MUCATAR Project
First-Year Progress Report

Daniel Gatica-Perez [1] Jean-Marc Odobez [1]
Sileye Ba [1] Kevin Smith [1]

IDIAP–RR

July 2003

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

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

[1] IDIAP, Martigny, Switzerland

IDIAP Research Report

MUCATAR PROJECT
FIRST-YEAR PROGRESS REPORT

Daniel Gatica-Perez Jean-Marc Odobez Sileye Ba Kevin Smith

JULY 2003

Abstract. This document briefly describes the progress on the MUCATAR (MUltiple CAmera Tracking and Activity Recognition) IM2 White Paper Project. We have concentrated on the problem of robust single-person tracking with Sequential Monte Carlo (particle filtering) techniques, on single-camera scenarios. In particular, we focused on three areas: the implementation and evaluation of state-of-the-art single-feature trackers, the investigation of data fusion mechanisms (considering both visual and audio-visual features) to improve tracking, and the initial investigation of models for joint tracking and activity recognition. Achievements on each of these areas are described.

1 Introduction

This report presents the progress of the MUCATAR (MUltiple CAmera Tracking and Activity Recognition) project at IDIAP. We have concentrated on the problem of robust single-person tracking on single-camera scenarios with Sequential Monte Carlo (SMC) techniques. Our research work has focused on the implementation and evaluation of single feature trackers, the investigation of fusion mechanisms of multiple features (both visual and audio-visual) to improve tracking, and the initial investigation of joint tracking and recognition of human activities. This report is a brief summary of results. The reader is referred to the following publications for details, and to a website (`www.idiap.ch/˜gatica/mucatar−report−D1.html`) where videos with results are available:

J-M. Odobez, S. Ba, and D. Gatica-Perez, “An Implicit Motion Likelihood for Tracking with Particle Filters,” in *Proc. British Machine Vision Conference (BMVC),* Norwich, Sep. 2003.

D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore, “Audio-visual speaker tracking with importance particle filters,” in *Proc. IEEE Int. Conf. on Image Processing (ICIP),* Barcelona, Sep. 2003.

The report is organized as follows. For sake of completeness, Section 2 reviews the basic SMC concepts. Section 3 describes the implementation and evaluation of two single-feature trackers. Section 4 presents our work regarding data fusion for tracking, using both multiple visual features and audio-visual features. Section 5 describes our starting work towards joint tracking-recognition. Section 6 presents some conclusions and discusses future work.

2 Tracking with Particle Filters

Sequential Monte Carlo methods a.k.a. particle filters (PFs) represent a principled methodology for tracking [1, 2]. Given a discriminative object representation and a Markov state-space model, with hidden states $\{\mathbf{x}_t\}$ that represent object configurations, and observations $\{\mathbf{y}_t\}$ extracted from one or more data streams, a PF recursively approximates the filtering distribution of states given observations $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ by

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

where $\mathbf{y}_{1:t} = \{\mathbf{y}_1, ..., \mathbf{y}_t\}$. The integral in Eq. 1 represents the prediction step, in which the dynamical model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and the previous distribution $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$ are used to compute a prediction distribution, which is then used as prior for the update step, and multiplied by the likelihood $p(\mathbf{y}_t|\mathbf{x}_t)$ to generate the current filtering distribution. Except for a few special cases, exact inference in this model is intractable. SMC methods are usually employed to approximate Eq. 1 for non-linear, non-Gaussian problems, using random sampling by (i) predicting candidate configurations, and (ii) measuring their likelihood, in a process that amounts to random search in a configuration space. The filtering distribution is first defined by a set of weighted samples or particles $\{(\mathbf{x}_t^{(i)}, \pi_t^{(i)}), i = 1, ..., N\}$, where $\mathbf{x}_t^{(i)}$ and $\pi_t^{(i)}$ denote the i-th sample and its importance weight at the current time. The point-mass approximation is given by $\hat{p}_N(\mathbf{x}_t|\mathbf{y}_{1:t}) = \sum_{i=1}^N \pi_t^{(i)} \delta(\mathbf{x}_t - \mathbf{x}_t^{(i)})$. The prediction step propagates each particle according to the dynamics, and the updating step reweights them using their likelihood, $\pi_t^{(i)} \propto \pi_{t-1}^{(i)} p(\mathbf{y}_t|\mathbf{x}_t^{(i)})$. A resampling step using the new weights is necessary to avoid degradation of the particle set.

The above methodology is general, allowing for a multitude of variants: tracking with multiple sources of observations, tracking in mixed-state configuration spaces composed of continuous and discrete components, or tracking multiple objects can be all defined by the same formulation. Designing a PF therefore involves defining a specific state-space, an object representation, dynamical and observation models, and sampling mechanisms. MUCATAR has investigated several of these issues, as discussed in the following subsections.

3 Single-feature SMC trackers

Two single-feature observation models were first implemented as baseline methods, one based on a contour-based object representation [1], and another based on a color distribution object model [10, 3].

3.1 Shape-based tracking

We use a very simple object model, where objects are represented by their silhouettes (contours) in the image plane [1]. In particular, we have used a parameterized vertical ellipse to represent the basic shape. The observation model further assumes that shapes are embedded in clutter [1]. Edge-based measurements are computed along L normal lines to a hypothesized contour, resulting in a vector of candidate positions for each line, $\mathbf{y}_t^l = \{\nu_m^l\}$ relative to the point lying on the contour ν_0^l. With some usual assumptions, the observation likelihood for L normal lines can be expressed as $p(\mathbf{y}_t|\mathbf{x}_t) \propto \prod_{l=1}^{L} p(\mathbf{y}_t^l|\mathbf{x}_t) \propto \prod_{l=1}^{L} \max\left(K, \exp(-\frac{\|\hat{\nu}_m^l - \nu_0^l\|^2}{2\sigma^2})\right)$, where $\hat{\nu}_m^l$ is the nearest edge detected on the l^{th} line, and K is a constant introduced when no edges are detected.

3.2 Color-based tracking

In this case, an object is represented by a region R enclosed by its initial configuration $\mathbf{x}_0$. The object color distribution is modeled by a normalized histogram in the HSV space. Spatial information is added by splitting the region into N_r sub-regions $R_r(\mathbf{x}_0)$, and a multidimensional histogram $\mathbf{b}(\mathbf{x}_0)$ is then computed. The observation likelihood is defined by comparing the candidate color model $\mathbf{b}(\mathbf{x}_t)$ to the reference color model $\mathbf{b}(\mathbf{x}_0)$, using the Bhattacharyya distance measure, denoted by D_{BT}^2, and assuming that the probability distribution of the square of this distance follows an exponential law [10, 3]. It is thus given by $p(\mathbf{y}_k|\mathbf{x}_k) \propto \exp\{-\lambda D_{BT}^2(\mathbf{b}(\mathbf{x}_k), \mathbf{b}(\mathbf{x}_0))\}$, where λ is the parameter of the exponential pdf.

3.3 Evaluation of algorithms

The two baseline methods were evaluated in two different datasets. The first dataset was captured at the IDIAP meeting room [8]. The second dataset is more general, and was captured in both indoor and outdoor scenarios with hand-held cameras. Additionally, parameters for the models (object models, observation likelihoods, and dynamical models) were set by hand but not exhaustively tuned.

Examples of shape-based tracking results are shown in Fig. 1. Results using color-based tracking are shown in Fig. 2. The results were evaluated only qualitatively. For both cases, results in the meeting room are of good quality when no occlusion occurs, as the trackers can deal with natural and fast human motion. However, the object shape- and color-based models can still be unspecific, and the lack of explicit occlusion models make them usually fail in cases of total occlusion, although the histogram-based tracker can better tolerate both occlusion and visual clutter. Regarding initialization, the shape-based tracker uses canonical shapes that need not be tuned to the specific object to achieve good performance. In contrast, the color-based tracker needs to be initialized to build the appearance model. Refer to the website for the example videos. Regarding the hand-held camera videos, the two trackers also deal well with a small amount of visual clutter, and little or no occlusion. Again, the histogram-based tracker is more robust to occlusion and clutter.

4 Feature fusion for improved tracking with particle filters

Three serious limitations of the baseline methods are:

- Object modeling. First, the object models that we use are quite unspecific. Thus, the trackers can fail because of the presence of visual distractors. By learning the color object models from the first frame of the sequence, the color tracker becomes more specific than the shape tracker and therefore more robust to clutter. However, this learning step depends on a precise initialization of the tracker. Secondly, in view of the latter, the object models we use might not be able to track objects over very long periods of time. Better object representations and learning mechanisms are needed to build models that precisely capture the variation in object appearance.

- Lack of (re)initialization mechanisms.

- Lack of robustness to occlusion.

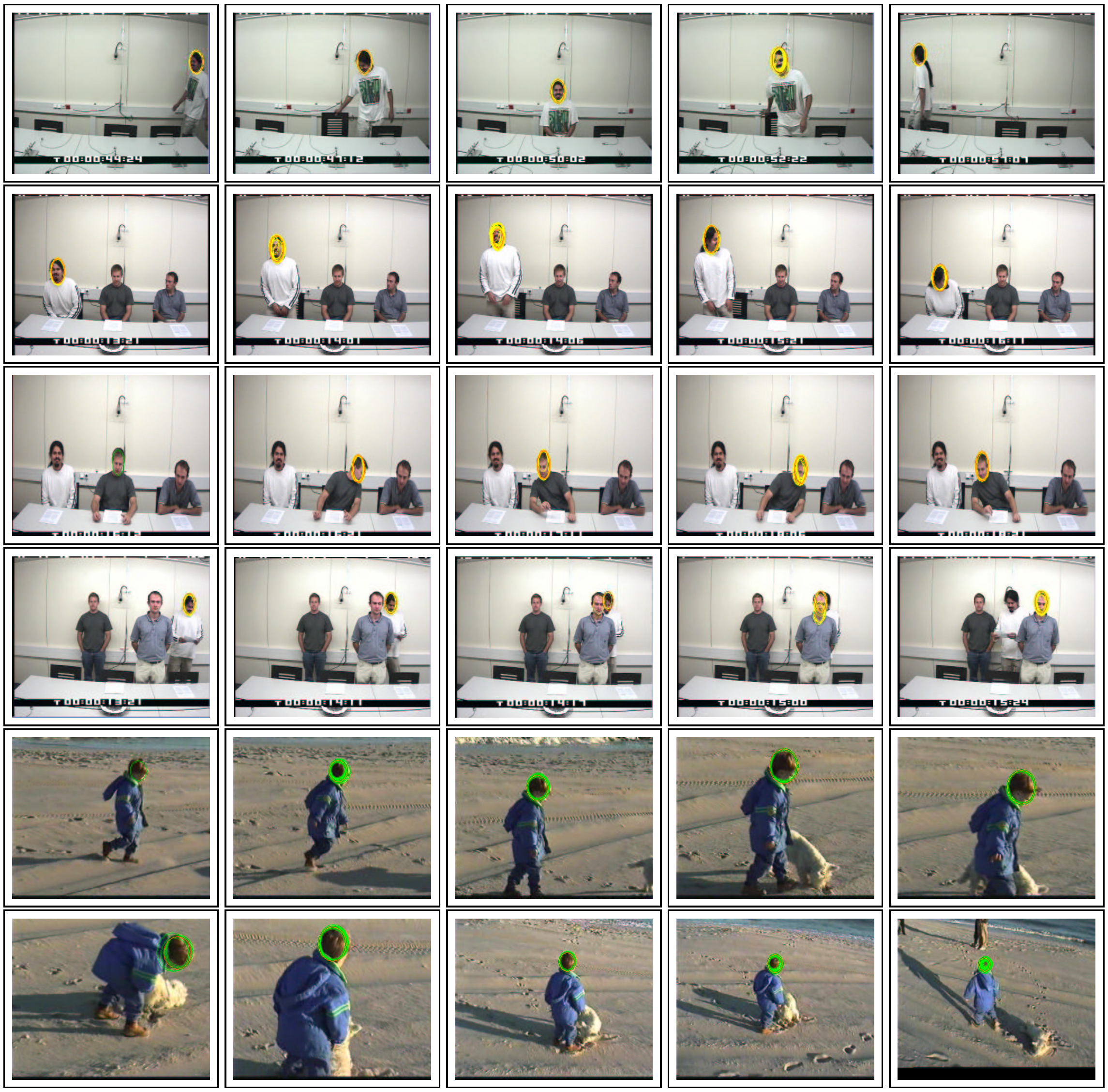

Figure 1: Shape-based tracking results. Rows 1-4: Tracking in the meeting room. Row 1: Single-object tracking, frames 0, 65, 130, 200, and 310. Row 2: Fast object tracking (standing up), frames 30, 35, 40, 80, and 85. Row 3: Fast object tracking (moving on a seat), frames 0, 10, 25, 45, and 50. Row 4: The tracker cannot handle total object occlusion, frames 0, 17, 23, 31, and 55. Rows 5-6: Tracking a running boy in a home video sequence (hand-held camera), frames 0, 25, 50, 200, 220, 250, 300, 325, 340, and 445. In all cases, the largest particle is shown in red, with other large particles shown in yellow or green.

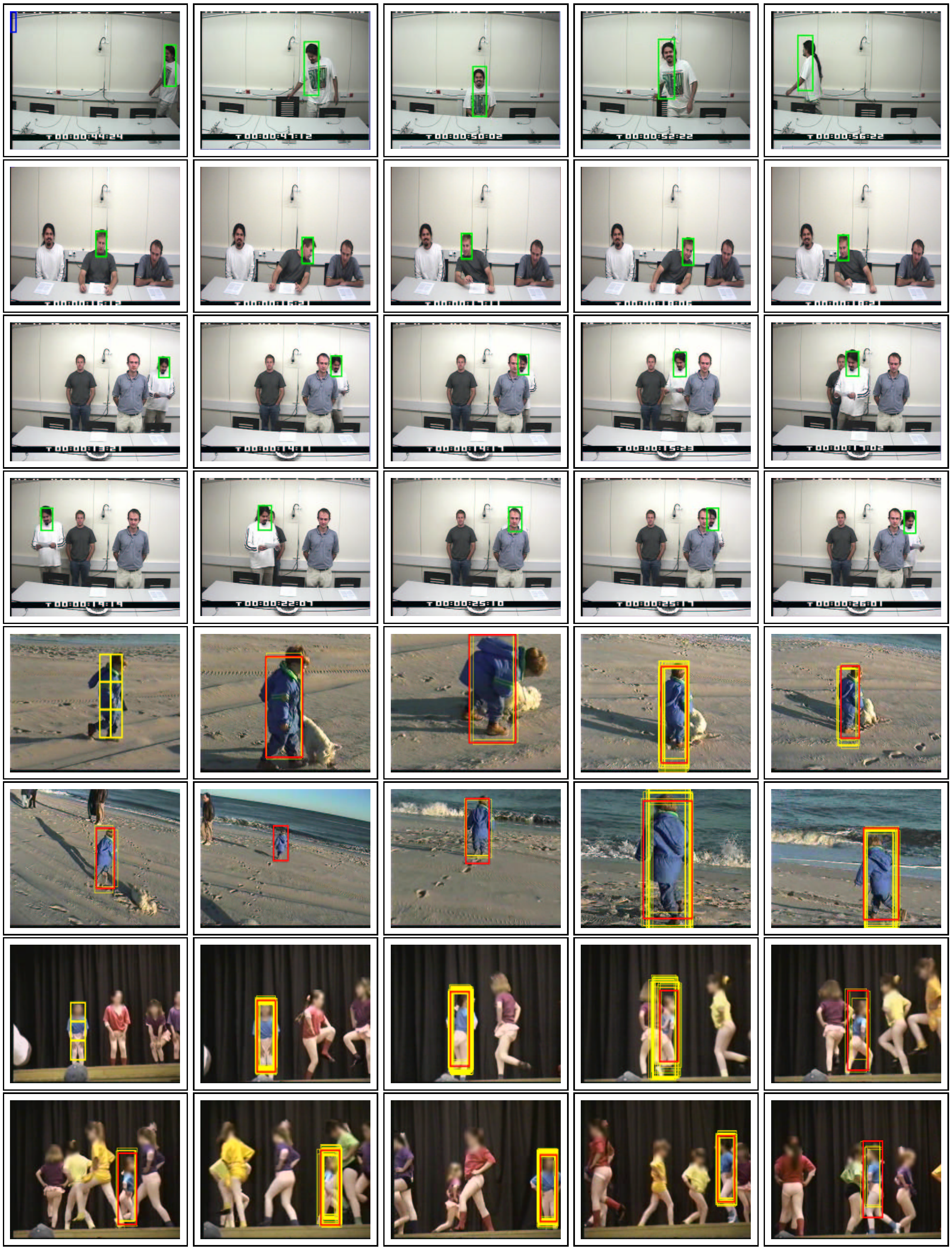

Figure 2: Color distribution-based tracking results. Rows 1-4: Tracking in the meeting room. Mean configuration in green. Row 1: Single-object tracking, frames 0, 65, 130, 200, and 310. Row 2: Fast object tracking (moving on a seat), frames 0, 10, 25, 45, and 50. Rows 3-4: The tracker handles total object occlusion, frames 0, 17, 23, 55, 83, 150, 213, 291, 198, and 307. Rows 5-8: Tracking objects in home video sequences (hand-held camera). The largest particle is shown in red, other large particles in yellow. Rows 5-6. Tracking a running boy, frames 0, 100, 150, 225, 250, 375, 450, 495, 525, and 620. Rows 7-8. Tracking a dancing girl.

We have investigated solutions for the two first problems, based on the exploitation of multiple features, both visual and audio-visual, as described in the following.

4.1 Fusing visual features: an implicit motion likelihood [9]

As we have discussed, generic shape-based or color-based object representations are rather unspecific. One way to improve the robustness of a tracker consists of combining low-level measurements such as shape and color [16]. A further step to render the target more discriminative is the use of appearence-based models such as templates [11, 12], leading to reliable trackers in stable conditions. However, such representations do not allow for large changes of appearence, unless adaptation is performed or more elaborate global appearence models are used [13]).

Another issue regards the modeling of dynamics, which are used in the prediction step of the PF, and which implicitly define a search range. The difficulty of modeling the dynamics arises from the two contradictory objectives. On one hand, the search space should be small enough to avoid the tracker being confused by distractors in the vicinity of the true object configuration, a situation that is likely to happen for unspecific object representations such as generic shapes or color distributions. On the other hand, it should be large enough so that it can cope with abrupt motion changes.

We have proposed a new tracking method based on the PF algorithm, arguing that the standard hypothesis of independence of observations given the state sequence is inappropriate in the case of visual tracking. In practice, all motion estimation and compensation algorithms like MPEG implicitly (and reasonably) assume that consecutive measurements are correlated. Our model thus assumes that current observations depends on the current and previous object configurations as well as on the past observation, which introduces in practice an implicit object motion likelihood. The benefits of this new model are two-fold. First, by exploiting a type of template correlation between successive images, our model turns generic trackers like shape or color histogram trackers into more specific ones, without resorting to complex appearence based models. Second, as a consequence, our model reduces the sensitivity of the algorithm to the size of the search range, as potential distractors are filtered out by the introduced correlation factor when using a large search range.

The new observation likelihood takes the form $p(\mathbf{y}_k|\mathbf{y}_{k-1}, \mathbf{x}_k, \mathbf{x}_{k-1}) = p_c(\mathbf{y}_k|\mathbf{y}_{k-1}, \mathbf{x}_k, \mathbf{x}_{k-1}) \times p_o(\mathbf{y}_k|\mathbf{x}_k)$, where $p_c(\cdot)$ models the correlation between the two consecutive observations, and $p_o(\cdot)$ is an object likelihood as described in Section 3. This choice decouples the model of the correlation existing between two images, whose implicit goal is to insure that the object trajectory follows the optical flow field implied by the sequence of images, from the shape or appearance object model. In turn, the implicit motion likelihood is defined as $p_c(\mathbf{y}_k|\mathbf{y}_{k-1}, \mathbf{x}_k, \mathbf{x}_{k-1}) \propto \exp(-\lambda_c \mathrm{d}_c(\tilde{z}_{\mathbf{x}_k}, \tilde{z}_{\mathbf{x}_{k-1}}))$, where d_c denotes a distance function between two (transformed) image patches at consecutive times, $\tilde{z}_{\mathbf{x}_{k-1}}, \tilde{z}_{\mathbf{x}_k}$. The distance function should satisfy a number of conditions [9]; we have selected the L_2 norm as a reasonable choice. Our model is different from other formulations [11], as no object template for correlation comparison is learned off-line or defined at the begining of the sequence, and the tracker does not maintain a single template object representation at each instant of the sequence. Thus, the correlation term is not object specific.

Results for hand-initialized trackers are shown in Figs. 3-5. In Fig. 3, the two hands are active, grasping or displacing objects, or waving from left to right over a cluttered background; the camera is moving as well. Tracking the right hand with the histogram-only (H) and histogram plus correlation (HC) trackers produced similar results. However, when tracking the left hand, the H tracker was several times confused by the presence of the right hand, as illustrated in Fig. 3, top row. This confusion usually lasted for several frames, but most of the time, the H tracker was able to resume. The HC tracker did never undergo this ambiguity problem. An objective evaluation is available in [9]. Fig. 4 further illustrates the benefit of the method in the presence of ambiguities. Despite the presence of a textured background, the camera and head motion, the change of appearence of the head, and partial occlusion, the head is correctly tracked using our method. The shape tracker alone is unable to perform a correct tracking after a few frames. The histogram-based tracker fails as well but a joint histogram and shape tracker was successful (not shown here). Finally, in Fig. 5, histogram models are inappropriate, as the tracked person undergoes 360^o head turns. The shape tracker is rapidly perturbed and gets lost due to the abrupt vertical camera motion and increasing visual clutter. On the other hand, the shape and correlation tracker successfuly tracks the head turn over the entire sequence, remaining less influenced by the lack of contour measurements and abrupt motion changes.

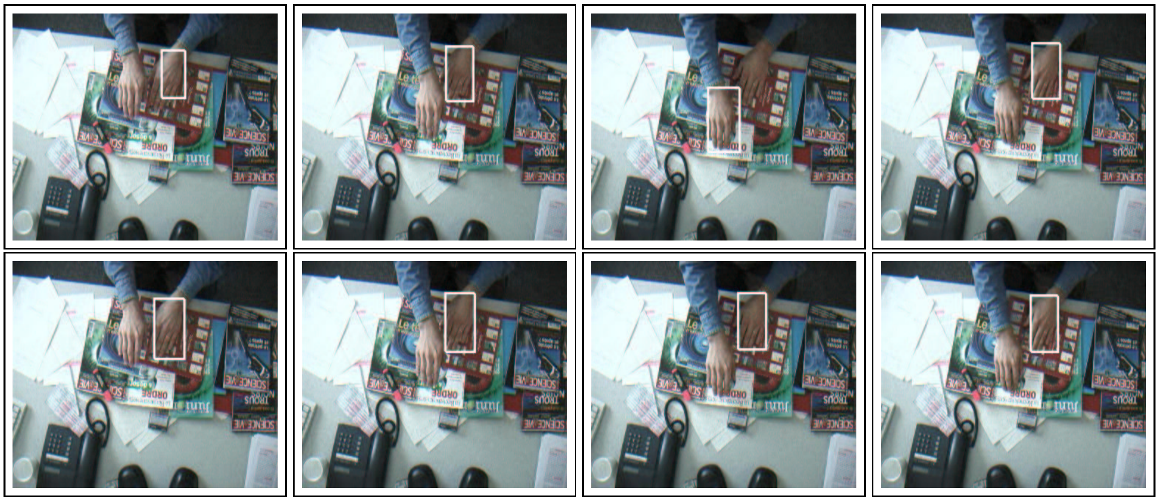

Figure 3: Tracking the left hand, frames 840, 845, 850, and 855. Top: histogram tracker. Bottom: histogram +correlation. The white box corresponds to the most likely particle.

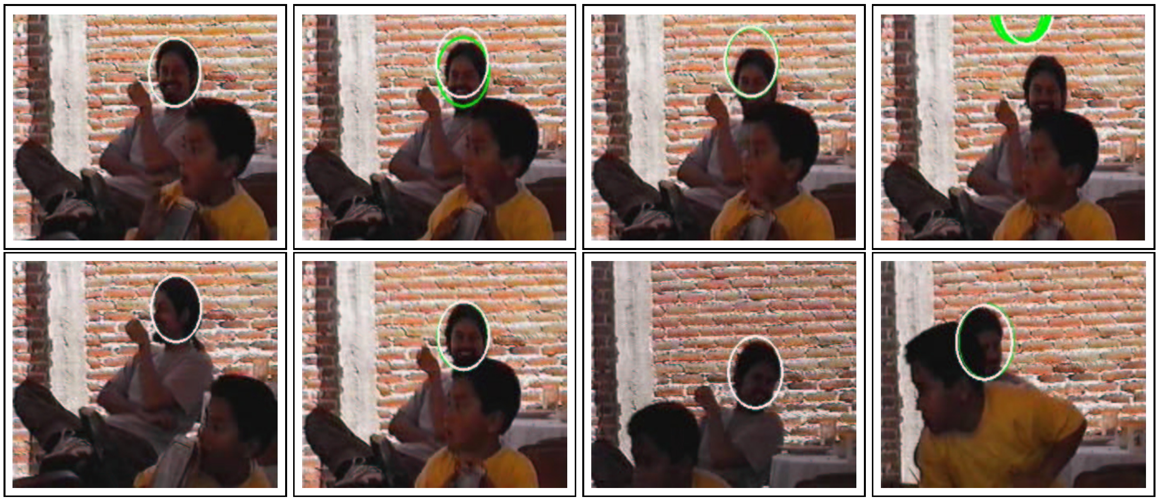

Figure 4: Head tracking 1. Top row: shape tracker, frames 9, 11, 13, and 15. Bottom row: shape+correlation tracker, frames 1, 15, 39, and 63.

4.2 Fusing audio and video for speaker tracking [4]

Speaker tracking constitutes a relevant task for applications that include automatic meeting analysis and remote conferencing [15]. In the context of meetings, speaker turn patterns convey a rich amount of information about the dynamics of a group and the individual behaviour of its members, as documented by a solid body of literature in social psychology [7]. The use of audio and video as separate cues for tracking are classic problems in signal processing and computer vision. However, although audio-based speaker localization offers very valuable information about speaker turns, sound and visual information are jointly generated when people speak, and provide complementary advantages for speaker tracking if their dependencies are jointly modeled.

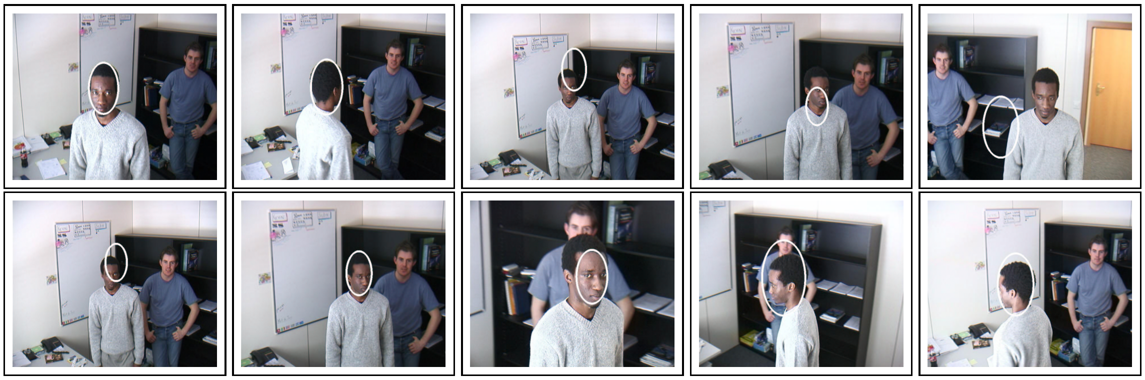

Figure 5: Top row: shape tracker, frames 6, 26, 66, 76, and 110. Bottom row: shape+correlation tracker, frames 66, 76, 90, 134, and 146. 800 particles

As stated earlier, initialization and recovery from failures are bottlenecks in visual tracking that can be robustly addressed with audio. However, precise object localization is better suited to visual processing.

Current SMC formulations for AV speaker tracking [14, 17] fuse audio and video only at the measurement level, thus leading to symmetrical models in which each modality accounts for the same relevance, and depending on the dynamical model to generate candidate configurations. Furthermore, cameras and microphones are independently (and carefully) calibrated for state modeling and measuring in 2-D or 3-D. Such formulations tend to overlook several important features of AV data. First, audio is a strong cue to model discontinuities that clearly violate usual assumptions in dynamics (including speaker turns across cameras), and (re)initialization. Its use for prediction would therefore bring benefits to modeling realistic situations. Second, although audio might be imprecise, and visual calibration can be erroneous due to distortion in wide-angle cameras, their joint occurrence tends to be more consistent, and can be robustly learned from data.

We have developed a methodology for AV speaker tracking using PFs at IDIAP meeting room, using one uncalibrated wide-angle camera, and an eight-element circular microphone array. Our work introduced novel aspects on data fusion and AV calibration. Regarding data fusion, given a 2-D configuration space, audio information is used both for sampling and measuring. In the first place, for sampling, basic PFs are limited by the dynamical model to generate candidate configurations, which cannot model abrupt changes of configuration, due for instance to a speaker turn. In contrast, in our work, 3-D audio localization estimates are computed at each frame and integrated into the PF formulation via importance sampling (IS), by defining an audio importance function that emphasizes the most informative regions of the space. An I-Particle Filter can then be defined by a mixture model in which samples are drawn from the original dynamical model, the dynamics (with IS), and a reinitialization prior defined also by audio information. In the second place, for measuring, audio and video are jointly used to compute the likelihood of candidate configurations. We use the shape-based object representation described in Section 3.1, but the approach is applicable to other visual cues. Finally, regarding AV calibration, we proposed a simple, yet robust AV calibration procedure that estimates a direct 3-D-to-2-D mapping from an audio localization estimate onto the image plane. The procedure uses training videos of people speaking while performing typical activities in the meeting room (walking, sitting and standing, moving on their seats), and does not require of precise geometric calibration of camera and microphones. The result is an algorithm that can robustly initialize and track a moving speaker, switch between multiple speakers, tolerate visual clutter, and recover from total AV object occlusion. Other AV speaker tracking methods would find limitations in these settings.

Results of single-person tracking are shown in Fig. 6 (top row) (see [4] for an objective evaluation). The AV tracker is automatically instantiated when the person enters the scene and starts talking, and remains on track when speech ceases. The largest errors during tracking are due to the detection of footsteps as the sound source, but the tracker copes with these short-term distractions. An example of speaker initialization/tracking in presence of visual clutter is shown in Fig. 6 (second row). A second person repeatedly passes behind the speaker without distracting the tracker. A third example illustrating speaker tracking in cases AV occlusion is shown in Fig. 6 (third row). The sequence is challenging due to physical obstacles between the speaker and the microphones. However, when a direct path between the sound source and the microphone array appears, the tracker locks onto the speaker. Furthermore, the AV tracker recovers from AV occlusion in repeated cases thanks to the audio modality as IS function, while video provides finer localization. Finally, Fig. 6 (last row). illustrates tracking a sequence of switching speakers. Three people seated at the table speak in turn (center, right, left). Our algorithm has shown to be adequate for accurately tracking speakers in conversations over extended periods of time.

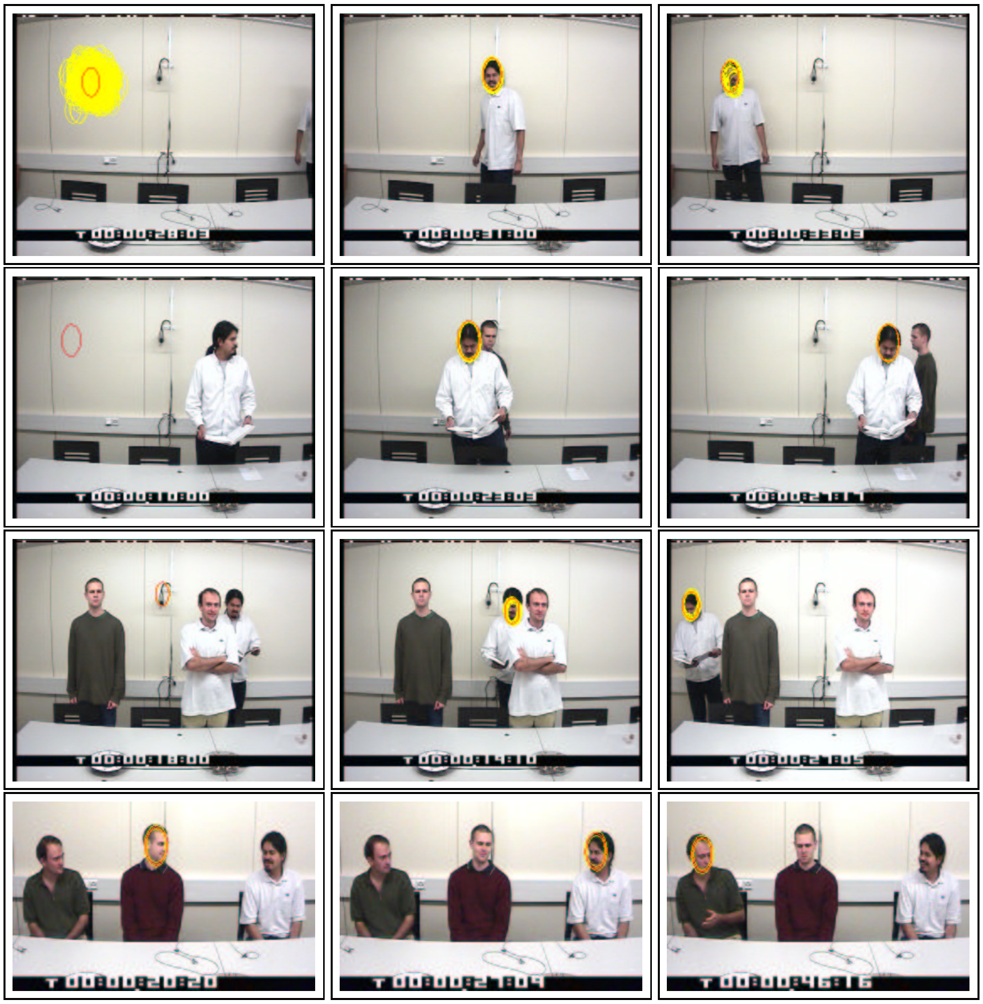

Figure 6: AV tracking. Mean in read, standard deviation from the mean in yellow. First row: *Walk*. Second row: *Visual Clutter*. Third row: *AV occlusion*. Fourth row: *Chat*.

5 Initial work on joint tracking and recognition: examplars and mixed-state particle filters

We have started work towards joint tracking and recognition using mixed-state models, that combines continuous-valued motion parameters with discrete labels that index a set of exemplars of appearance in a joint pdf [5, 13]. Recovering the discrete components of a state constitutes a basic form of recognition. In the Bayesian formulation for tracking (Eq. 1), each state now corresponds to a recognized action, and the motion parameters of an object, and the particle filter algorithm should now account for the different nature of the latent variables.

Exploratory work was done on person-dependent, joint head tracking and coarse head orientation recognition, using video taken with a free-motion camera. A configuration of the state space for the system can be described by $\mathbf{x}_t = (k_t, x_t)$, where x_t are elements of a subspace of affine transformations, and k_t is a discrete index over a set of head orientations. As initial features, we extended the color histogram representation [10] by defining multiple template color histograms in HSV space, defined by a color histogram mask (Fig. 7). The procedure has shown promising results. After training the templates on one sequence, the tracker showed accurate performance on other data sequence with the same subject in different lighting at a different zoom. A few examples of joint tracking and recognition results are shown in Fig. 7. We are currently extending the work to person-independent models, using the same exemplar-based representation. The current head model is based on a discrete grid representation, and the observations are extracted from filtering the image on the elements of the grid with rotation-invariant Gabor wavelets (Fig. 8).

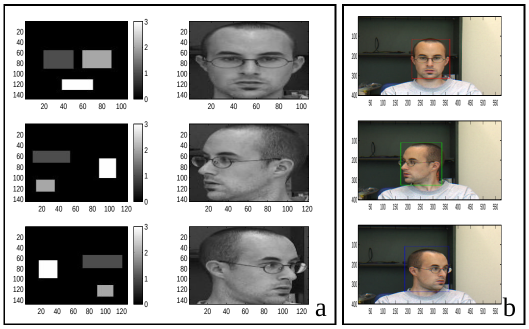

Figure 7: (a) Template masks for head facing experiments, and associated images. (b) Tracking results. Each color represents a diferent head orientation.

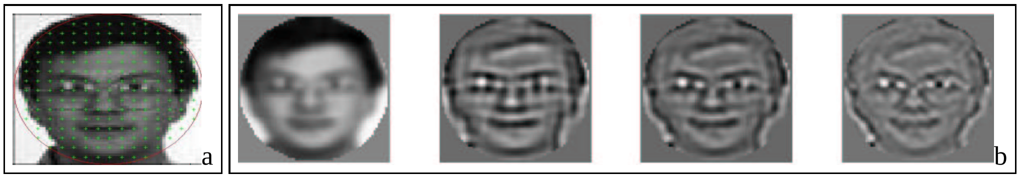

Figure 8: (a) Grid for feature computation. (b) Left-to-right: filtered images by a Gaussian and three Gabor wavelets.

6 Conclusions and future work

In the first year of the project, we investigated solutions to two limitations of the standard particle filter tracker in a single-camera setting: object model unspecificity and tracker (re)initialization. Currently and in the future, we will investigate other aspects of the project. More precisely, we are addressing the following issues:

- Multi-camera tracking. The audio-video tracking method is currently being extended to handle speaker tracking over three cameras in the meeting room.
- Joint tracking and recognition. The work described in the Section 5 will be applied to meeting room scenarios. More specifically, it will be used to recognize head actions, like looking in a specific direction (e.g. to the white-board, to the notes, or to other persons/speaker in the room), or approval/disapproval gestures.
- Among the limitations pointed out in Section 4, occlusion has been investigated only indirectly, through the importance sampling reinitialization mechanism in the AV tracking method. Noticing that object occlusion is often due to occlusion by other objects, we plan to address this problem by investigating the

multi-object tracking issue. Two applications will be considered for this research : multi-person tracking in the meeting room and surveillance/traffic monitoring applications.

Results along all of these directions will be reported in the future.

7 Acknowlegments

This work was funded by the Swiss NCCR on Interactive Multimodal Information Management (IM)2. We thank several of our colleagues at IDIAP for their collaboration on audio-visual tracking (Guillaume Lathoud, Iain McCowan, and Darren Moore) and for technical support (Thierry Collado and Sebastien Marcel). We also thank the Eastman Kodak Company for providing the hand-held video sequences, which belong to the Kodak Home Video Database ©.

References

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

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

[3] D. Gatica-Perez and M.T. Sun, Linking Objects in Videos by Importance Sampling. In *Proc. IEEE ICME*, Lausanne, Aug. 2002.

[4] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M. Odobez, and D. Moore, Audio-visual speaker tracking with importance particle filters. In *Proc. IEEE ICIP*, Barcelona, Sep. 2003.

[5] I. Isard and A. Blake. A Mixed-State CONDENSATION Tracker with Automatic Model-Switching. In *Proc. IEEE ICCV*, 1998.

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

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

[8] D. Moore, “The IDIAP smart meeting room,” *IDIAP Communication 02-07*, 2002.

[9] J-M. Odobez, S. Ba, and D. Gatica-Perez, An Implicit Motion Likelihood for Tracking with Particle Filters. In *Proc. BMVC*, Norwich, Sep. 2003.

[10] P. Perez, C. Hue, J. Vermaak, and M. Gagnet. Color-based probabilistic tracking. In *Proc. ECCV*, May 2002.

[11] J. Sullivan and Rittscher J. Guiding random particles by deterministic search. In *Proc. IEEE ICCV*, Vancouver, July 2001.

[12] H. Tao, H. S. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Trans. on PAMI*, 24(1):75–89, 2001.

[13] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Proc. IEEE ICCV*, Vancouver, July 2001.

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

[15] 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, UT, May 2001.

[16] Y. Wu and T. Huang. A co-inference approach for robust visual tracking. In *Proc. IEEE ICCV*, Vancouver, July 2001.

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