin{smallmatrix} 2 & -1 \\ 1 & 0 \end{smallmatrix}\right]$, $B = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, and ω_t is a white noise process with standard deviations for translation and scaling equal to 4 and 10^{-4}, respectively. The TPM parameters for speaking activity were set to $\beta_{00} = \beta_{11} = 0.8, \beta_{01} = \beta_{10} = 0.2$. For the interaction model, $\lambda_\phi = 3$. In the audio observation model, $R^a = 50$ pixels, and $K_1^a = 1 = 10K_2^a$. For the shape-based observations, the number of measurement lines $L = 16$, each with length $a = 20$ pixels, $\sigma^{sh} = 5$, and $K^{sh} = \exp(-\frac{(a/2)^2}{2(\sigma^{sh})^2}) = e^{-2}$. A scaling procedure was applied as the various likelihood terms have a different dynamic range. Finally, we assume a uniform prior for the proposal $q(i)$ in the MH sampler.

8.3 Performance evaluation measures

We evaluate both the tracking quality and the ability to infer speaking status. For the first criterion, a semi-automatic head bounding-box ground truth (GT) is generated for each person at each video frame, using a color-based single-person tracker [35]. We then compute precision and recall between the GT and our tracker estimates (represented by bounding boxes) for each person at each frame, and define four person-based measures:

1. *Track state* (TS). A frame-level binary variable, which is unity if both precision and recall are greater than a threshold T_{TS}, and zero otherwise. In the experiments, $T_{TS} = 0$.

2. *Track F-measure* (F_T). The precision/recall combination ($F_T = \frac{2\nu\rho}{\nu+\rho}$) is computed for those frames with $TS = 1$.

3. *Success rate* (SR). A sequence-level variable, defined as unity if $TS = 1$ for the entire sequence, and zero otherwise.

4. *Tracking rate* (TR). A sequence-level variable, defined as the ratio between the number of frames where $TS = 1$ and the number of frames in the sequence.

Jointly, F_T, SR, and TR provide an indication of the quality and stability of the tracker, including eventual recovery from failures. Note that SR is a much stricter measure than TR. All results are computed over multiple runs of the particle filter, to account for its stochastic nature. SR and TR are then reported as averages. Finally, an overall average, over the number of people, is also reported.

Regarding speaking activity, a binary GT of speaker turns was manually generated for each person. Precision and recall are then computed between the GT and the tracker estimate for each person at each video frame, defining:

5. *Speaking status F-measure* (F_S). For each person, a measure computed as in F_T.

F_S is also reported as averages over multiple runs and people, as with the other measures.

8.4 Results and discussion

We first evaluated the specific abilities of our framework to estimate location and speaking activity, conducting experiments under two controlled conditions: (1) the number of tracked people was known and kept fixed for the duration of the sequence, and (2) the tracker was hand-initialized in the first frame. The source localization algorithm was originally evaluated in [14]. Details of such evaluation are not included here for space reasons, but in summary, the objective evaluation of the short-term clustering algorithm confirmed its superiority over energy-based methods for the speech detection task, as further developed in [24]. Results of the algorithm handling varying numbers of objects are discussed at the end of this section. The mean configuration is displayed, at each time-step for each person, as an ellipse of distinct color. Inferred speaking activity is shown as a double ellipse with contrasting tones. The number of particles in all cases was $B + N = 500$, with 30% of the particles being discarded in the burn-in period of the MH sampler. All the results were obtained using 20 runs of the MCMC-PF. People are given an object identifier (O1,O2,...) with respect to the position they first occupy in the video, from left to right. The results are best appreciated by watching the videos acccompanying this submission at $\texttt{www.idiap.ch/} \sim \texttt{gatica/publications/pub05/av} - \texttt{tracking} - \texttt{multiperson.html}$.

8.4.1 Meeting1

The results on the *meeting1* sequence are shown in Fig. 4 and Table 1. In this sequence, recorded with no visual background clutter, four seated speakers are engaged in a conversation and talk at a relaxed pace, taking turns with little overlap, which occurs for instance when people laugh. The last row in Table 1 (S_{GT}) indicates the proportion of time during which each person spoke in the sequence, as labeled in the speaking activity GT.

Regarding visual tracking, the four objects were tracked with good quality and stably throughout the sequence for all runs (see SR, TR, and F_T rows in Table 1, and video `meeting1_mcmc_500.avi` in the website). The algorithm can handle partial visual self-occlusion (e.g. person O3 touches his chin and rests his head on his right hand in Fig. 4(a), (e), and (g)), and variations of head pose (from frontal to side views), which confirms the advantages of combining visual cues.

With respect to speaking activity, our source localization method, combined with the AV calibration procedure, has shown to be able to estimate location reasonably well, and detect speaker turns with good accuracy and low latency, when people talk at the meeting table [14]. The audio activity inferred by the MCMC-PF preserves these properties for those segments where only one speaker takes the turn, while smoothing out very short speaker turns with the dynamical model (see F_S row in Table 1). Although we use a single-source localization algorithm, the MCMC-PF can sometimes infer simultaneous speaking activity for multiple participants (see Fig. 4(d)). In general, however, a "dominant speaker" effect is observed in overlapping speech segments.

To study the efficiency of the MCMC-PF, we compare it with a traditional joint multi-object PF, which uses IS instead of MCMC, while all other aspects and parameters of the filter remain fixed. Results are also computed using 20 runs, and are shown in Table 1, Fig. 4, and video `meeting1_pf_500.avi`. Clearly, our approach outperforms the traditional PF in both ability to track and estimation of the speaking status. With the classic PF, a loss of track occurred for all of the objects (see SR) at some point in the sequence (especially poor for O2, who was tracked successfully only in 60% of the runs). The tracker also has high visual jitter. Furthermore, the inference of speaking activity is degraded considerably. The performance loss is explained by the way in which the multi-object state-space is explored in a traditional joint PF. Many particles are effectively wasted: a bad candidate configuration for one object will produce a low multi-object likelihood value, even though the candidates for all other objects are good. Additionally, relatively bad geometric configurations combined with correct speaking activity values might be common, given the joint audio-visual observation model. A combination of these factors with e.g. motion changes might eventually drift one or more objects away from the correct configuration and result in tracking loss (Fig. 4(j-l)). A standard proportion test indicates that the difference in performance between MCMC-PF and PF for O_{avg} is statistically significant with confidence greater than 99.99% for both SR and TR. As a final remark, the number of particles that is required with the traditional PF to perform as well as the MCMC-PF is prohibitively high.

8.4.2 Meeting2

The results on the *meeting2* sequence are shown in Fig. 5, Table 2, and in the video `meeting2_mcmc_500.avi`. This sequence depicts four seated speakers in a more animated conversation (see S_{GT} row in Table 2), with many turns and cases of overlapped speech. There are also two sources of visual clutter: the textured background, and a fifth walking person (not tracked) who enters and leaves the scene creating visual distraction by approaching the speakers. Our algorithm performs quite satisfactorily with respect to quality of tracking and speaker activity detection. Although the tracker gets momentarily distracted by the walking person or the background, it recovers in almost all cases, as shown by the SR, TR, and F_T rows in Table 2. The com-

Figure 4: Multi-speaker tracking results, *meeting1*. Both location and speaking status (double ellipse if a person speaks) are inferred for each participant. (a-h): MCMC-PF, images correspond to frames 10, 370, 420, 526, 760, 1120, 1420, and 1690, respectively. (i-l): Traditional PF: images correspond to frames 760, 905, 908, and 910, respectively. Tracking is lost for O1 for the rest of the sequence.

method	*measure*	O1	O2	O3	O4	O_{avg}
MCMC-PF	SR	1.00	1.00	1.00	1.00	1.00
	TR	1.00	1.00	1.00	1.00	1.00
	F_T	0.89	0.87	0.85	0.92	0.88
	F_S	0.71	0.75	0.77	0.75	0.75
PF	SR	0.80	0.60	0.85	0.90	0.79
	TR	0.95	0.79	0.94	0.98	0.92
	F_T	0.87	0.86	0.83	0.87	0.86
	F_S	0.60	0.62	0.55	0.61	0.59
	S_{GT}	0.19	0.19	0.10	0.16	0.16

Table 1: Tracking results for *meeting1*, for our approach and a traditional multi-object PF. Results are shown for individual people, and averaged over all people.

bination of visual cues renders the tracker more robust: On one hand, the spatial structure observations help in cases of uncertainty with respect to edge information (e.g. textured background), On the other hand, the shape observations refine the spatial structure model, which consistently drives the tracker to skin-blob areas, but sometimes without much accuracy. A limitation of the likelihood model can be seen for O3, for whom we can observe a combined effect of edge-related clutter, and head spatial structure (i.e., less hair) that might not have been represented accurately in the training data (Fig. 5(e) and (h)), also evident when playing the video). Regarding speaking activity, our approach can correctly infer some cases of simultaneous speech (Fig. 5 (d) and (g)). Finally, our approach is again significantly more effective than a traditional joint multi-person PF, which in this case shows an even more severe performance degradation (see Fig. 5(i-n), Table 2 and video `meeting2_pf_500.avi`). For the same reasons discussed for *meeting1*, and challenged by the distractions introduced by the walking person, the traditional PF is unable to track all of the objects consistently, and the quality of the estimation of speaking activity is considerably degraded. SR can be as low as 35-40%, F_S as low as 38%, and with the exception of F_T, all other figures are considerably lower than the ones obtained for *meeting1*. The standard proportion test shows a statistically significant difference in performance between our approach and a traditional PF for O_{avg} with confidence greater than 99.99% , for both SR and TR. Overall, the results on this sequence suggest that the MCMC-PF is more robust to realistic conditions than the traditional approach.

method	*measure*	O1	O2	O3	O4	O_{avg}
MCMC-PF	SR	0.95	0.95	1.00	1.00	0.98
	TR	0.99	0.97	1.00	1.00	0.99
	F_T	0.88	0.88	0.87	0.86	0.87
	F_S	0.72	0.72	0.71	0.79	0.74
PF	SR	0.60	0.40	0.35	0.75	0.52
	TR	0.68	0.45	0.50	0.79	0.61
	F_T	0.86	0.85	0.81	0.85	0.84
	F_S	0.55	0.38	0.42	0.65	0.50
	S_{GT}	0.26	0.24	0.19	0.35	0.26

Table 2: Tracking results for *meeting2*, for our approach and a traditional multi-object PF.

Fig. 6 presents the results regarding speaking activity detection for each person as a function of video frame number. In the first place, we display the results obtained with audio-source localization mapped onto the image plane, assuming that the location of the meeting participants is known in the ground-truth, i.e., dividing the image into four regions, assigning regions to people, and associating speaking activity estimates to individuals based on the region the estimates fall in (A). We also present the results for our approach (AV), where the results correspond to the median of the inferred speaking activity estimated over 20 runs of the MCMC-PF. We finally display the manually labeled GT. Two cases are displayed, for different values of the parameters of the symmetric TPM: $\beta_{00} = 0.8, \beta_{01} = 0.2$ (Fig. 6(a-d)), and $\beta_{00} = 0.95, \beta_{01} = 0.05$ (Fig. 6(e-h)). We first observe that the source localization algorithm detects speak/non-speak changes with low delay, and classifies most of the correct location estimates as speech frames [14]. Furthermore, given the lack of a temporal smoothness constraint, long speaker turns are often broken into smaller turns. We also observe that short spurious turns due to noise are sometimes detected and incorrectly assigned to speakers. In turn,

Figure 5: Multi-speaker tracking results, *meeting2*. (a-h): MCMC-PF, images correspond to frames 50, 200, 360, 546, 575, 630, 860, and 909, respectively. (i-n): Traditional PF: images correspond to frames 150, 170, 190, 195, 200, and 210, respectively. Tracking is lost for O2 and O4 for the rest of the sequence.

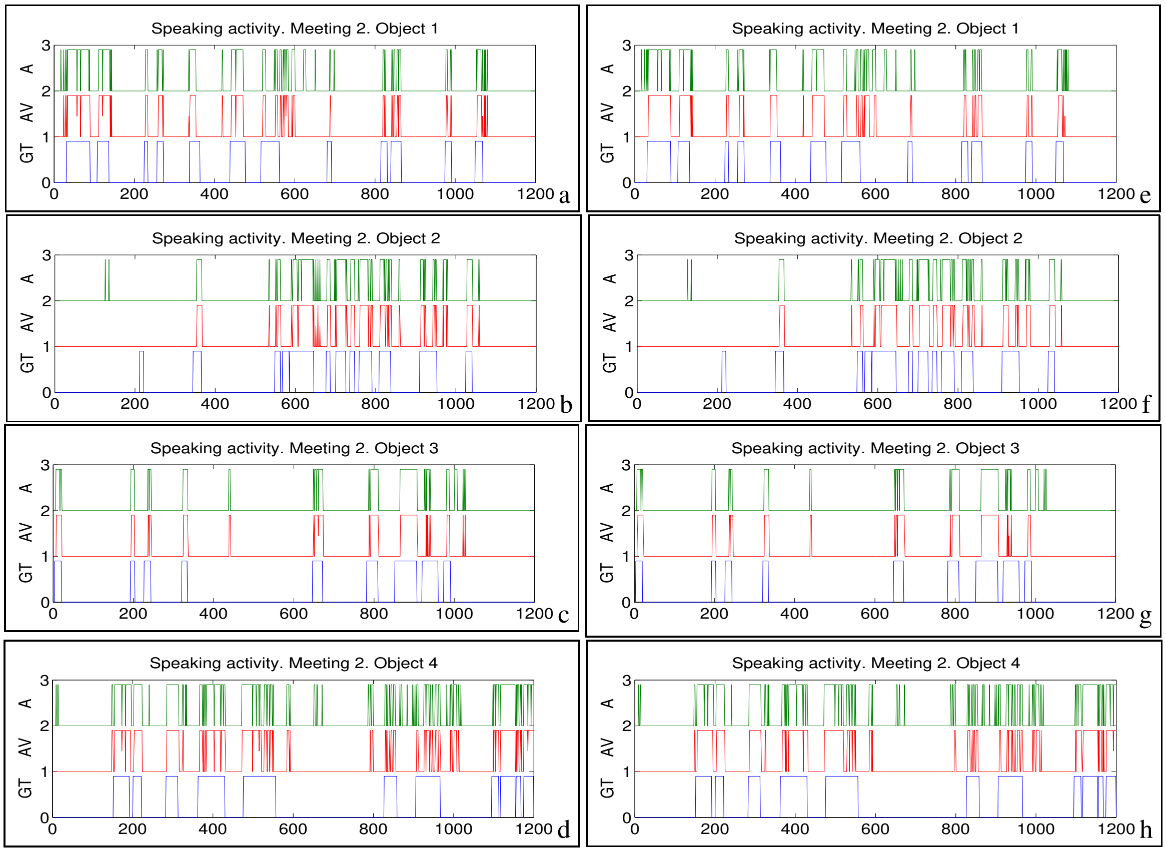

Figure 6: Speaker activity inference results, *meeting2*. In each figure, results are shown, as a function of frame number, for audio+av mapping, with manually labeled location (A), the MCMC-PF approach (AV), and for the manually labeled ground truth (GT). Figures (a-d) correspond to O1-O4 , respectively, for TPM values $\{0.8, 0.2\}$. Figures (e-h) are the corresponding results for TPM values $\{0.95, 0.05\}$

the MCMC-PF introduces a temporal dynamical process that serves as a smoothing filter. Short turns are often eliminated, which also tends to merge broken turns into longer units. The degree of smoothness clearly depends on the TPM parameters and their relation to the parameters in the switching audio observation likelihood (Eqs. 13-14). As expected, the filtering is higher in Fig. 6(e-h) compared to 6(a-d). However, the larger the smoothing effect, the longer the delay to detect beginnings and ending of speaker turns, thus defining a filtering tradeoff.

8.4.3 Effect of the MRF prior

The influence of the MRF prior cannot be appreciated in the previous two cases. To analyze the effect of this term, we conducted experiments with a five-person tracker on two excerpts of *meeting2*, where the walking silent person (O5) is significantly occluded by two seated participants several times. The first excerpt (*meeting2-oc1*) is 140 frames long (frames 550-690). The second one (*meeting2-oc2*) is 170 frames long (frames 830-1000). Performance is computed over 20 runs of the MCMC-PF, without and with the MRF prior.

Results for *meeting2-oc1* are shown in Table 3, Fig.7, and videos `meeting2_o1_mcmc_500_no_int.avi` (without MRF) and `meeting2_o1_mcmc_500_int.avi` (with MRF). Without the MRF prior, tracking is of high quality for objects O2-O4 ($SR = 1$, $TR = 1$, and $F_T \geq 0.87$). The results for O1 and O5 are shown in Table 3. For O1, tracking was lost once in 20 runs, locking onto O5. More importantly, for O5, tracking was lost in 25% of the cases, locking sometimes onto O1, or drifting away. In contrast, the use of the MRF prior produced high quality tracking for the five objects ($SR = 1$, $TR = 1$, $F_T \geq 0.87$ for O2-O4, see Table 3 for O1 and O5).

For *meeting2-oc2*, the results obtained with and without the use of the MRF prior are shown in Table 3. The occlusion affects objects O3 and O5, which got lost or confused with each other. On one hand, an

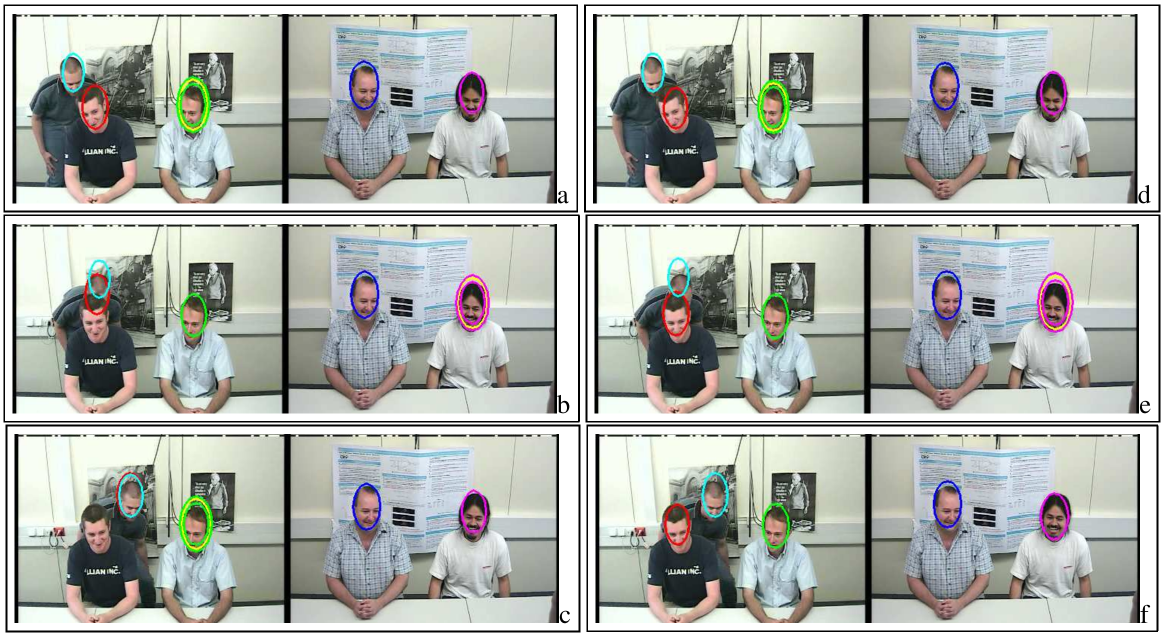

Figure 7: Five-object tracker, effects of the MRF model. (a-c): *meeting2-oc1*, frames 560, 585, and 594, without MRF prior. (d-f): same frames, with MRF prior.

improvement was obtained for O5 with the use of the interaction term (O5 was correctly tracked in the 20 runs). On the other hand, handling the occlusion for O3 represented the main challenge, given that his spatial structure model, as described previously, was not so accurate. Although not perfect, the performance for O3 was of good quality, and equivalent with or without the use of the MRF prior ($SR = 0.9$, $TR = 0.92$ in both cases). The performance for the other three objects (O1,O2, and O4) is of high quality in both cases ($SR = TR = 1$, $F_T \geq 0.87$). An example of the performance with the MRF prior can be seen in video `meeting2_o2_mcmc_500_int.avi`.

Overall, the results indicate that the interaction term model is effective, but its impact depends on the model of each object being accurate. One could expect that the occlusion model's advantage would diminish in case the involved objects did not fit the learned visual models well.

measure	*meet2-oc1*				*meet2-oc2*			
	O1wo	O5wo	O1w	O5w	O3wo	O5wo	O3w	O5w
SR	0.95	0.75	1.00	1.00	0.90	0.95	0.90	1.00
TR	0.96	0.89	1.00	1.00	0.92	0.96	0.92	1.00
F_T	0.87	0.74	0.88	0.74	0.77	0.72	0.76	0.74

Table 3: Results for sequences with person occlusion. Performance was affected for O1 and O5 in *meeting2-oc1*, and for O3 and O5 in *meeting2-oc2*. wo / w indicates without / with MRF prior.

8.4.4 Meeting3

Results for this sequence are shown in Table 4, Fig. 8, and video `meeting3_mcmc_500.avi`. In this case, a person (O3) makes a presentation and uses the whiteboard, while the others remain seated and mostly silent (see S_{GT} row in Table 4). Due to the FOV overlap, one person (O2) appears in two views. We only track this person in the frontal view. Our algorithm correctly tracks the location of the four people across the sequence, although tracking is more challenging for O3 (the presenter) due to his size and distance from the array. As one would expect, the audio localization algorithm detects a speaker at the table better than at a whiteboard/presentation, given their shorter distance to the microphone array. Due to this fact, when O2 makes noise (in frames 200-225), or when the presenter and a seated person speak simultaneously (e.g., O1 in frames 660-700, Fig. 8(c)), the tracker infers speaking activity only for the person at the table. Additionally, when people clearly face away from the array, the audio estimates degrade, and so does the inference of the speaking status. Overall, the speaking activity of the presenter is inferred with reasonable quality, although some of his turns are missed (e.g. around frame 1130, Fig. 8(e)). The activity of O1 and O4 is estimated with good quality, comparable to the obtained with the previous sequences. The comparatively low F_S-value for O2 is explained by the fact that he spoke briefly only once, and made noise that was identified as speaking activity. In comparison, as can be seen from Table 4 and video `meeting3_pf_500.avi`, a traditional joint PF shows again a much degraded performance. On one hand, such approach is not able to track all of the objects through the entire sequence in many occasions. The SR is especially bad for O2 who, although seated, had a relatively sudden change of dynamics, and for O3, the presenter. On the other hand, the quality of speaking activity inference is much worse, the situation being aggravated by the sparseness of speaker turns for three of the participants. Once again, the standard proportion test indicates a statistically significant performance difference between our method and the traditional PF for O_{avg} with confidence greater than 99.99% , for both SR and TR.

method	*measure*	O1	O2	O3	O4	O$_{avg}$
MCMC-PF	SR	1.00	1.00	1.00	1.00	1.00
	TR	1.00	1.00	1.00	1.00	1.00
	F_l	0.88	0.88	0.87	0.90	0.88
	F_s	0.81	0.57	0.57	0.79	0.69
PF	SR	0.95	0.15	0.45	0.95	0.62
	TR	0.95	0.27	0.87	0.96	0.76
	F_l	0.87	0.83	0.84	0.90	0.86
	F_s	0.38	0.05	0.15	0.05	0.16
	S_{GT}	0.07	0.02	0.57	0.03	0.17

Table 4: Tracking results for *meeting3* for our approach and a traditional multi-object PF.

8.4.5 Auto-initialization

A first example is shown in Figure 9(a-c) and `meeting1_mcmc_500_autoinit.avi`. The tracker is initialized at frame 0 with $\mathcal{I}_0 = \emptyset$ (zero objects). The birth-likely area for this example is the entire scene, roughly above the chest of the seated participants. At frame 1, O1, O2, and O4 are automatically initialized, while O3 is initialized once he moves his hand away from his face at frame 24. From this frame on, the performance is equivalent to the one obtained with manually initialized objects. A second example is shown in Figure 9(d-f) and video `meeting2_mcmc_500_autoinit.avi`. In this case, the tracker is initialized at frame 800 with $\mathcal{I}_{800} = \emptyset$. At frame 801, the algorithm detects O1-O4 and initializes them correctly. O5 is detected at frame 825.

An objective evaluation was conducted on *meeting1* and *meeting2* to test the sole effect of automatic initialization. We consider for evaluation only those frames after which the objects have been automatically initialized, which occurs around frame 25 for *meeting1*, and around frame 1 for *meeting2*. To be able to provide a fair comparison with the results obtained for *meeting2* in Table 2 (i.e. keeping equivalent conditions), we set $m_t^{max} = 4$, allowing a maximum of four objects along the sequence. This implies that only the four first detected objects are tracked, while the fifth person is excluded from this process, as in the experiments reported in Table 2. The results are shown in Table 5. For *meeting1*, we observe, once the objects are automatically initialized, the performance essentially remains the same as the case of manual initialization (Table 1). For *meeting2*, we observe that, comparing with Table 2, the performance decreased for O2 (SR decreased from 0.95 to 0.85, all other figures remain approximately the same), and improved slightly for O1 (SR increased from 0.95 to 1.00). The performance for O3-O4 remained the same. Considering all objects, the performance degradation is marginal (compare the O$_{avg}$ columns in Tables 2 and 5). Overall, these results suggest that the algorithm is adequate for our application, although as stated in Section 7, it could be improved by the use of a specialized face detector algorithm [44]. Finally, a more thorough evaluation of the algorithm for varying number of people would require the adaptation of our evaluation protocol to handle multi-object configuration and identification issues. Designing good evaluation procedures for such cases constitutes a research topic on its own [40].

9 Conclusion and future work

In this paper, we presented a probabilistic framework for the joint tracking of multiple people and their speaking activity in a multi-sensor meeting environment. Our framework integrates a novel AV observation model, a

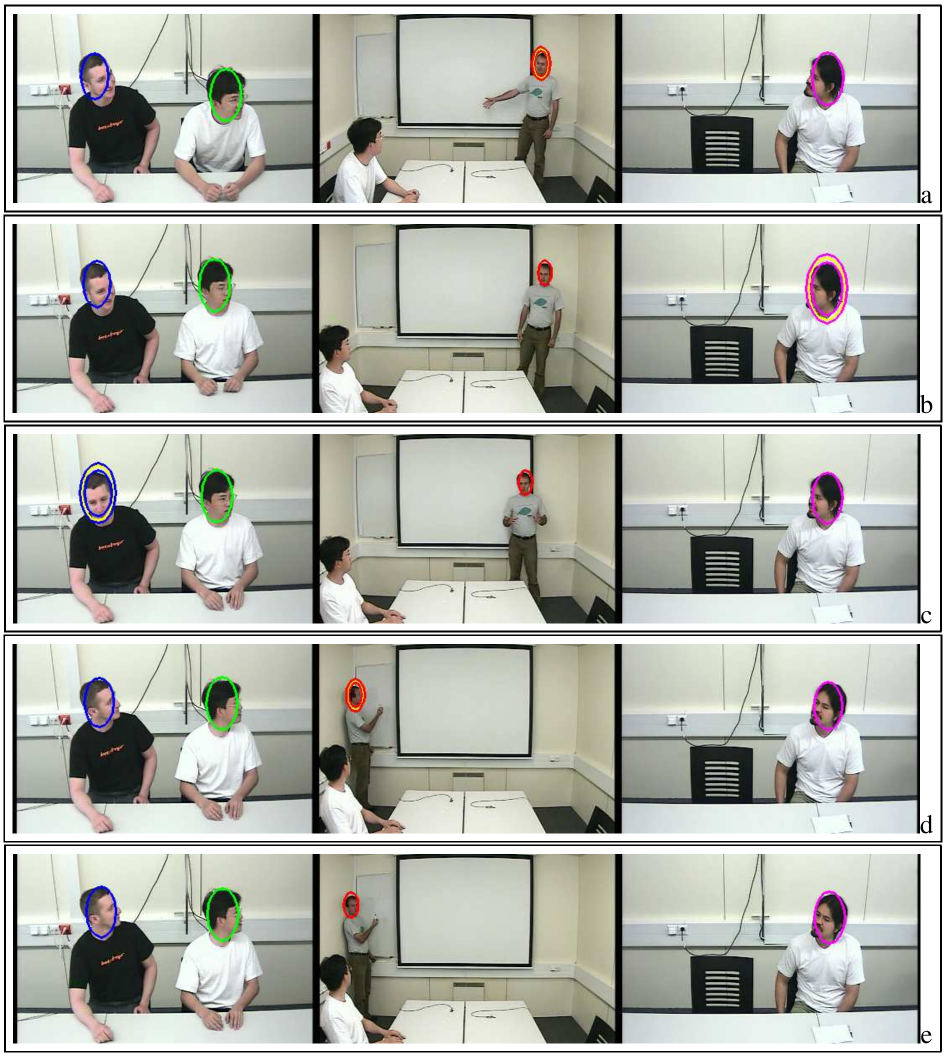

Figure 8: Multi-speaker tracking results, *meeting3*. Figures (a-e) correspond to frames 50, 490, 690, 945, and 1130, respectively. The presenter is speaking in all the displayed frames, excepting (b).

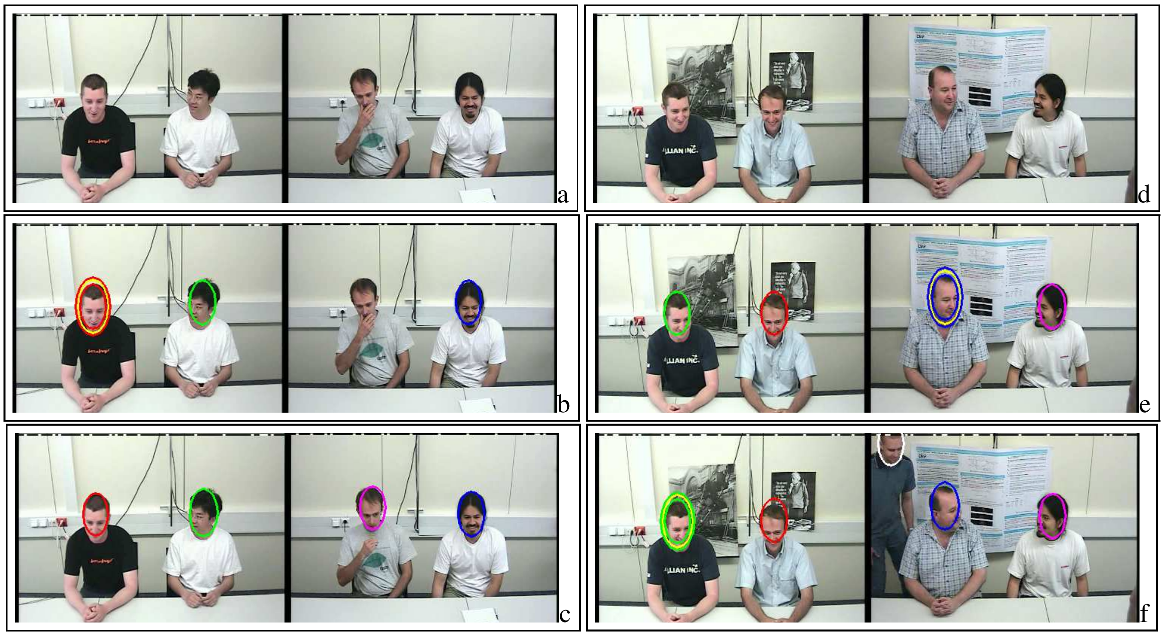

Figure 9: Auto-initialization. The initial configuration contains no objects. *meeting1*: (a-c) frames 0, 1, and 24. *meeting2*: (d-f) correspond to frames 800, 801, and 825.

method	*measure*	O1	O2	O3	O4	O_{avg}
meeting1	SR	1.00	1.00	1.00	1.00	1.00
	TR	1.00	1.00	1.00	1.00	1.00
	F_l	0.89	0.87	0.84	0.92	0.88
	F_s	0.72	0.74	0.78	0.75	0.75
meeting2	SR	1.00	0.85	1.00	1.00	0.96
	TR	1.00	0.98	1.00	1.00	0.99
	F_l	0.89	0.85	0.88	0.86	0.87
	F_s	0.72	0.73	0.71	0.80	0.74

Table 5: Tracking results for *meeting1* and *meeting2* for the MCMC-PF with auto-initialization.

principled mechanism to represent simple, proximity-based interactions (occlusion), and an efficient sampling strategy that overcomes some of the problems faced by traditional PFs in high-dimensional state-spaces. In principle, the sensor calibration algorithm we defined sets few constraints on the sensors' location, so cameras and microphones could potentially be placed in various configurations. We have shown that our framework can localize and track multiple people and their speaking status with good accuracy, tolerating visual clutter, and outperforming a traditional PF. Several issues remain open for improvement. First, more refined interaction models could be proposed, making use of the speaking status variable in the MRF prior, and introducing an occlusion variable in the state-space, which could explicitly define a set of switching occlusion modes. Second, although our model can reflect simultaneous speaking activity from multiple people, it is based on a limiting single-audio-source assumption. We are currently developing truly multi-speaker detection techniques [25] and plan to integrate them in our framework in the future. Third, the auto-initialization mechanism could be enhanced by using audio-based localization and/or face detection, whose integration in the MCMC-PF is conceptually direct. Finally, the evaluation on more dynamic data, including more complex cases of object birth/death, are also part of future work.

10 Acknowledgements

This work was supported by the Swiss National Center of Competence in Research on Interactive Multimodal Information Management (IM2), and the EC projects Multimodal Meeting Manager (M4), Augmented Multi-party Interaction (AMI), and Hearing Organization and Recognition of Speech in Europe (HOARSE), through the Swiss Federal Office for Education and Science. We thank Kevin Smith for discussions about various aspects of this work.

References

[1] P. Aarabi and S. Zaky, "Robust sound localization using multi-source audiovisual information fusion," *Information Fusion*, vol. 3, no. 2, pp. 209–223, Sep. 2001.

[2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, 1999.

[3] M. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *Proc. European Conf. on Computer Vision (ECCV)*, May 2002.

[4] C. Berzuini, N. Best, W. Gilks, and C. Larizza, "Dynamic conditional independence models and Markov Chain Monte Carlo Methods," *Journal of the American Statistical Association*, vol. 92, no. 440, pp. 1403–1412, 1997.

[5] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, May 2004.

[6] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *Proc. of the IEEE*, vol. 92, no. 3, pp. 485–494, Mar. 2004.

[7] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *Proc. IEEE Int. Conf. on Multimedia (ICME)*, New York, July 2000.

[8] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: a meeting capture and broadcasting system," in *Proc. ACM Int. Conf. on Multimedia (MM)*, Juan les Pins, Dec. 2002.

[9] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., chapter 8, pp. 157–180. Springer, 2001.

[10] A. Dielmann and S. Renals, "Dynamic Bayesian networks for meeting structuring," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, May 2004.

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

[12] P. Fearnhead, "MCMC, sufficient statistics and particle filters," *Journal of Computational and Graphical Statistics*, vol. 11, pp. 848–862, 2002.

[13] J. Fisher, T. Darrell, W.T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. Neural Information Processing Systems (NIPS)*, Denver, Dec. 2000.

[14] D. Gatica-Perez, G. Lathoud, I. McCowan, and J.-M. Odobez, "A mixed-state i-particle filter for multi-camera speaker tracking," in *Proc. IEEE Int. Conf. on Computer Vision, Workshop on Multimedia Technologies for E-Learning and Collaboration (ICCV-WOMTEC)*, Nice, Oct. 2003.

[15] W. Gilks and C. Berzuini, "Following a moving target - Bayesian inference for dynamic Bayesian models," *Journal of the Royal Statistical Society, Series B*, vol. 63, pp. 127–146, 2001.

[16] S. M. Griebel and M. S. Brandstein, "Microphone array source localization using realizable delay vectors," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz NY, Oct. 2001.

[17] J. Hershey and J. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," in *Proc. Neural Information Processing Systems (NIPS)*, Denver, Nov. 1999.

[18] M. Isard, *Visual Motion Analysis by Probabilistic Propagation of Conditional Density*, D.Phil. Thesis, Oxford University, 1998.

[19] M. Isard and J. MacCormick, "BRAMBLE: A Bayesian multi-blob tracker," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, Vancouver, Jul. 2001.

[20] B. Kapralos, M. Jenkin, and E. Milios, "Audio-visual localization of multiple speakers in a video teleconferencing setting," *Int. J. Imaging Systems and Technology*, vol. 13, pp. 95–105, 2003.

[21] Z. Khan, T. Balch, and F. Dellaert, "An MCMC-based particle filter for tracking multiple interacting targets," in *Proc. European Conf. on Computer Vision (ECCV)*, Prague, May 2004.

[22] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[23] F. Kubala, S. Colbath, D. Liu, and J. Makhoul, "Rough'n'ready: a meeting recorder and browser," *ACM Computing Surveys*, vol. 31, no. 2es, Jun. 1999.

[24] G. Lathoud, I. McCowan, and J.-M. Odobez, "Unsupervised Location-Based Segmentation of Multi-Party Speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), NIST Meeting Recognition Workshop*, Montreal, May 2004.

[25] G. Lathoud and M. Magimai.-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Mar. 2005.

[26] S.Z. Li, *Markov Random Field Modeling in Computer Vision*, Springer, 1995.

[27] J. S. Liu and R. Chen, "Sequential Monte Carlo methods for dynamic systems," *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 1032–1044, 1998.

[28] J.S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, 2001.

[29] L. Lu and H. J. Zhang, "Content analysis for audio classification and segmentation," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, 2002.

[30] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[31] J.E. McGrath, *Groups: Interaction and Performance*, Prentice-Hall, 1984.

[32] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proc. Human Language Technology Conf. (HLT)*, San Diego, CA, March 2001.

[33] D. Moore, "The IDIAP smart meeting room," IDIAP-COM 07, IDIAP Research Institute, Martigny, 2002.

[34] V. Pavlovic, A. Garg, and J. Rehg, "Multimodal speaker detection using error feedback dynamic bayesian networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Hilton Head Island, SC, 2000.

[35] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based Probabilistic Tracking," in *Proc. European Conf. on Computer Vision (ECCV)*, Copenhagen, May 2002.

[36] G.S Pingali, G. Tunali, and I. Carlbom, "Audio-visual tracking for natural interactivity," in *Proc. ACM Int. Conf. on Multimedia (MM)*, Orlando, Oct. 1999.

[37] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 5, Sep. 2004.

[38] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation," in *Proc. European Conf. on Speech Comm. and Tech. (Eurospeech)*, Aalborg, Sep. 2001.

[39] M. Siracusa, L.-P. Morency, K. Wilson, J. Fisher, and T. Darrell, "A multi-modal approach for determining speaker location and focus," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, Vancouver, 2003.

[40] K. Smith, D. Gatica-Perez, J.-M. Odobez, and S. Ba, "Evaluating multi-object tracking," in *Proc. IEEE CVPR Workshop on Empirical Evaluation Methods on Computer Vision (CVPR-EEMCV)*, San Diego, Jun. 2005.

[41] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE IEEE Trans. on Neural Networks*, vol. 13, no. 4, pp. 928–938, 2002.

[42] D. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone array measurements," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Munich, Apr. 1997.

[43] J. Vermaak, M. Gagnet, A. Blake, and P. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, Vancouver, July 2001.

[44] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Kawaii, Dec. 2001.

[45] B. Vo, S. Singh and W.K. Ma, "Tracking Multiple Speakers with Random Sets," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, May 2004.

[46] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, UT, May 2001.

[47] B. Wrede and E. Shriberg, "The relationship between dialogue acts and hot spots in meetings," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Virgin Islands, Dec. 2003.

[48] D. Zotkin, R. Duraiswami, and L. Davis, "Multimodal 3-D tracking and event detection via the particle filter," in *IEEE ICCV Workshop on Detection and Recognition of Events in Video (ICCV-EVENT)*, Vancouver, Jul. 2001.