COLOR-BASED 3D PARTICLE FILTERING FOR ROBUST TRACKING IN HETEROGENEOUS ENVIRONMENTS

Carlos R. del-Blanco, Raúl Mohedano, Narciso García, Luis Salgado and Fernando Jaureguizar

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

ABSTRACT

Most multi-camera 3D tracking and positioning systems rely on several independent 2D tracking modules applied over individual camera streams, fused using both geometrical relationships across cameras and/or observed appearance of objects. However, 2D tracking systems suffer inherent difficulties due to point of view limitations (perceptually similar foreground and background regions causing fragmentation of moving objects, occlusions, etc.) and, therefore, 3D tracking based on partially erroneous 2D tracks are likely to fail when handling multiple-people interaction. In this paper, we propose a Bayesian framework for combining 2D low-level cues from multiple cameras directly into the 3D world through 3D Particle Filters. This novel method (direct 3D operation) allows the estimation of the probability of a certain volume being occupied by a moving object, using 2D motion detection and color features as state observations of the Particle Filter framework. For this purpose, an efficient color descriptor has been implemented, which automatically adapts itself to image noise, proving able to deal with changes in illumination and shape variations. The ability of the proposed framework to correctly track multiple 3D objects over time is tested on a real indoor scenario, showing satisfactory results.

Index Terms— Multi-camera, Particle Filter, 3D Tracking, Color Descriptor, Visual Surveillance

1. INTRODUCTION

Visual surveillance and monitoring of indoor and outdoor environments has become a field of very active research in computer vision due to its applicability to surveillance systems, security and restricted access area control, intelligent rooms, etc. Although 2D tracking has been largely addressed in the literature [1], it shows itself unable to describe complex scenarios where multiple target objects interact. To overcome inherent limitations of single camera 2D tracking algorithms, it is necessary to develop 3D positioning and tracking systems based on several cameras with overlapping fields of view.

Different approaches have been proposed for tracking multiple people in multi-camera environments. The most common one assumes a ground-plane restriction, presuming that objects of interest move on a visible principal plane (the real ground), thus allowing to establish homographies relating different views, and finally combining the trajectories observed in different cameras onto that common plane [2, 3]. Although this approach proves effective in many situations, the ground plane assumption is too restrictive and does not hold for many interesting environments.

Ground plane assumption can be avoided if the calibration of the cameras is known, allowing more sophisticated processing. The direct approach for fusing information from multiple calibrated cameras consists in performing 2D tracking on each camera independently, and subsequently fusing 2D decisions into the 3D world using either only geometrical considerations [4], or geometry along with appearance consistency [5]. As this approach relies on 2D tracks performed individually, it depends decisively on the decisions made at 2D tracking level. This method is thus prone to fail when dealing with complex situation, since 2D tracking shows inherent limitations due to camera point of view, occlusions, etc.

The loss of information due to hard decisions made at 2D level can be avoided by fusing directly all the information captured by the cameras of the system. This fusion or combination is usually performed in a probabilistic framework. So, in [6], fusion is performed through the estimation of an occupancy probability projection onto the ground plane using background subtraction on individual cameras. However, in [7], the 3D space is discretized using voxelization, and combination of multi-camera information is performed at voxel level, allowing 3D segmentation, positioning, and tracking. Although the latter approach provides excellent information for scene understanding, voxelization represents an excessive computational and storage cost, as it divides uniformly the space.

In this work, we address the problem of tracking multiple people in environments monitored using multiple fully-calibrated cameras with overlapping fields of view, handling

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).

978-1-4244-2665-2/08/$25.00 ©2008 IEEE

successfully complex situations like multiple people interaction and severe occlusions. Fusion of the information acquired independently by each camera is performed directly in the 3D world, following a probabilistic framework based on a 3D Particle Filter [8]. These 3D Particle Filters aim to estimate the volumetric occupancy probability density for each tracked person over time, by using a sample representation based on a finite set of weighted samples. The main advantage of this method is that it allows robust 3D positioning and tracking of moving objects.

Two different types of information are extracted from images acquired by each camera: movement detection (using a standard motion-region binary segmentation), and color information. Both are integrated in the proposed probabilistic framework as observations of the 3D particles composing the 3D Particle Filter. As for color information, a novel 3D color descriptor has been implemented. It characterizes a 3D point by computing a multi-camera color distribution robust to noise, shape variations and changes in illumination. This approach shows a better performance than those ones that address the color appearance independently in each camera [9, 10].

The paper is organized as follows. Sec. 2 describes the proposed 3D Particle Filter framework, showing also the dynamic model of the system and the observation model based on motion segmentation. Sec. 3 presents the 3D color descriptor implemented, describing the color-based observation likelihood model for the 3D particles. Finally, Sec. 4 shows experimental results of the proposed system, and Sec. 5 outlines the achievements reached by the proposed approaches.

2. 3D PARTICLE FILTERS FOR 3D TRACKING

Bayesian tracking [8] has become the main approach for visual tracking of moving objects over time. It models dynamic systems as sequences of hidden states $\mathbf{x}_t$ that cannot be seen directly, but only as noisy observations $\mathbf{z}_t$. The Bayesian approach aims to estimate the posterior probability density function (pdf) of $\mathbf{x}_t$ based on all the available information up to time step t.

The posterior pdf of $\mathbf{x}_t$, given every available observation $\mathbf{Z}^t = (\mathbf{z}_1, \ldots, \mathbf{z}_t)$, can be expressed in terms of them from time step $t-1$ through t

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

assuming that, on one hand, the dynamic model governing state evolution does not depend on previous measurements $\mathbf{Z}^{t-1}$ and, on the other hand, observations at time t only depend on the hidden state $\mathbf{x}_t$. The predicted probability density function $p(\mathbf{x}_t|\mathbf{Z}^{t-1})$ is obtained using the prior distribution $p(\mathbf{x}_{t-1}|\mathbf{Z}^{t-1})$ (available, as it has been estimated in the previous time step $t-1$) and the dynamic model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ of the system. The $p(\mathbf{x}_t|\mathbf{Z}^{t-1})$ 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 [11], which deal with sampled versions of distributions using importance sampling, making it possible to handle non-Gaussianity and/or non-linearity in both dynamic and observation models. This approach is known as Particle Filter [8], as it deals with sets of weighted particles (or samples) $\{\mathbf{x}_t^{(i)}, w_t^{(i)}\}_{i=1}^{N_S}$, where N_S is the number of samples, and $w_t^{(i)}$ represents the weight associated to the particle $\mathbf{x}_t^{(i)}$.

Particle Filter-based systems differ, as well as in their importance sampling policies (a critical subject in practice), in the meaning of the particles themselves. The simplest situation aims to estimate 2D trajectories from varying 2D positions (for example, object centroids over time), resulting in $\mathbf{x}_t$ representing 2D positions. However, most 2D tracking systems consider multi-dimensional states $\mathbf{x}_t$ having abstract and global meaning, describing both position and shape/appearance of objects: two excellent examples are [12], where $\mathbf{x}_t$ represents control points of parameterised image curves, and [13], which addresses 3D tracking of people with a single fully-calibrated camera, suffering problems when handling multiple people interaction. The latter system assumes ground plane movement, and represents humans using cylinders: in that context $\mathbf{x}_t$ stands for both ground position and cylinder parameters. All the above cited systems coincide on the fact that a particular point $\mathbf{x}_t$ describes completely a possible global situation of the tracked object. This is the common approach for Bayesian tracking, and it can be considered a Bayesian point estimation problem, where the final point decision is carried out using the posterior pdf $p(\mathbf{x}_t|\mathbf{Z}^t)$.

The presented work also intends to estimate both 3D position and shape of moving objects, but proposing a different approach. We assume monitored environment using two or more overlapping, fully-calibrated cameras. Instead of considering that a particular point $\mathbf{x}_t$ represents a global situation of an object, we consider that it describes the situation of a particular spatial point (whose position is $\mathbf{x}_t$ itself). Assuming that a person H_k is being tracked, it would be very interesting to establish the likelihood of H_k being contained in a certain volume V. This probability distribution can be supposed absolutely continuous, having an associated *volumetric occupancy* probability density function. Both are related through

$$P\big(H_k \subseteq V \mid H_k\big) = \iiint_V p\big(H_k \subseteq \mathrm{d}V \mid H_k\big)\, \mathrm{d}V. \qquad (2)$$

Evidently, integration over the whole space must be 1, as $p\big(H_k \subseteq \mathrm{d}V \mid H_k\big)$ is a volumetric pdf. This is consistent with the likelihood of H_k being somewhere in the scene is 1, given that it is present. The core idea of this work is to estimate the volumetric occupancy pdf of a person H_k using 3D Particle Filters, and then to use this to define a *3D bounding volume*

V_k for H_k (which could be defined as the minimum volume V that contains H_k with probability greater or equal to P_H). In this framework, the event $\mathbf{x}_t$ represents that the $\mathbf{x}_t$ position is contained in the volume occupied by the person H_k, giving

$$p\big(H_k \subseteq \mathrm{d}V \mid H_k\big) = p\big(\mathbf{x}_t | \mathbf{Z}^t, H_k\big), \tag{3}$$

where the differential of volume $\mathrm{d}V$ is centered at $\mathbf{x}_t$ spatial position. This approach allows to track 3D objects without assuming *a priori* shape models, which usually is a problematic task, specially for non-rigid objects (*e.g.* people).

2.1. 3D Particle Filter dynamic model

The discussed approach requires hidden states $\mathbf{x}_t$ considering both spatial coordinates and appearance information, given by:

$$\mathbf{x}_t = \big(\mathbf{P}_t, \dot{\mathbf{P}}_t, C_t\big), \tag{4}$$

where $\mathbf{P}_t$ is the 3D spatial position, defined by the coordinates x_t, y_t and z_t, $\dot{\mathbf{P}}_t$ is the 3D spatial velocity, and C_t is the normalized histogram modeling the multi-view color descriptor of that spatial point (as described in detail in Sec. 3).

As shown in Eq. (1), Bayesian recursive tracking requires both a dynamic model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ (to predict $p(\mathbf{x}_t|\mathbf{Z}^{t-1})$) and an observation model to perform tracking over time. As for the dynamic model of particles in the proposed approach, given that both position and appearance are reflected in $\mathbf{x}_t$, mechanics (spatial position evolution) and color evolution should be addressed separately.

System mechanics is usually described using a simple linear velocity dynamic model: this hypothesis, applied to the 3D particles, results in states $\mathbf{x}_t$ containing both 3D position $\mathbf{P}_t$ (x, y and z) and velocity $\dot{\mathbf{P}}_t$ ($\dot{x}$, $\dot{y}$ and $\dot{z}$). Particle mechanical evolution could thus be written, using matrix notation, as

$$\left[\begin{array}{c} \mathbf{P}_t \\ \hline \dot{\mathbf{P}}_t \end{array}\right] = \left[\begin{array}{c} x_t \\ y_t \\ z_t \\ \hline \dot{x}_t \\ \dot{y}_t \\ \dot{z}_t \end{array}\right]^T = \left[\begin{array}{c|c} I_3 & I_3 \\ \hline 0_3 & I_3 \end{array}\right] \left[\begin{array}{c} \mathbf{P}_{t-1} \\ \hline \dot{\mathbf{P}}_{t-1} \end{array}\right] + \mathbf{n}_t, \tag{5}$$

where $\mathbf{n}_t$ is a 6×1 matrix representing the prediction error of the system [8], and where I_3 and 0_3 are the 3×3 identity and zero matrices, respectively. If the video acquisition frame-rate is uniform (which actually is the usual situation), the velocity components directly represent position differences between consecutive time steps.

However, appearance of objects must ideally remain approximately constant over time. To reflect this behavior, the dynamic model for color follows the expression

$$C_t = \alpha\, A_{t-1} \,+\, (1-\alpha)\, C_{t-1}, \tag{6}$$

where C_{t-1} is the normalized histogram of the particle $\mathbf{x}_{t-1}^{(i)}$, A_{t-1} is the normalized histogram observed at position $\mathbf{P}_{t-1}$, and α is the forgetting factor. This model must allow certain adaptation to slow minor changes in appearance, but must avoid erroneous adaptations to static foreground objects or other moving objects appearance. That can be achieved by taking two different measures: α must be close to zero, and particle resampling [8] should be performed at each time step. Frequent resampling discards particles that have quickly changed their appearance, and then avoids erroneous adaptation of color histograms.

The importance density of the system has been chosen to be the dynamic model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, simplifying enormously the particle updating process [8].

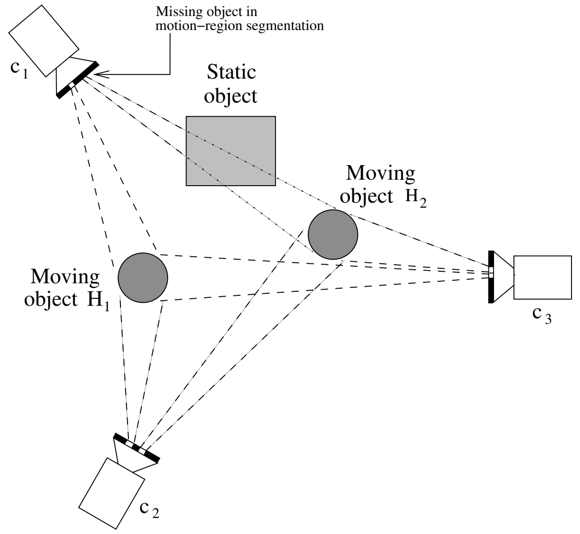


Fig. 1. Two moving objects H_1 and H_2 in a scene monitored with three different cameras (c_1, c_2 and c_3). Note that H_2 cannot be seen directly from c_1 point of view due to a static occluding object.

2.2. Motion segmentation-based observation model

As for the observation of the hidden states, it is necessary to discuss how moving objects are captured and detected in the different cameras. Let us suppose that 2D motion-region segmentation is performed on images acquired independently by every camera c_j of the system (*e.g.* using background subtraction [14]), and that segmentation is errorless. Bearing in mind that the event $\mathbf{x}_t$ represents that the $\mathbf{x}_t$ position is contained in the volume occupied by the person H_k, we can say that if $\mathbf{x}_t$ is actually contained in H_k, it would then be seen as part of a moving object by camera c_j. In other words, its projection onto the image plane of camera c_j will lie in one of the detected moving regions of that camera, unless a static object occludes $\mathbf{x}_t$ from that particular point of view, as shown in Fig. 1. Since the latter situation is possible (that is, it is possible that a person actually present in the scene is not detected in some of the cameras), it is necessary to allow a certain uncertainty even if no evidence of movement is detected in some of the cameras.

Let us suppose that there are M different cameras in the system, and that a moving-region binary segmentation in each camera is available. Using the binary segmentations as visual cues to decide if a 3D particle is contained into an actual moving object or not, observation $\mathbf{z}_t$ could be written as

$$\mathbf{z}_t = \left(m_t^{c_1}, m_t^{c_2}, \ldots, m_t^{c_M}\right). \tag{7}$$

where $m_t^{c_j}$ is the resulting binary mask from motion segmentation performed on the image acquired by camera c_j at time step t, so

$$m_t^{c_j}(\mathbf{u}) = \begin{cases} 1 & \forall\, \mathbf{u} \in \mathrm{R} \\ 0 & \text{otherwise} \end{cases} . \tag{8}$$

where R is the set of pixels where movement has been detected in camera c_j, and $\mathbf{u}$ is the 2D image coordinate (pixels). We are interested in the observation likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$, key in the updating step of Particle Filtering. Although masks $m_t^{c_j}$ in the different cameras are clearly statistically dependent (*e.g.* projections of a common 3D point onto image planes of different cameras must lie on epipolar lines [15]), we have noticed that, in the studied environment, observations in other cameras cannot add significant extra information (this is specially true when $\mathbf{x}_t$ is seen directly from camera c_j point of view). Using this assumption, cameras can be treated independently and finally combined through

$$p(\mathbf{z}_t|\mathbf{x}_t) = p(M_t|\mathbf{x}_t) = \prod_{j=1}^{M} p\left(m_t^{c_j}|\mathbf{x}_t\right). \tag{9}$$

Let us focus our attention on a particular camera c_j at time step t. We aim to estimate $p\left(m_t^{c_j}|\mathbf{x}_t\right)$ for each posible state $\mathbf{x}_t$, assuming the motion-region segmentation mask. Let $\mathbf{v}_{c_j} = P_{c_j}(\mathbf{x}_t)$ be the projection of the 3D position of $\mathbf{x}_t$ onto the camera c_j image plane. Having into account the uncertainty caused by the possibility that a person actually present in the scene is not detected in camera c_j due to occlusions, and expressing it as a certain *background probability* p_B (where p_B is a positive value close to 0), the contribution of camera c_j to the observation likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$ can be expressed as

$$p\left(m_t^{c_j}|\mathbf{x}_t\right) = \begin{cases} 1 - p_B & \text{if } m_t^{c_j}\left(\mathbf{v}_{c_j}\right) = 1 \\ p_B & \text{otherwise} \end{cases} . \tag{10}$$

The shape characteristics of 3D tracked object H_k are implicitly included in the presented probabilistic approach, as they can be inferred from the volumetric occupancy pdf of H_k (by defining a 3D bounding volume V_k, for example). Appearance characteristics, however, must be included explicitly in the hidden states $\mathbf{x}_t$, along with the considered 3D position.

3. COLOR-BASED OBSERVATION MODEL

Although the motion segmentation-based observation model discussed previously can be enough to track isolated moving objects, it shows itself insufficient to handle multiple target interaction correctly. Thus it is necessary to consider appearance information to provide tracking system with robustness to complex tracking scenarios. Particularly, the considered appearance model is based exclusively on color information.

Considering that the observation of the hidden state $\mathbf{x}_t$ at time step t consists of both motion M_t and appearance A_t information, and assuming they are conditionally independent, the observation likelihood can be written as

$$p(\mathbf{z}_t|\mathbf{x}_t) = p(M_t|\mathbf{x}_t) \cdot p(A_t|\mathbf{x}_t) \tag{11}$$

The conditional independence assumption is motivated, once again, by the fact that both motion and appearance observations of a hidden state $\mathbf{x}_t$ are influenced by $\mathbf{x}_t$ to a great extent, so that no significant statistical information can be added. The factor $p(M_t|\mathbf{x}_t)$ of Eq. (11), which represents the motion observation model, is described in the previous section.

The color-based observation model uses a descriptor based on histograms of the HSV space, that efficiently characterizes the appearance of points in the 3D space, and it is robust to the image noise and shape variations of the moving objects. Using this color measurement model, the appearance probability $p(A_t|\mathbf{x}_t)$ of a particle $\mathbf{x}_t$ is modeled by a Gaussian distribution that depends on a complex similarity measure between the color descriptor predicted for the particle and the color observed, which is robust to changes in illumination and shadows.

3.1. Noise analysis in the HSV color space

The acquisition and display devices usually use the RGB color model to encode the image information, since it is closely related to the hardware implementation. However, HSV color model is more suitable to measure perceptual distances between colors, since it is more similar to the human perception of the color. Therefore, the acquired images are converted from RGB to HSV color space. This color conversion, which is based on non-linear transformations, produces a different noise level in each dimension of the HSV space, which is dependent on the initial RGB value. The equations of the RGB to HSV color conversion are given by:

$$V = \text{max} \tag{12}$$

$$S = \begin{cases} 0, & \text{if max} = 0 \\ 1 - \frac{\text{min}}{\text{max}}, & \text{otherwise} \end{cases} \tag{13}$$

$$H = \begin{cases} 0, \text{ if max} = \text{min} \\ \frac{1}{6}\cdot\frac{G-B}{\text{max}-\text{min}}, \text{ if max} = R\text{, and G} \geq \text{B} \\ 1 + \frac{1}{6}\cdot\frac{G-B}{\text{max}-\text{min}}, \text{ if max} = R\text{, and G} < \text{B} \\ \frac{1}{3} + \frac{1}{6}\cdot\frac{B-R}{\text{max}-\text{min}}, \text{ if max} = G \\ \frac{2}{3} + \frac{1}{6}\cdot\frac{R-G}{\text{max}-\text{min}}, \text{ if max} = B \end{cases} \tag{14}$$

where max and min are respectively the maximum and minimum of the RGB components.

Consequently, to correctly compute perceptual color distances it is necessary to characterize the level of noise in each HSV dimension. To achieve this task, it is assumed that the image noise in each RGB channel can be modeled by an independent additive Gaussian distribution [16], as shown in:

$$(R_n, G_n, B_n) = (N(\mu_R, \sigma_n), N(\mu_G, \sigma_n), N(\mu_B, \sigma_n)) \tag{15}$$

where μ_R, μ_G, and μ_B are respectively the RGB values related to an image pixel; and σ_n is the level of noise, which is assumed the same for each dimension. According to this, HSV values can be modeled as a combination of these Gaussian distributions as shown in Eq. (16), where it has been assumed, without loss of generality, that $\mu_R \geq \mu_G \geq \mu_B$:

$$\begin{array}{l} V_n = N(\mu_R, \sigma_n) \\ S_n = 1 - \frac{N(\mu_B,\sigma_n)}{N(\mu_R,\sigma_n)} \\ H_n = \frac{N(\mu_G,\sigma_n)-N(\mu_B,\sigma_n)}{N(\mu_R,\sigma_n)-N(\mu_B,\sigma_n)} \end{array} \tag{16}$$

The resulting level of noise of each component, expressed as their variance ($var(V_n)$, $var(S_n)$, and $var(H_n)$), is different and dependent on the RGB value. Therefore, to correctly compute measurements in HSV space, the H and S channels are equalized respect to V channel (which has the same level of noise as the RGB channels) as shown in:

$$\begin{array}{l} \alpha_S = \left(\frac{var(V_n)}{var(S_n)}\right)^2 \\ \alpha_H = \left(\frac{var(V_n)}{var(H_n)}\right)^2 \end{array} \tag{17}$$

Note that α_S and α_H depend on μ_R, μ_G, μ_B, and σ_n, where σ_n is estimated from two consecutive images by:

$$\sigma_n = \sqrt{\frac{\sum_{i=1}^{N_{pix}} (V_{t-1} - V_t)^2}{3N_{pix}}} \tag{18}$$

where $t-1$ and t are two consecutive time steps; and N_{pix} is the number of pixels per channel in one image.

In the case of considering an image patch instead of a unique image pixel, μ_R, μ_G, μ_B are obtained as the mean values of the patch for each color channel. This approach is adopted in Sec. 3.2, where α_S and α_H are used to properly compute the perceptual distance between color descriptors in the HSV space.

3.2. Color descriptor for 3D points

Let $\mathbf{P}$ be a point in the 3D space located on the surface of an object, and $\mathbf{N_P}$ the set of points belonging to its 3D spherical neighborhood, given by:

$$\mathbf{N_P} = \{\mathbf{Q} \in \mathbb{R}^3 | E(\mathbf{Q}, \mathbf{P}) < r_N\} \tag{19}$$

where $E(\cdot)$ is the Euclidean distance and r_N is the radius that determines the size of the 3D neighborhood. The projections of $\mathbf{P}$ and $\mathbf{N_P}$ over the image plane of each camera are respectively represented by $\mathbf{p}^i$ and $\mathbf{n_p}^i, i = 1, ..., N_c$, where N_c is the number of cameras. However, it is not possible to ensure that all cameras are observing $\mathbf{P}$ and $\mathbf{N_P}$ due to point of view limitations. Figure 2 depicts this situation, where $\mathbf{P}$, $\mathbf{N_P}$ and the camera locations are represented from a zenithal view. Camera 1 and 2 observe $\mathbf{P}$ and $\mathbf{N_P}$, and therefore they are successfully projected onto their respective image planes. However, the camera 3 does not observe them because the object, to which $\mathbf{P}$ belongs, is occluding them. In this case, $\mathbf{O}$ and $\mathbf{N_O}$, that also belong to the object, are projected onto the image plane of the camera 3 instead of $\mathbf{P}$ and $\mathbf{N_P}$. Under the common hypothesis that the color appearance must be the same for all $\mathbf{n}_\mathbf{p}^i, i = 1, ..., N_c$ to consider that they belong to the same object, the previous situation may significantly reduce the performance of 3D tracking, since if the color appearance of $\mathbf{N_P}$ and $\mathbf{N_O}$ are different, the corresponding projections are incorrectly interpreted as belonging to different objects. The proposed 3D color descriptor overcomes this problem by characterizing a point in the 3D space by means of all its 2D projections, taking advantage of the different color appearance related to each projection to be more distinctive, instead of using the previous restriction that assumes that objects have the same appearance independent of the point of view.

The computation of the color descriptor involves several steps: all 2D projections $\mathbf{n}_\mathbf{p}^i, i = 1, ..., N_c$ related to the 3D point $\mathbf{P}$ are jointly used to computed three different weighted histograms, one per dimension of the HSV color space, by means of the expression:

$$D^X(P) = \frac{\sum_{i=1}^{N_c} \sum_{\mathbf{u}_k \in \mathbf{n}_\mathbf{p}^i} G(\|\mathbf{p}^i - \mathbf{u}_k\|) h(\mathbf{u}_k)}{\sum_{i=1}^{N_c} \sum_{\mathbf{u}_k \in \mathbf{n}_\mathbf{p}^i} G(\|\mathbf{p}^i - \mathbf{u}_k\|)} \tag{20}$$

$$X \in \{H, S, V\}$$

where $\mathbf{u}_k$ is the k^{th} pixel of $\mathbf{n}_\mathbf{p}^i$; G is a Gaussian kernel with mean $\mu = (0, 0)$ and covariance matrix $\sum_\sigma = \mathbf{I} r_\mathbf{n}^i$ where $\mathbf{I}$ is the identity matrix, and $r_\mathbf{n}^i$ is the radius of $\mathbf{n}_\mathbf{p}^i$, that under the assumption that its shape is approximately circular (since $\mathbf{N_P}$ is a sphere) can be computed as $r_\mathbf{n}^i = \sqrt{A_p/\pi}$, where A_p is the area in pixels of $\mathbf{n}_\mathbf{p}^i$; and $h(\mathbf{u}_k)$ is a unidimensional N_b-bin histogram function that computes the contribution of $\mathbf{u}_k$ to the corresponding bins. The contribution is performed through a linear interpolation, that avoids large changes in the

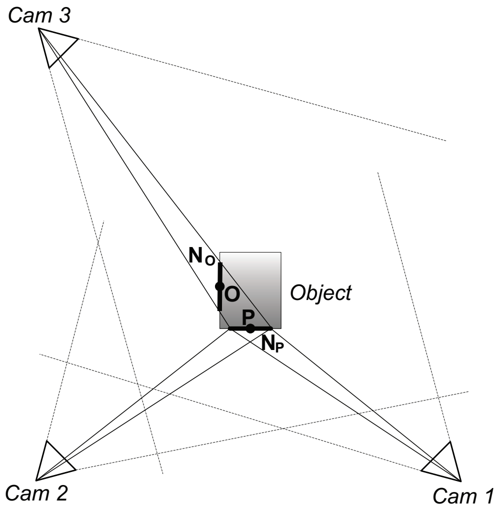


Fig. 2. $\mathbf{P}$, $\mathbf{N_P}$ and the camera locations from a zenithal view. Camera 3 can not observe $\mathbf{P}$ and its neighborhood due to the occlusion of the own object.

histogram when the pixel contribution oscillates between two different bins due to slight variations induced by the image noise. Thanks to this histogram-based characterization, the color descriptor is robust to image noise and shape variations due to changes in the 3D camera perspective and motions of deformable objects. In addition, the scale invariant property of the histograms allows to deal with 2D projections of different sizes due to the relative distance of the object to each camera.

The Gaussian kernel gives less relevance to the pixels located far from $\mathbf{p}^i$, since small variations in the location of those pixels could produce that they contribute or not to the histogram, and thus avoiding large variations in the histogram. Note, that the denominator of Eq. (20) is a normalization factor because of the Gaussian weighting.

Figure 3 depicts the computation of each weighted histogram using the 2D projections $\mathbf{n}_p^i, i = 1, 2, 3$ related to the three camera configuration shown in Fig. 2. The color dimensions H, S and V related to each $\mathbf{n}_p^i$ are shown separately. Note that the size of $\mathbf{n}_p^3$ is smaller than the rest due to the camera 3 is further away to the object, but this is not a problem thanks to the histogram-based approach. Also, notice that the hue histogram has two different modes and the saturation and value histogram only one, since the camera 3 observes $\mathbf{N_O}$ instead of $\mathbf{N_P}$, which has the same saturation and value, but different hue. Thus, the descriptor is able to handle objects with arbitrary shapes and colors.

The weighted histograms $D^H(\mathbf{P})$, $D^S(\mathbf{P})$, and $D^V(\mathbf{P})$ can be considered as specific descriptors for each color dimension, respectively the hue descriptor, the saturation descriptor, and the value descriptor, which jointly represent the color descriptor of the 3D point $\mathbf{P}$:

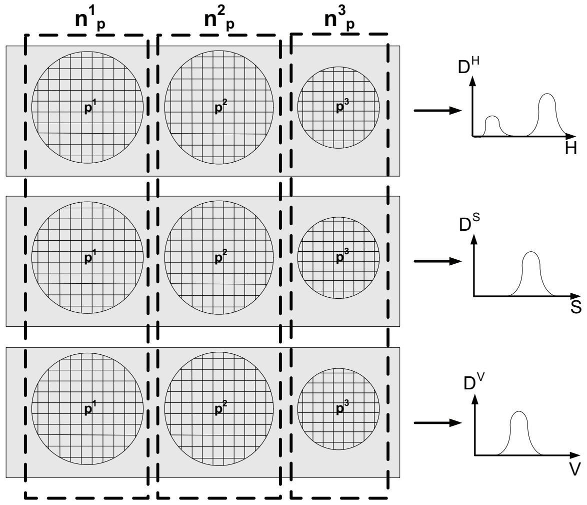


Fig. 3. Computation of the weighted histograms $D^H(\mathbf{P})$, $D^S(\mathbf{P})$, and $D^V(\mathbf{P})$. The 2D projection related to the camera 3 is smaller than the others.

$$D(\mathbf{P}) = \{D^H(\mathbf{P}), D^S(\mathbf{P}), D^V(\mathbf{P})\} \tag{21}$$

3.3. Appearance Probability

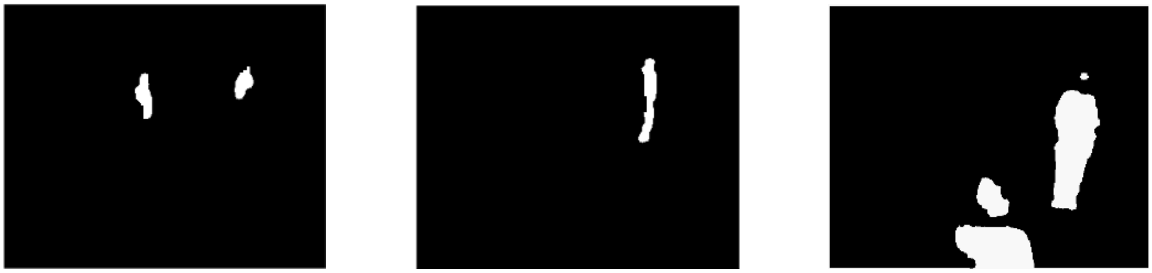
Taking into account that each particle $\mathbf{x}_t^{(i)}$ contains the location $\mathbf{P}_t^{(i)}$ of a 3D point belonging to a moving object, and also the predicted appearance $C_t^{(i)}$ of that spatial point, the observed appearance A_t for a particle $\mathbf{x}_t^{(i)}$ is defined by the color descriptor related to its location $\mathbf{P}_t^{(i)}$:

$$A_t = D(\mathbf{P}_t^{(i)}). \tag{22}$$

The appearance probability $p(A_t|\mathbf{x}_t)$ determines the likelihood of the appearance descriptor A_t measured at the position $\mathbf{P}_t^{(i)}$ being observed, given that the "real" appearance at that position is supposed to be $C_t^{(i)}$. This is accomplished by measuring the similarity between the color descriptor observed A_t and the predicted appearance $C_t^{(i)}$ at that position. The similarity measurement addresses each component of the color descriptor (D^H, D^S, and D^V) in a specific way to compensate the different level of noise in each HSV channel (as a result of the non-linear conversion between the RGB and HSV color spaces), as shown in:

$$\begin{aligned} d^H &= \frac{E\left(\, C_{t,H}^{(i)},\, D^H(\mathbf{P}_t^{(i)})\,\right)\cdot\alpha_H}{\sqrt{2}} \\ d^S &= \frac{E\left(\, C_{t,S}^{(i)},\, D^S(\mathbf{P}_t^{(i)})\,\right)\cdot\alpha_S}{\sqrt{2}} \\ d^V &= \frac{E\left(\, C_{t,V}^{(i)},\, D^V(\mathbf{P}_t^{(i)})\,\right)}{\sqrt{2}} \end{aligned} \tag{23}$$

where $C_{t,H}^{(i)}$, $C_{t,S}^{(i)}$ and $C_{t,V}^{(i)}$ are the three components of the appearance descriptor $C_t^{(i)}$, $E(\cdot)$ is the Euclidean distance,



(a) Camera 1

(b) Camera 2

(c) Camera 3

Fig. 4. Motion-based segmentation of the cameras views corresponding to the Fig. 5.

and both α_S and α_H are the weights explained in the Sec. 3.1 used to equalize the image noise in each HSV channel. Note that d^H, d^S, and d^V have been normalized respect to the maximum possible distance, i.e. $\sqrt{2}$ since each component of the color descriptor is already normalized.

The similarity distance between the observed A_t and the predicted appearance $C_t^{(i)}$ is then computed by combining d^H, d^S, and d^V as shown in:

$$M\big(A_t, C_t^{(i)}\big) = \frac{d^H + d^S + \beta_V(\sigma_n) d^V}{\alpha_H + \alpha_S + \beta_V(\sigma_n)} \qquad (24)$$

where the denominator is used to normalized the measure; and $\beta_V(\sigma_n)$ is a factor to make the similarity measure robust to changes in the illumination and shadows, and it is computed as:

$$\beta_V(\sigma_n) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(- \frac{(d^V)^2}{2\sigma_n^2} \right) \qquad (25)$$

Note that the illumination variations can produce large changes in d^V, and thus determining incorrectly that the appearance of two 3D points are very dissimilar. $\beta_V(\sigma_n)$ overcomes this problem by penalizing d^V in these cases, so that the similarity measure depends more on the distances d^S and d^H.

Finally, the appearance probability is computed as a Gaussian function of the similarity measure $M\big(A_t, C_t^{(i)}\big)$:

$$p(A_t|\mathbf{x}_t) = \frac{1}{\sqrt{2\pi}\sigma_p} \exp\left(- \frac{\Big(M\big(A_t, C_t^{(i)}\big)\Big)^2}{2\sigma_p^2} \right) \qquad (26)$$

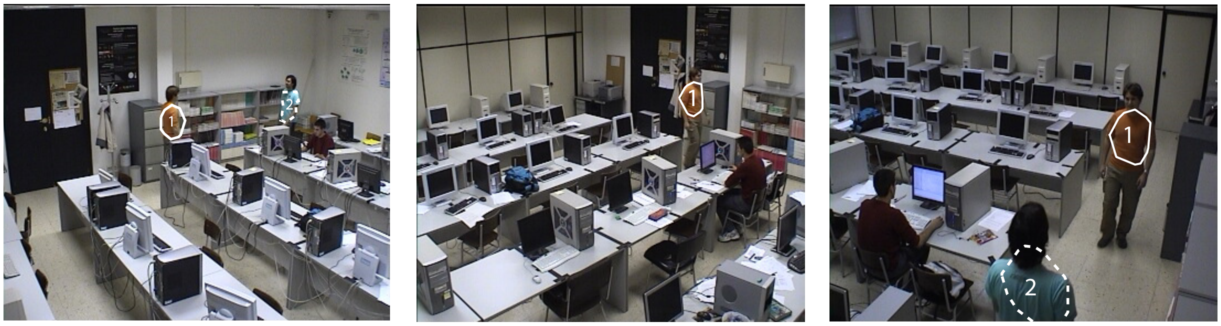
This probability takes a high value when $M\big(A_t, C_t^{(i)}\big)$ is less than σ_p, and it quickly decreases when it is higher. According to this, σ_p determines the allowed variations in the appearances of two 3D points due to the noise, the 3D perspective of the cameras, the object deformations and changes in illumination.

The appearance observation likelihood $p(A_t|\mathbf{x}_t)$ along with $p(M_t|\mathbf{x}_t)$ (explained in the Sec. 2.2) compose the observation likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$ of the proposed Particle Filter framework.

4. RESULTS

The proposed 3D tracking system has been tested on a office room equipped with a large number of pieces of office furniture and hardware equipment (tables, chairs, computers, etc.), as shown the Fig. 5. The spatial configuration of these objects makes that people are almost always partially occluded, especially from the feet up to the waist. In this situation, the motion segmentation strategies, mainly based on background substraction techniques, give poor results, making the tracking extremely complicated. Figure 4 shows the motion segmentation of images presented in Fig. 5, where it is obvious the low quality of the segmentation, which is incomplete and fragmented. Moreover, this highly cluttered scenario prevents using the information coming from the floor plane (it is hardly visible), frequently used to improve the segmentation and the tracking. The office room is monitored by three fully-calibrated static cameras, which are placed in the top corners of the room, in such a way that the field of views of the cameras are partially overlapped.

An especially complicated sequence has been used to show the performance of the 3D tracking system. The sequence starts with two people coming into the office room from opposite sides and walking towards each other ('coming situation'). Then, they meet each other, shaking their hands ('occlusion situation with interaction'), and finally they leave the office room by different ways. The sequence has been processed using $N_S = 500$ particles for each tracked target. Figure 5 shows the views of each camera for the 'coming situation', and the people segmentations marked by the numbers 1 and 2, enclosed respectively by a solid line and a dashed line. These segmentations have been calculated by estimating the V_k 3D bounding volume for each H_k as the 3D convex hull of the minimum set of particles that accumulate a probability greater or equal than a threshold P_H. The experimentation has proved that $P_H = 0.95$ provides satisfactory results, although its final value is pending to future optimization. In spite of a non accurate segmentation, the tracking performance does not decrease since the parts that are not segmented correspond to those that are frequently occluded, and therefore are more prone to tracking errors. Figure 6 shows



(a) Camera 1 (b) Camera 2 (c) Camera 3

Fig. 5. Views of each camera for the ‘coming situation’, and the resulting particle-based segmentation.

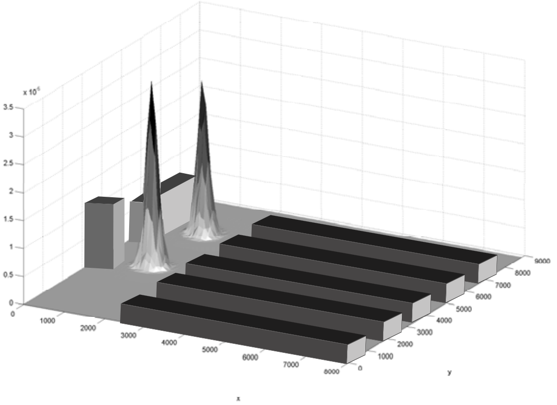
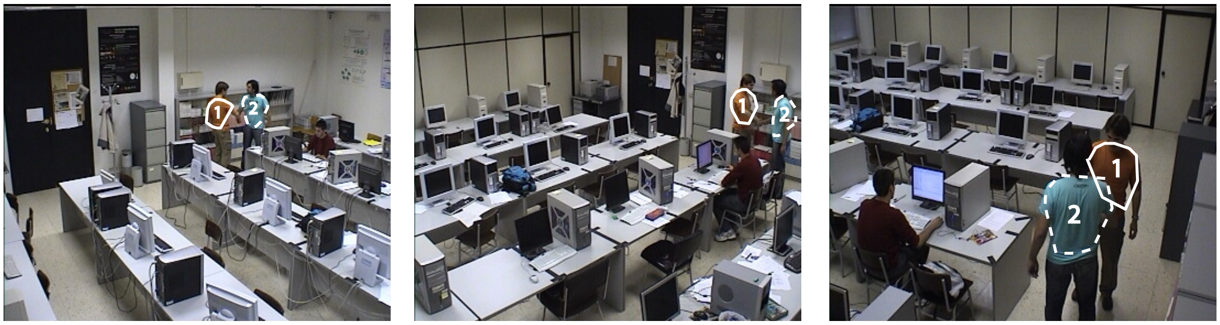


Fig. 6. Volumetric occupancy probability for two different people related to the cameras views corresponding to the Fig. 5.

a schematic 3D representation of the office room (the rectangular objects are the tables and others pieces of furniture with a significative size), and the result of the 3D tracking for the Fig. 5, where the 3D peaks are the probabilities $p(\mathbf{x}_t|\mathbf{Z}^t, H_1)$ and $p(\mathbf{x}_t|\mathbf{Z}^t, H_2)$. These have been computed by applying a 3D Gaussian Kernel to each particle, and represent the spatial location of the tracked people.

Figure 7 shows the views of each camera for the ‘occlusion situation with interaction’, along with the corresponding people segmentations. Despite the occlusion in the camera 3 and the people-interaction in the rest of the cameras (hand shaking), the proposed system segments correctly each person. The 3D tracking is also successfully accomplished, as shown in Fig. 8, where $p(\mathbf{x}_t|\mathbf{Z}^t, H_1)$ and $p(\mathbf{x}_t|\mathbf{Z}^t, H_2)$ can be distinguished from each other. The system efficiently performs the tracking along all the occlusion and people interaction, keeping correctly the tag numbers of each person.

The proposed tracking system has been also tested in situations in which people wear multi-color clothes, and the color distribution acquired for each camera is different. In this context, the proposed 3D color descriptor outperforms the approaches that demand the coherence of the object color distribution in each camera. This is demonstrated using a sequence that contains a person wearing a green sweater and a blue t-shirt, in such a way that the camera 1 mainly watches the blue t-shirt, and the camera 2 the green sweater. Figure 9 shows four different views of the sequence in the upper part. The two views in the left, (a) and (b), have been captured by the cameras 1 and 2 in the time step t_1, while the two views in right, (c) and (d), correspond to the same cameras in the time step t_2. In the bottom part, Fig. 9 shows an array of 3×4 normalized color histograms related to the person. The first column represents respectively from top to bottom: the hue histogram computed from the camera 1, the camera 2, and the combination of both cameras (representing the hue component of the 3D color histogram) in the time step t_1. These histograms have been computed according to Sec. 3.2, taking into account that for the two first histograms only the pixels corresponding to a specific view have been used. The second column shows the same information, but using the saturation channel. And the two last columns depict the same information as the two first, but for the time step t_2. As it can be observed, the normalized histograms computed from the combination of both cameras are very similar in both time steps. This means that the 3D color descriptor (i.e. the appearance model) robustly describes the person, allowing to track him



(a) Camera 1 (b) Camera 2 (c) Camera 3

Fig. 7. Views of each camera for the ‘occlusion situation with interaction’, and the resulting particle-based segmentation.

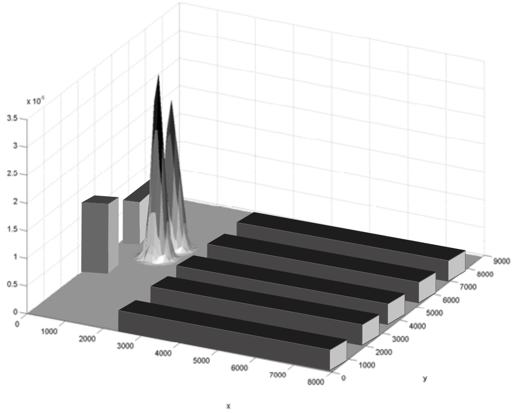


Fig. 8. Volumetric occupancy probability for two different people related to the cameras views corresponding to the Fig. 7.

satisfactorily, as shown in the different views by means of the white line and the corresponding tag.

5. CONCLUSIONS

This paper presents a novel Bayesian framework for performing 3D tracking of multiple interacting people in complex environments monitored using multiple cameras. This framework aims to estimate the volumetric occupancy probability density of a person over time using 3D Particle Filters, according to camera observations. Once the volumetric occupancy pdf has been obtained, 3D positioning and tracking can be performed by establishing a volume with a high probability of containing the target.

Volumetric occupancy densities are evolved over time by using both 2D motion detection and color based observation models to update the weights of the 3D particles. Both contributions are handled separately using conditional independence assumptions. For this purpose, a new 3D color descriptor has been developed, which allows to track successfully moving objects in complex environments and challenging situations such as severe occlusions and interaction of spatially close targets.

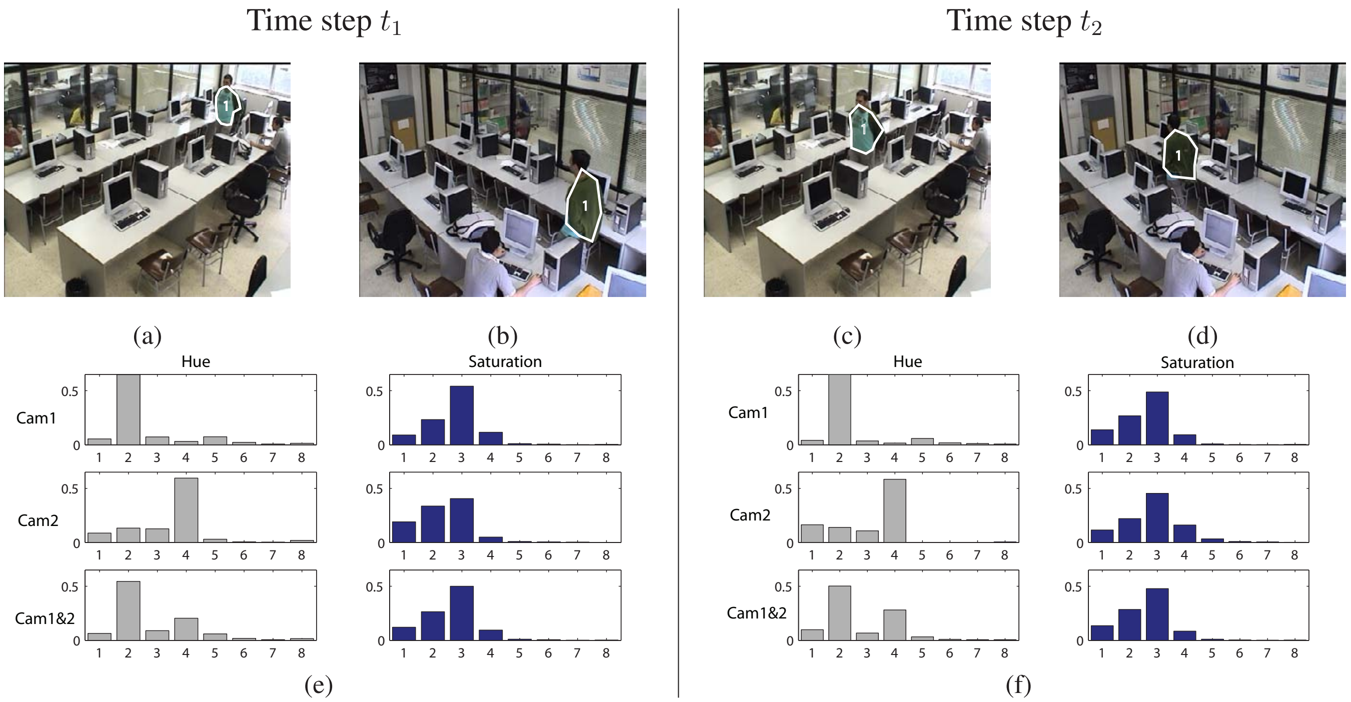
6. REFERENCES

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, “A survey on visual surveillance of object motion and behaviors,” *IEEE Trans. Systems, Man & Cybernetics, Part C*, vol. 34, no. 3, pp. 334–352, 2004.

[2] J. Black, T. Ellis, and P. Rosin, “Multi view image surveillance and tracking,” in *Proc. IEEE Workshop on Motion and Video Computing*, 2002, pp. 169–174.

[3] G. Kayumbi and A. Cavallaro, “Robust homography-based trajectory transformation for multi-camera scene analysis,” in *Proc. ACM/IEEE Int. Conf. Distributed Smart Cameras*, 2006, pp. 59–66.

[4] R. Mohedano, C. R. del Blanco, F. Jaureguizar, L. Salgado, and N. García, “Robust 3d people tracking and positioning system in a semi-overlapped multi-camera environment,” in *Proc. IEEE Int. Conf. Image Processing*, 2008 (in press).



(a) (b) (c) (d)

(e) (f)

Fig. 9. Sequence with a bi-color person, in which each camera is watching a different color. (a)(b) Views of each camera in the time step t_1. (c)(d) Views of each camera in the time step t_2. For each view, the resulting particle-based segmentation is shown. (e)(f) Hue and saturation histograms from the camera 1, the camera 2 and the combination of both cameras for each time step.

[5] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for easy living," in *Proc. IEEE Int. Workshop on Visual Surveillance*, 2000, p. 3.

[6] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multi-camera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, 2007.

[7] J. L. Landabaso and M. Pardás, "Foreground regions extraction and characterization towards real-time object tracking," in *Proc. Int. Workshop on Machine Learning for Multimodal Interaction*. 2005, Lecture Notes in Computer Science, pp. 241–249, Springer.

[8] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 174–188, February 2002.

[9] C. Nigam, R. V. Babu, S. K. Raja, and K. R. Ramakrishnan, "Feature fusion for robust object tracking with fragmented particles," in *Proc. ACM/IEEE Int. Conf. Distributed Smart Cameras*, 2006, pp. 283–290.

[10] K. Nummiaro, E. Koller-Meier, T. Svoboda, D. Roth, and L. Van Gool, "Color-based object tracking in multi-camera environments," in *Symposium for Pattern Recognition of the DAGM*. 2003, Lecture Notes in Computer Science, pp. 591–599, Springer.

[11] A. Doucet, S. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, July 2000.

[12] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *Int. Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[13] M. Isard and J. Maccormick, "Bramble: a bayesian multiple-blob tracker," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2001, vol. 2, pp. 34–41.

[14] M. Piccardi, "Background subtraction techniques: a review," in *Proc. IEEE Int. Conf. Systems, Man & Cybernetics*, 2004, vol. 4, pp. 3099–3104.

[15] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[16] C. Liu, W. T. Freeman, R. Szeliski, and S. B. Kang, "Noise estimation from a single image," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2006, vol. 1, pp. 901–908.