This chapter appeared in:

Multimodal Signal Processing: Human Interactions in Meetings. S. Renals, H. Bourlard, J. Carletta and A. Popescu-Belis editors. Cambridge University Press, 2012.

Chapter 6

Sampling techniques for audio-visual tracking and head pose estimation

Jean-Marc Odobez (Idiap Research Institute)
Oswald Lanz (Fondazione Bruno Kessler)

6.1 Introduction

Analyzing people behaviors in smart environment using multimodal sensors requires to answer a set of typical questions: *who* are the people, *where* are they, *what* activities are they doing, *when*, with *whom* are they interacting, and *how*. In this view, locating people or their faces and characterizing them (e.g. extracting their body or head orientation) allows to address the first two questions (who and where), and is usually one of the first steps before applying higher level multimodal scene analysis algorithms that address the other questions. In the last ten years, tracking algorithms have experienced considerable progresses, particularly in indoor environment or for specific applications, where they have reached a maturity allowing their deployment in real systems and applications. Nevertheless, there are still several issues that can make the tracking difficult: background clutter, potentially small object size; complex shape, appearance, and motion, and their changes over time or across camera views; inaccurate/rough scene calibration or inconsistent camera calibration between views for 3D tracking; real-time processing requirements. In what follows, we discuss some important aspects of tracking algorithms, and ultimately introduce the remaining of the chapter content.

Scenarios and Set-ups. Scenarios and application needs strongly influence the considered physical environment, and therefore the set-up (where, how many, and what type of sensors are used) and choice of a tracking

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis

method. A first set of scenarios commonly involves the tracking of people in the so-called *smart spaces* [Singh et al., 2006]. These are indoor environments -homes, offices, classrooms- equipped with multiple cameras located on room sides, along with microphone-arrays and potentially networked pervasive devices that can perceive ongoing human activities and respond to them. These settings usually involve the tracking of few people. Cameras usually provide good image quality, and people sizes in the images are relatively high and of the same value across camera views. In this context, robust and accurate tracking results have been demonstrated, e.g. [Bernardin et al., 2006, Fleuret et al., 2008], and current goals consists of improving tracking robustness under higher crowding levels, recovering the pose of objects in addition to their localization, exploiting other modalities such as audio [Bernardin and Stiefelhagen, 2007], and characterizing people activities.

Meetings and teleconferences -whether for professional of for families- are other specific but nonetheless important scenarios that require tracking technologies, as highlighted in other chapters of this book in the meeting case. In these scenarios, localizing people in the room is less the issue than assessing their presence and image position, and understanding their activities and interactions. They therefore rely on different and potentially lighter settings than the smart room cases. Cameras (and microphones) are often located on a table, displays (e.g. showing remote participants) or walls, and focus at people upper body and faces. Good performances have been reported as well in this case, although tracking the head whatever under any pose can remain challenging. For instance, in meetings, people can look down for quite a long time, resulting in head tilts that face detectors can not cope with, and thus require robust tracking techniques to be handled.

Tracking problem formulation. Several approaches can be used to formulate the tracking problem. In a simple case, tracking can be done by detecting objects at each frame and matching them across time using motion heuristics to constrain the correspondence problem. Other deterministic approaches can also be exploited, such as the popular Mean-Shift approach that is quite appropriate for tracking faces due to their distinctive skin-colored faces. However, in general, such methods have difficulties with short term ambiguities and for performing robust long term tracking.

In past years, Bayesian state-space formulations have been shown to be very successful to address the single or multi-person tracking problem [Isard and MacCormick, 2001, Khan et al., 2005, Smith et al., 2005, Yao and Odobez, 2008a]. As will be described in the next section, it offers a principled and intuitive way of introducing dependencies between variables of the state and observation spaces, and general tools for inference and model parameter learning. Note that while the probabilistic tracking framework is appealing, it does not solve all the problems by itself. For instance, due

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012

to the curse of dimension in multi-object tracking, solving the inference problem is not a straightforward issue. The use of a plain particle filter will quickly fail when more than 3 or 4 people need to be tracked. In recent years, several tools such as reversible-jump MCMC stochastic optimization have been introduced and shown to be more effective at handling the large dimensional state.

Chapter organization. In this chapter, our goal will be to highlight some important aspects of tracking that we believe have been shown to be successful in past years, as well as current limitations, focusing on the two set-ups of interest (smart rooms, meetings). The list of reviewed works will therefore not be exhaustive, both in breadth and in depth, but illustrative of approaches and of developments that have been performed in the context of the EU AMI and related projects.

The next section reviews the main principles and elements of Bayesian tracking and of particle filters which has been identified as one of the main successful frameworks for tracking. We then review and illustrate with more details multi-person tracking techniques for smart environments, followed by a section more dedicated to face and head pose tracking, which is the main task in conference and meeting scenarios. The complementary exploitation of audio information for these tasks is described in a separate section before the conclusion.

6.2 State-space Bayesian tracking

The Bayesian formulation of the tracking problem is well known. Denoting the hidden state representing the object configuration at time t by $\mathbf{x}_t$ and the observation extracted from the image by $\mathbf{z}_t$, the objective is to estimate the filtering distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ of the state $\mathbf{x}_t$ given the sequence of all the observations $\mathbf{z}_{1:t} = (\mathbf{z}_1, \ldots, \mathbf{z}_t)$ up to the current time. In order to solve the problem recursively, standard assumptions are usually made: the state follows a first order Markovian process, i.e. $p(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})$, and the observations are conditionally independent given the state sequence, i.e. $p(\mathbf{z}_t|\mathbf{x}_{1:t}, \mathbf{z}_{1:t-1}) = p(\mathbf{z}_t|\mathbf{x}_t)$. Bayesian tracking then amounts to solving the following prediction and update equations:

$$p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}) \mathrm{d}\mathbf{x}_{t-1} \quad (6.1)$$

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{z}_{1:t-1}). \quad (6.2)$$

which involve two important terms: the process dynamics $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, and the data likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$. Methods to solve the above equations depend on how these terms are modeled. [Arulampalam et al., 2002] provides a good and detailed review of these. Below, we only discuss two of them.

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis

6.2.1 The Kalman filter

The main assumption of this model is linearity and Gaussianity: the dynamics is given by $\mathbf{x}_t = f(\mathbf{x}_{t-1}, \mathbf{v}_t) = \mathbf{A}_t\mathbf{x}_{t-1} + \mathbf{v}_t$ and similarly for the observation model: $\mathbf{z}_t = g(\mathbf{x}_t, \mathbf{w}_t) = \mathbf{C}_t\mathbf{x}_t + \mathbf{w}_t$, where $\mathbf{v}_t$ and $\mathbf{w}_t$ are zero-mean Gaussian noise with covariances $\Sigma^{\mathbf{v}_t}$ and $\Sigma^{\mathbf{w}_t}$, respectively. Under these assumptions, all probability distributions of the model (the joint and marginals) are known to be Gaussian, including the predictive and filtering ones. The main advantages are that a closed-form solution to the above equations can be found, with a principled way to fuse multiple observations, e.g. from different modalities, and account for both process and measurement uncertainties at each time step. The model suffers from important drawbacks however: it has difficulties in handling more complex dynamics due to non-linearity or state-dependencies (i.e. when a state component at time t depends on another state component at the same time). But in vision, the main issue of the KF is the measurement model. First the KF modeling requires the tracker to extract observations of a similar nature to the state like localization observations. That is, if one wants to exploit *image observations* and therefore powerful object representations *directly* in the tracking framework, such as color histograms or shape features, defining $\mathbf{C}_t$ or more generally the g measurement function is very complex if not impossible. Secondly, due to clutter and local ambiguities, likelihood distributions (and hence the predicted and filtering distributions) in vision are often multimodal, something that is not accounted for by the Gaussian assumption.

6.2.2 Monte-Carlo methods

In non-Gaussian and non linear cases, the recursive equations can be solved using sampling approaches, also broadly known as Particle Filter (PF). The idea behind the PF approach consists in representing the filtering distribution in a non parametric way, using a set of N_s weighted samples (particles) $\{\mathbf{x}_t^n, w_t^n, n = 1, ..., N_s\}$, and updating this representation when new data arrives. The standard PF relies on the importance sampling principle. Given the particle set of the previous time step, $\{\mathbf{x}_{t-1}^n, w_{t-1}^n, n = 1, ..., N_s\}$, configurations of the current step are drawn from a proposal distribution $\mathbf{x}_t^i \sim q(\mathbf{x}_t|\mathbf{x}_{t-1}^i, \mathbf{z}_t)$. The weights are then updated as $w_t^i \propto w_{t-1}^i \frac{p(\mathbf{z}_t|\mathbf{x}_t^i)p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)}{q(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i, \mathbf{z}_t)}$. In addition, to avoid sample impoverishment, a resampling step need to be regularly applied [Arulampalam et al., 2002].

The PF frameworks offers several advantages: handling of non-linear, non-Gaussian, and multimodal distributions; easy accounting of probabilistic relationships and dependencies between variable, much more opportunity and diversity for likelihood modeling, ease of programming. Four main element enter in the definition of a PF tracker. We discuss below general

issues regarding them, while subsequent sections will provide more specific examples.

The state space. The state space defines the object parameters one wishes to recover. In its simplest case, one might only be interested in the location of an object in the 2D image plane or in the 3D space. However, depending on the scenario, one is often interested in recovering object-centric information, such as its size, orientation, or pose. In general, the selection of an adequate state space is a compromise between two goals: on one hand, the state space should provide the richest information to further higher level analysis modules, and be precise enough so as to model as well as possible the information in the image and video. In other words (and even if such information is not requested by the application), adding relevant auxiliary variables in the state space that simplifies the modeling of other components (dynamics, appearance) is often useful [Perez and Vermaak, 2005]. On the other hand, the state has to remain simple enough and appropriate to the quality level of the data (i.e. the impact of changing the state variable value should be observable on the data) in order to obtain reliable estimates and keep the computation time low.

The dynamical model. Defined by $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, it provides the prior on the state sequence and governs the temporal evolution of the state. Often, for simplicity, the dynamics of each state components are defined independently, e.g. using auto-regressive models for continuous variables. However, simple but efficient dependencies can be specified in the dynamics. The graphical model in Fig. 6.6a shows an example related to head pose tracking. Another typical example is switching dynamical models, where the state $\mathbf{x}_t = (\mathbf{s_t}, \mathbf{a_t})$ is characterized by object parameters $\mathbf{s_t}$ (e.g. a person's location) and an activity index $\mathbf{a_t}$ (e.g. the person is either static or moves). Then we can have $p(\mathbf{x}_t|\mathbf{x}_{t-1}) = p(\mathbf{s_t}|\mathbf{s_{t-1}}, \mathbf{a_t})\mathbf{p}(\mathbf{a_t}|\mathbf{a_{t-1}})$, where in the first term the activity $\mathbf{a_t}$ controls which dynamics is applied to $\mathbf{s_t}$, and the second term models the sequence of activities (and thus implicitely the transitions between activities that will introduce discontinuities in the dynamics of $\mathbf{s_t}$).

The likelihood model $p(\mathbf{z}_t|\mathbf{x}_t)$ measures the adequacy of data given the proposed configuration of the tracked object. This is probably the most important term for the tracking success. Ideally, the likelihood should always have its maximum at the right state value. It should be broad so that the closer a particles is from the true state, the higher its likelihood, and peaky enough to separate the object from the clutter and provide a precise localization. [Deutscher et al., 2000] addressed this issue by proposing an annealed PF, where the likelihood is progressively changed from a broad distribution into the final peaky one by controlling a temperature coefficient as in simulated annealing optimization. Broadly speaking, likelihood based on color-histograms or patches are often broad (they are robust to imprecise

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012

localization or scale), while those based on contours are more peaky.

Proposal and sampling scheme. The proposal $q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ specifies the sampling mechanism responsible for exploring the state space in an intelligent way, i.e. it has to propose new state configurations in high probability regions of the filtering distribution. In the CONDENSATION seminal paper [Isard and Blake, 1998], the Bootstrap filter was used, in which the proposal is defined as the state dynamics, which simplifies the particle weight expression (weights are directly proportional to the likelihood). However, despite the representation with samples, which can handle temporally localized ambiguities, the Bootstrap PF is sensitive to drift and can not recover from tracking failures. Ways to overcome this issue as well as more advanced sampling strategies will be introduced in the next Sections.

6.3 Person Tracking in Rooms

In the following we give an overview of main Bayesian methods proposed in the literature, grouping them into two categories according to whether they represent and sense objects in 2D or 3D space. Some were specifically designed to operate in a Smart Room type environment while others address tracking in a general setting and are important to get a more comprehensive overview of the problem. We will zoom into some technical detail of one representative paper per approach, will explicitly refer to the specific aspects of applying them in a Smart Room settings.

6.3.1 Specific issues

Object tracking is a well studied topic in Computer Vision (see e.g. [Yilmaz et al., 2006] for a survey) and proposed solutions differ largely in the type of environment they are designed to operate on (indoor vs. outdoor, single camera vs. multi-camera) and adopted methodology (model-based vs. data-driven, distributed vs. centralized). When the targeted application requires the tracking of a number of interacting people in an indoor setting, the main challenges posed to video analysis can be summarized as follows:

High variability in pose and appearance. There is a great amount of variability in the way people appear in images. This is true even when observing a single subject over a short period of time, from a fixed viewpoint, in a controlled environment. It is therefore difficult to characterize a human subject in the image domain and, as a consequence of that, both detecting such a target and subsequently re-localizing it from frame to frame become ambiguous processes.

Occlusions. Interacting people are usually located close to each other, and they may remain at the same location, e.g. as long as a conversation goes on among them. From a tracking perspective this means that people's

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012

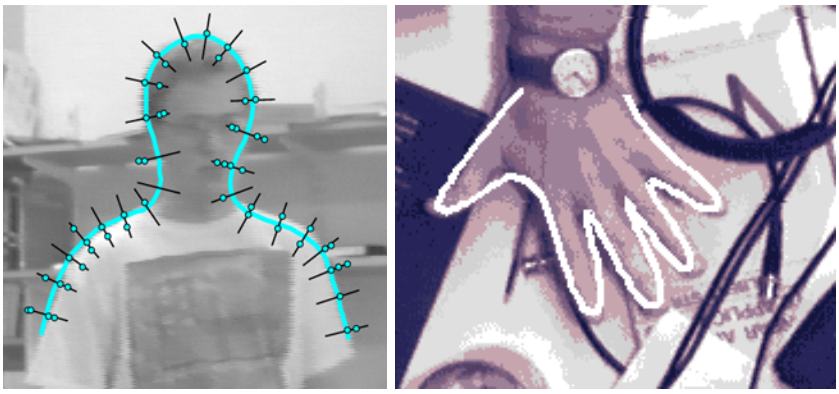

Figure 6.1: B-spline head-shoulder (a) and articulated hand (b) implicit shape models to track in 2D on contours with CONDENSATION. (Images taken from *Active Contours*. Courtesy by Andrew Blake and Michael Isard.)

bodies may be rendered partially or even completely occluded into one or more cameras, simultaneously, for an unpredictable amount of time. The likelihood of occurrence of such events is high in our envisaged scenarios, thus reliable tracking can only be achieved by methods that explicitly reason about occlusions.

Clutter and uneven illumination. It is not always possible to describe the appearance of an object with discriminative features. This is of particular concern when tracking people in populated scenes, where due to incomplete modeling the visual signature of one person may match that of another, or where a similar pattern may appear in the background. Additional difficulties arise during occlusion, where only a partial measurement of it, if any, may be available. In environments where the lighting conditions cannot be controlled (like a windowed scene) color based measurements may be distorted by local variations in illumination. These sources of uncertainty are likely to cause failures if their effects are not properly handled (technically speaking, major failure modes are *drift* and *coalescence*).

In summary, our envisaged scenario encompasses factors out of our control that unavoidably induce intrinsic uncertainty in the measurements upon which a tracking process is instantiated. This motivates our choice to focus on Bayesian sampling methods for tracking which, either implicitly or explicitly, propagate estimates in form of distributions, and are thus able to represent and maintain uncertainty which is inherently present in the measurements.

6.3.2 Tracking in the image plane

Contour tracking. The particle filter was first proposed as a tracking

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012

framework in the late 90ies. In their seminal work [Isard and Blake, 1998] proposed a sampling scheme for contour tracking, namely CONDENSATION, which was found to be outstanding (w.r.t. state-of-the-art at that time) in coping with the complexity of tracking agile motion in clutter. Authors proposed parametric B-splines to model the head-shoulder shape of a person when captured from a near-horizontal view (see Fig. 6.1). The state $\mathbf{x}_t$ of a person is hereby encoded by the B-spline parameter vector, thus an implicit representation was chosen. The adopted likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$ measures the degree of overlap between the spline at $\mathbf{x}_t$ and image edges $\mathbf{z}_t$ detected along a set of spline normals (the probes), assuming that the edge pixels are normally distributed along the probe normals. To account for missing edges (e.g. due to low contrast w.r.t. background) an outlier probability term was added to the model. State evolution was modeled as a linear 2nd order process whose parameters were learned from labeled data.

Color based tracking. Contour based methods attempt to re-localize a target using information only about its shape, thus with cues that may not allow to discriminate between subjects when tracking multiple people (the key factor causing *coalescence*). To exploit a more discriminative characterization of a subject, a color tracker can be designed, whose state $\mathbf{x}_t$ is typically chosen to be the center, the apparent motion, and the scale and aspect ratio of the bounding box or ellipse enclosing the object's extent in the image [Perez et al., 2002]. The dynamical model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is a Gaussian with diagonal covariance matrix whose entries define the amount of change that each component of $\mathbf{x}_t$ may undergo from one frame to the next. The likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$ is defined in terms of a distance between color features (e.g. histograms) extracted from the bounding box at $\mathbf{x}_t$ and a reference model of the target. The performance depends strongly on how well the reference model matches the target's appearance at its current state. To handle pose changes and uneven illumination, either additional cues invariant to those factors such as motion, depth, and sound are used [Perez et al., 2004, Badrinarayanan et al., 2007], or the external part of the object (the local background) is included in the likelihood definition [Lehuger et al., 2006]. To avoid *drift*, i.e. adaptation to background clutter and subsequent locking, the weights associated to the fusion scheme must be carefully selected and, ideally, updated on-line from data evidence [Badrinarayanan et al., 2007].

6.3.3 Tracking in 3D space with calibrated camera(s)

A multitude of 2D approaches find their analogs in 3D methods. An important difference is that with a 3D approach the correlation between measurements from different cameras can be explicitly modeled in the likelihood and conditioned to a unique state and appearance characterization of the object, which is then more appropriate as it simplifies multi camera integration and presents several additional advantages. Parameter setting, in most cases,

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012

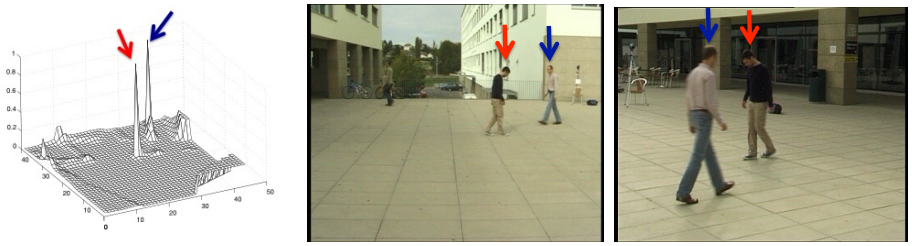

Figure 6.2: The ground plane POM computed with a variational minimization technique from blobs extracted by background subtraction on the images to the right (only two of the three images that where used are shown) [Fleuret et al., 2008].

will have a physical meaning (e.g. the standard height of a person, the average speed of a walking person [Yao and Odobez, 2008a]). Similarly, prior information about the state values will be more easy to specify, as they are somewhat 'built-in': for instance, according to the 3D position, we automatically know what should be the size of a person in the image plane. Finally, occlusion reasoning -when tracking multiple people- is simplified when using the 3D position. It is important to note that some approaches perform early fusion [Fleuret et al., 2008, Canton-Ferrer et al., 2008] to obtain 3D observations and track in a 3D appearance space and thus absolutely need multiple cameras, while others interpret images as projections of the 3D world [Isard and MacCormick, 2001, Lanz, 2006, Yao and Odobez, 2008b] and can track multiple targets even with a single calibrated camera.

Tracking on voxels. In indoor environments one can often assume that the background is stable, and that a robust model of it can be initialized and updated over time, either independently on each camera, or more robustly by exploiting redundancy in a calibrated setting [Tzevanidis and Argyros, 2011]. A discrete occupancy representation of the 3D space can then be generated at each iteration via *space carving*: a voxel (a element of the 3D grid representing the scene) is labelled as foreground by checking the consistency of its projection with the model across all view's. In [Canton-Ferrer et al., 2008] a PF in the voxel space propagates particles with a likelihood measuring the fraction of foreground voxels in its 26-connected neighborhood. Although this way the particles do not really represent the state of a target but rather volume elements that may be associated to its body, their expectation can be used to approximate the center of mass of an isolated object. To avoid merging when multiple targets are being tracked with independent filters a blocking scheme is used to down-weight particles ending up in the envelop (an ellipsoid) of the estimated previous state of another target. Voxel based tracking is effective with many cameras and when a top-down view of the scene is available [Stiefelhagen et al., 2008].

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012

Tracking by projection on the ground plane. Voxel based techniques rely on a discretized representation of the 3D scene, which reach their limits when the monitored space becomes large (explosion in the number of voxels). Since the movements of people are constrained to the ground plane, a more effective approach is representing and estimating the state of people on the 2D reference plane. The Probabilistic Occupancy Map (POM) represents a virtual top-down view of the scene where each entry in the POM corresponds to the probability of a location in the 2D plane being occupied. In [Fleuret et al., 2008] such probabilities are inferred from background subtraction images computed from multiple synchronized views (Fig. 6.2). By representing humans as simple rectangles to create synthetic ideal foreground images we can evaluate if people are at a given location. Such probabilities of occupancy are approximated at every location as the marginals of a product law minimizing the Kullback-Leibler divergence from the true conditional posterior distribution. This allows to evaluate the probabilities of occupancy at every ground location as the fixed point of a large system of equations, avoiding combinatorial explosion (curse of dimension) in the number of targets and utilized views. POMs are powerful representations for detection, and can be combined with a color and motion model to compute the trajectories of a variable number of people going through complex occlusions, by searching for individual trajectories using the Viterbi algorithm over frame batches of a few seconds, and using heuristics to decide on the order in which such trajectories are processed.

Integrating detection and target interaction in the Bayesian model. Visual interactions among targets, scene structure and the sensing geometry (occlusions, shadows, reflections, etc.) induce dependencies in the appearance of the targets. In addition, there are physical constraints (two targets cannot occupy the same location in space) and behavior patterns (people look to each other during a conversation, or move in groups) that relate the state of a target to that of the others. The processes governing such interactions are known and can be modeled in the Bayesian framework when the joint state of the scene ($\mathbf{x}_t$ is a multi-dimensional vector with an entry for each of the interacting entities) is being tracked. Such interactions can be implicitly exploited when defining the likelihood term as done in the BraMBLe system [Isard and MacCormick, 2001]. There, generalized-cylinder shape model and perspective projection are used to map the joint multi-target ground plane state into an image partition (Fig. 6.3) that allows to define an occlusion robust likelihood based on learned color features in which object states 'interact' to best explain the data. [Khan et al., 2005] use a more direct approach and define a dynamical model which includes an explicit interaction term (modeled through a Markov Random Field, MRF) to enforce the spatial exclusion principle within an efficient sampling scheme.

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012

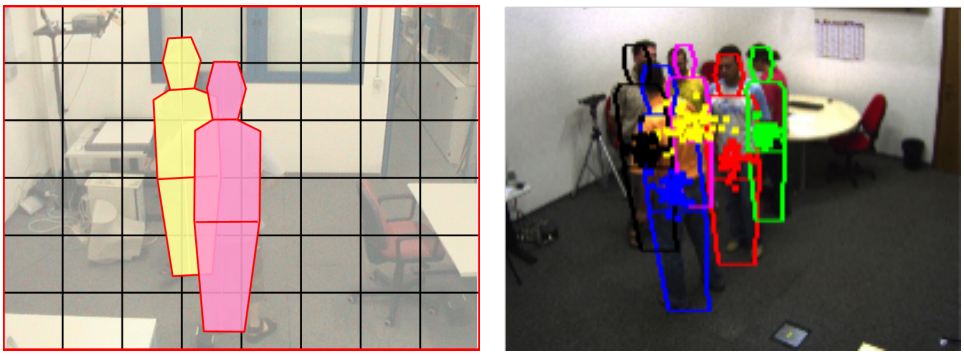

Figure 6.3: Image partition induced by the projection of a joint 3D multi-target hypothesis using part-based generalized-cylinder shape model [Isard and MacCormick, 2001, Yao and Odobez, 2008b], and real time estimates with HJS-PF implementing a color based likelihood built over it [Lanz, 2006], using one calibrated camera.

6.3.4 Multi-object tracking inference

While Bayesian tracking in the joint space is a powerful framework, the computational burden induced by the curse of dimension (see Fig. 6.4) is a major concern when it comes to implementation. Particle filtering (i.i.d. sampling) attempts to populate the full support of the proposal density, whose volume increases exponentially with the dimension of the state space, thus requiring the number of particles to increase exponentially as well. To address this issue, MCMC methods attempt to sample the state space more intelligently, by generating particles in a sequence that is controlled by a Markov chain (MC). MCMC is most effective if the chain peaks towards the modes of the posterior which may be obtained by including the current measurement in the design of the chain. [Yao and Odobez, 2008b] define the chain in such a way by designing MC moves from a mixture with one component integrating the output of a detector and the other emulating blind propagation to account for the case where detection has not succeeded or is not reliable (e.g. during occlusion, or in the presence of clutter). To track a variable number of targets, where joint particles of different dimensions have to compete due to uncertainty in the detection process (i.e. the dimension of the state space is itself a random variable), the RJ-MCMC framework is adopted [Smith et al., 2005, Khan et al., 2005, Yao and Odobez, 2008b]. Additional moves are implemented (i.e. the MC chain is expanded) to allow for track initialization (a *birth* move, driven by detection on locations that are not covered by active tracks), track termination (a *dead* move, upon thresholding the likelihood), and identity exchange (a *swap* move, where the appearance models of targets are exchanged). Gibbs sampling [Hue et al., 2002] is another technique for sampling in high-dimensional space.

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012

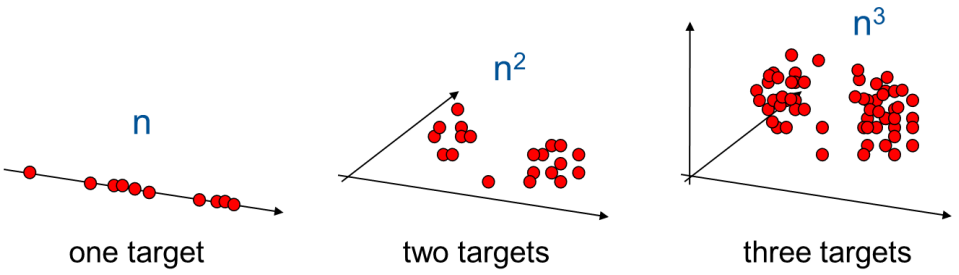

Figure 6.4: The curse of dimension: independent sampling (particle filtering) in the joint space requires the number of particles increase exponentially with the dimension of the state space, i.e. with the number of tracked targets.

Approximating the Bayesian model. To avoid the curse of dimension inherent in the joint formulation, the underlying Bayesian model can be simplified. In the Mean Field approach [Medrano et al., 2009] a factored representation $\prod_k q(\mathbf{x}_t^k|\mathbf{z}_{1:t})$ of the multi-target posterior is updated over time by iterating over a set of fixed point equations minimizing the approximation error introduced by the factored representation (its KL divergence to the joint posterior). This leads to efficient inference over a less complex but much flexible representations of the posterior if dependencies are of generic form but defined pairwise, an assumption that in general does not hold for occlusions.

A factored representation of the posterior is also propagated in the Hybrid Joint Separable (HJS) tracking framework of [Lanz, 2006]. Differently to Mean Field, the assumptions here are on the q's which are chosen to be the marginals of the joint multi-target posterior (shown to be the a priori optimal choice for q in the KL divergence sense). While tracking with the HJS model does, per-se, not induce any savings (explicit marginalization with generic dynamical and likelihood model is still exponential), it is shown that with a joint likelihood implementing the occlusion process (Fig. 6.3) the marginals can be updated with quadratic upper bound in the number of targets, without the curse of dimension, leading to a framework that scales. To do so, HJS-PF exploits occlusion dependencies to avoid exponential blow-up in the likelihood update (i.e. explicit marginalization) and is therefore specifically designed to handle visual interactions. MCMC sampling in the joint space, on the other hand, can be applied to tracking under generic interactions in the state space. Note however that it is possible to embed, via Belief Propagation, MRF dynamics in the HJS model to effectively account for pair-wise state-space dependencies and introduce, for instance, a model of social behavior for each target that is learned from tracks collected during natural interactions [Zen et al., 2010].

To further alleviate the computational burden in multi target tracking

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Editors. Cambridge University Press, 2012.

the number of particles can be adapted to tracking uncertainty, which may vary significantly over time due to occlusion, clutter, illumination, etc. In [Lanz, 2007] an information theoretic rule is derived which uses entropy estimation to decide on-the-fly on how many particles are needed to maintain uncertainty on the estimates of the chosen representation (on the joint, its factors, or on each marginal independently). This way the trade-off between robustness and efficiency is self managed by the multi target PF in a consistent manner.

6.4 Head tracking and pose estimation

Due to the crucial role that faces play in human communication, the real-time visual detection and tracking of faces and estimation of their head pose has been a topic of particular interest in meeting and video-conferencing applications or in the analysis of social interaction. In the following, we present some techniques that have been used for these tasks, introducing first simpler head tracking algorithms, and then focus on approaches that also address head pose estimation.

6.4.1 Head tracking

For many applications of interest, head tracking can be conducted in the 2D plane. Hence the Bayesian techniques described in Section 6.3.2 relying either on shape or color histogram information have been widely applied for this task. Below we complement these on three specific points: skin color detection, tracking-by-detection, tracking failure detection.

Skin detection. Locating skin-coloured regions is an obvious approach when dealing with faces, and is useful not only for face detection and tracking due to its low computation cost, but also for head pose estimation. [Kakumanu et al., 2007] present a recent and comprehensive survey of the field. Much of the existing literature on skin colour modelling is about building a *general* colour likelihood model — i.e. a model across all possible ethnicities and illumination conditions. However, such a general model can still be distracted by objects that are approximately skin-toned, like wood and T-shirts. There are two main keys for improving skin detection. The first one is to build a color model of the background, allowing to compute *likelihood-ratios* thus avoiding the problem of setting a skin likelihood threshold. The second one is to perform automatic adaptation of the colour model, e.g. by updating some model parameters (e.g. of a Gaussian). The selection of appropriate data (pixels) for adaptation is the critical point. For a tracking task, this can be done recursively by using the output of the tracker at the previous step. However, this is subject to drift issues, e.g. in case of track loss or if the person does not face the camera. A better

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012.

approach consists of using side information independent of color. This can be achieved by fusing skin with shape information for instance, or, more robustly, by relying on prior on the skin location inside the bounding-box returned by a face detector.

Tracking by detection. It is commonly believed that face tracking can be solved using a face detector. However, despite much progress on multi-view face detection, even in "simple" scenarios where people predominantly look towards the camera 30 to 40% of faces are missed [Duffner and Odobez, 2011]. Unfortunately, the missed detections do not happen at random time, since they are often due to common head poses that people naturally take to look at other people or to look down (at a table, or if they are bored) and can last for long periods. In practice, this means that face detection algorithms have to be complemented by robust tracking approaches; not only to interpolate detection results or filter out spurious detection, but also to allow head localisation over extended periods of time.

There are two principled ways (that can be combined) in which detections $\mathbf{z}_t^{det,j}$ can be used in a PF tracker. The first one is to consider them as observations and hence define an appropriate likelihood term that typically $p(\mathbf{z}_t^{det}|\mathbf{x}_t) \propto \exp\left(-\|\mathbf{x}_t - \mathbf{z}_t^{det,clo}\|\right)$) drives the particles towards their closest detection and thus prevents drift. The second one is to use them in the proposal, by defining it as a mixture according to:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) \propto \beta p(\mathbf{x}_t|\mathbf{x}_{t-1}) + (1-\beta)\sum_{j=1}^{N_{det}} p(\mathbf{x}_t|\mathbf{z}_t^{det,j}) \tag{6.3}$$

with the first term favoring temporal continuity, and the second favoring automatic (re-)initialization. Note that particles sampled around detections that are unlikely according to the dynamics will receive a very low weight (cf the weight update rule in Section 6.2.2), and will be discarded quickly with high probably. Handling these particles which are useless result in a loss of computation resources. This can be avoided by allowing in the dynamical term random jumps in the whole space with a low (but non zero) probability, or by only considering the closest detections for sampling if such jumps are not desirable.

Track creation and failure detection. Few works address these issues, although they are essential for real system applications. It is often assumed that a face detector is used for initialization, but how to rely on a face detector? When a *high* detection threshold is used, there is a risk of missing an early track initialisation, while with a *low* threshold false track alarms are likely to occur. Even fewer works address track termination. Indeed, how do we know that a tracker is doing fine or that there is a failure? This is an important issue in practice, since a false failure detection may mean losing a person track for a long period until the detector finds the face again.

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Editors. Cambridge University Press, 2012

Most algorithms work recursively and assess tracking failure from the (sudden) drop of likelihood measures but these are not always that easy to control in practice. Principled methods like the RJ-MCMC methodology described in 6.3.3 exist to integrate track creation and termination, but usually rely on appropriate global scene likelihood models that are difficult to build in multi-face tracking. [Duffner and Odobez, 2011] show that it is simpler and more efficient to address this issue by designing a side tracking failure detector. Such a detector can rely on multiple features characterizing the tracker status, some of which would be difficult to formally integrate in the tracking framework: likelihood, estimated state mean and spread, as well as observations about abrupt change detection in these values. In addition, when the camera is static, the above method automatically learns over time the usual face locations or person behavior, an information that greatly helps in improving the failure detection accuracy.

6.4.2 Joint head tracking and pose estimation

Broadly speaking, head pose tracking algorithms differentiate themselves according to whether or not the tracking and the pose estimation are conducted jointly. In the first case, a generic tracker is used to locate the head, and then features extracted at this location are used to estimate the pose using any regression or classification tools. For example, [Stiefelhagen et al., 2002] used neural networks for such a task. Decoupling the tracking and pose estimation results in a computational cost reduction. However, since pose estimation is very sensitive to head localization as shown in many studies, head pose results are highly dependent on the tracking accuracy.

A better alternative consists in modeling tracking and pose recognition as two paired tasks in a single framework. In this way the tracking robustness is improved by defining a pose-dependent observation model while the pose estimation accuracy is increased due to a better localization of the target. This is the approach taken by [Lozano and Otsuka, 2009]. They adapt a generic 3D Active Shape model (ASM) to a specific face in the first frame, learn the corresponding texture, and then track the resulting 3D template with a PF. This methods require that high enough resolution images are used. In addition, tracking failures often occur when the pose reaches profile views, due to the small visual area covered by the face and the large uncertainty in pose estimation that results from this. In such cases, tracking only resumes when the face is detected. These are common issues shared with approaches that rely on the tracking of facial features.

Appearance-based head pose tracking using a Rao-Blackwellized PF (RBPF). To still benefit from the joint location and pose tracking in low to mid-resolution face images and achieve more robust and continuous tracking, pose specific appearance models can be exploited. More precisely, in our work, the state-space comprised continuous parameters allowing for

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012.

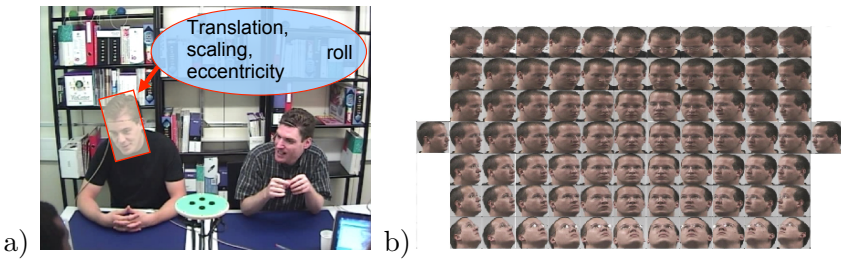

Figure 6.5: Mixed continuous and discrete head state space [Ba and Odobez, 2005]. Continuous parameters (S_t for position, box size and eccentricity, γ_t for in plane-rotation) specify where is the head in the image plane, while a discrete index l_t denotes the out-of-plane head orientation.

the localization of the head in the image plane (including the in-plane pose angle γ_t), and an index l_t denoting one of the discretized out-of-plane head poses, as illustrated in Fig. 6.5. Then, appearance models were built for each possible value of l. Usually, two types of observations are used: skin features, which provide little pose information but are important for tracking robustness; and texture features that are more discriminative with respect to pose. For instance, as illustrated in Fig. 6.6, in [Ricci and Odobez, 2009] we relied on skin binary masks and on Histogram of Oriented Gradients (HOG) features to allow for real-time processing, and on a large-margin approach to learn the pose specific likelihood models.

From the inference viewpoint, a RBPF method was proposed in [Ba and Odobez, 2005]. It is an approach that can be applied when the filtering pdf of some state components can be computed exactly given the samples of the remaining components. In our mixed-state approach, this is the case of the discrete head pose labels l_t: given the sequence of head positions, the inference of l_t can be performed as with a standard HMM. In other words, the sample representation of the filtering pdf is:

$$\text{In PF: } \{S^i_{1:t}, \gamma^i_{1:t}, l^i_{1:t}, w^i_t\}_{i=1}^{N_s} \text{ and in RBPF: } \{S^i_{1:t}, \gamma^i_{1:t}, \pi^i_t(l_t), \tilde{w}^i_t\}_{i=1}^{N_s} \text{ with}$$

$$w^i_t \propto p(S^i_{1:t}, \gamma^i_{1:t}, l^i_{1:t}|\mathbf{z}_{1:t}), \tilde{w}^i_t \propto p(S^i_{1:t}, \gamma^i_{1:t}|\mathbf{z}_{1:t}), \pi^i_t(l_t) = p(l_t|S^i_{1:t}, \gamma^i_{1:t}, \mathbf{z}_{1:t}).$$

These highlights the main differences between the PF and RBPF. In the PF weight w^i_t, the probability of a sample location is always tied to a given sample pose, whereas in RPBF, the probability $\tilde{w}^i_t$ of the location *whatever the pose* is estimated: the pose component has been marginalized in the RB process, and its distribution w.r.t. sample i is maintained in $\pi^i_t(l_t)$. In concrete terms, this marginalization will help in tracking rapid head pose changes by testing all poses at each time step, not only those that are the most likely according to the dynamics, and fewer samples will be needed (e.g. 50 instead of 200 in [Ba and Odobez, 2005]) for the same performance[1].

[1]Note however that the computation per sample is higher for the RBPF.

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012

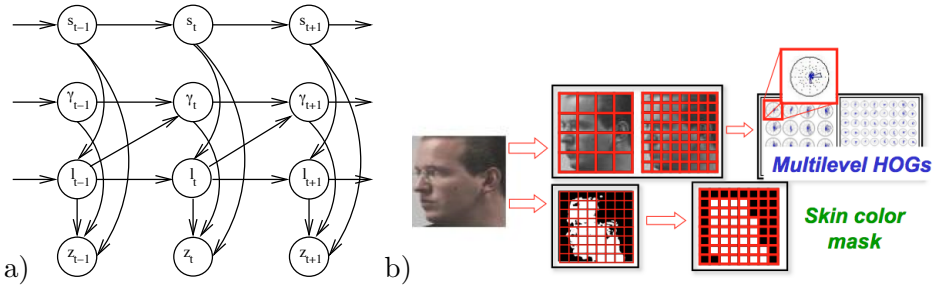

Figure 6.6: Graphical model of [Ba and Odobez, 2005]. Example of features for head representation [Ricci and Odobez, 2009].

6.4.3 Head pose estimation in smart rooms

Head pose estimation in smart rooms or open spaces is very challenging. Due to the size of the space covered by cameras field-of-view, heads and faces are often seen with low resolution. In addition, people usually have more freedom to move so that less prior on the pose can be exploited: heads seen from the side or the back as well as down-looking viewpoints are more common. This differs from more specific applications (e.g. HCI, meetings) in which cameras can be placed to face people's expected gaze directions. Nevertheless, despite these difficulties, in order to move beyond position-based activity analysis, and due to its importance for behavior understanding, gaze and its head pose surrogate are becoming a new important research topic.

Several probabilistic methods have been proposed to address this issue, several of which have been evaluated in the Clear07 contest (`http://isl.ira.uka.de/clear07/`). The method of [Ba and Odobez, 2005] has been extended to the multi-camera smart room scenario Ba and Odobez [2007]. Independent 2D joint location and pose head trackers are run on each view (although tracking in the 3D space might be better), and the pose estimates from the different views are fused using the amount of skin pixels in the head region as reliability factor (pose estimators are more accurate for frontal than profile or back faces). A similar scheme is used by [Segura et al., 2007], but with an approach directly embedded in the 3D space and relying on a multi-camera set-up. There, given a person position, the head localisation is further refined by fitting an ellipse to voxel information, and the positions of the skin blobs within the head region in each view are used to infer the head pose. The approach of [Lanz and Brunelli, 2006] uses a similar color-based approach, where measured color histograms are compared to protocol-based initialized histogram templates sensitive to head orientation to evaluate the data likelihood. In addition, the measured histograms are collected in an image region sensitive to the head orientation w.r.t. the body (when a person is looking to the left in an image, his/her head center is often located to the left of the body axis) whereby achieving a joint body and head orientation

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
editors. Cambridge University Press, 2012

tracking. This latter method suggests that in such low resolution conditions, using external information like the coupling with body information is indeed often necessary to obtain more robust performance.

6.5 Audio-visual tracking

Acoustic data provided by microphone-arrays can also be used to localize people in room, and are of course of primary importance when analyzing interaction behaviors in meeting situations or smart rooms without requesting people to wear head-sets or lapels. Thus, many audio-visual (AV) tracking algorithms have been considered for the multi-person tracking task, and more recently, for head pose estimation. Below, we summarize the main points related to this issue.

6.5.1 Audio-visual tracking

Audio and video signals are the result of different person activities, and measure different physical quantities. Because of this, their role for tracking is not symmetric, and provide some good complementarity.

The time difference of arrival (TDOA) of a sound source signal at a pair of microphones, which depends on the location of the source, is the main characteristic exploited for sound localization. TDOA observations can for instance be obtained by finding peaks in the Generalized Cross-Correlation (GCC) function between the acquired signals, and 3D location can be obtained by combining the TDOA information from several microphone pairs. Note that more robust observations (to noise, to the energy of the signals, to their frequency content -impulse sounds vs speech voiced sounds) exist, and the reader can refer to the audio chapter of this book for more information on the topic.

Audio is interesting for providing cheap instantaneous observations which can be conveniently exploited for tracking initialization and tracking failure recovery. It is of course the primary cue for inferring speaking activity as far as a good speech/non-speech segmentation can be conducted. On the other hand, audio source localization can be quite inaccurate due to several reasons: low signal-to-noise ratio when speakers are far from the microphone array, room reverberations, overlapped speech, presence of secondary audio sources -computers, doors, footstep sounds. Importantly, as people are not speaking all the time, audio localization information is discontinuous and might not be available for long periods. In such conditions, keeping track of the identity of potentially moving people from audio signals will essentially rely on some form of biometric approaches, i.e. on the extraction of acoustic signatures characterizing people voices and on their association over time, which is a difficult task in practice.

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
editors. Cambridge University Press, 2012

Video, on its side, provides continuous localization information, usually with a higher accuracy than audio, but is subject to drift and failure, unless face/person detectors are regularly used. And in many cases, video resolution is not high enough and much more cumbersome to be used for speaking activity detection.

The Bayesian approach described in Section 6.2 with its sampling approximation proved to be particularly adapted to exploit the complementary of the cues and their specificities. We explain below how this can be achieved.

State space and AV calibration. Associating audio and video observations requires the existence of some mapping function between both modalities. When tracking two or a group of persons in front of a single camera, the state space is often defined in the image domain, and complemented with a discrete index indicating the speaking status of a person. In such cases, special configurations have often been assumed, allowing to relate in a simple way the azimuth angle of a single microphone array (placed as close as possible to the camera center) to the column number in the image plane. In [Gatica-Perez et al., 2007], which addresses multi-speaker tracking in non-overlapping cameras, a more general yet simple data-driven method is used. There, audio estimates are mapped into the (camera index, 2D location) state space using a nearest-neighbor approach exploiting (audio,video) state pairs gathered during a training phase. This non-parametric approach is efficient, allows to easily handle distortions, but does not allow for a precise audio-visual mapping.

When multiple cameras or microphone arrays are available, a 3D location state space can be more conveniently used. It assumes a jointly calibrated audio-visual system able to associate with a 3D location the corresponding image and audio measurements. Note that the sound signal produced by a human is located more or less 20cm in front of the mouth. Hence, a state space characterizing a person should be able to locate this point if one wants to exploit audio.

Data fusion, likelihood models, and proposals. The transient and cheap localization information nature provided by the audio signal can be assimilated to a simple additional detection. Therefore, its integration within an AV PF framework can be done following the two methods described in tracking-by-detection paragraphs of Section 6.4.1: either as an additional likelihood term[2] or in the proposal to potentially recover from failure or initiate a new track [Gatica-Perez et al., 2007]. Note that the second option is

[2]A point of attention here is that, as with all multi cue likelihoods for PF, acoustic and visual likelihood should have comparable sensitivity to small state changes, i.e. they should both be equally broad. GCC from microphone pairs with short baseline is, in general, irregular and spiky, and must be smoothed to obtain an acoustic likelihood good for PF with, e.g., color likelihood.

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
editors. Cambridge University Press, 2012

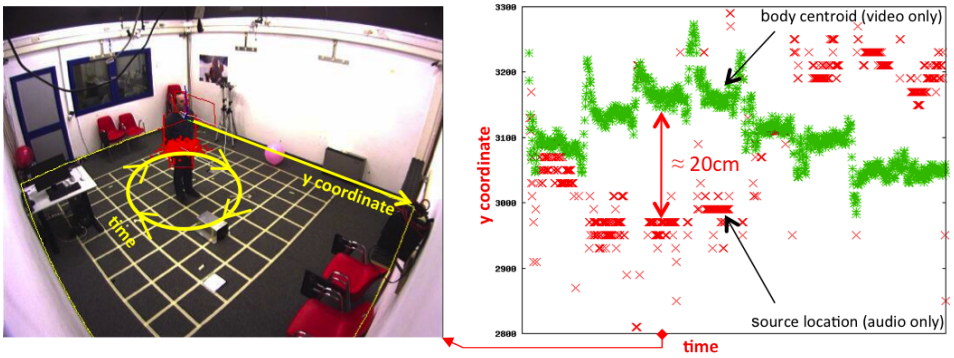

Figure 6.7: The plot shows the y coordinate (in mm) of the estimated source location (crosses - audio only) and body centroid (stars - video only) of a person turning around while speaking (left image). There is an evident offset of about 20cm among the two. The plot shows also that the acoustic and visual likelihoods from [Brutti and Lanz, 2010] provide location estimates that are sufficiently accurate to exploit this offset for audio-visual pose tracking.

only possible if (i) the audio localization information is sufficiently reliable to sample the full state or part of it; and (ii) synchronization between the processed signals can be achieved. If such conditions are not met in real time systems, data fusion then consists in associating sound sources (possibly separated into speech/non speech categories) with the visually tracked person or head in a late fusion process.

Also, as with the video case, one advantage of the PF is that audio information can be treated usefully without requiring to first perform 3D reconstruction through triangulation or optimization. For instance, when using a single microphone array, azimuth and elevation angles can be measured with some precision, but depth is often very unreliable. Still, audio can be used for 3D tracking (using a 3D stereo or multi-camera system) to check that the 3D localization of a person head (or mouth, cf above) is consistent with some audio TDOA or GCC measurements which taken alone would not be able to provide good 3D estimates.

Inference. Multi-speaker tracking can be considered as a specific instance of multiple person tracking. Thus, all considerations w.r.t. complexity, curse of dimension, and inference discussed in Section 6.3.4 are valid, as well as the Rao-Blackwellized PF methodology described in Section 6.4.2, that often could be used to perform an exact inference on the speaking status variable.

6.5.2 Head pose tracking with audio information

Due to the geometry of the human vocal tract, the acoustic emission pattern of a speaking person is not uniform but uni-directional. Furthermore,

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012

the location of the speech source (the mouth) has a horizontal offset from the body center along the head orientation (see Fig. 6.7). While these two factors, if ignored, complicate acoustic localization in reverberant environments and late fusion with visual estimates for multi modal tracking, they can be conveniently exploited to estimate the joint 3D position and orientation state of a person with audio-visual particle filtering.

In [Brutti and Lanz, 2010] the color histogram based PF described at the end of Section 6.4.3 is integrated with an additional orientation sensitive acoustic likelihood for 3D pose tracking with multiple cameras and distributed microphone pairs. For a given state (representing body centroid and horizontal head orientation) GCC based TDOA likelihoods are evaluated for each pair at the mouth position (i.e. at the 3D point shifted horizontally by 20cm from the body axis along the state direction). Then, a joint acoustic likelihood is constructed as a weighted sum of individual contributions. The weights are hereby computed from the state, taking into account the spatial distribution of microphone pairs in the environment: higher weights are given to pairs which are located in the direction of the state (where the direct wavefront is expected to form the dominant peak in the GCC function), while others that are placed lateral or opposite to it (where reverberations may suppress the contribution of the direct wavefront) are assigned a low weight. Acoustic measurements integrated this way, if available (this is decided upon joint likelihood thresholding), stabilize head pose estimates while speaking, especially when the color model of the target is either weak (as e.g. for a bald person) or noisy (if acquired via detection).

6.6 Conclusion and further reading

In this chapter, we have shown that the Bayesian framework is a powerful yet flexible formalism to address the tracking of single or multiple persons, and of their characteristics (head pose, speaking status). It allows to easily introduce appropriate state variables and observations and model their relationships in order to get a good representation of their probabilistic dependencies. Thanks to the use of sampling inference methods, we can exploit appropriate likelihood models (exhibiting multiple modes, providing finer information about the observations) in both the visual and audio domain. Generic methodological points have been illustrated using examples from working systems that have been developed within the AMI or the CHIL European projects.

To learn more about sampling methods for sequential Bayesian state estimation the interested reader is referred to [Isard and Blake, 1998, Arulampalam et al., 2002] for particle filtering and to [Gilks et al., 1996] for an introduction to MCMC. Recommended surveys of state-of-the-art methods are [Yilmaz et al., 2006] for object tracking, [Kakumanu et al., 2007] on skin

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Cambridge University Press, 2012.

modeling and detection, and [Murphy-Chutorian and Trivedi, 2009] for head pose estimation. For a comprehensive presentation of sampling techniques as a framework for sequential data fusion of multiple cues and modalities we indicate [Perez et al., 2004, Gatica-Perez et al., 2007].

Bibliography

S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, February 2002.

S. Ba and J.-M. Odobez. Probabilistic head pose tracking evaluation in single and multiple camera setups. In *Proc. second Workshop on Classification of Events, Activities and Relationships (CLEAR'07)*, Baltimore, May 2007.

S. O. Ba and J-M. Odobez. A rao-blackwellized mixed state particle filter for head pose tracking. In *ACM-ICMI Workshop on Multi-modal Multi-party Meeting Processing (MMMP), Trento Italy*, pages 9–16, 2005.

V. Badrinarayanan, P. Pérez, F. Le Clerc, and L. Oisel. Probabilistic color and adaptive multi-feature tracking with dynamically switched priority between cues. In *Proc. Int. Conf. Comp. Vis.(ICCV)*, 2007.

K. Bernardin and R. Stiefelhagen. Audio-visual multi-person tracking and identification for smart environments. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, 2007.

K. Bernardin, T. Gehrig, and R. Stiefelhagen. Multi- and single view multiperson tracking for smart room environments. In *Proc. Workshop on Classification of Events, Actions and Relations (CLEAR)*, 2006.

A. Brutti and O. Lanz. A joint particle filter to track the position and head orientation of people using audio visual cues. In *Proc. of European Signal Processing Conference (EUSIPCO)*, 2010.

C. Canton-Ferrer, J. Salvador, J. Casas, and M. Pards. Multi-person tracking strategies based on voxel analysis. In *Multimodal Technologies for Perception of Humans*, LNCS. Springer-Verlag, Berlin, Heidelberg, 2008.

J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *IEEE Comp. Vis Pattern Recognition*, 2000.

S. Duffner and J.M. Odobez. Exploiting long-term observations for track creation and deletion in online multi-face tracking. In *IEEE Conference on Automatic Face and Gesture Recognition*, mar 2011.

This chapter appeared in:
Multimodal Signal Processing: Human Interactions in Meetings.
Steve Renals, Herv Bourlard, Jean Carletta and Andrei Popescu-Belis
Editors. Cambridge University Press, 2012

F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multi-camera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, February 2008.

D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audio-visual probabilistic tracking of multiple speakers in meetings. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(2):601–616, February 2007.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in practice*. Chapman-Hall, 1996.

C. Hue, J.-P. Le Cadre, and P. Perez. Sequential monte carlo methods for multiple target tracking and data fusion. *Signal Processing, IEEE Transactions on*, 50(2):309 –325, feb 2002.

M. Isard and A. Blake. Condensation—conditional density propagation for visual tracking. *Int. Journal of Comp. Vision*, 29(1):5–28, 1998.

M. Isard and J. MacCormick. BRAMBLE: A Bayesian multi-blob tracker. In *Proc. IEEE ICCV*, Vancouver, Jul. 2001.

P. Kakumanu, S. Makrogiannis, and N. Bourbakis. A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3):1106–1122, 2007.

Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *IEEE Trans. Pattern Anal. Machine Intell.*, 27:1805–1819, 2005.

O. Lanz. Approximate bayesian multibody tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9), December 2006.

O. Lanz. An information theoretic rule for sample size adaptation in particle filtering. In *Int. Conf. Image Analysis and Processing (ICIAP)*, 2007.

O. Lanz and R. Brunelli. Dynamic head location and pose from video. In *Int. Conf. Multi-sensor Fusion and Integration (MFI)*, 2006.

A. Lehuger, P. Lechat, and P. Perez. An adaptive mixture color model for robust visual tracking. In *Proc. Int. Conf. on Image Processing (ICIP'06)*, pages 573–576, Atlanta, USA, October 2006.

O.M. Lozano and K. Otsuka. Real-time visual tracker by stream processing simultaneous and fast 3d tracking of multiple faces in video sequences by using a particle filter. *JOURNAL OF SIGNAL PROCESSING SYSTEMS*, 57(2):285–295, 2009.

C. Medrano, J.E. Herrero, J. Martnez, and C. Orrite. Mean field approach for tracking similar objects. *Computer Vision and Image Understanding*, 113(8):907 – 920, 2009.

E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009. ISSN 0162-8828.

P. Perez and J. Vermaak. Bayesian tracking with auxiliary discrete processes. application to detection and tracking of objects with occlusions. In *IEEE ICCV Workshop on Dynamical Vision*, 2005.

P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *Proc. European Conference on Computer Vision (ECCV)*, Copenhagen, May 2002.

P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. IEEE*, 92(3):495–513, 2004.

E. Ricci and J.M Odobez. Learning large margin likelihood for realtime head pose tracking. In *IEEE Int. Conf. on Image Processing (ICIP)*, 2009.

C. Segura, C. Canton-Ferrer, A. Abad, J.R. Casas, and J. Hernando. Multimodal head orientation towards attention tracking in smartrooms. In *Int. Conf. Acoustic, Speech and Signal processing (ICASSP)*, April 2007.

R. Singh, P. Bhargava, and S. Kain. State of the art smart spaces: application models and software infrastructure. *Ubiquity*, 37(7):2–9, 2006.

K. Smith, D. Gatica-Perez, and J.-M. Odobez. Using particles to track varying numbers of interacting people. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Diego, June 2005.

R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4):928–938, 2002.

R. Stiefelhagen, R. Bowers, and J. Fiscus, editors. *Multimodal Technologies for Perception of Humans*. Springer-Verlag, Berlin, Heidelberg, 2008.

K. Tzevanidis and A. Argyros. Unsupervised learning of background modeling parameters in multicamera systems. *Comput. Vis. Image Underst.*, 115:105–116, January 2011. ISSN 1077-3142.

J. Yao and J-M. Odobez. Multi-camera 3D person tracking with particle filter in a surveillance environment. In *The 16-th European Signal Processing Conference (EUSIPCO-2008)*, August 2008a.

J. Yao and J.-M. Odobez. Multi-camera multi-person 3d space tracking with mcmc in surveillance scenarios. In *ECCV 2008 Workshop on Multi Camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2)*, October 2008b.

A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38, December 2006. ISSN 0360-0300.

G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space speaks - towards socially and personality aware visual surveillance. In *ACM MPVA Workshop*, 2010.