3D Tracking Using Multi-view Based Particle Filters

Raúl Mohedano, Narciso García, Luis Salgado, and Fernando Jaureguizar

Grupo de Tratamiento de Imágenes, Universidad Politécnica de Madrid,
28040, Madrid, Spain
{rmp,narciso,lsa,fjn}@gti.ssr.upm.es
http://www.gti.ssr.upm.es

Abstract. Visual surveillance and monitoring of indoor environments using multiple cameras has become a field of great activity in computer vision. Usual 3D tracking and positioning systems rely on several independent 2D tracking modules applied over individual camera streams, fused using geometrical relationships across cameras. As 2D tracking systems suffer inherent difficulties due to point of view limitations (perceptually similar foreground and background regions causing fragmentation of moving objects, occlusions), 3D tracking based on partially erroneous 2D tracks are likely to fail when handling multiple-people interaction. To overcome this problem, this paper proposes a Bayesian framework for combining 2D low-level cues from multiple cameras directly into the 3D world through 3D Particle Filters. This method allows to estimate the probability of a certain volume being occupied by a moving object, and thus to segment and track multiple people across the monitored area. The proposed method is developed on the basis of simple, binary 2D moving region segmentation on each camera, considered as different state observations. In addition, the method is proved well suited for integrating additional 2D low-level cues to increase system robustness to occlusions: in this line, a naïve color-based (HSI) appearance model has been integrated, resulting in clear performance improvements when dealing with complex scenarios.

1 Introduction

Tracking multiple people in both indoor and outdoor environments is a very active research topic due to its applicability to surveillance systems, security and restricted area control, intelligent rooms, etc. Whilst many works have largely addressed 2D tracking [8], many potential capabilities of 3D positioning and tracking of interest targets in multi-camera environments are not yet been studied.

A frequent approach for tracking multiple people in multi-camera environments assumes a ground plane restriction: interest objects move on a visible plane (ground), making it possible to establish homographies relating different views [2], and then combining different trajectories onto the ground plane. Although this method shows effective in several situations, ground plane restriction does not hold for many interesting environments (specially for indoor scenarios).

Calibration of cameras allows more sophisticated processing. It allows, for instance, to perform 2D tracking on each of the cameras independently, and subsequently combining 2D tracks into 3D world using only geometrical considerations [5], or both geometry and appearance consistency [10]. This approach relies on 2D tracking, which suffers from certain limitations due to camera point of view, occlusions, etc. Mistakes derived from erroneous decisions at 2D tracking level translate into 3D tracking failures.

Probabilistic combination of multiple cameras observations avoids the loss of valuable information due to hard decisions at 2D tracking level. In [4], occupancy probability projection over ground plane is estimated using background subtraction on individual cameras. More oriented towards accurate 3D segmentation, [11] proposes space voxelization, shape from silhouette and voxel grouping for 3D positioning and tracking. This approach provides powerful information for robust tracking and scene understanding, but voxelization represents a great computational cost, specially if detailed 3D segmentation is desired. A non-uniform volume partition could overcome this limitation, paying attention only to interesting areas.

The 3D positioning and tracking system proposed in this paper also intends to monitor a limited area, covered with several cameras, where multiple objects of interest can enter, interact, and exit. It relies on a set of 2 or more fully-calibrated cameras, each performing independently a standard motion-region (binary) segmentation subject to problems such as fragmentation of objects and false detections, and considering also color information. Loss of 2D information in hard decisions at 2D level is avoided, as it fuses information directly into 3D world using 3D Particle Filters [1]. Proposed 3D Particle Filters are specially suited for 3D tracking over time, and they can even provide accurate 3D description of shapes, avoiding the computational cost of voxelization.

This paper introduces a complete framework for combining different cues from multiple cameras into a single, consistent measure of the observation likelihood of the 3D Particle Filter. This measure allows updating a volumetric occupancy probability density function for each tracked person over time, making 3D segmentation and tracking posible, as described in Section 2. Subsection 2.1 presents the proposed multi-camera 3D tracking method using motion-region (binary) segmentations in each sensor as system measurement, while Subsection 2.2 extends the system including color as an additional cue for 3D tracking. Subsection 2.3 addresses the dynamic models for the 3D Particle Filters of both systems, whose results are discussed in Section 3. Finally, Section 4 outlines the achievements reached by the proposed methods.

2 State-Space Models for 3D Segmentation and Tracking

The state-space approach for modeling dynamic systems has proved suitable for robust estimation from several noisy information sources. This approach forms the basis of the different Bayesian Tracking methods [1] (e.g. Kalman Filter,

Particle Filters), that are broadly used and specially interesting for real-time tasks as they are well suited for recursive implementation.

The Recursive Bayesian tracking has traditionally been applied for on-line probabilistic estimation of trajectories and shapes of isolated targets, and mainly for 2D objects. Most problems are modeled as a state vector $\mathbf{x}_t$ at each time step (e.g. position-velocity, position-rotation angles, or control points of a parametrized contour [9]). This $\mathbf{x}_t$ is estimated accurately bearing in mind every observation up to time step t ($\mathbf{Z}^t$) by means of its posterior likelihood $p(\mathbf{x}_t|\mathbf{Z}^t)$.

The posterior likelihood at time step t can be expressed in terms of that from time step $t-1$ through

$$p(\mathbf{x}_t|\mathbf{Z}^t) = \frac{p(\mathbf{z}_t|\mathbf{x}_t)\; p(\mathbf{x}_t|\mathbf{Z}^{t-1})}{p(\mathbf{z}_t|\mathbf{Z}^{t-1})}, \tag{1}$$

where $\mathbf{z}_t$ represents observations at time step t. The dynamic model governing state evolution is expressed in $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, and the prior distribution $p(\mathbf{x}_{t-1}|\mathbf{Z}^{t-1})$ is considered available as it has been estimated in the previous time step $t-1$. The predicted probability density function $p(\mathbf{x}_t|\mathbf{Z}^{t-1})$ is updated through (1) using $p(\mathbf{z}_t|\mathbf{x}_t)$, which shows the likelihood of the observation $\mathbf{z}_t$ given the state $\mathbf{x}_t$. The posterior likelihood $p(\mathbf{x}_t|\mathbf{Z}^t)$ can be approximated using Monte Carlo methods [3], which deal with sampled versions of distributions. This approach is known as Particle Filter [1].

The core idea of this paper is to estimate the volumetric occupancy probability density function $p(\mathbf{x}_t|\mathbf{Z}^t, H_k)$ of a person H_k (given that H_k is present in the scene) using a 3D Particle Filter, and considering different views from a set of fully-calibrated cameras as system observations. Using this pdf, the probability that H_k is contained in a volume V is

$$P(H_k \subset V \mid H_k) = \iiint_V p(\mathbf{x}_t|\mathbf{Z}^t, H_k)\, \mathrm{d}V \tag{2}$$

According to (2), H_k spatial position could be identified with *the minimum volume V_k that contains H_k with probability greater or equal to P_H*, given that H_k is present in the scene (where, evidently, $0 < P_H < 1$). The volume V_k should then be considered as a *3D bounding volume* for H_k, analogous to the classical bounding box concept but much more general. The volumetric occupancy pdf evolves over time according to the dynamic model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and the observations at time step t, allowing 3D tracking. Following the same reasoning, it is possible to estimate $p(\mathbf{x}_t|\mathbf{Z}^t, H_k)$, $k = 1, \ldots, N_P$ using one individual 3D Particle Filter for each moving object H_k, and thus track several people simultaneously.

Combining different cues from multiple cameras into a single and consistent measure of the observation likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$ is essential for updating $p(\mathbf{x}_t|\mathbf{Z}^t, H_k)$ over time. Supposing that the scene is monitored using M different cameras, the state observation at time step t can break down into camera contributions according to

$$\mathbf{z}_t = \left(\mathbf{z}_t^{c_1}, \mathbf{z}_t^{c_2}, \ldots, \mathbf{z}_t^{c_M}\right). \tag{3}$$

Therefore, the observation likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$ can be rewritten as

$$p(\mathbf{z}_t|\mathbf{x}_t) = p\big(\mathbf{z}_t^{c_1}, \mathbf{z}_t^{c_2}, \ldots, \mathbf{z}_t^{c_M}|\mathbf{x}_t\big). \tag{4}$$

For computation purposes, it would be desirable to decompose $p(\mathbf{z}_t|\mathbf{x}_t)$ into factors concerning observations from each camera independently. It would be possible if we assume that the cameras are conditionally independent given $\mathbf{x}_t$. Besides, although the F different features $(f_1, f_2, \ldots, f_F)$ in each camera c_j are statistically dependent, they could be considered nearly conditionally independent given $\mathbf{x}_t$. Therefore, this reasoning justifies processing each feature on each camera view independently using individual likelihood measurement models. The complete observation likelihood measure $p(\mathbf{z}_t|\mathbf{x}_t)$ can be finally obtained by simply multiplying all the contributions considered.

Next subsections propose measurement models for two different cues obtained from cameras: moving-region (binary) segmentation and color. The former expresses whether a particular camera has detected movement in the environment of the projection of a particle onto the camera plane, indicating that this particle is likely to be contained into a moving object. As for color, it provides a simple but effective appearance description of objects. Two different systems have been implemented to prove the abilities of the proposed method: the first one is based only on moving-region segmentations, and the second incorporates also color cues into the measurement model. The performance of both methods is compared in Section 3.

2.1 Measurement Model from 2D Motion Segmentation of Views

A common starting point for tracking systems is motion-region segmentation. Detecting moving regions in video sequences usually provides a focus of attention for latter processing, as it aims to discover image areas corresponding to interest objects. Conventional methods for motion segmentation are temporal differencing, optical-flow computation and background subtraction [8]. The quality of the motion-region segmentation shows a strong dependency on the complexity of the monitored scene, and it conditions clearly the performance of subsequent processes: 2D tracking can be seriously damaged as a result of inaccurate segmentations, and thus 3D tracking derived from 2D tracking in several cameras is prone to fail.

The proposed method aims to extract as much information as possible from (generally) inaccurate motion-region segmentations from several overlapping cameras by combining cues directly in the 3D world. The key point to do it is to interpret motion segmentation information as a measurement of the state $\mathbf{x}_t$, and then to integrate them into the 3D Particle Filter framework discussed previously.

Let us consider, first of all, only one camera (c_j). Let $M_t^{c_j}$ be the binary mask (image) obtained as a result of the motion-region segmentation performed on the image acquired by camera c_j at time step t. This mask has an associated domain R representing image areas where movement has been detected, and can be expressed as

$$M_t^{c_j}(\mathbf{r}) = \begin{cases} 1 & \forall \mathbf{r} \in \mathrm{R} \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

Let $\mathbf{y}_{c_j} = P_{c_j}(\mathbf{x}_t)$ be the projection of the 3D position of $\mathbf{x}_t$ onto the camera c_j image plane. Thus $\mathbf{y}_{c_j}$ is a 2D vector expressed in pixel units. If $\mathbf{x}_t$ is contained in a 3D moving object, we might expect $M_t^{c_j}(\mathbf{y}_{c_j})$ to be 1. Although it is intuitive that $p(\mathbf{z}_t^{c_j}|\mathbf{x}_t)$ is closely related to the value of $M_t^{c_j}$ in its projected position, some considerations are needed.

If a spatial position $\mathbf{x}_t$ is actually part of the tracked person H_k, then certain neighborhood of it should be also part of H_k. So, $p(\mathbf{z}_t^{c_j}|\mathbf{x}_t)$ should be determined not only by the value of $M_t^{c_j}$ at the projected point $\mathbf{y}_{c_j}$, but also by its 2D neighborhood. As some pixels contain more information about $\mathbf{x}_t$ than others, it is clear that their contribution to $p(\mathbf{z}_t^{c_j}|\mathbf{x}_t)$ should be weighted bearing in mind their distance to the ideal projection of $\mathbf{x}_t$. The proposed measurement model sets weights according to a bivariate normal distribution $\mathcal{N}(\mathbf{r}; \mathbf{y}_{c_j}, \Sigma)$, whose covariance matrix is $\Sigma = \sigma^2 I_2$ (where I_2 is the 2×2 identity matrix). All considered, and using a discrete version of the discussed normal distribution with probability mass function $g(\mathbf{r})$ centered at $\mathbf{r} = 0$, camera c_j contribution to the observation likelihood can be written as

$$p(\mathbf{z}_t^{c_j}|\mathbf{x}_t) = \sum_{\forall \mathbf{r}} g(\mathbf{r} - \mathbf{y}_{c_j})\, M_t^{c_j}(\mathbf{r}). \tag{6}$$

As $g(\mathbf{r})$ is symmetrical with respect to the origin of coordinates, $g(\mathbf{r}) = g(-\mathbf{r})$ and then $p(\mathbf{z}_t^{c_j}|\mathbf{x}_t)$ can be rewritten using the convolution operator as a simple image filtering, yielding

$$p(\mathbf{z}_t^{c_j}|\mathbf{x}_t) = \sum_{\forall \mathbf{r}} g(\mathbf{y}_{c_j} - \mathbf{r})\, M_t^{c_j}(\mathbf{r}) = \Big[\, g(\mathbf{r}) * M_t^{c_j}(\mathbf{r}) \Big]_{\mathbf{r}=\mathbf{y}_{c_j}}. \tag{7}$$

Although it is not strictly necessary due to the normalization performed at weight particle updating, expression (7) can be proved a probability distribution given $\mathbf{x}_t$.

Additionally, it is not possible to assure that person H_k is not at 3D position $\mathbf{x}_t$ even if no trace of it can be found in mask $M_t^{c_j}$. The system must let a certain uncertainty to handle these situations in which moving objects are not seen from a particular point of view (due to occlusions), and also possible segmentation errors. To express this constant uncertainty, $p(\mathbf{z}_t^{c_j}|\mathbf{x}_t)$ is limited to a minimum *background probability* p_B. Thus binary images $M_t^{c_j}$ provided by the motion-region segmentation module can be transformed into *motion likelihood images* $D_t^{c_j}$ according to

$$D_t^{c_j}(\mathbf{r}) = \max \begin{cases} g(\mathbf{r}) * M_t^{c_j}(\mathbf{r}) \\ p_B \end{cases}. \tag{8}$$

Finally, $p(\mathbf{z}_t^{c_j}|\mathbf{x}_t)$ can be directly taken from $D_t^{c_j}(\mathbf{y}_{c_j})$. Fig. 1 shows a particular example of motion likelihood image estimated from a motion-region segmentation.

(a) (b) (c)

Fig. 1. (a) Original image. (b) Binary motion-region segmentation of (a). (c) Motion likelihood image , using a filtering mask with $\sigma = 4$ pixels and background probability $p_B = 0.2$ on (b) (note the greyish background tone due to p_B).

Complete observation likelihood measure $p(\mathbf{z}_t|\mathbf{x}_t)$ using motion information from all cameras can be combined by multiplying individual contributions. Results based only on moving-region segmentation from four different cameras are discussed in Section 3.

2.2 Measurement Model from Color Cues

Motion detection in several overlapping cameras proves to be an effective cue for performing 3D tracking. However, an essential limitation should be pointed out: due to projective geometry fundamentals, assumptions made for a particular pixel apply equally to every 3D point along its back-projected ray [7]. This leads to serious ambiguities when tracking two or more interacting people, yielding 3D tracking failures (see Section 3). To overcome this drawback, it is essential to disambiguate between binary silhouettes using appearance modeling (*e.g.* color, texture,...). This subsection describes a simple color-based appearance model that, combined with the motion segmentation-based model presented in the previous subsection, results in a clear improvement of the system in robustness to occlusions.

Color is usually described according to the RGB color model, as it is closely related to hardware implementations. However, it is not suited for describing perceptual proximity of colors and, in addition, is strongly dependent on illumination conditions. This paper proposes thus a color model based on the HSI (Hue-Saturation-Intensity) [6], as it decouples intensity information (I channel) from the color-carrying information (contained in both H and S channels) and provides a perceptual-oriented description of colors.

Ignoring the resulting I channel, and ignoring possible conversion problems arising from low intensity levels, any RGB color can be robustly represented as two values: H and S. Hue and saturation describe a circular color model where H can be regarded as angular information (contained in the range 0°–360°, or between 0 and 1 after normalization) and S as distance from the origin (also normalized between 0 and 1). Consequently, appearance A is characterized using color according to the discussed color model can break down into

$$A = (H, S). \tag{9}$$

Initially, let us consider a single camera (c_j). The observed appearance at each pixel is described in terms of hue and saturation. It is clear that the H and S values of a certain pixel are not independent. However, once again, they can be considered approximately conditionally independent given $\mathbf{x}_t$ (assuming that state $\mathbf{x}_t$ conveys color information as well as spatial position), as the likelihood of measuring H (or S) is strongly conditioned by $\mathbf{x}_t$ appearance, and is completely determined when $\mathbf{x}_t$ is seen directly in camera c_j (*i.e.* it is not occluded). Therefore the color contribution to observation likelihood in camera c_j can be written as

$$p\big(A_t^{c_j}|\mathbf{x}_t\big) \approx p\big(H_t^{c_j}|\mathbf{x}_t\big)\, p\big(S_t^{c_j}|\mathbf{x}_t\big). \tag{10}$$

Let h_t and s_t be hue and saturation of $\mathbf{x}_t$, respectively, and let $\mathbf{y}_{c_j} = P_{c_j}(\mathbf{x}_t)$ be once again the projection of the 3D position of $\mathbf{x}_t$ onto camera c_j image plane. The hue (saturation) of pixel $\mathbf{y}_{c_j}$ should be approximately h_t (s_t), taking into account a certain Gaussian measurement noise with standard deviation σ_h (σ_s). Since both H and S have been normalized between 0 and 1, the standard deviation of measurement noise must be $\sigma_h, \sigma_s \ll 1$. Additionally, as discussed for the 2D motion segmentation cues, it is necessary to set a certain background level (h_B and s_B) to consider the possibility of occlusions. All considered, hue and saturation contributions to the observation likelihood measurement can be expressed as

$$p\big(H_t^{c_j}|\mathbf{x}_t\big) = h_B + (1 - h_B)\frac{1}{\sigma_h\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{d_h}{\sigma_h}\right)^2\right\}, \tag{11}$$

with

$$d_h = \min\left\{\left|H_t^{c_j}(\mathbf{y}_{c_j}) - h_t\right|,\ \frac{1}{2} - \left|H_t^{c_j}(\mathbf{y}_{c_j}) - h_t\right|\right\}, \tag{12}$$

and

$$p\big(S_t^{c_j}|\mathbf{x}_t\big) = s_B + (1 - s_B)\frac{1}{\sigma_s\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{S_t^{c_j}(\mathbf{y}_{c_j}) - s_t}{\sigma_s}\right)^2\right\}. \tag{13}$$

The use of this appearance measurement model, along with the motion segmentation-based model described in Subsection 2.1, proves effective for performing 3D tracking of multiple people.

2.3 Dynamic Model

The dynamic model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ represents the evolution of the volumetric occupancy probability density function $p(\mathbf{x}_t|\mathbf{Z}^t, H_k)$ over time. As the importance density $q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$ has been selected so that it is equal to $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, it means

that, in practice, the dynamic model itself governs the evolution of particles $\mathbf{x}_t^{(i)}$ over time.

Two different systems for 3D tracking have been developed: the first one considers only motion-region segmentation from multiple cameras as state observation, and the latter integrates color modeling along with visual motion detection.

- The system considering only motion-region segmentation works on $\mathbf{x}_t$ containing 3D position (x, y and z) and velocity ($\dot{x}$, $\dot{y}$ and $\dot{z}$). Dynamic model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ predicts position assuming constant velocity. Using matrix notation, prediction can be expressed as

$$\mathbf{x}_t = [x_t, y_t, z_t, \dot{x}_t, \dot{y}_t, \dot{z}_t]^T = \left[\begin{array}{c|c} I_3 & I_3 \\ \hline 0_3 & I_3 \end{array}\right] \mathbf{x}_{t-1} + \mathbf{n}_t, \tag{14}$$

  where I_3 and 0_3 are the 3×3 identity and zero matrices, respectively, and $\mathbf{n}_t$ represents the process noise vector [1]. Both position and velocity noise components follow normal distributions. Every direction is treated equally, having so identical power noise.
- The system considering both motion-region segmentation and color works on $\mathbf{x}_t$ containing 3D position and velocity, and also two color dimensions: h and s. Dynamic model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ predicts position assuming constant velocity (exactly as explained above), considering constant color dimensions. In this case, prediction can be written as

$$\mathbf{x}_t = [x_t, y_t, z_t, \dot{x}_t, \dot{y}_t, \dot{z}_t, h_t, s_t]^T = \left[\begin{array}{c|c|c} I_3 & I_3 & 0_{3,2} \\ \hline 0_3 & I_3 & 0_{3,2} \\ \hline 0_{2,3} & 0_{2,3} & I_2 \end{array}\right] \mathbf{x}_{t-1} + \mathbf{n}'_t, \tag{15}$$

  where $0_{n,m}$ is the $n \times m$ zero matrix, and $\mathbf{n}'_t$ the process noise vector for position, velocity and color.

3 Experiments and Results

This section shows and compares results of both proposed systems for 3D segmentation and tracking (the former using only motion-region segmentation in each camera, MS, and the latter considering also color cues, MS+C) on a highly cluttered scenario: a typical office room. This working environment has been monitored using four fully-calibrated static cameras placed in the four top corners of the room, having in addition overlapping fields of view (see Fig. 2). The complex working environment produces frequent occlusions (due to static, foreground objects) in some of the cameras: however, both MS and MS+C systems proves effective for tracking non-interacting targets. Both systems use $N_S = 1000$ particles to track each of the targets. Segmentation has been simply performed by estimating the V_k 3D bounding volume for each H_k as the 3D convex hull of the minimum set of particles of the H_k particle filter accumulating a probability

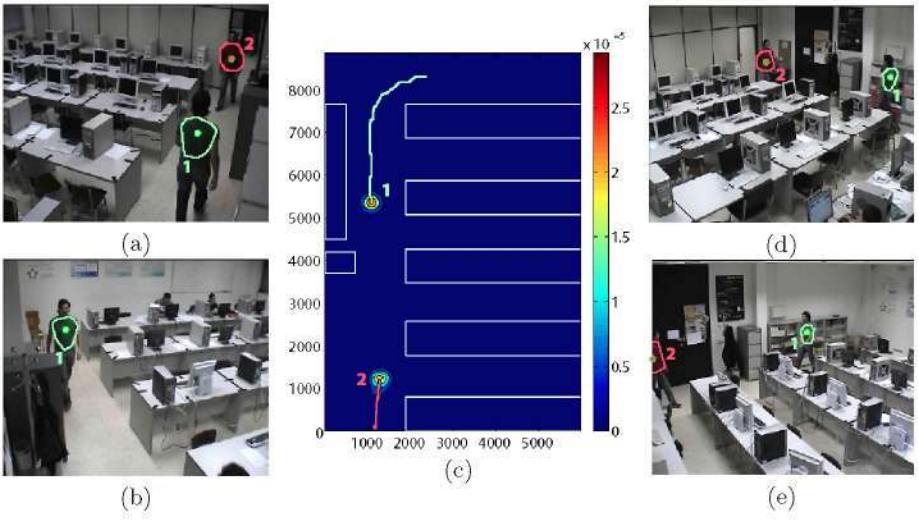

Fig. 2. (a)(b)(d)(e) Four different views of the monitored office environment, with convex hull 3D segmentation superimposed (using both motion segmentation and color). (c) Volumetric occupancy probability density for two different people (bird's eye view).

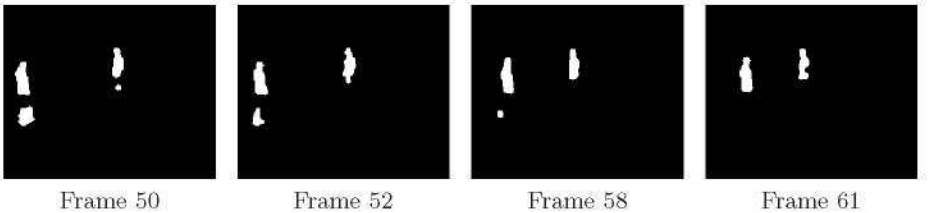

Fig. 3. Binary motion-region segmentation performed on images acquired by camera 1. Note fragmentation and inaccuracy of regions.

greater or equal to a certain limit P_H. The experimentation has proved that $P_H = 0.95$ provides good segmentation results.

Fig. 2 depicts 3D segmentation and tracking results of the MS+C system, with two different people in the scene. Fig. 2(c) shows level curves for both $p(\mathbf{x}_t|\mathbf{Z}^t, H_1)$ and $p(\mathbf{x}_t|\mathbf{Z}^t, H_2)$ projection into the ground plane. The volumetric probability density $p(\mathbf{x}_t|\mathbf{Z}^t, H_k)$ has been estimated by applying a 3D Gaussian kernel on the H_k particle set $\{\mathbf{x}_t^{(i)}, w_t^{(i)}\}_{i=1}^{N_S}$. The white lines represent the layout of the room furniture (office desks and bookcases), and have been included only for displaying purposes. Fig. 2(a)(b)(d)(e) show the four camera views, with estimated V_k 3D bounding volumes (convex hulls) for both present people superimposed. Tracking has been performed using a real motion detection method on images from the cameras, resulting in erroneous splitting and merging of regions due to the environment characteristics, as show in Fig. 3. The consistent results presented prove system robustness to defective motion segmentations.

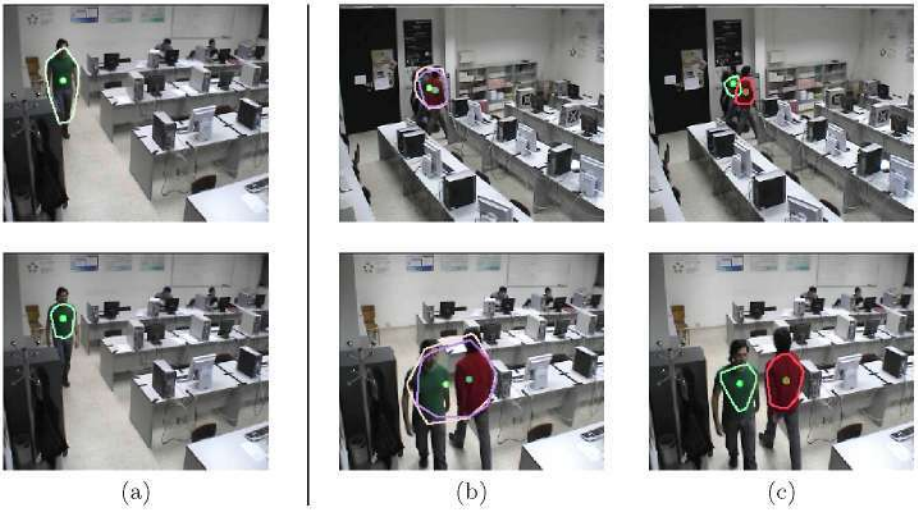

Fig. 4. System using only motion segmentation (MS) versus system using both motion and color (MS+C). (a) Spatial segmentation using MS (above) and MS+C (below). (b) Failed multiple people 3D tracking using MS. (c) Multiple people tracked correctly using MS+C.

Fig. 4(a) compares 3D segmentation precision for both systems. System using only motion-region segmentation shows better behavior as it proves able to distribute particles across targets. Motion segmentation and color-based system, however, tends to undergo particle degeneracy into one single dominant color. This tendency translates usually into characterization and tracking of people torsos (see Fig. 2 and 4), as they represent the higher uniform-color visible areas.

Although the MS system performs more accurate 3D segmentations of tracked objects, it is prone to fail when handling occlusions between two or more spatially close people (Fig. 4(b)). This is a result of an erroneous weight updating of particles due to proximity. However, Fig. 4(c) shows MS+C system ability to handle occlusions. Color consistency ensures correct weight updating of particles from both people with no external help, resulting in excellent robustness to people interaction.

4 Conclusions

This paper proposes a Bayesian framework for performing 3D segmentation and tracking of multiple people using multiple cameras, allowing consideration of several features or cues in each camera. This features can be similar for the whole set of cameras, or even different, allowing thus the utilization of different type of sensors (CCD, IR, thermal imaging,...). The framework uses independent 3D particle filters to fuse directly in the 3D world cues observed in each sensor, avoiding losing information across intermediate hard decisions. It is based on

the estimation of a volumetric occupancy probability density of a moving target over time, sampled using 3D particles, and updated according to camera observations. Thus 3D segmentation can be performed by setting a volume with a high probability (0.95) of containing the target.

Multi-camera and multi-feature observation likelihood has been simplified using conditional independence assumptions. Working on a set of CCD cameras, two different cues or features have been proposed: motion-region segmentation on each camera, and color characterization. Using exclusively motion region segmentation allows accurate 3D segmentation and tracking of non-interacting targets, but fails to overcome interaction of spatially close objects. However, integrating both motion segmentation and (HSI based) color results in excellent 3D tracking robustness, even in situations involving multiple interacting people.

Acknowledgements

This work has been partially supported by the Ministerio de Ciencia e Innovación of the Spanish Government under project TEC2007-67764 (SmartVision) and by the Comunidad de Madrid under project S0505/TIC-0223 (Pro-Multidis).

References

1. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on Particle Filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Processing 50(2), 174–188 (2002)
2. Black, J., Ellis, T., Rosin, P.: Multi view image surveillance and tracking. In: Proc. Workshop on Motion and Video Computing, pp. 169–174 (2002)
3. Doucet, A., Godsill, S.J., Andrieu, C.: On sequential Monte Carlo sampling methods for Bayesian filtering. Statistics and Computing 10(3), 197–208 (2000)
4. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. IEEE Trans. Pattern Analysis and Machine Intelligence 30(2), 267–282 (2008)
5. Focken, D., Stiefelhagen, R.: Towards vision-based 3-D people tracking in a smart room. In: Proc. IEEE Int. Conf. Multimodal Interfaces, pp. 400–405 (2002)
6. Gonzalez, R.I., Woods, R.E.: Digital Image Processing, 3rd edn. Prentice Hall, Englewood Cliffs (2008)
7. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004)
8. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Systems, Man, and Cybernetics 34(3), 334–352 (2004)
9. Isard, M., Blake, A.: CONDENSATION - Conditional Density Propagation for visual tracking. Int. J. Computer Vision 29(1), 5–28 (1998)
10. Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M.: Multi-camera multi-person tracking for easy living. In: IEEE Int. Workshop on Visual Surveillance (2000)
11. Landabaso, J.L., Pardás, M.: Foreground regions extraction and characterization towards real-time object tracking. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 241–249. Springer, Heidelberg (2006)