MULTI-PERSON 3D TRACKING WITH PARTICLE FILTERS ON VOXELS

A. López, C. Canton-Ferrer, J.R. Casas

Image Processing Group
Technical University of Catalonia
Barcelona, Spain

ABSTRACT

This paper presents a new approach to the problem of simultaneous tracking of several people in low resolution sequences from multiple calibrated cameras. Spatial redundancy is exploited to generate a discrete 3D binary representation of the scene. A particle filtering scheme adapted to the incoming 3D discrete data is proposed. A volume likelihood function and a discrete 3D re-sampling procedure are introduced to evaluate and drive particles. Multiple targets are tracked by means of multiple particle filters and interaction among them is modeled through a 3D blocking scheme. Test over annotated databases yield quantitative results showing the effectiveness of the proposed algorithm in the indoor scenarios.

Index Terms— Multi-target tracking, particle filtering, 3D processing, multi-camera analysis, human-computer interfaces

1. INTRODUCTION

The current paper addresses the problem of detecting and tracking a group of people present in an indoor scenario in a multiple camera setup. Robust, multi-person tracking systems are employed in a wide range of applications, including SmartRoom environments, surveillance for security, health monitoring, as well as providing location and context features for human-computer interaction.

A number of methods for camera based multi-person 3D tracking has been proposed in the literature [1]. A common goal in these systems is robustness under occlusions created by multiple objects present in the scene when estimating the position of a target. Single camera approaches [2] have been widely employed but are more vulnerable to occlusions, rotation and scale changes of the target. In order to avoid these drawbacks, multi-camera tracking techniques [3] exploit spatial redundancy among different views and provide 3D information as well. Integration of features extracted from multiple cameras has been proposed in terms of image correspondences [4], multi-view histograms [5] or voxel reconstructions [6].

Filtering techniques are employed to add temporal consistency to tracks. Kalman filtering approaches have been extensively used to track a single object under Gaussian uncertainty models and linear dynamics [2]. However, these methods do not perform accurately when facing noisy scenes or rapidly maneuvering targets. Particle filtering has been applied to cope with these situations since it can deal with multi-modal *pdf*s and is able to recover from lost tracks [7, 8].

We propose a method for 3D tracking of multiple people in a multi-camera environment. Redundancy among cameras is exploited to obtain a binary 3D voxel representation of the scene as the input of the tracking system. A particle filter is employed to track a target estimating its 3D centroid and no motion model has been assumed to keep a reduced state space. Particle weights are evaluated through a volume likelihood function measuring whether a particle falls inside or outside a volume. A 3D discrete re-sampling technique is introduced to propagate particles and to capture object shifts. Multiple targets are tracked assigning a particle filter to every one. In order to achieve the most independent set of trackers, we consider a 3D blocking method to model interactions. Finally, effectiveness of the proposed algorithm is assessed by means of objective metrics defined in the framework the CLEAR06 [9] multi-target tracking database.

This material is based upon work partially supported by the IST programme of the EU through the IP IST-2004-506909 CHIL and by TEC2004-01914 project of the Spanish Government.

2. SYSTEM OVERVIEW

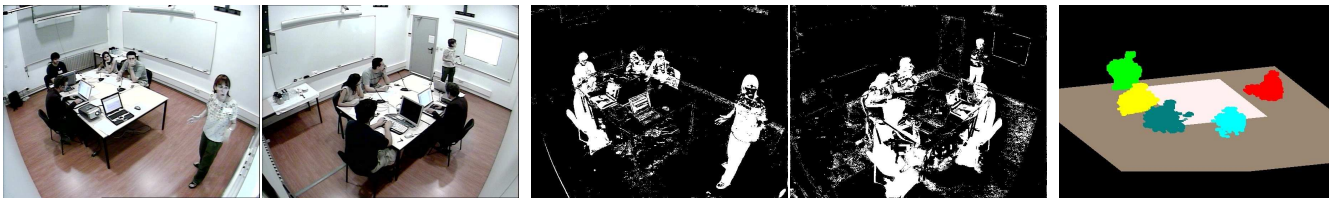
For a given frame in the video sequence, a set of N images are obtained from the N cameras (see a sample in Fig.1a). Each camera is modeled using a pinhole camera model based on perspective projection with camera calibration information available. Foreground regions from input images are obtained using a segmentation algorithm based on Stauffer-Grimson's background learning and substraction technique [10] as shown in Fig.1b.

Redundancy among cameras is exploited by means of a Shape-from-Silhouette (SfS) technique [6]. This process generates a discrete occupancy representation of the 3D space (voxels). A voxel is labelled as foreground or background by checking the spatial consistency of its projection on of the N segmented silhouettes. The data obtained with this 3D reconstruction is corrupted by spurious voxels introduced due to wrong segmentation, camera calibration inaccuracies, etc. A connectivity filter is introduced in order to remove these voxels by checking its connectivity consistency with its spatial neighbors. An example of the output of the whole 3D processing module is depicted in Fig.1c.

The resulting binary 3D scene reconstruction is fed to the proposed tracking system that assigns a particle filter to each target. Finally, a higher semantic analysis is performed over the resulting tracks. Information about the environment (dimensions of the room, furniture, etc.) allow discarding tracks that are clearly wrong.

3. 3D TRACKING ALGORITHM

Particle Filtering (PF) is an approximation technique for estimation problems where the variables involved do not hold Gaussianity uncertainty models and linear dynamics. The current tracking scenario can be tackled by means of this algorithm to estimate the 3D position of a person $\mathbf{x}_t = (x, y, z)_t$ at time t, taking as observation a set of binary voxels representing the 3D scene up to time t denoted as $\mathbf{z}_{1:t}$.



(a) (b) (c)

Fig. 1. In (a), a sample of multiview original images. In (b), foreground segmentation of the input images employed by the SfS algorithm. In (c), example of the binary 3D voxel reconstruction used in this paper (false colors are employed to depict various people).

Multiple people might be tracked assigning a PF to each target and defining an interaction model to ensure track coherence.

For a given target $\mathbf{x}_t$, PF approximates the posterior density $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ with a sum of N_s Dirac functions:

$$p\left(\mathbf{x}_t|\mathbf{z}_{1:t}\right) \approx \sum_{j=1}^{N_s} w_t^j \delta(\mathbf{x}_t - \mathbf{x}_t^j), \tag{1}$$

where w_t^j are the weights associated to the particles and $\mathbf{x}_t^j$ their positions. For this type of tracking problem, a Sampling Importance Re-sampling (SIR) PF is applied to drive particles across time [7]. Assuming importance density to be equal to the prior density, weight update is recursively computed as:

$$w_t^j \propto w_{t-1}^j \; p(\mathbf{z}_t|\mathbf{x}_t^j). \tag{2}$$

SIR PF avoids the particle degeneracy problem by re-sampling at every time step. In this case, weights are set to $w_{t-1}^j = 1/N_s, \forall j$, therefore

$$w_t^j \propto p(\mathbf{z}_t|\mathbf{x}_t^j). \tag{3}$$

Hence, the weights are proportional to the likelihood function that will be computed over the incoming volume $\mathbf{z}_t$ as defined in Sec.3.1. The re-sampling step derives the particles depending on the weights of the previous step, then all the new particles receive a starting weight equal to $1/N_s$ which will be updated by the next volume likelihood function.

Finally, the best state at time t of target m, $\mathbf{X}_t^m$, is derived based on the discrete approximation of Eq.1. The most common solution is the Monte Carlo approximation of the expectation as

$$\mathbf{X}_t^m = \mathbb{E}\left[\mathbf{x}_t|\mathbf{z}_{1:t}\right] \approx \frac{1}{N_s} \sum_{j=1}^{N_s} w_t^j \mathbf{x}_t^i. \tag{4}$$

The major limit of PF, and specially SIR ones, is the capability of the particle set of representing the *pdf* when the sampling density of the state space is low. Scenarios with high number of degrees of freedom require a large number of particles to perform an efficient estimation with the consequent increase in terms of computational cost. An unnecessary computational load could appear with a number of particles larger than required.

Up to authors knowledge, the novelty of the proposed of scheme is to employ the minimum unit of the scene, the voxel, to redefine state space sampling. Being our volume a discrete representation, particles are constrained to occupy a single voxel and move with displacements on the 3D discrete orthogonal grid.

3.1. Likelihood Evaluation

Function $p(\mathbf{z}_t|\mathbf{x}_t)$ can be defined as the likelihood of a particle belonging to the volume corresponding to a person. For a given particle j occupying a voxel, its likelihood may be formulated as

$$p(\mathbf{z}_t|\ \mathbf{x}_t^j) = \frac{1}{|\mathcal{C}(\mathbf{x}_t^j, q)|} \sum_{\mathbf{p}\in\mathcal{C}(\mathbf{x}_t^j, q)} d(\mathbf{x}_t^j, \mathbf{p}), \tag{5}$$

where $\mathcal{C}(\cdot)$ stands for the neighborhood over a connectivity q domain on the 3D orthogonal grid and $|\mathcal{C}(\cdot)|$ represents its cardinality. Typically, connectivity in 3D discrete grids can be 6, 14 and 26 and in our research $q = 26$ provided accurate results. Function $d(\cdot)$ measures the distance between a foreground voxel $\mathbf{p}$ in the neighborhood and the particle.

Ideally, particles placed inside the volume of the target achieve maximum likelihood while those being on the surface of the volume attain a non-zero value. Volumes belonging to people would be completely solid but, in practice, there are holes introduced as the effect of segmentation inaccuracies during the SfS reconstruction.

3.2. 3D Discrete Re-sampling

The re-sampling step has been defined according to the condition that every particle is assigned to a foreground voxel. In other words, re-sampling has usually been defined as a process where some noise is added to the position of the re-sampled particles according to their weights [7]. The higher the weight, the more replicas will be created. In our current tracking scenario, re-sampling adds some *discrete* noise to particles only allowing motion within the 3D discrete positions of adjacent foreground voxels as depicted in Fig.2a. Then, non populated foreground voxels are assigned to re-sampled particles. In some cases, there are not enough adjacent foreground voxels to be assigned, then a connectivity search finds closer non-empty voxels to be assigned as shown in Fig.2b.

(a) (b)

Fig. 2. Discrete re-sampling example (in 2D).

No motion model has been assumed in the space state in order to keep a reduced dimensionality of our estimation problem. However, object translations are captured within the re-sampling step by means of this particle set expansion leading to satisfactory results.

(a) (b)

Fig. 3. Particles from the tracker A (yellow ellipsoid) falling into the exclusion zone of tracker B (green ellipsoid) will be penalized by a multiplicative factor $\alpha \in [0, 1]$.

3.3. Multi-Person PF Tracking

Challenges in 3D multi-person tracking from volumetric scene reconstruction are basically twofold. First, finding an interaction model in order to avoid missmatches and target merging. The second is filtering spurious objects that appear in scene reconstruction and discarding non-relevant objects such as chairs or furniture. This last problem is managed by the last module of the system that performs a higher semantic analysis of the scene.

Several approaches have been proposed [3, 5] but the joint PF presented in [8] is the optimal solution to multi-target tracking using PFs. However, its computational load increases dramatically with the number of targets to track since every particle estimates the location of all targets in the scene simultaneously. The proposed solution is to use a split PF per person, which requires less computational load at the cost of not being able to solve some complex cross-overs. However, this situation is alleviated by the fact that cross-overs are restricted to the horizontal plane in our scenario (see Fig.3a).

Let us assume that there are M independent PF trackers, being M the number of humans in the room. Nevertheless, they are not fully independent since each PF can consider voxels from other tracked targets in both the likelihood evaluation or the 3D re-sampling step resulting in target merging or identity missmatches. In order to achieve the most independent set of trackers, we consider a blocking method to model interactions. Many blocking proposals can be found in 2D tracking related works [8, 11] and we extend it to our 3D case. Blocking methods penalize particles that overlap zones with other targets. Hence, blocking information can be also considered when computing the particle weights as:

$$w_t^j = \frac{1}{N_s}\, p(z_t|x_t^j) \prod_{\substack{k=1 \\ k\neq m}}^{M} \beta\left(X_{t-1}^m, X_{t-1}^M\right), \qquad (6)$$

where M is the total number of trackers, m the index of the evaluated tracker and X is the estimated state. Term $\beta(\cdot)$ is the blocking function defining exclusion zones that penalize particles that fall into them. For our particular case, considering that people in the room are always sitting or standing up (this is a meeting room so we assume that they never lay down), a way to define an exclusion region modeling the human body is by using an ellipsoid with fixed x and y axis. Axis in z is a function of the estimated centroid height. An example of this exclusion technique is depicted in Fig.3. Tracked objects that come very close can be successfully tracked even though their volumes have partially merged.

Num.Particles	MOTP	$\overline{m}$	$\overline{fp}$	$\overline{mme}$	MOTA
50	222	27.7%	14.7%	47.5%	9.9%
100	206	64.9%	14.4%	8.5%	65.0%
150	193	74.9%	15.1%	6.7%	74.9%
300	187	81.4%	24.2%	9.7%	81.4%
600	**185**	**81.1%**	**9.4%**	**18.1%**	**81.2%**
1000	188	79.8%	9.9%	16.0%	80.0%

Table 1. Quantitative results for a tracking experiment in the better case with voxel size of 2 cm. Legend: misses ($\overline{m}$), false positives ($\overline{fp}$) and mismatches ($\overline{mme}$).

4. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed algorithm, we collected a set of multi-view scenes in an indoor scenario involving up to 6 people, for a total of approximately 25 min. The analysis sequences were recorded with 5 fully calibrated and synchronized wide angle lense cameras in the SmartRoom at UPC with a resolution of 720x576 pixels at 25 fps (see a sample in Fig.1). The test environment is a 5m by 4m room with occluding elements such as tables and chairs. Groundtruth data was labelled manually allowing a quantitative measure of tracker's performance. It should be noted that the employed test database has been included in the CLEAR06 Evaluation [9].

Metrics proposed in [12] for multi-person tracking evaluation have been adopted. These metrics, being used in international evaluation contests [9] and adopted by several research projects such as the European CHIL [13] or the U.S. Vace [14] allow objective and fair comparisons. Two metrics employed are: the **M**ultiple **O**bject **T**racking **P**recision (*MOTP*), which shows tracker's ability to estimate precise object positions, and the **M**ultiple **O**bject **T**racking **A**ccuracy (*MOTA*), which expresses its performance at estimating the number of objects, and at keeping consistent trajectories. *MOTP* scores the average metric error when estimating multiple target 3D centroids, while *MOTA* evaluates the percentage of frames where targets have been missed, wrongly detected or mismatched.

Two parameters drive the performance of the algorithm: the voxel size ν and the number of particles. Experiments carried out explore the influence of these two variables on the *MOTP* and *MOTA* scores as depicted in Fig.4. This plot shows how scenes reconstructed with a large voxel size do not capture well all spatial details and may miss some objects thus decreasing performance of the tracking system. Furthermore, the larger the number of particles the more accurate the performance of the algorithm; however, no substantial improvement is achieved for more than 600 particles due to the restriction imposed that every particle occupies the size of one voxel. Visual results of these effects are depicted in Fig.5.

Quantitative results given by the best set of parameters are summarized in Table 1 where the best performance is achieved when $\nu = 2$cm and 600 particles are employed for each target. It should be noted that for a large number of particles, i.e. 1000, re-sampling is not able to find enough foreground voxels where to place all particles thus its performance decreases.

5. CONCLUSION AND FUTURE WORK

This paper presented a multi-person tracking system in a multiple camera view environment. Redundant information among cameras is exploited to produce 3D information that is employed by the proposed tracker. A volume likelihood function and a discrete 3D re-

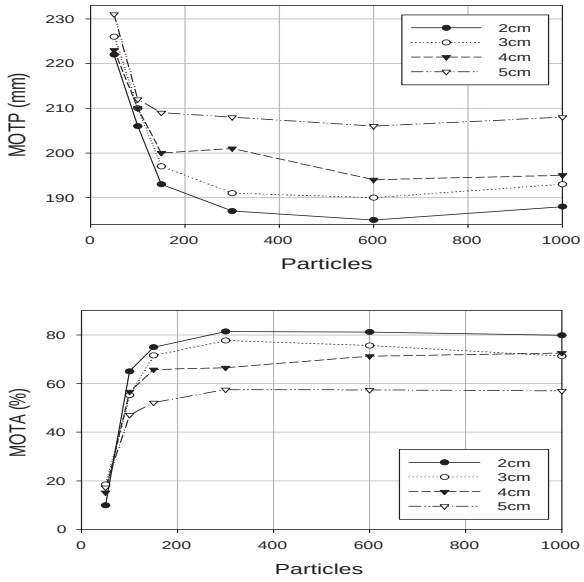
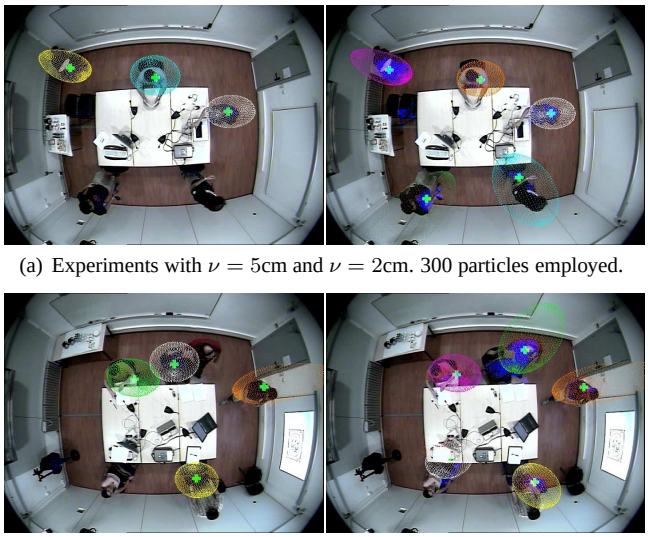


Fig. 4. *MOTP* and *MOTA* scores for various voxels sizes and number of particles. Low MOTP and high MOTA scores are preferred indicating low metric error when estimating multiple target 3D positions and high tracking performance.



(a) Experiments with $\nu = 5$cm and $\nu = 2$cm. 300 particles employed.

(b) Experiments with 100 and 300 particles. Voxel size set to $\nu = 2$cm.

Fig. 5. Zenital view of two comparative experiments. In (a), two tracking runs showing that large voxel reconstructions miss some objects. In (b), two tracking runs in a scene involving sudden motion showing how a reduced number of particles filter lose track of one target.

sampling have been introduced as an effective approach for this 3D PF implementation. Target interactions modeled through the blocking technique employed in this work allowed tracking multiple objects and resolving cross-overs and missmatches among targets.

Promising results obtained over a large test database proved the effectiveness of our technique. Compared to the results provided by the CLEAR Evaluation [9], our system would had been ranked on the 2nd position over 10 participants.

Future research within this topic involve the incorporation of additional modalities such as color to the obtained binary voxel reconstruction of the scene towards increasing system robustness. Color information might be helpful in maintaining the identity of every moving object when they get very close thus resolving mismatch of targets. Real-time implementations of the presented algorithm are under study.

6. REFERENCES

[1] N. Checka, K.W. Wilson, M.R. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2004, vol. 5, pp. 881–884.

[2] D. Focken and R. Stiefelhagen, "Towards vision-based 3D people tracking in a smart room.," in *IEEE Int. Conf. on Multimodal Interfaces*, 2002, pp. 400–405.

[3] K. Bernardin, T. Gehrig, and R. Stiefelhagen, "Multi and single view multiperson tracking for SmartRoom environments," in *CLEAR Evaluation Workshop*, 2006.

[4] C. Canton-Ferrer, J. R. Casas, and M. Pardàs, "Towards a Bayesian approach to robust finding correspondences in multiple view geometry environments," in *Lecture Notes on Computer Science*, 2005, vol. 3515, pp. 281–289.

[5] O. Lanz, "Approximate Bayesian multibody tracking," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 28, no. 9, pp. 1436–1449, 2006.

[6] G.K.M. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler, "A real time system for robust 3D voxel reconstruction of human motions," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2000, vol. 2, pp. 714–720.

[7] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *Signal Processing, IEEE Tran. on*, vol. 50, no. 2, pp. 174–188, 2002.

[8] Z. Khan, T. Balch, and F. Dellaert, "Efficient particle filter-based tracking of multiple interacting targets using an MRF-based motion model," in *Int. Conf. on Intelligent Robots and Systems*, 2003, vol. 1, pp. 254–259.

[9] "CLEAR Evaluation Campaign," http://www.clear-evaluation.org, 2006.

[10] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 1999, pp. 252–259.

[11] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," *Int. Journal of Computer Vision*, vol. 39, no. 1, pp. 57–71, 2000.

[12] K. Bernardin, A. Elbs, and R. Stiefelhagen, "Multiple object tracking performance metrics and evaluation in a smart room environment," in *IEEE Int. Workshop on Vision Algorithms*, 2006, pp. 53–68.

[13] "CHIL-Computers In the Human Interaction Loop," http://chil.server.de.

[14] "VACE-Video Analysis and Context Extraction," http://www.ic-arda.org.