MULTIPERSON 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 representation of scene. We apply a particle filter. Interactions among people are modeled though a blocking scheme thus allowing multiple target tracking. Test over large annotated databases are performed thus obtaining quantitative results showing the effectiveness of the proposed algorithm in the indoor scenarios.

Index Terms— Falten paraules

1. INTRODUCTION

The current paper addresses the problem of detecting and tracking a group of people present in an indoor scenario in the framework of multiple view geometry. Robust, multi-person tracking systems are a basic functionality that have been employed in a wide rangle 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 to be robust under occlusion created by moving and fixed 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 circumvent these drawbacks, multi-camera tracking techniques [3] exploit spatial redundancy among different views and provide 3D information as well. Integration of information and feature extraction coming from multiple cameras has been proposed in terms of multi-view histograms [4], image correspondences [5] or voxel reconstructions [6].

Filtering techniques are employed to grant temporal consistency to tracks. Kalman filter based solutions have been extensively used to perform tracking of a single object under Gaussian uncertainty models and linear dynamics [7]. However, these methods do not perform accurately when facing noisy scenes or rapidly manouvering targets. Particle filtering have been applied to cope with these situations since it can deal with multi-modal *pdf*s and is able to recover from lost tracks [8, 9].

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 that is the input of the tracking system. A multi-target tracking scheme based on multiple interacting particle filters is introduced. Finally, efectiveness of the proposed algorithm is shown by means of objective metrics when applied to the CLEAR06 [10] 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.1(a)). Each camera is modeled using a pinhole camera model based on perspective projection. Accurate calibration information is available. Foreground regions from input images are obtained using a segmentation algorithm based on Stauffer-Grimson's background learning and substraction technique [11] as shown in Fig.1(b).

Redundancy among cameras is exploited by means of a Shape-from-Silhouette technique [6]. This process generates a discrete occupancy representation of the 3D space (voxels) deciding whether a voxel is foreground or background by checking the spatial consistency 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.1(b)

The resulting unlabeled 3D scene reconstruction is fed to a tracker that assigns a particle filter to each target.

Finally, a higher semantical analysis is performed over the resulting tracks. Information about the environment (dimensions of the room, furniture,etc.) allow discarding tracks that are not likely to be human people.

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, denoted as $\mathbf{z}_t$. Multiple people might be tracked asigning a PF for 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:

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

(a) (b)

Fig. 1. Example of the outputs from the 3D processing module in the SmartRoom scenario. In (a), multiview original and foreground images. In (b), example of the 3D voxel reconstruction used in this paper.

where w_t^j are the weights associated to the particles. For this type of tracking problem, a Sampling Importance Resampling (SIR) PF is applied to drive particles across time [8]. 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 $w_{t-1}^j = 1/N_s, \forall i$, therefore

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

Hence, the weights are proportional to the likelyhood function that will be computed over the incoming volume $\mathbf{z}_t$. 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 likelyhood 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

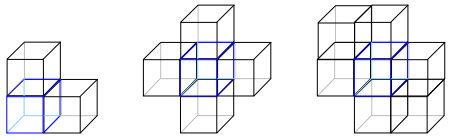
$$\mathbf{X}_t^m = \mathbb{E}\left[\mathbf{x}_t|\mathbf{z}_{1:t}\right] \approx \frac{1}{N_s}\sum_{j=1}^{N_s} w_t^j \mathbf{x}_t^i. \tag{4}$$

3.1. Likelyhood evaluation

Function $p(\mathbf{z}_t|\mathbf{x}_t)$ can be defined as the likelyhood of a particle belonging to the volume corresponding to a person. For a given particle j, its likelyhood 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 neighbourhood over a connectivity q domain on the 3D orthogonal grid and $|\mathcal{C}(\cdot)|$ its cardinality. Function $d(\cdot)$ measures the distance between a neighbourhood voxel



(a) (b) (c)

Fig. 2. Manca.

This means that the available weight at every frame is 1/Ns and the weights are proportional to the likelihood. In other words, weights are basically computed through voxel density analysis. Ideally, volume reconstruction would be completely solid, but in practice there are holes. Thats why reconstruction is seen as voxel clouds. For every particle a voxel neighbourhood is evaluated, counting full and empty voxels. The maximum likelihood is obtained when full voxels are equal to empty voxels.

3.2. Particle Evaluation

The limit of PF, and specially SIR ones, is the capability of the set of particles to represent the pdf when the sampling density of the sate 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. In our case, to avoid over-sampling, we use the minimum unit of the scene to redefine sampling: the voxel. Once the resampling step has been performed, every particle is assigned to a voxel. No motion model has been used. Human motion is very restricted in smart room environments and the particle set gets expanded enough in the resampling step to track a target. To summarize, the basic proposed PF algorithm include the following steps:

- Particle Redrawing: Every particle is set into a voxel by searching the particles neighbourhood.
- Weight computation: Surrounding voxels are evaluated for every single particle in order to estimate its weight. Then weights are normalized to compute particles centroid. In other words, human positions X_k are the mean of all their associated particles:

$$X_k = \sum_{i=1}^{N_s} w_k^i x_k^i \tag{6}$$

- Resampling: Particles are re-sampled according to their weights. The higher weight the more replicas will be created. A uniform distribution has been proposed to expand the particles.

3.3. Multi-person PF Tracking

Challenges in 3D multi-person tracking with 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. However,

since filtering step belongs to data acquisition we will focus this section on interaction model.

The Joint Particle Filter is the optimal solution to PF multi-target tracking, but its computational load increases dramatically with the state space dimension. In a joint PF every particle estimates the location of all targets in the scene. The proposed solution is to use an Split PF per person., which requires less computational computations. The initial assumption is that we have M independent trackers, being M the number of humans in the room, but in fact they are not fully independent because each PF can consider other targets to track. In order to achieve the most independently set of trackers, we consider a blocking method to model interactions. Many blocking proposals can be found in 2D tracking related works [12, 9]. Blocking methods penalize particles that overlap zones with other targets. In other words, we also consider blocking information to compute the final weights:

$$w_k^j = \frac{1}{N_s} p\left(z_k | x_k^j\right) \prod_{j=1; j \neq m}^{M} \beta\left(X_{k-1}^m, X_{k-1}^M\right) \qquad (7)$$

where M is the number of trackers, m the index of the evaluated tracker, X the estimated state and $\beta(\cdot)$ is the blocking function. To penalize particles we define exclusion zones. Considering that people in the room are always sitting or standing up, they never lay down, the easiest way to define a region to model the human body is by using an ellipsoid with fixed X-axis and Y-axis. Z-axis is a function of the estimated centroid height. Fixed axis wont propagate estimation errors thus blocking becomes more robust. Any particle from the tracker A into the ellipsoid of a tracker B will be penalized by , with 0¡1, which can be defined as a constant or as a function of the distance between the particle and the A PF estimated centroid. This technique is very interesting in our scenario because when two people are very near, their volumes merge and become indistinguishable for their respective trackers.

4. 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 chair. Groundtruth data was labelled manually allowing a quantitative measure of tracker's performance.

Metrics proposed by [13] for multi-person tracking evaluation have been adopted. These metrics, being used in international evaluation contests [10] and adopted by several research projects such as the European CHIL [14] or the U.S. Vace [15] allow objective and fair comparisons. Two employed metrics 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 targets 3D centroid, 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 and the number of particles. Experiments carried out explored the influence of these two variables on the *MOTP* and *MOTA* scores as depicted in Fig.3. This plot shows how scenes reconstructed with a large voxel size do not capture well all spatial details and may miss some small 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 substancial improvement is acchieved for more than 300 particles due to the restriction imposed that every particle occupies the size of one voxel. Quantitative results for are shown in Table 1.

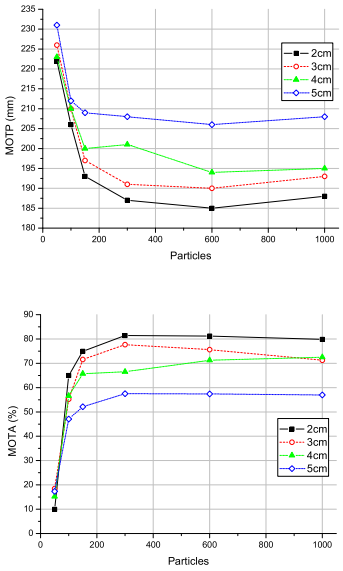


Fig. 3. *MOTP* and *MOTA* scores for various voxels sizes and number of particles.

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%

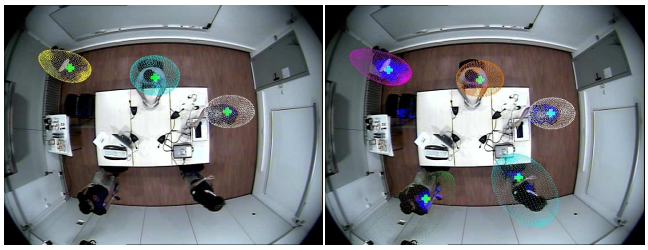
Table 1. Quantitative results for a tracking experiment in the better case with voxel size of 2 cm^3.

5. CONCLUSION AND FUTURE WORK

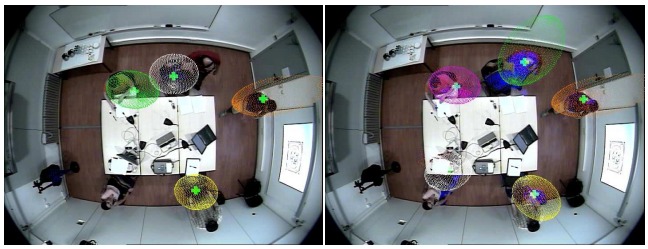
This paper presented a multi-person tracking system in a multiple camera views environment. Redudant information among cameras is exploited to generate a 3D reconstruction of the scene described by voxels.

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.



(a) Experiments with $\nu = 5\text{cm}^3$ and $\nu = 2\text{cm}^3$. 300 particles employed.



(b) Experiments with 100 and 300 particles. Voxel size set to $\nu = 2\text{cm}^3$.

Fig. 4. Zenital view of two comparative experiments. In (a), two tracking runs showing that large voxel reconstructions miss small 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.

[2] J.L. Landabaso, L.Q. Xu, and M. Pardàs, "Robust tracking and object classification towards automated video surveillance," in *Int. Conf. on Image Analysis and Recognition*, 2004.

[3] Keni Bernardin, Tobias Gehrig, and Rainer Stiefelhagen, "Multi- and Single View Multiperson Tracking for Smart Room Environments," in *CLEAR Evaluation Workshop*, 2006.

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

[5] 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.

[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] 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.

[8] 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 Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, vol. 50, no. 2, pp. 174–188, 2002.

[9] 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.

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

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

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

[13] 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.

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

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