TRACKING PEOPLE IN MEETINGS WITH PARTICLES

Daniel Gatica-Perez, Jean-Marc Odobez, Sileye Ba, Kevin Smith, and Guillaume Lathoud

IDIAP Research Institute
Rue de Simplon 4, CH-1920 Martigny, Switzerland
{gatica, odobez, sba, smith, lathoud}@idiap.ch

ABSTRACT

Automatic meeting analysis is an emerging research field. In this paper, we present stochastic algorithms for tracking people in multi-sensor meeting rooms, for a number of relevant tasks, including tracking multiple people, tracking head pose towards analysis of visual focus-of-attention, and tracking speaker activity using audio-visual information. A Bayesian framework based on Sequential Monte Carlo methods is used in all cases. We discuss the advantages and limitations of our approach, illustrate it with results, and highlight a number of open issues.

1. INTRODUCTION

The automatic analysis of human interaction constitutes a rich research field. In particular, meetings exemplify the multimodal nature of human communication, and the complex patterns that emerge from the interaction between multiple people [10]. In view of the amount of relevant information in meetings suitable for automatic extraction, this domain has attracted attention in fields spanning computer vision, speech processing, human-computer interaction, and information retrieval [16].

Localizing and tracking people play important roles in meeting analysis. As a data source, meetings recorded in multi-sensor rooms consist of unedited streams of audio and video, captured with multiple cameras and microphones covering participants and workspace areas. In such setups, tracking is useful to determine the number and location of participants, to provide accumulated information for person identification, to select a fixed camera or to steer a motorized one as part of a visualization or production model, to enhance the audio stream for speech recognition using microphone arrays, and to provide cues for detection of location-based events. In all of these cases, the availability of multiple views and modalities represents an advantage.

Tracking people and their activity is also relevant for higher-level multimodal tasks that relate to the communicative goal of meetings. Experimental evidence in social psychology has highlighted the role of non-verbal behavior (e.g. gaze and facial expressions) in interactions [12], and the power of speaker turn patterns to capture information about the behavior of a group and its members [10, 12]. Identifying such multimodal behaviors requires reliable people tracking.

In this paper, we discuss algorithms to track people in meetings using a consistent Bayesian framework, namely sequential Monte Carlo (SMC) methods or particle filters (PF). SMC methods approximate the Bayesian solution to the tracking problem using sampling techniques, and have gained popularity in recent years to deal with non-linear and non-Gaussian state-space models, due to their versatility, ease of implementation, and success in challenging applications. We present PFs to track multiple interacting people, with occlusion as the typical problem in meeting rooms, to track location and head pose, as a surrogate for gaze, and to track location and speaking activity using audio-visual data. While the SMC formulation is general, each of the addressed problems pose specific challenges, and call for a number of specific choices. We highlight each of them in the following sections. Our work is an ongoing effort towards building probabilistic models of multimodal human interaction.

The paper is organized as follows. Section 2 summarizes the SMC framework. Section 3 briefly describes our multi-sensor meeting room. Section 4 describes our work on multi-object visual tracking. Section 5 describes our progress on head-pose tracking. Section 6 presents our work on audio-visual tracking. Videos with results for all sections can be found in the paper's companion website [17]. Section 7 provides some final remarks.

2. SEQUENTIAL MONTE CARLO FRAMEWORK

The Bayesian formulation of the tracking problem is well known. Denoting by X_t the hidden state representing the object configuration at time t, and by Y_t the observation extracted from the image, the filtering distribution $p(X_t|Y_{1:t})$ of X_t given all the observations $Y_{1:t} = (Y_1 \dots Y_t)$ up to the current time can be recursively computed by [3]:

$$p(X_t|Y_{1:t}) = Z^{-1} p(Y_t|X_t) \times \int_{X_{t-1}} p(X_t|X_{t-1}) p(X_{t-1}|Y_{1:t-1}) dX_{t-1} \qquad (1)$$

where Z is a normalizing constant. A PF is a numerical approximation to the above recursion in the case of non-linear and non-Gaussian models. The basic idea behind PF consists of representing the filtering distribution using a weighted set of samples $\{X_t^n, w_t^n\}_{n=1}^{N_s}$, and updating this representation as new data arrives. With this representation, Eq. 1 can be approximated by :

$$p(X_t|Y_{1:t}) \approx Z^{-1} p(Y_t|X_t) \sum_{n=1}^{N_s} w_{t-1}^n p(X_t|X_{t-1}^n) \qquad (2)$$

using importance sampling. Given the particle set at the previous time step $\{X_{t-1}^n, w_{t-1}^n\}$, configurations at the current time step are drawn from a proposal distribution $q(X_t) = \sum_n w_{t-1}^n p(X_t|X_{t-1}^n)$. The weights are then computed as $w_t^n \propto p(Y_t|X_t^n)$.

Four elements are important in defining a PF:

1. The *state space*. We use mixed-spaces, where the state is the conjunction of continuous variables specifying the spatial object configuration (e.g. position, scale) and discrete variables labeling the object state (e.g. whether a person is occluded or not).

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, pub. AMI-43), and Hearing Organization and Recognition of Speech in Europe (HOARSE).

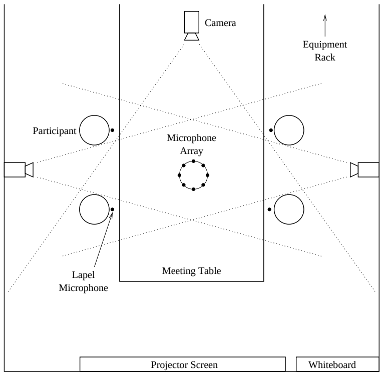

Figure 1: Multi-sensor meeting room configuration.

2. The *dynamical model* $p(X_t|X_{t-1})$ defines the temporal evolution of the state.

3. The *observation likelihood* $p(Y_t|X_t)$ measures the adequacy between the observation and the state.

4. The *sampling mechanism* places new samples as close as possible to regions of high likelihood.

These elements, along with specific issues and proposed solutions, will be described in each of the following three sections.

3. MULTI-SENSOR MEETING ROOM

Our algorithms are tested on data captured 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 CCTV cameras [11]. 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. Sample images can be seen in the following sections. 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. A diagram in shown in Fig. 1.

4. TRACKING MULTIPLE PEOPLE

Challenges. The long-term, reliable tracking of multiple people in meetings is a challenging task. Meeting rooms pose a number of issues for visual tracking including occlusion, clutter, variation of illumination, and variation of appearance arising from changing pose. On the other hand, multi-sensor meeting rooms offer some unique advantages that ease the task of tracking. These can include constraints on the working space and group dynamics, and redundancies in video data from cameras with overlapping FOVs.

Our approach. We define a joint multi-object state space, which constitutes a rigorous implementation of the problem. The state X_t contains the configuration for every person in the scene $X_t = (x_{1,t}, ..., x_{M,t})$, where M denotes the number of people, and $x_{i,t}$ contains translation and scaling parameters for person i.

Tracking a significant number of objects in a joint-object framework becomes increasingly difficult as adding new objects to the scene increases the search space exponentially. A sampling strategy known as Partitioned Sampling (PS) [9] helps reduce the dimensionality problem by handling one object at a time, but introduces problems with bias and impoverishment of the particle representation, dependent on the object ordering. We propose sampling using Distributed Partitioned Sampling (DPS), which redefines the distribution as a mixture model composed of subsets of particles, each of which performs PS in a different ordering [15]. In DPS, we re-express Eq. 1 as

$$p(X_t|Y_{1:t}) = \sum_{c=1}^{C} \pi_{c,t}\; p_c(X_t|Y_{1:t}) \tag{3}$$

where p_c is a mixture component and $c = 1, ..., C$ is the subset index. PS is performed using a different ordering for each subset to fairly distribute the bias and impoverishment effects between each object. The subsets are then reassembled and evaluated normally.

seq	PF	PS (1 → 2 → 3)	PS (2 → 3 → 1)	PS (3 → 1 → 2)	DPS
1	32	18	40	34	100
2	10	0	12	0	78

Table 1: Tracking success rate for an occluded object for different sampling methods on two meeting room data sequences. For PS, the numbers correspond to different object orderings.

The observation model used in this work consisted of 8-bin color-space (HS) histograms with spatial components [13]. The resulting multi-dimensional histogram consists of a concatenation of 2-D HS histograms, each built from pixels taken from different areas of the head (eyes, mouth, hair, etc) according to a template. The observation likelihood is defined as $p(Y_t|X_t) = \prod_i p(Y_{i,t}|x_{i,t})$, where $Y_{i,t}$ is the image region enclosed by $x_{i,t}$, and each object likelihood is defined as $p(Y_{i,t}|x_{i,t}) \propto e^{-\lambda d_i^2(Y_{i,t})}$ where λ is a hyper-parameter and $d_i(Y_{i,t})$ is the distance based on the Bhattacharyya coefficient between the observation $Y_{i,t}$ and the specific object template histogram.

Results. Head tracking experiments were conducted in the meeting room to test the ability of DPS to overcome impoverishment problems associated with PS. Specifically, DPS and PS were tested for their ability to recover from occlusion (impoverishment hinders this ability) over 50 runs per method, to account for the stochastic nature of the tracker, with $N_S = 200$ particles. Performance is measured by the *success rate* (SR), the percentage of successful runs (a successful run occurs when the tracking estimate overlaps the ground truth throughout the entire sequence). As seen in Table 1, DPS signficantly outperformed both a simple multi-object PF (denoted by PF) and a PS tracker. Some results can be seen in Fig. 2 and [17].

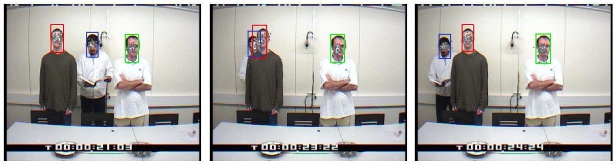

Figure 2: Tracking multiple heads through occlusion with DPS sampling in the multi-sensor meeting room.

Open issues. Some relevant issues currently being pursued include alternative sampling strategies, and handling variable numbers of objects, including automatic initialization.

5. TRACKING HEAD POSE

Challenges. Head pose estimation is often used as a first step for other higher level tasks such as facial expression recognition or gaze direction estimation. In meetings, head pose can be reasonably used as a proxy for gaze (which usually calls for close views), and can thus be useful for determination of visual focus-of-attention and addressees in conversations. Most of the existing work for head tracking and pose estimation defines the task as two sequential and separate problems: the head is tracked, its location is extracted, and the head pose is estimated from the head location. As a consequence, the estimated head pose totally depends on the tracking accuracy. This formulation misses the fact that knowledge about head pose could be used to improve head modeling and thus improve tracking accuracy.

Our approach. We couple head tracking and pose estimation using a mixed-state PF [1]. The state $X_t = (x_t, l_t)$ is a mixed variable. The continuous variable $x = (\mathbf{T}, s)$ specifies the head location and scale. The discrete variable l specifies an element of the head pose exemplars set. The pose at given time is obtained by marginalizing over the spatial configuration part of the state. In the following paragraph, we describe the head pose models, the dynamical model, and the observation model.

Head pose exemplars are learned using the PIE database. A total of N_θ head poses are defined by a pan angle ranging from -90 to 90 degrees discretized with 22.5-degree steps. For each head pose θ, Gaussian and Gabor features are extracted from training images, concatenated into a single feature vector, and clustered with K-means into L_θ clusters $\{e^\theta_l = (e^\theta_{l,j}), l \in \mathcal{L}_\theta\}, |\mathcal{L}_\theta| = L_\theta$. The cluster centers are taken to be the head pose exemplars. The number of elements of each cluster are used to define prior distributions π^θ_l, and the diagonal covariance matrix of the features $\sigma^\theta_l = diag((\sigma^\theta_{l,j}))$ is used to define pose probability models. The pose of an head image is estimated by extracting its feature vector $Y = (Y_j)$, and finding the pose MAP estimate by $p(Y|\theta) = \sum_{l \in \mathcal{L}_\theta} \pi^\theta_l p(Y|l)$, with

$$p(Y|l) = \prod_j \frac{1}{\sigma^\theta_{l,j}} \max(\exp -\frac{1}{2}\left(\frac{Y_j - e^\theta_{l,j}}{\sigma^\theta_{l,j}}\right)^2, T) \qquad (4)$$

where T is a bound introduced to tolerate modeling errors.

The dynamical model is a second order autoregressive process $p(X_t|X_{t-1}, X_{t-2})$. Assuming that the two components x_t and l_t are independent, and that head pose depends only on the previous pose give, the dynamics factorize as $p(x_t|x_{t-1}, x_{t-2})p(l_t|l_{t-1})$.

Finally, the observations are obtained by extracting the features $Y(x)$ from the image region specified by the spatial configuration x. The observation likelihood is given by $p(Y_t|X_t) = p_T(Y_t(x_t)|l_t)$, with p_T defined in Eq. 4.

Results. Head pose estimation was tested on PIE database. The best result was obtained with two exemplars per pose, with a recognition rate of 94.8% while the state-of-the-art obtains around 90% [2]. More details about evaluation can be found in [1]. The joint tracking algorithm was also tested on video sequences from our meeting room. An example with $N_S = 100$ particles is shown in Fig. 3. Tracking and head pose estimation are visually quite satisfactory. Other results can be found to our website [17]. However, in view of the limitations of visual evaluation, and the inaccuracy obtained by manually labeling head pose in real videos, we have recently recorded a set of meetings with four participants, with head pose ground truth produced by a flock-of-birds device. An objective evaluation of our algorithm is in process.

Open issues. The current features are obtained using gray-level information. While our head tracking and pose estimation system works well in general, some problems might occur when the background is highly textured. The use of color information for more robust tracking is under investigation.

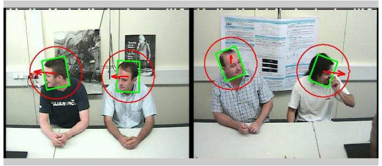

Figure 3: Joint tracking and head pose estimation in meeting room. The green box and red arrow specify the estimated head location and head pose, respectively. The red circle gives information about the pose value; its radius corresponds to 90 degrees. The participants are looking at the room entrance.

6. TRACKING SPEAKERS

Challenges. Sound and visual information are jointly generated when people speak, and provide complementary advantages. Initialization and recovery from failures are tasks for which audio is convenient; precise localization is better suited for vision. In addition to the problems for visual tracking described in previous sections, the challenge on the audio side is to detect individual speaker turns over time. This is a difficult task in spontaneous multi-party speech, since the various speakers often talk for very short durations and overlap significantly [14]. Speaker turns are therefore highly dynamical and often concurrent temporal events.

Our approach. We use an approach in which a person's head is represented by its silhouette in the image plane. In one formulation, the state-space is defined over only one person, the target being the current speaker at each instant, in single- or multi-camera setups [4]. In the second formulation, states are defined as a joint multi-object representation, where both the location and the speaking activity of each participant are tracked [5]. In both cases, we employ mixed-states. In addition to continuous variables for head motion, discrete variables are included to model speaker switching across cameras in the single-object case, and to model the speaking status of each participant in the multi-object case.

Our methodology exploits the complementary features of the AV modalities, taking advantage of the fact that data fusion can be introduced in both the sampling and the measuring stages of a PF. In [4], we asymmetrically handle audio and video. Audio localization information in 3-D space is first estimated by an algorithm that reliably detects speaker changes with low latency, while maintaining good estimation accuracy. Audio and skin-color blob information are then used for prediction, and introduced in the PF via importance sampling, a technique which guides the search process of the PF towards regions of the state space likely to contain the true configurations. Additionally, audio, color, and shape information are jointly used in the observation likelihood. We also use an AV calibration procedure to relate audio estimates in 3-D and visual information in 2-D. The procedure uses easily generated training data, and does not require precise geometric calibration of cameras and microphones. In [5], we have dealt with the dimensionality of the multi-object state space by combining Markov Chain Monte Carlo (MCMC) and PF, which provides efficient sampling in a formalism that is naturally suitable for interaction modeling.

Results. On real data, the audio source localization system provided the direction of the active speaker within a decent but not too precise $(\pm 6^o)$ margin. Range estimation is not reliable. On the other hand, audio source detection is quite precise, with a false alarm rate of only 1.6%, for a false rejection rate of 23.4% (details in [4]). For the single-object case, our tracking framework can initialize and track a moving speaker, and switch between multi-

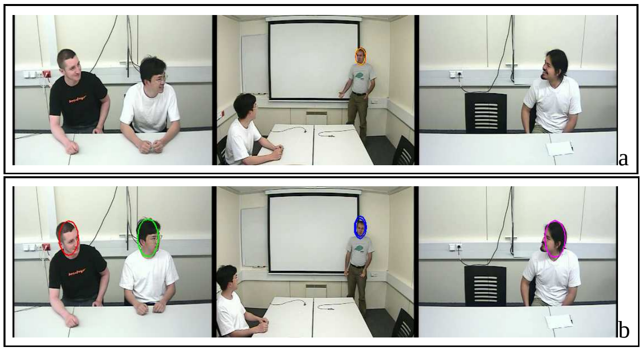

Figure 4: (a) Single-object speaker tracker in the meeting room. The tracker locks onto the speaker. (b) Multi-object speaker tracker. Both location and speaking status (double ellipse if a person speaks) are inferred for each participant.

method	SR	F_x	F_s
PF	78.8	0.85	0.59
MCMC	100.0	0.88	0.75

Table 2: Tracking success rate, and F-measures for location (F_x) and speaking status (F_s), averaged over the four objects in the meeting video sequence (initial 1715 frames). PF denotes a basic PF multi-object tracker. MCMC denotes the approach in [5].

ple people across cameras with low delay, while tolerating moderate visual clutter. An example is shown in Fig. 4(a), for a two-minute sequence, using $N_S = 500$ particles. Given a ground-truth of speaker segments, camera index and speaker head location, an objective evaluation procedure showed that the error on the estimated camera indices is small for the close-view cameras ($< 2\%$), but much larger for the wide-view case (25%), due to the larger distance of the speaker at the whiteboeard compared to the seated participants. The localization error in the image plane also remains small. For the multi-object case, an example using $N_S = 500$ particles is shown in Fig. 4(b). After manual initialization, the four participants are simultaneously tracked, and their speaking status is inferred at each time. An objective evaluation procedure involves the computation for each participant of the success rate measure mentioned in Section 4, and the *F-measure* (which combines precision and recall) for location and speaking status, over a number of runs of the trackers. Results for the first 1715 frames are shown in Table 2, comparing the proposed method with a basic multi-object PF over 20 runs. They show that MCMC sampling outperforms the basic PF in both ability to track and estimation of the speaking status. Other examples can be found in [17].

Open issues. Our audio observation model can already reflect activity from multiple people at the same time [6]. However, it is based on a limiting single-audio-source assumption. We are currently developing truly multi-speaker detection techniques with a sector-based approach [8], and plan to integrate them in the SMC framework. We also plan to improve the multi-object speaker tracker for automatic initialization, and to deal with a multi-camera scenario with overlapping fields of view. Finally, an audio-visual corpus for the localization and tracking tasks has been collected, and its annotation is in progress [7].

7. CONCLUSION

We presented three different algorithms for people tracking in multi-sensor meeting environments, each focusing on a specific task. They all rely on a Bayesian framework implemented via SMC, and produced good results. While the improvement of each algorithm constitutes a research topic in itself, the integration of all of them into a unique process, which we are targeting, raises some important issues. For instance, as meeting participants often look at the current speaker, head orientation and speaker localization are two correlated processes. Hence, jointly performing both tasks could lead to performance gain w.r.t. a sequential system first performing multiple people tracking and speaker identification, and then head pose estimation. However, in practice, the significance of such a gain has to be balanced against other considerations, such as the complexity of an integrated system, and the difficulties in modeling and learning the interactions. These issues also apply to the recognition of other high-level processes, like focus-of-attention, person behavior, or group actions. In these cases, the use of layered approaches might be an appropriate alternative.

Acknowlegments. We thank several of our collegues at IDIAP for technical support and discussions.

8. REFERENCES

[1] S. Ba and J.-M. Odobez. A probabilistic framework for joint head tracking and pose estimation. In *Proc. ICPR*, Cambridge, Aug. 2004.

[2] L. Brown and Y. Tian. A study of coarse head pose estimation. In *IEEE Workshop on Motion and Video Computing*, Orlando, Dec. 2002.

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

[4] 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 ICCV WOMTEC*, Nice, Oct. 2003.

[5] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audio-visual tracking of multiple speakers in meetings. IDIAP Research Report RR-04-66, Dec. 2004.

[6] G. Lathoud, I. McCowan, and J.-M. Odobez. Unsupervised location-based segmentation of multi-party speech. In *Proc. NIST Meeting Recognition Workshop*, Montreal, May 2004.

[7] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez. AV16.3: an audio-visual corpus for speaker localization and tracking. In *Proc. MLMI*, Martigny, Jun. 2004.

[8] G. Lathoud and M. Magimai.-Doss. A sector-based, frequency-domain approach to detection and localization of multiple speakers. In *Proc. IEEE ICASSP*, Philadelphia, Mar. 2005.

[9] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *Proc. IEEE ICCV*, pages 572–578, 1999.

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

[11] D. Moore. The IDIAP Smart Meeting Room. IDIAP Communication 02-07, Nov. 2002.

[12] K.C.H. Parker. Speaking turns in small group interaction: a context sensitive event sequence model. *Journal of Personality and Social Psychology*, 54(6):965–971, 1988.

[13] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-based Probabilistic Tracking. In *Proc. ECCV*, Copenhagen, May 2002.

[14] E. Shriberg, A. Stolcke, and D. Baron. Observations on overlap: findings and implications for automatic processing of multi-party conversation. In *Proc. Eurospeech*, Aalborg, Sep. 2001.

[15] K. Smith and D. Gatica-Perez. Order matters: a distributed sampling method for multi-object tracking. In *Proc. BMVC*, London, Sep. 2004.

[16] 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 ICASSP*, Salt Lake City, May 2001.

[17] `http://www.idiap.ch/mucatar/WIAMIS`.