IDIAP RESEARCH REPORT

MULTIMODAL MULTISPEAKER PROBABILISTIC TRACKING IN MEETINGS

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

IDIAP–RR 04-66

DECEMBER 2004

PUBLISHED IN
Proc. Int. Conf. on Multimodal Interfaces (ICMI), Trento, Oct. 2005

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 04-66

MULTIMODAL MULTISPEAKER PROBABILISTIC TRACKING IN MEETINGS

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

DECEMBER 2004

PUBLISHED IN
Proc. Int. Conf. on Multimodal Interfaces (ICMI), Trento, Oct. 2005

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 [4]. In this domain, 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. 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 [12]. 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). Approaches addressing some of these issues have begun to appear [2, 3].

In this paper, we address the problem as one of approximate inference in a dynamical graphical model, using PF techniques [7, 11], building on recent advances in the field [9]. 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 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 MCMC techniques with the PF framework, allowing for efficient sampling [11, 9]. Our work integrates data captured by a small microphone array and multiple cameras with mostly non-overlapping fields of view by 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 previous work. Section 3 summarizes our framework. Section 4 describes the multi-person dynamical model. Section 5 presents the AV observation model. Section 6 describes the sampling technique. Section 7 discusses the case of varying numbers of people. Section 8 describes experiments and results. 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 [13, 15, 1, 6, 3, 2]. 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 [13] or multiperson scenes where only the location of the current speaker needs to be tracked [15, 1, 6]. Many of these works used simple sensor configurations (one camera and a microphone pair) [13, 15, 1]. Among the existing techniques, probabilistic generative models based on exact [13] or approximate inference methods (both variational [1] and sampling-based [15]) appear to be the most promising, given their principled formulation and demonstrated performance.

Although multiobject visual tracking is a classic field in computer vision [8, 14, 9], the AV multispeaker tracking problem has been studied relatively recently, making use of more complex sensor configurations to cover an entire workspace (table, whiteboards, etc.) [4, 3, 2]. The work in [4] described a system based on a

device that integrates a small circular microphone array and several calibrated cameras. The system, in which each person is tracked independently, consists of three modules: AV auto-initialization, HMM-based visual tracking, and tracking verification. The closest works to ours are [2, 3], both based on PF techniques. The work in [2] used two calibrated cameras and four linear sub-microphone arrays on a wall, and was based on the model by [8], defining a multi-person state-space in which the number of people can vary over time. An observation model was defined by two terms: one for video, derived from background substraction, and one for audio, derived from short-time Fourier transforms computed on each microphone's signal. The PF relied on importance sampling (IS), and is thus likely to become rapidly inefficient as the number of objects increases. The work in [3] used the same setup as [4], 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.

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 [9], our model has two advantages over [2, 3]. First, unlike [2, 3], 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 [2], we use 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 our algorithm in more detail than [2, 3].

3 Model specification

We use a generative approach to model the tracking problem [7, 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 up to time-step t, $\mathbf{Y}_{1:t} = (\mathbf{Y}_1, ..., \mathbf{Y}_t)$, 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]. Following [7], 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,

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

using IS (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 in the image plane and the speaking activity are tracked. In this work, 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. Each of these issues is 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 [10, 9]. 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

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

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

$$\begin{aligned} p(\mathbf{X}_t|\mathbf{Y}_{1:t}) \quad &\approx \quad Z^{-1}p(\mathbf{Y}_t|\mathbf{X}_t)\prod_{(i,j)\in\mathcal{C}}\phi(\mathbf{X}_{i,t},\mathbf{X}_{j,t})\cdot \\ &\qquad \sum_n w_{t-1}^{(n)}\prod_{i\in\mathcal{I}_t} p(\mathbf{X}_{i,t}|\mathbf{X}_{i,t-1}^{(n)}), \end{aligned} \tag{4}$$

where the interaction term can be moved out of the sum over all particles, as it does not depend on past information [9]. 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 [7], and $p(k_{i,t}|k_{i,t-1})$ is a 2×2 transition probability matrix (TPM).

The interaction model we adopt takes into account visual information, and penalizes large visual overlaps between objects, 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 $\mathbf{x}_{j,t}$, that is, the application of the continuous transformation to the object template in 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 from information retrieval. Assuming that $\mathcal{S}_{i,t}$ is the reference, the two above measures are respectively defined by

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

Note that, 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, and their minimum (zero) if the objects have no overlap.

5 Audio-visual observation model

A person is represented by the elliptical silhouette of the head in the image plane. 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. 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 image plane. We first use an audio speaker localization approach consisting of two steps: finding candidate source locations, and classifying them as speech or non-speech [6]. In the first step, a single-source localization algorithm based on the Steered Response Power - Phase Transform (SRP-PHAT) measure [5] is used to generate candidate 3-D speaker locations. In practice, the estimated range is imprecise, and only azimuth and elevation are significant. The algorithm exhaustively searches for the location with maximum SRP-PHAT on a fixed concentric grid of points. SRP-PHAT is known to be suitable for reverberant environments. 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 estimates. This technique is effective, unlike most energy-thresholding algorithms, at detecting the low-energy beginning of utterances, and short speaker turns [6]. Using these filtered 3-D location estimates, the frame rates of audio and video are matched, generating between zero and three audio location estimates per video frame. Finally, we use a nearest-neighbor approach to project 3-D audio estimates on the 2-D image plane [6]. This requires an off-line calibration procedure between the AV sensors. The procedure uses training data collected by having a person talking while moving in typical areas of the meeting room. The correspondences between 3-D and 2-D points are obtained from the audio estimates and the output of a single-person visual tracker, respectively. A precise sensor calibration procedure is not required in our case [6].

The audio observation likelihood is defined on the image plane, relating the Euclidean distance between the 2-D audio location estimates and the candidate particles. Let $x_{i,t}^a = (u_{i,t}^a, v_{i,t}^a)$ denote the audio estimate closest to the translation components of a configuration $x_{i,t} = (u_{i,t}, v_{i,t})$ (i.e., the ellipse center). We define a distribution for each value the speaking status variable can take,

$$p(\mathbf{Y}_{i,t}^a|\mathbf{x}_{i,t}, k_{i,t} = 1) \propto \begin{cases} K_1^a, & ||x_{i,t} - x_{i,t}^a|| \leq R^a, \\ K_2^a, & \text{otherwise}, \end{cases} \tag{8}$$

$$p(\mathbf{Y}_{i,t}^a|\mathbf{x}_{i,t}, k_{i,t} = 0) \propto \begin{cases} K_1^a, & ||x_{i,t} - x_{i,t}^a|| \geq R^a, \\ K_2^a, & \text{otherwise}, \end{cases} \tag{9}$$

where $||\cdot||$ denotes the Euclidean distance, R^a defines a radius around the translation components of $\mathbf{X}_{i,t}$, and $K_1^a > K_2^a$ 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. In case no audio location estimates exist, $u_{i,t}^a$ and $v_{i,t}^a$ are set to an arbitrarily large number.

5.2 Spatial structure observations

We propose an observation model of spatial structure of human heads, based on a parametric representation of the overlap between skin-color blobs and head configurations. The model is based on the fact that the presence of skin pixels in a typical head blob (e.g. face and neck) is usually confined to specific regions inside and outside

a proposed head configuration. Skin-color blobs are first extracted at each frame as described in [6]. 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 is further represented by 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 [16], 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 a Gaussian Mixture Model (GMM) with diagonal covariance matrices, and a uniform distribution $U(\cdot)$ used to limit the effect of very low likelihood values,

$$p(\mathbf{Y}^{st}_{i,t}|\mathbf{X}_{i,t}) \propto \omega_0 U(\mathbf{Y}^{st}_{i,t}) + \sum_{l=1}^{N_{st}} \omega_l \mathcal{N}(\mathbf{Y}^{st}_{i,t}, \mu_l, \Sigma_l), \tag{10}$$

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 using the spatial support of the configuration parts and the blob.

5.3 Shape observations

Edge-based observations are computed, based on a classic model, along L lines normal to a hypothesized contour [7]. 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

$$p(\mathbf{Y}^{sh}_{i,t}|\mathbf{X}_{i,t}) \propto \prod_{l=1}^{L} \max\left(K^{sh}, \exp(-\frac{\|z^l_{i,\hat{c}} - z^l_{i,0}\|^2}{2(\sigma^{sh})^2})\right), \tag{11}$$

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

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 [11]. In order to efficiently place samples as close as possible to regions of high likelihood, we build on recent work and propose to sample from Eq. 4 with MCMC techniques [11], using a Metropolis-Hastings (MH) sampler at each time-step [9]. 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 [11]. 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)

$$\alpha = \min\left(1, \frac{p(\mathbf{X}^*)q(\mathbf{X}|\mathbf{X}^*)}{p(\mathbf{X})q(\mathbf{X}^*|\mathbf{X})}\right), \tag{12}$$

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 $\frac{1}{N}$ [11].

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. 12 involves the evaluation of Eq. 4, i.e., a sum over all particles. We use a mixture model formulation over all objects, where one object is chosen at each step in the chain to attempt a move,

$$q(\mathbf{X}_t^*|\mathbf{X}_t) = \sum_i q(i) q(\mathbf{X}_t^*|\mathbf{X}_t, i), \tag{13}$$

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

$$q(\mathbf{X}_t^*|\mathbf{X}_t, i) = \begin{cases} \frac{1}{N}\sum_n p(\mathbf{X}_t^*|\mathbf{X}_{t-1}^{(n)}) & i = i^*, \\ \frac{1}{N}\sum_n p(\mathbf{X}_t^*|\mathbf{X}_{t-1}^{(n)}) & i \neq i^*, \mathbf{X}_t^* = \mathbf{X}_t, \\ 0 & i \neq i^*, \mathbf{X}_t^* \neq \mathbf{X}_t, \end{cases} \tag{14}$$

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. 14 is

$$\begin{aligned} q(\mathbf{X}_t^*|\mathbf{X}_t, i) \quad &= \tfrac{1}{N}\textstyle\sum_n p(\mathbf{X}_{i,t}^*|\mathbf{X}_{i,t-1}^{(n)}) \cdot \\ &\textstyle\prod_{l \in \mathcal{I}_t - \{i\}} p(\mathbf{X}_{l,t}|\mathbf{X}_{l,t-1}^{(n)}) \delta(\mathbf{X}_{l,t}^* - \mathbf{X}_{l,t}). \end{aligned} \tag{15}$$

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

$$\alpha = \min\left(1, \frac{p(\mathbf{Y}_t|\mathbf{X}_t^*) \prod_{(i^*,j)\in\mathcal{C}_{i^*}} \phi(\mathbf{X}_{i^*,t}^*, \mathbf{X}_{j,t}^*)}{p(\mathbf{Y}_t|\mathbf{X}_t) \prod_{(i^*,j)\in\mathcal{C}_{i^*}} \phi(\mathbf{X}_{i^*,t}, \mathbf{X}_{j,t})}\right), \tag{16}$$

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

$$\alpha = \min\left(1, \frac{p(\mathbf{Y}_t|\mathbf{X}_{i^*,t}^*) \prod_{(i^*,j)\in\mathcal{C}_{i^*}} \phi(\mathbf{X}_{i^*,t}^*, \mathbf{X}_{j,t}^*)}{p(\mathbf{Y}_t|\mathbf{X}_{i^*,t}) \prod_{(i^*,j)\in\mathcal{C}_{i^*}} \phi(\mathbf{X}_{i^*,t}, \mathbf{X}_{j,t})}\right), \tag{17}$$

which only involves the evaluation of single-object AV likelihood terms. MH 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.

Finally, the mean estimate is approximated by the marginal mean estimates for each object, $\bar{\mathbf{X}}_t = (\bar{\mathbf{X}}_{i,t}), i \in \mathcal{I}_t$. Each $\bar{\mathbf{X}}_{i,t}$ 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 [7]. The full algorithm is summarized in Fig. 2.

7 Varying number of people

Although the MCMC-PF could formally integrate birth-death processes as part of the filtering recursion, this would require a multi-person observation model that allowed for the comparison between configurations containing varying number of people [8, 2]. The factorized observation model in Eq. 7 is not suitable for such a

---

`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. 13).
  - `Compute acceptance ratio` α (Eq. 17).
  - `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` $\bar{\mathbf{X}}_t$.

Figure 2: MCMC-PF algorithm.

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 $\mathcal{I}_t$.

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 small 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. Needless to say, more robust people detectors could be integrated in our approach [16]. 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 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. Two cameras on opposite walls record frontal views of the 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 the skin-color and spatial structure models, 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 frames, respectively), and one three-camera sequence (*meeting3*, 1200 frames). The sequences are composed by merging the different views. The first two sequences have non-overlapping FOVs, while in the third one there is some overlap.

8.2 Parameter setting

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 for speaking activity was set to $p(k_{i,t}|k_{i,t-1}) = \left[\begin{smallmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{smallmatrix}\right]$. 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 semiautomatic head bounding-box ground truth (GT) is generated using a color-based single-person tracker [14]. 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 precision and recall are greater than zero, and zero otherwise.

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. 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 frame, defining:

5. *Speak F-measure* (F_S). 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. 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 in the companion website[1].

[1] `www.idiap.ch/ ~ gatica/av-tracking-multiperson.html`.

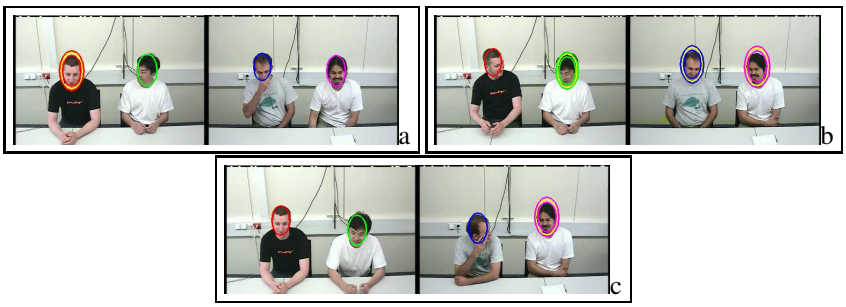

Figure 3: Multispeaker tracking results, *meeting1*. Location and speaking status (double ellipse if a person speaks) are inferred for each person. Figures (a-c) correspond to frames 10, 526, and 1420, respectively.

Meeting1. The results on the *meeting1* sequence are shown in Fig. 3 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`). 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. 3(a),(c)), 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 [6]. 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 infer simultaneous speaking activity for multiple participants (see Fig. 3(b)). 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 computed using 20 runs, and are shown in Table 1 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. Finally, the inference of speaking activity is degraded considerably. The number of particles that is required with the basic PF to perform as well as the MCMC-PF is prohibitively high.

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.

Meeting2. The results are shown in Fig. 4, Table 2, and 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

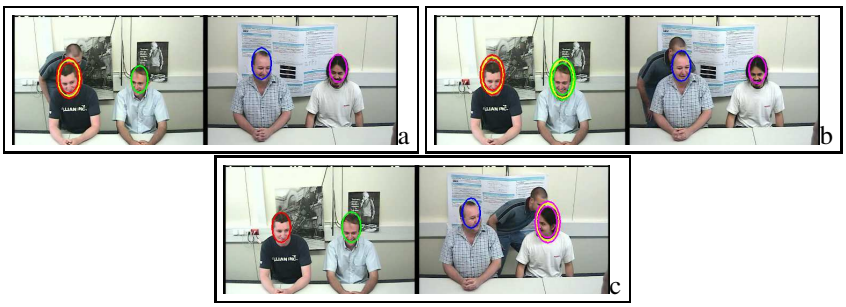

Figure 4: Results, *meeting2*. Figures (a-c) correspond to frames 575, 860, and 909, respectively.

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 combination of visual cues renders the tracker more robust: while the spatial structure observations help in cases of uncertainty with respect to edge information (e.g. textured background), 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. 4(a),(c)). Regarding speaking activity, our approach can correctly infer some cases of simultaneous speech (Fig. 4(b)). Finally, our approach is significantly more effective than a traditional joint PF, which shows an even more severe performance degradation: SR can be as low as 35-40% and, with the exception of F_T, all other figures are considerably lower than the ones obtained for *meeting1*. This result suggests that the MCMC-PF is more robust to realistic conditions than traditional approaches.

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 joint multi-object PF.

Testing the interaction prior. We conducted experiments with a five-person tracker on an excerpt of the *meeting2* sequence (frames 550-690), where the walking person (O5) is partially occluded by two seated participants. Performance is computed over 20 runs of the MCMC-PF, without and with the MRF prior. Results are shown in Table 3, Fig. 5, `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). The performance with the MRF prior on a second excerpt of *meeting2* (frames 830-1000) is shown in `meeting2_o2_mcmc_500_int.avi`.

Meeting3. Results for this sequence are shown in Table 4, Fig. 6, 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 due to his size and distance from the array. As one would expect, the single-source audio localization algorithm detects a speaker at the table better than at

measure	O1w	O5w	O1p	O5p
SR	0.95	0.75	1.00	1.00
TR	0.96	0.89	1.00	1.00
F_T	0.87	0.74	0.88	0.74

Table 3: Results for an excerpt of *meeting2* with partial occlusion. Performance was affected for O1 and O5; w (resp. p) indicates without (resp. with) MRF prior.

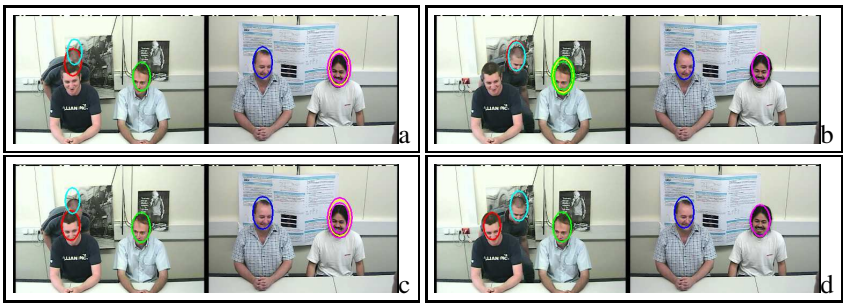

Figure 5: Five-person tracker, effects of the interaction model. (a-b): frames 585 and 594, without MRF prior. (c-d): same frames, with MRF prior.

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. 6(b)), the tracker infers speaking activity only for the person at the table. Overall, the speaking activity of the presenter is inferred with reasonable quality, although some of his turns are missed. The activity of O1 and O4 is estimated with good quality, comparable to the obtained with 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.

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	0.88	0.88	0.87	0.90	0.88
	F_S	0.81	0.57	0.57	0.79	0.69
	S_{GT}	0.07	0.02	0.57	0.03	0.17

Table 4: Tracking results for *meeting3* for our approach.

Auto-initialization. The performance of the algorithm is illustrated in video `meeting1_mcmc_500_autoinit.avi`. The tracker is initialized at frame 0 with $\mathcal{I}_0 = \emptyset$. 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 Fig. 7 and `meeting2_mcmc_500_autoinit.avi`. 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. Overall, 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 [16].

9 Conclusions

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 principled mechanism to represent simple, proximity-based interactions (occlusion), and an efficient sampling strategy that overcomes some of the problems faced by traditional PFs. 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

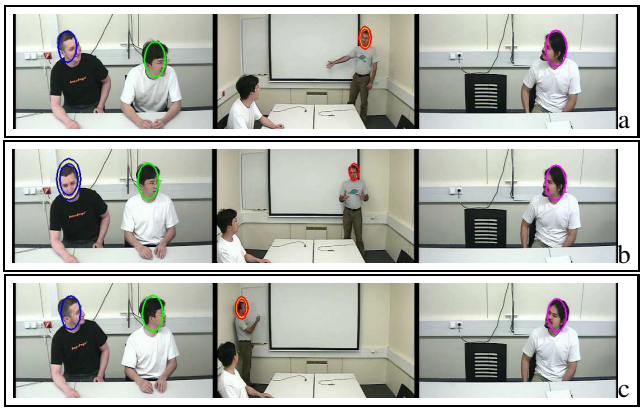

Figure 6: Results, *meeting3*. Figures (a-c) correspond to frames 50, 690, and 945, respectively.

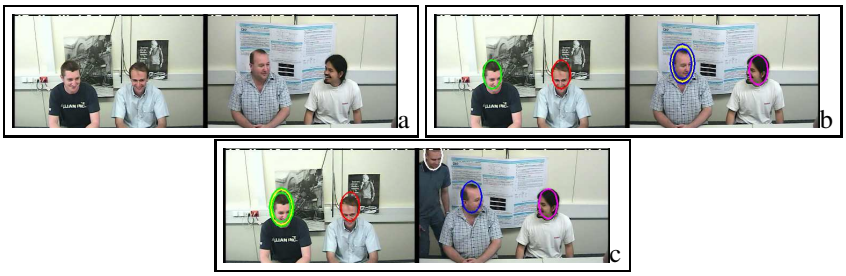

Figure 7: Handling varying number of objects, *meeting2*; (a-c) correspond to frames 800, 801, and 825, resp.

models could be proposed, making use of the speaking status variable in the MRF prior, and introducing an occlusion variable in the state-space. 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. 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 of our approach on more dynamic data will also be 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 Multiparty Interaction (AMI, pub. AMI-115), 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] M. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *Proc. ECCV*, May 2002.

[2] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proc. ICASSP*, May 2004.

[3] 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.

[4] R. Cutler, Y. Rui, A. Gupta, JJ Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: A meeting capture and broadcasting system," in *Proc. ACM MM*, Oct. 2002.

[5] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, Ch. 8, pp. 157–180. Springer, 2001.

[6] D. Gatica-Perez, G. Lathoud, I. McCowan, and J.-M. Odobez, "A mixed-state i-particle filter for multi-camera speaker tracking," in *Proc. ICCV-WOMTEC*, Oct. 2003.

[7] M. Isard, *Visual Motion Analysis by Probabilistic Propagation of Conditional Density*, PhD Thesis, 1998.

[8] M. Isard and J. MacCormick, "Bramble: A Bayesian multi-blob tracker," in *Proc. ICCV*, Jul. 2001.

[9] Z. Khan, T. Balch, and F. Dellaert, "An MCMC-based particle filter for tracking multiple interacting targets," in *Proc. ECCV*, May 2004.

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

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

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

[13] V. Pavlovic, A. Garg, and J. Rehg, "Multimodal speaker detection using error feedback dynamic Bayesian networks," in *Proc. CVPR*, Jun. 2000.

[14] P. Perez, C. Hue, and J. Vermaak, "Color-based Probabilistic Tracking," in *Proc. ECCV*, May 2002.

[15] J. Vermaak, M. Gagnet, A. Blake, and P. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. ICCV*, July 2001.

[16] P. Viola and M. Jones, "Rapid object detection by boosted cascade of simple features," in *Proc. CVPR*, Dec. 2001.