International Journal of Approximate Reasoning 50 (2009) 732–749

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Multi-camera people tracking using evidential filters

Rafael Muñoz-Salinas *, R. Medina-Carnicer, F.J. Madrid-Cuevas, A. Carmona-Poyato

Department of Computing and Numerical Analysis, University of Córdoba, 14071 Córdoba, Spain

ARTICLE INFO

Article history:
Received 15 October 2008
Received in revised form 2 February 2009
Accepted 2 February 2009
Available online 12 February 2009

Keywords:
Particle filters
Dempster–Shafer
Transferable belief model
People tracking
Multiple-views

ABSTRACT

This work proposes a novel filtering algorithm that constitutes an extension of Bayesian particle filters to the Dempster–Shafer theory. Our proposal solves the multi-target problem by combining evidences from multiple heterogeneous and unreliable sensors. The modelling of uncertainty and absence of knowledge in our approach is specially attractive since it does not require to specify prior nor conditionals that might be difficult to obtain in complex problems.

The algorithm is employed to propose a novel solution to the multi-camera people tracking problem in indoor environments. For each particle, the evidence of finding the person being tracked at the particle location is calculated by each sensor. Sensors also provide a degree of evidence about their reliability. The reliability is calculated based on the visible portion of the targets and their occlusions. Evidences collected from the camera set are fused considering their reliability to calculate the best hypothesis. The experiments conducted in several environments show the validity of the proposal.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

People tracking is a topic that has attracted the interest of researchers in several fields such as ambient intelligent systems [12,29,49,58], visual servoing [3,43], human–computer interaction [10,25,68], video compression [45,67] or robotics [6,50]. Although most of the research has focused on single camera approaches, this configuration is not the ideal solution for multi-people tracking due to the occlusion problem and because the space covered by a single camera might be too small for certain applications. Multiple camera tracking approaches aim to solve these problems.

Among the numerous approaches proposed for tracking, Bayesian filtering is the most frequently employed. In particular, particle filters [24,30,31,38] have gained popularity in the vision community because of their advantages over the Kalman filter [26]. First, particle filters are able to carry multiple hypotheses simultaneously. And second, they can deal with non-linear and non-Gaussian systems. Nevertheless, the main obstacle to applying Bayesian filters is the effort some problems require for finding precise probabilistic models that describe the process or the sensors employed. In many real scenarios, it may be difficult to obtain complete knowledge of the problem due to high occlusion, background clutter, illumination or camera calibration errors. In order to fuse data handling uncertainty, the Bayesian theory [55] requires prior probability, likelihoods, and posterior probabilities to be defined. Without precise information, these three elements might not be defined properly, thus leading to assumptions and restrictions of the problem. Furthermore, in real tracking problems a large number of cameras are required. Then, it is likely to find situations in which a target is not visible in some of the cameras. In that case, proposing an observation model using particle filters is complicated, i.e., which likelihood $(p(z|x))$ must be assigned to a state that cannot be observed? How do we fuse information from multiple cameras using a Bayesian approach if one of them does not observe the target?

* Corresponding author.
E-mail address: rmsalinas@uco.es (R. Muñoz-Salinas).

0888-613X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2009.02.001

An alternative to deal with the above mentioned difficulties is Dempster–Shafer (DS) theory of evidence [2,5,8,17,27,57]. DS theory is able to model systems assuming that the knowledge about the problem is not completely precise, thus allowing the natural manipulation of ignorance and uncertainty. The above formulated questions are be answered by the DS theory using the "unknown" subset representing the absence of knowledge about an state. This work proposes a reformulation of the classical particle filtering algorithm from the basis of the DS theory of evidence. The evidential filter proposed is specially designed for fusing data from multiple sensors and is applied in this work to solve the multiple camera people tracking problem.

1.1. Related work

The multi-camera people tracking problem has been addressed from different perspectives. In [34], Kang et al. propose a method for tracking multiple objects employing a homography that registers the cameras on top of a known ground plane. Multi-camera tracking is formulated as the maximisation of a joint probability model based on the colour of the blobs detected after background subtraction. The motion of the models is estimated using Kalman filters and the 3D positions of the objects are obtained from the observation of the positions of people's feet. Data association is carried out using the Joint Probabilistic Data Association Filter (JPDAF). In [37], Kim and Davis propose a method for tracking people in multiple-views using a particle filtering approach. After background subtraction, the foreground pixels are classified into blobs that are assigned to the people being tracked. The information from the multiple cameras is then combined to determine the ground plane position. To do so, the centre vertical axes of each person across views are mapped to the ground plane and their intersection point on the ground is estimated. The method requires that people's feet are visible and the ground plane homography. Similarly, the work in [4] employs the floor homography for tracking the 3D positions of objects using a Kalman filter. Ref. [35] also uses the ground plane homography to track people using a look-ahead technique that combines information from multiple frames in order to detect people's paths.

The works revised above require the whole silhouette of the people being tracked to be visible. For this purpose, the cameras must be placed at elevated positions and relatively far from the people. Although this restriction could be feasible in outdoor scenarios, it might be impossible in indoor scenarios where the areas to be covered are small and cameras must be placed nearer to the people. A solution to the tracking problem in indoor environments can be found in Ref. [21]. In that work, Fleuret et al. present a tracking approach using multiple cameras placed at eye level. They employ a generative model to determine the ground locations of people at each frame. For that purpose, the monitored area is discretized into cells to create a probabilistic occupancy map. At each frame, they employ an iterative process in order to determine the locations of the people. Although the authors claim that the computing time is improved by the use of integral images, their applicability imposes restrictions on the camera positions, i.e., the cameras must be placed in such a manner as to prevent people from appearing to be inclined in the images. Furthermore, the complexity of their approach grows exponentially with the size of the area monitored. In [22], the authors describe a distributed tracking system using multiple cameras. At each frame, independent blobs are detected at each camera and passed to a centralised tracker that estimates the 3D people locations. They test both a best hypothesis heuristic tracking approach and a probabilistic multi-hypothesis tracker, reporting similar performance for both methods.

Other authors have employed stereo information in order to enhance tracking. The authors of this work have proposed several approaches for people detection and tracking using a single stereo camera. While Ref. [51] proposes a tracking approach combining colour and stereo extracted directly from camera image, the work in [52] proposes the use of plan-view maps to represent stereo information more efficiently. However, using a single stereo camera still limits the area of surveillance. Therefore, some authors have proposed tracking approach using multiple stereo cameras. In [47], Mittal and Davis present a probabilistic approach for tracking people in cluttered scenes using multiple monocular cameras. They employ a fine camera calibration and compute epipolar lines for each camera pair. People are defined using a cylindrical colour model that registers the colour of their clothes. After background subtraction, the foreground pixels are assigned to the people being tracked. The stereo information of the people being tracked is then extracted by matching foreground segments across adjacent cameras. This information is then projected onto a ground map to detect the positions of the people being tracked. The main drawback of their approach is that their algorithm requires several iterations per frame in order to achieve convergence. In [40], Krumm et al. show a people tracking system for a smart room. In their work, a pair of stereo cameras with a short base line is employed to monitor the area of interest. The extrinsic camera parameters are roughly estimated by matching the paths of people walking in the room in an initial stage. People are detected by grouping 3D blobs extracted from the stereo information. Tracking is performed on ground plane coordinates merging past observations and colour information. Three-dimensional information is also employed in [70] for localising people. In that case, the volumetric information is obtained with a standard visual hull procedure.

An important aspect of the works reviewed above is that tracking is performed only in the area where the cameras field of view (fov) intersect, i.e., the area visible by all cameras. This is a limitation for many applications, specially when the number of cameras grows in order to cover large areas. In that case, there might be a group of cameras with overlapping fovs but not all of them might share a common visible area. Applying a particle filter for fusing information in that case is not straightforward. Each particle could be evaluated independently in each camera and then fuse evidences using a joint approach. However, which likelihood must be assigned to a particle that cannot be even observed? We propose a novel evidential filter based in the Dempster–Shaffer theory of evidence to solve that problem.

Previous works have proposed the use of evidential filters as an alternative to traditional Bayesian filters. In Ref. [32], Kagiwada and Kalaba formulates a theoretical non-linear evidential filter based on dynamic programming and fuzzy sets. Fuzzy sets are employed to compute a degree of belief about the presence of the target in each of the cells in which the space is divided. The belief of a cell is considered the maximum one provided by all sensors thus discarding possibly important information provided by the rest of sensors. Another problem of their approach is that a discretization of the space is required. Moreover, the proposal requires that each cell be estimated, thereby reducing the scalability of the method. In Ref. [42], Mahler designs an evidential filter using the Dempster–Shafer (DS) theory of evidence [57]. This filter is an extension of the Kalman filter which is applicable when the measurement of uncertainty is modelled in the DS domain. However, it can only be employed to model linear systems with Gaussian noise. In Ref. [63], Smets and Ristic develop a novel solution to the tracking and classification problem using the Transferable Belief Model (TBM) as an extension to traditional approaches based on the Kalman filter. In Ref. [64], the authors propose an evidential filter using a particle filtering approach where observations are measured using the Dezert-Smarandache Theory (DSmT). The DSmT is an extension of the DS theory for modelling the paradoxical interpretation of conflicting sources of information. In their work, DSmT is employed to fuse the colour and position features while tracking two people using a single camera. In their approach, the degree of evidence for all targets are calculated at each particle, i.e., a joint configuration is employed. Thus, the main problem of their approach is that the complexity of their filter grows exponentially with the number of targets. Furthermore, they do not deal with the problem of fusing information in case of partial or total occlusion of the targets. Another interesting piece of work related to ours is in Ref. [39]. The authors employ a traditional particle filter scheme for tracking vehicles in roads. To do so, several features are extracted from image patche and combined using the TBM. Then, the fused evidences at each particle are transformed into probabilities using the pignistic transform. The main drawback of their method is that it is specifically designed for the car tracking problem and for a single target.

1.2. Proposed contribution

This work proposes a novel evidential particle filter for tracking multiple targets using a set of heterogeneous and possibly unreliable sensors. The problem is formulated in terms of the DS theory of evidence. The DS theory is a generalisation of the Bayes theory of subjective probability for the mathematical representation of uncertainty. It has been applied to several disciplines such as fraud detection [54], classification [16], risk analysis [13], clustering [15,44], image processing [7,5,28,56], autonomous robot mapping [72], human–computer interaction [69], land mine detection [46] and diagnosis [71], amongst others.

The proposed algorithm models, the possible states of the dynamic system being tracked as a set of particles. For each particle, sensors estimate a degree of evidence that particles will represent the true target state. But the sensors also provide a degree of evidence about their own reliability. In a final data fusion step, data collected from all the sensors are fused to provide the best location hypothesis taking uncertainty into account. The modelling of uncertainty and absence of knowledge of our approach is especially attractive since it does not require specifying priors or conditionals that might be difficult to obtain in complex problems. Since joint particle filters suffer from the curse of dimensionality [66], our algorithm employs a multiple particle filtering approach [18,19,36,51,53,65], i.e., an independent particle filter is employed for each target and possible interactions are considered.

The proposed tracking algorithm, called the Multiple Evidential Particle Filter (MEPF), is a general tracking algorithm for multiple targets using multiple sensors. This algorithm is employed to provide a novel solution to the multi-camera people tracking problem. For each camera, our approach computes a degree of evidence about the possibility of finding the tracked person at the particle location. For that purpose, a generative-based approach that analyses the projection of a 3D person model in the camera images is employed. Foreground, colour and shape information are used to compute a degree of evidence for each camera. Using a depth-ordering scheme, occlusion is calculated separately in each camera. Occlusion is treated by our algorithm as the absence of knowledge about the locations of the people being tracked. In the data fusion step, the evidence collected from all the cameras is fused in order to obtain the best estimation of the target location. Information from unreliable cameras (those with high occlusion or that only partially see the target) is weakly considered.

This paper makes two main contributions. First, an evidential filtering algorithm is proposed for tracking multiple targets by fusing information from multiple unreliable sensors. Since independent trackers are employed instead of a joint configuration, the complexity of the algorithm grows linearly with the number of targets instead of exponentially as in [64]. Moreover, it does not require assuming the Gaussian and linear conditions imposed by the Kalman filter [42]. Second, a novel solution to the multi-camera people tracking problem in indoor environments is proposed. Our approach does not require the whole silhouette of the targets to be visible but undergoes a process of reasoning using the visible portion of the targets while considering the uncertainty associated to the lack of visibility and occlusion. Additionally, it is not necessary to explicitly compute stereo information as in Refs. [47,40,70]. Instead, the locations of people are estimated by intersecting evidence collected from multiple cameras. Furthermore, the proposed approach does not require a discretization of the space as in [21,32], meaning that the scalability of the algorithm is better.

The remainder of this paper is structured as follows. In Section 2 the basis of the DS theory of evidence is explained, while the proposed MEPF algorithm is explained in Section 3. Section 4 shows the proposed multiple camera people tracking solution using the MEPF algorithm. Finally, the experiment is shown in Section 5 and conclusions are drawn in Section 6.

2. Dempster–Shafer theory of evidence

The DS theory, which is also known as the evidence theory, is a generalisation of the Bayes theory of subjective probability. It includes several models of reasoning under uncertainty such as the Smets' Transferable Belief Model (TBM) [59]. The DS approach employs degrees of evidence that are a weaker version of probabilities. The management of uncertainty in the DS theory is especially attractive because of its simplicity and because it does not require specifying priors or conditionals that might be unfeasible to obtain in certain problems. In the DS domain, it is possible to set a degree of ignorance to an event instead of being forced to supply prior probabilities adding to unity.

Let us consider a variable ω taking values in the frame of discernment Ω and let us denote to the set of all its possible subsets by 2^Ω (also called power set). A basic belief assignment (bba)

$$m : 2^\Omega \to [0,1]$$

is a function that assign masses of belief to the subsets A of the power set, verifying:

$$\sum_{A \in \Omega} m(A) = 1. \tag{1}$$

While the evidence assigned to an event in the Bayesian approach must be a probability distribution function, the mass $m(A)$ of a power set element can be a subjective function expressing how much evidence supports the fact A. Furthermore, complete ignorance about the problem can be represented by $m(\Omega) = 1$.

The original Shafer's model imposes the condition $m(\emptyset) = 0$ in addition to that expressed in Eq. (1), i.e., the empty subset should not have mass of belief. However, Smets' TBM model relaxes that condition so that $m(\emptyset) > 0$ stands for the possibility of incompleteness and conflict (see Ref. [60]). In the first case, $m(\emptyset)$ is interpreted as the belief that something out of Ω happens, i.e., accepting the *open-world assumption*. In the second case, the mass of the empty set can be seen as a measure of conflict arising when merging information from sources pointing towards different directions.

Nonetheless, a renormalisation can transform a Smets' bba m into a Demspter's bba m^* as:

$$\begin{aligned} &m^*(\emptyset) = 0, \\ &m^*(A) = \frac{m(A)}{1 - m(\emptyset)} \text{ if } A \neq \emptyset. \end{aligned} \tag{2}$$

One of the most attractive features of DS theory is the set of methods available to fuse information from several sources. Let us consider two bbas m_1 and m_2 representing distinct pieces of evidences, the standard way of combining them is using the conjunctive sum operation [61] defined as:

$$(m_1 \textcircled{\cap} m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \subseteq \Omega. \tag{3}$$

The Dempster's rule of combination can be derived from Eq. (3) by imposing normality (i.e., $m(\emptyset) = 0$) as:

$$(m_1 \otimes m_2)(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \subseteq \Omega, \ A \neq \emptyset, \tag{4}$$

with

$$K = (m_1 \textcircled{\cap} m_2)(\emptyset). \tag{5}$$

The above rules assume that the sources manage independent pieces of information. However, if information is correlated, the cautious rule should be employed [14].

In some applications it is necessary to make a decision and choose the most reliable single hypothesis ω. To do so, Smets and Kennes [62] proposed the use of the pignistic transformation that is defined for a normal bba as:

$$\mathrm{Bet}P(\omega) = \sum_{A \subseteq \Omega, \omega \in A} \frac{m(A)}{|A|}, \tag{6}$$

where $|A|$ denotes the cardinality of A.

MEPF Iteration
From the sample set $C(t-1)$ construct a new sample set $C(t)$ as:

(i) **Resampling and propagation:**

– Calculate the normalised cumulative relevance of each sample cr^i_{t-1} as

$$cr^i_{t-1} = \begin{cases} \dfrac{\mathbf{r}^i_{t-1}}{sum(\mathbf{r}_{t-1})} & \text{if } i = 1; \\ \\ \dfrac{cr^{i-1}_{t-1} + \mathbf{r}^i_{t-1}}{sum(\mathbf{r}_{t-1})} & \text{otherwise,} \end{cases} \tag{7}$$

where $sum(\mathbf{r}_{t-1}) = \sum_{j=1}^{N} \mathbf{r}^j_{t-1}$

– Select the new particles from the set $C(t-1)$ repeating for $i = 1 \dots N$:

· Generate a random number $r, \mathcal{U}(0,1)$ and by binary subdivision find the smallest j for which $cr^j_{t-1} \geq r$.

· Set $c^i_t = c^j_{t-1}$ and propagate the particle state according to the target's movement model:

$$\overrightarrow{x}^i_t = \mathsf{A}\overrightarrow{x}^j_{t-1} + w(t-1) \tag{8}$$

(ii) **Belief Estimation:** For each particle i and each sensor v, calculate the masses $M^v(\overrightarrow{x}^i_t) \in \mathbb{M}^i_t$.

(iii) **Data fusion:** Fuse the data of each particle obtaining $\mathcal{M}^i_t$ and calculate its relevance taking into account interactions with other targets as:

$$\mathbf{r}^i_t = \mathcal{R}(\mathcal{M}^i_t) * \mathcal{I}^i_t. \tag{9}$$

(iv) **State Estimation:** Obtain the best state estimation $\mathbb{E}[C(t)] = \overrightarrow{x}^b_t$ where

$$b = \underset{i}{\operatorname{argmax}}(\mathbf{r}^i_t).$$

Fig. 1. MEPF algorithm.

3. Multiple evidential particle filtering

Using the DS theory, and also inspired in particle based algorithms, in this section we explain the Multiple Evidential Particle Filtering (MEPF) algorithm proposed in this work. The goal of tracking is to estimate the state of a dynamic system. The system might be comprised of a set of n subsystems, each of which has its own dynamics such that

$$\mathbf{X}_t = \{\overrightarrow{x}^1_t, \dots, \overrightarrow{x}^n_t\}.$$

The underlying idea of our algorithm is similar to that employed in particle filtering approaches. The true target state is estimated from a set of possible states (called *particles*). The main difference with regard to particle filtering approaches is that our proposal does not evaluate the likelihood of particles (in the Bayesian sense), but their degree of evidence (in the DS sense). The algorithm is specifically conceived to simultaneously deal with multiple sensors. Hence, the evidence of particles is evaluated using all the available sensors and finally fused. Let us denote the total number of available sensors by V.

To avoid the *curse of dimensionality* that arises when a joint state configuration is employed, a separate tracker is employed for each target. Nevertheless, target interactions are considered by using an interaction factor to maintain multimodality and avoid the *coalescence problem*, as explained below. Each independent tracker keeps a set of N particles. For each particle, each sensor is asked: *Is the target at the particle location?* Using symbols, we define the facts to be evaluated for a sensor at each particle as:

$$\mathscr{S} = \{present, \neg present\},$$

that define the power set:

$$\mathbb{P}(\mathscr{S}) = \{\emptyset, \{present\}, \{\neg present\}, \{unknown\}\}.$$

Then, for each type of sensor, a bba must be defined for the elements of $\mathbb{P}(\mathscr{S})$. By

$$M^v(\overrightarrow{x}) = \{m^v(present), m^v(\neg present), m^v(unknown)\},$$

we shall denote the bba provided by the vth sensor about the subsets in the power set. Mass $m^v(present)$ represents the degree of evidence assigned by the vth sensor to the fact that the target is at $\overrightarrow{x}$. On the other hand, mass $m^v(\neg present)$ represents the evidence that the target is not at $\overrightarrow{x}$. Finally, $m^v(unknown)$ represents the degree of evidence of the sensor itself, i.e., high values of $m^v(unknown)$ denote that the sensor is not reliable for that particle.

On the basis of the power set just defined, let us denote the set of particles of each tracker by:

$$C(t) = \{c^i_t = \langle \overrightarrow{x}^i_t, \mathbb{M}^i_t, \mathscr{M}^i_t, \mathbf{r}^i_t \rangle | i = 1, \dots, N\}. \tag{10}$$

The parameter

$$\mathbb{M}^i_t = \{M^v(\overrightarrow{x}^i_t) | v = 1, \dots, V\}$$

represents all the bbas provided by the V sensors about the state $\overrightarrow{x}_t^i$, and

$$\mathscr{M}_t^i = M^{1,\dots,V}\left(\overrightarrow{x}_t^i\right) \tag{11}$$

represents the bba resulting from fusing the evidence in $\mathbb{M}_t^i$ using the most appropriate combination rule. Its selection depends on the nature of the data manipulated. If the pieces of evidence to be fused are uncorrelated, then either the conjunctive sum operation (Eq. (3)), or the Dempster's rule of combination (Eq. (4)) might be employed. However, if they are correlated, the cautious rule should be employed [14].

The relevance of a particle $\mathbf{r}_t^i$ is a single value indicating the likelihood that the particle will represent the true target's state. It is computed using the mapping function $\mathscr{R}$ and an interaction factor $\mathscr{I}_t^i$. The function

$$\mathscr{R}(M) = \mathrm{Bet}P(\mathit{present}) \tag{12}$$

is calculated as the pignistic probability (Eq. (6)) of the *present* event in M. The interaction factor $\mathscr{I}_t^i$ models target interactions in order to maintain multi-modality and avoid the coalescence problem [9,36,51,65]. The coalescence problem occurs when two (or more) targets with similar characteristics are close to each other. In that case, the target that obtains a higher relevance might "hijack" the particles of the rest of the trackers. Imagine for example the problem of tracking people based on the colour of their clothes. In this case, two people wearing the same clothes might be indistinguishable when they come close to each other. If any of the people are severely occluded in all the cameras, the particles of their tracker will move towards the position of the visible target. The interaction factor is defined such that it tends to 0 when the particles of a tracker are near the positions of other targets and tends to 1 when the particle is far from other targets. Therefore, the relevance of particles near other targets diminishes. The interaction factor of a particle can be defined as a function that is inversely proportional to the distance from the nearest target. Since the positions of the targets in time t are not known, the position estimated by the algorithm in the previous time step is employed. For further information on the role of the interaction factor the reader is referred to [9,36,48,51].

The outline of the proposed algorithm is shown in Fig. 1. At the beginning, the algorithm is provided by an initial sample set $C(0)$ of N particles. The particles in $C(0)$ might be sampled around the initial target position using any suitable distribution. At each iteration, the algorithm uses the particle set $C(t-1)$ to create a new set $C(t)$ by selecting, with replacement, N particles from $C(t-1)$. For that purpose, the cumulative normalised relevance of the particles is calculated first. Using binary subdivision, the new particles are then selected by finding the particle whose cumulative normalised relevance is nearer a selected random number r. This resampling mechanism permits particles with a high relevance to be selected a greater number of times than particles with a low relevance that are rapidly discarded from one iteration to another. As can be observed, this is the approach employed in the CONDENSATION algorithm [31]. Afterwards, for each selected particle, the algorithm computes its next state $\overrightarrow{x}_t^i$ according to a dynamic model of the system. (Eq. 8) propagates the state using a transition model A affected by some noise $w(t-1)$.

Once the new particle set is obtained, all the sensors are employed to calculate the masses of power set $\mathbb{M}_t^i$. Afterwards, all the evidence collected for each particle is fused into $\mathscr{M}_t^i$ and the particle relevance $\mathbf{r}_t^i$ is computed using the mapping function $\mathscr{R}$ and an interaction factor $\mathscr{I}_t^i$.

Finally, the algorithm provides the best state estimation $\mathbb{E}[C(t)]$ as the state of the particle with a higher relevance $\overrightarrow{x}_t^b$. The main difference with respect to traditional particle filters is that the Bayesian conditions are relaxed, thus allowing non-probabilistic distributions to be used when estimating the particle evidence. Furthermore, the use of the DS theory allows the reliability of the sensors to be modelled easily.

4. MEPF for tracking people in multiple cameras

This section explains how the MEPF algorithm is employed for tracking people using several cameras. First, we provide a brief overview of the algorithm. A detailed explanation of the elements required to implement the proposed algorithm is then given.

4.1. Algorithm overview

The purpose of our people tracking problem is to estimate the ground plane positions

$$\mathbf{X}_t = \{\overrightarrow{x}_t^1, \dots, \overrightarrow{x}_t^{n_t}\}$$

of a set of people in the area of analysis. Let n_t represent the number of people being tracked at time t and $\overrightarrow{x}_t$ a position in the ground plane. Let us assume that there is a set of V heterogeneous cameras sharing a common reference system obtained by calibration, thus making it possible to know the projection of a three-dimensional point in each of the cameras. Please notice that a fine camera calibration is not required since epipolar lines are not employed in this work. Let us also assume that people are mostly seen in a standing position and that there is a people detector mechanism which indicates the positions of the people entering the area under surveillance in an initial time step.

The outline of the proposed algorithm can be seen in Fig. 2. In an initial stage, a background model for each camera is created. The background modelling technique proposed in Ref. [20] has been employed in this work. Afterwards, the tracking

```
MEPF Iteration for People Tracking
From the sample set $\mathcal{P}(t-1)$ construct the new set $\mathcal{P}(t)$ as:

  (i) Background Subtraction: perform background subtraction and obtain the foreground map $\mathcal{F}^v$ for each camera image.
 (ii) Resample and propagate: the particle sets $C_p(t-1)$.
(iii) Evidence Estimation: Repeat for each camera $v = 1 \dots V$
      – Reset the occlusion map $\mathcal{O}^v$
      – Sort people incrementally by their distance to the camera
      – For each person $p = 1 \dots n_t$
            · Calculate the masses $M^v(\overrightarrow{x}^i_{p,t}) \in \mathbb{M}^{i,p}_t$ of the particles.
            · Obtain the particle with higher relevance $r^{i,v}_{p,t} = \mathcal{R}(M^v(\overrightarrow{x}^i_{p,t}))$ for this camera.
            · Project the 3D person model in $\mathcal{O}^v$ using the position of the particle with higher relevance.
 (iv) Data fusion: Fuse the data of the particle in every particle set $C^p(t-1)$ obtaining $\mathcal{M}^i_{p,t}$. Then, calculate the relevance of particles
      taking into account interactions with other targets as:
            $\mathbf{r}^i_{p,t} = \mathcal{R}(\mathcal{M}^i_{p,t}) * \mathcal{I}^i_{p,t}.$          (13)
  (v) State Estimation: Obtain the best state estimation $\mathbb{E}[C_p(t)]$.
 (vi) Colour models and background update: Update the colour models $A_p(t-1)$ of each person using the estimation obtained and
      the background models for each camera.
```

Fig. 2. MEPF for people tracking.

process starts. Following this stage, the image set is captured and background subtraction is performed. By $\mathscr{F}^v$ let us denote the foreground map obtained for the vth camera. A pixel of the foreground map is 1 when it is classified as foreground and 0 otherwise. The trackers then iterate in order to estimate the new locations of the people being tracked.

By

$$\mathscr{P}(t) = \{P_p(t) | p = 1, \dots n_t\},$$

let us denote the information that each tracker keeps about its target p. $P_p(t)$ is defined as:

$$P_p(t) = \{C_p(t), \mathbb{E}[C_p(t)], A_p(t)\},$$

where $C_p(t)$ represents the particle set (Eq. (10)) and $\mathbb{E}[C_p(t)]$ the best estimation obtained from the particle set. Additionally, each tracker keeps a colour model of the clothes of its target

$$A_p(t) = \{a^v_{p,t} | v = 1, \dots, V\},$$

in each of the cameras. Since we consider that the scene might be analysed by cameras with different sensor characteristics and that illumination is not uniform, a point in the scene might be seen with a different colour in each of the cameras. Therefore, a different colour model $a^v_{p,t}$ is kept for each camera.

Particle propagation is performed using a random walk movement model because of the unpredictable behaviour of people. Therefore, the matrix modelling the dynamics of system A (Eq. (8)) is set to the identity. The random noise applied is assumed to follow a Gaussian distribution $N(0, \sigma^2_m(t-1))$ whose deviation is calculated as:

$$\sigma^2_m(t-1) = \frac{2\hat{\sigma}^2_m}{1 + \mathbf{r}^b_{t-1}}. \tag{14}$$

When the person is properly located, the relevance of the best particle $\mathbf{r}^b_{t-1} \simeq 1$. The deviation is then set to the minimum value, $\hat{\sigma}^2_m$, representing the distance walked by a person at normal speed. However, as the relevance decreases (indicating that the location of the person is not properly known) noise is increased. Consequently, particles are spread over a wider search area in an attempt to relocate the target in the next iteration. Value $\hat{\sigma}^2_m$ is calculated based on the fact that average human walking speed is about 1 m/s (3.6 km/h). Then, if the proposed system is able to operate at *fps* hertz, the parameter $\hat{\sigma}^2_m$ is calculated as:

$$\hat{\sigma}_m = \frac{1}{fps}. \tag{15}$$

After propagation, the algorithm proceeds with the particle evaluation. For each particle, each camera evaluates the evidence of the person at the particle position. Hence, the masses $M^v(\overrightarrow{x}^i_{p,t}) \in \mathbb{M}^i_{p,t}$ are calculated for each camera and particle. For that purpose, a generic 3D model of a person is rendered in each camera image, assuming that the model is placed at the particle position. Then, the number, shape and colour of the foreground points in the projections are analysed. The number and shape of the foreground points are evaluated in order to see if the model is projected in an occupied region of the space. The colour of the foreground points is employed to create a colour model that is compared against the person colour model $a^v_{p,t}$. The colour models $A_p(t)$ are initialised from the information of the first frame. Later, they are dynamically updated in order to be adapted to illumination changes and body movements.

One of the most attractive advantages to using multiple cameras is the management of occlusion. When a person is occluded in a camera, another camera might be employed to keep track of that person. The algorithm proposed in this work is specifically designed to deal with occlusions. For that purpose, an occupation map $\mathscr{O}^v$ is maintained for each camera. The occupation map has the same dimensions as the original camera image and indicates at each pixel if it is occupied by any of the people being tracked. The occupation map is calculated independently for each camera using a depth-ordered approach. First, targets are sorted according to their distance from the camera. Then, starting from the nearest person to the camera, the evidence of their particles is calculated. The particle with a higher relevance $\mathbf{r}^{i,v}_{p,t} = \mathscr{R}\left(M^v\left(\overrightarrow{x}^i_t\right)\right)$ in a camera is selected as the best position for that camera. Please notice that $\mathbf{r}^{i,v}_{p,t}$ does not represent the relevance of the fused masses but the relevance in camera v. Therefore, the selected position might not be the best global solution but is a good local solution that allows us to calculate the occlusion independently for each camera. Thus, step (iii) of the algorithm in Fig. 2 can be distributed in as many processes as cameras. The position of the particle with higher relevance in the camera is employed to project the 3D person model in the occupation map $\mathscr{O}^v$ (setting all the points inside the projection to 1). Afterwards, the particles of the next person are evaluated, but this time employing the occupation map to take occlusions into account (this is explained in greater depth in Section 4.4). In brief, particles projecting at image positions already occupied by other people are assigned high values of uncertainty, i.e., high values of $m^v(unknown)$. Therefore, in the data fusion step, cameras in which a person is occluded are not as relevant to determining the person's location as are the cameras in which the person is fully visible.

When the particle sets have been evaluated in all the cameras, data are fused in order to obtain a global estimation of people's locations. We have employed for this application the Dempster's combination rule (Eq. (4)) thus assuming independence between the cameras employed. This holds true while cameras see the target from separated points of view. However, as the number of cameras in the environment increases, so does the correlation between adjacent cameras. In that case, it would be appropriate to develop further fusion strategies considering the cameras degree of correlation (using for instance the cautious rule [14]).

The relevance of particles at this stage represent a global solution that takes into account both the information from all the sensors and the information about the rest of the targets via the interaction factor $\mathscr{I}^i_{p,t}$. In this work, the interaction factor is defined as a Gaussian function

$$\mathscr{I}^i_{p,t} = 1 - e^{-\frac{dm^2}{2\sigma^2_{dm}}}, \tag{16}$$

where dm is the Euclidean distance to the nearest target (excluding itself) and σ_{dm} the deviation. The deviation is set to $\sigma_{dm} = 0.5$ m, which corresponds to the width of the 3D model employed. The interaction factor tends to 0 for particles drawn near the location of other targets, and tends to 1 when it is far from other targets. Therefore, particles drawn near the location of other targets are considered inappropriate so they are provided with a low relevance value thus avoiding the *coalescense* problem.

Using the fused information, the best location hypothesis $\mathbb{E}[C_p(t)]$ is estimated. Afterwards, in step (vi) of the algorithm, the colour models $A_p(t-1)$ are updated using the information from the best hypothesis. The model of each view is updated according to its visibility, i.e., the colour models of views where the person is highly visible are more strongly modified than the colour models of views where the person is partially occluded (see Section 4.5 for further details). Finally, the background models are smoothly adapted to changes in the environment. To prevent people standing for long periods of time from becoming part of the background model, pixels marked as occupied in the occupancy maps $\mathscr{O}^v$ are not updated.

In the following sections, we give a detailed explanation of the elements required to implement the proposed algorithm. Section 4.2 shows the 3D geometric model employed to model people's appearance and the information extracted from its projection, while the people colour models are explained in Section 4.3. Section 4.4 shows how the masses of the particles are calculated. Finally, Section 4.5 explains how the colour models are updated to adapt them to illumination and body pose changes.

4.2. 3D model projection

The proposed method relies on the use of a geometric 3D model representing the shape of people. We have selected a basic model consisting of a box whose dimensions have been selected taking into account the dimensions of an average adult person. It has been assumed that the box is 0.5 m in width and 1.8 m in height. Although the model dimensions are fixed in this work, they can be adapted to the particular characteristics of the people being observed. Since the cameras are calibrated, it is possible to calculate the projection of the 3D model in a given position $\overrightarrow{x}$. Let us define by

$$pm(\overrightarrow{x})^v = \{p_i = (x_i, y_i)\},$$

the image pixels of the vth camera image that lies in the projection of a 3D model placed at $\overrightarrow{x}$. Fig. 3 shows the projection of the model employed in four different cameras. Although in practise a solid model is employed, Fig. 3 shows its wired version for viewing purposes.

Note that some of the pixels in $pm(\overrightarrow{x})^v$ might belong to background pixels and are therefore not relevant. But some of the pixels in $pm(\overrightarrow{x})^v$ might have already been set as belonging to another person in the occupancy map $\mathscr{O}^v$. Then, let us denote by

$$fpm(\overrightarrow{x})^v = \{p_i | \mathscr{F}_{p_i} = 1 \wedge p_i \in pm(\overrightarrow{x})^v\},$$

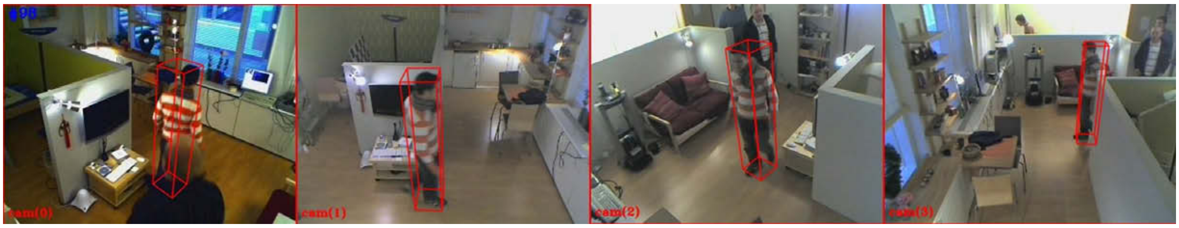

Fig. 3. Projection of the 3D geometric model employed for tracking people. Cameras are calibrated in order to calculate the model projection in each camera.

the pixels in $pm^v(x)$ that are foreground pixels $(\mathscr{F}_{p_i} = 1)$. Also, let us define by

$$\mathit{vpm}(\overrightarrow{x})^v = \{p_i | \mathscr{O}_{p_i} = 0 \wedge p_i \in \mathit{fpm}(\overrightarrow{x})^v\},$$

the pixels from fpm that have not yet been occupied by other people, i.e., $\mathscr{O}_{p_i} = 0$. Finally, let us also define $\mathit{Vis}(\overrightarrow{x})^v$ as a measure that indicates the visibility of the model projection in the vth camera. This measure accounts for the possibility that the model will not project entirely in the camera plane. The measure $\mathit{Vis}(\overrightarrow{x})^v$ is 1 when the whole model is projected in the camera image. However, it tends to 0 as the model projects outside the camera's field of view. So, $\mathit{Vis}(\overrightarrow{x})^v = 0$ means that the particle is not visible from the vth camera.

4.3. Person colour model

A colour histogram $a^v_{p,t}$ is maintained at each camera to model the colours of the clothes of the person being tracked. Colour histograms have often been used for modelling colour in tracking problems since they allow the global properties of objects to be captured with invariability to scale, rotation and translation [11]. In this work, histograms are created using the colour of the non-occluded foreground pixels in a model projection, i.e., $\mathit{vpm}(\overrightarrow{x})^v$. The *HSV* colour space [23] has been employed because it is relatively invariable to illumination changes. A histogram is comprised of $n_h n_s$ bins for the hue and saturation. However, as chromatic information is not reliable when the value component is too low or too high, pixels in that situation are not used to describe the chromaticity. Because these "colour-free" pixels might contain important information, histograms are also populated with n_v bins to capture their luminance information. Thus, histograms are composed of $m = n_h n_s + n_v$ bins. Let us define a function $b : \Re^2 \rightarrow \{1, \dots, m\}$ which associates a pixel p_i with the index of the histogram bin $b(p_i) = w$ corresponding to its colour. Then, the wth bin of a histogram is calculated as

$$a^v(w) = \frac{\sum_{p_i \in \mathit{vpm}(\overrightarrow{x})^v} k[b(p_i) - w]}{|\mathit{vpm}(\overrightarrow{x})^v|}, \tag{17}$$

were k is the Kronecker's delta function and $||$ denotes the cardinal. Please notice that the histogram bins are normalised:

$$\sum_{w=1}^{m} a^v(w) = 1.$$

4.4. Degrees of evidence calculation

This section explains the basic probability assignment for the masses $M^v(\overrightarrow{x}^i_{p,t})$ of each particle in each camera. For the sake of clarity, scripts i, p, v and t are omitted.

As previously explained, the masses of the power set elements are evaluated for each particle-camera pair:

$$M(\overrightarrow{x}) = \{m(\mathit{present}), m(\neg \mathit{present}), m(\mathit{unknown})\}.$$

The mass $m(\mathit{unknown})$ is the degree at which a sensor cannot provide a solution to the problem. This can be seen as the uncertainty of the sensor or as the inability of the sensor to decide between the two other subsets *present* and ¬*present*. The mass $m(\mathit{unknown})$ is modelled by two components: an occlusion measure and the visibility measure $\mathit{Vis}(\overrightarrow{x})$.

The occlusion measure, $\mathit{Oclu}(\overrightarrow{x})$, indicates the portion of points of the model projection that are occluded by other people. It is defined as:

$$\mathit{Oclu}(\overrightarrow{x}) = 1 - \frac{|\mathit{vpm}(\overrightarrow{x})|}{|\mathit{fpm}(\overrightarrow{x})| + \epsilon},$$

where ϵ is a small value to prevent dividing by 0. The mass of the *unknown* subset is then defined as:

$$m(\mathit{unknown}) = 1 - (\mathit{Oclu}(\overrightarrow{x}) * \mathit{Vis}(\overrightarrow{x})). \tag{18}$$

The value calculated in Eq. (18) tends to 0 for fully visible particles with low occlusion. However, $m(unknown)$ increases as the visibility is reduced or the portion of occlusion becomes greater.

The masses $m(present)$ and $m(\neg present)$ are calculated simultaneously since we consider that the former complements the latter. While $m(present)$ denotes the evidence of a particle to be placed at the person's location, $m(\neg present)$ means exactly the opposite. It has been assumed that a particle is likely to be at the person's location (thus obtaining high values of $m(present)$) if three conditions are met. Firstly, the particle should be placed at an occupied region of the space (i.e., empty regions are unlikely to be occupied by people). Secondly, the particle should be projected in the centre of the target instead of in its boundaries. Thirdly and finally, the colour distribution of the foreground points in the particle projection should be similar to the colour distribution of the target. The overall idea is that a particle is assigned high values of $m(present)$ if it projects in a region of the space with sufficient foreground points, they are centred, and their colour distribution is the same as the colour model of the person being tracked. These three conditions are evaluated by the three measures $Occ(\overrightarrow{x})$, $Centr(\overrightarrow{x})$ and $Cd(\overrightarrow{x})$ explained below.

The first measure, $Occ(\overrightarrow{x})$, indicates if the amount of foreground points in the image region where the model projects is appropriate to consider it occupied by a person. For that purpose, let us define

$$bckg(\overrightarrow{x}) = 1 - \frac{|fpm(\overrightarrow{x})|}{|pm(\overrightarrow{x})| + \epsilon}, \tag{19}$$

as the proportion of background points of the image region where the model projects. The measure $Occ(\overrightarrow{x})$ is calculated by applying a Butterworth filter to the previous measure. Butterworth filters are defined as:

$$B(f, f_c, n) = \frac{1}{1 + \left(\frac{f}{f_c}\right)^{2n}},$$

where parameter n is the order of the filter controlling the smoothness of the curve and parameter f_c is a cutoff value (see Fig. 4). The filter response is 1 when f is smaller than f_c and tends to 0 as f becomes greater than f_c. The occupancy measure is then defined as:

$$Occ(\overrightarrow{x}) = B(bckg(\overrightarrow{x}), \theta_{occ}, \gamma_{occ}). \tag{20}$$

Therefore, when the proportion of background points in the projection of the model is below θ_{occ}, $Occ(\overrightarrow{x})$ is 1, indicating that the region is properly occupied. However, as the proportion of background points increases, the value of $Occ(\overrightarrow{x})$ decreases, thus indicating that the region is empty.

The second measure, $Centr(\overrightarrow{x})$, indicates whether the mass centre of the foreground points coincides with the mass centre of the projected model. The goal is to assign higher degrees of evidence to particles projected in the centre of the target than to particles projected in its boundaries. For that purpose, let us define a function that calculates the centre of mass of a point set ps as:

$$Cent(ps) = \sum_{p_i \in ps} \frac{p_i}{|ps|}. \tag{21}$$

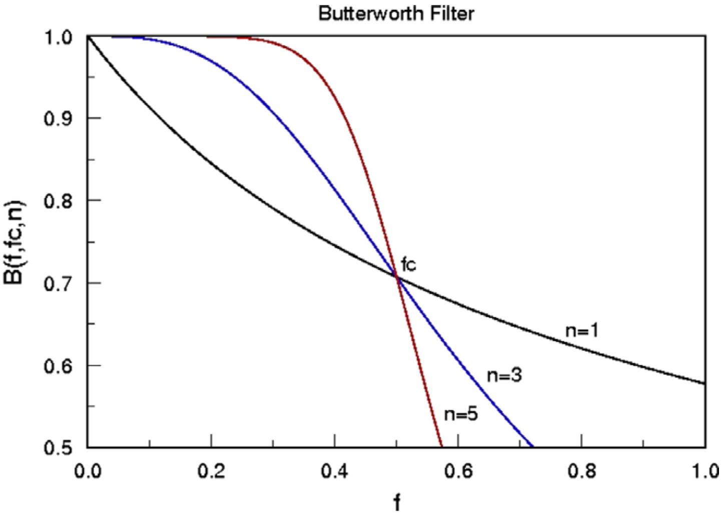

Fig. 4. Response of the Butterworth filter for different configurations.

Then, $Cent(pm(\overrightarrow{x}))$ denotes the mass centre of the model projection and $Cent(fpm(\overrightarrow{x}))$ the mass centre of the foreground points. We therefore define the distance

$$dn(\overrightarrow{x}) = \frac{\|Cent(pm(\overrightarrow{x})) - Cent(fpm(\overrightarrow{x}))\|}{\sqrt{pm(\overrightarrow{x})^2_{height} + pm(\overrightarrow{x})^2_{width}}}, \tag{22}$$

as a normalised distance between the two centres. In Eq. (22), $\| \|$ denotes the Euclidean distance, while $pm(\overrightarrow{x})_{height}$ and $pm(\overrightarrow{x})_{width}$ represent the height and width of the model projection, respectively. Therefore, the denominator represents the maximum possible distance between two points in the rectangle enclosed by the model projection. Normalisation is done in order to achieve independence from the distance of the particle to the camera. Finally, $Dn(\overrightarrow{x})$ is modelled as a Gaussian distribution

$$Dn(\overrightarrow{x}) = exp\left(-\frac{dn(\overrightarrow{x})^2}{2\sigma^2_{dn}}\right), \tag{23}$$

with $\sigma_{dn} = 1/3$ so that $Dn(\overrightarrow{x}) \simeq 0$ when $dn(\overrightarrow{x}) = 1$.

Finally, measure $Cd(\overrightarrow{x})$ represents the colour distance between the colour distribution of the pixels in $vpm(\overrightarrow{x})$ and the colour model of the person (see Section 4.3). A popular approach for measuring the similarity of two distributions is the Bhattacharyya distance [1,33]. The Bhattacharyya distance of two colour histograms ρ_1 and ρ_2 is calculated as:

$$cd(\rho_1, \rho_2) = \sqrt{1 - \sum_w \sqrt{\rho(w)_1 \rho(w)_2}}. \tag{24}$$

The distance $cd(\rho_1, \rho_2)$ is 0 when both colour histograms are identical and tends to 1 as they differ. Using the Bhattacharyya distance we define the measure

$$Cd(\overrightarrow{x}) = cd(a, \hat{a}) \tag{25}$$

indicating the colour distance between the colour histogram of the target (a) and the colour histogram of the points in the particle projection $vpm(\overrightarrow{x})$ (denoted as $\hat{a}$).

Using the three measures explained above, the mass of the *present* subset is defined as:

$$m(present) = (1 - m(unknown)) * Occ(\overrightarrow{x}) * Dn(\overrightarrow{x}) * Cd(\overrightarrow{x}). \tag{26}$$

As can be noticed, $m(present)$ has high values when the $m(unknown)$ is low, the number of foreground points in the projection of the 3D model is high, they are centred and their colour distribution is similar to the colour distribution of the person being tracked.

Finally, let us define:

$$m(\neg present) = 1 - (m(unknown) + m(present)) \tag{27}$$

so that the sum of masses is equal to one.

4.5. Colour model update

Changes in illumination conditions and body movements might alter the observed colour distribution of a person's clothes. It is therefore necessary to continuously update the people's models $a^v_{p,t}$. These are updated using the colour models of the best estimated hypothesis $\mathbb{E}[C_p(t)]$ at each iteration. Let us denote by $\hat{a}^{v,b}_{p,t}$ the colour model in the vth view of the best particle evaluated (the one with higher fused relevance $\mathbf{r}^b_t$). Then, the bins of the colour histograms are updated as:

$$a^v_{p,t+1}(w) = \left(1 - \lambda^v_{p,t}\right)a^v_{p,t}(w) + \lambda^v_{p,t}\hat{a}^{v,b}_{p,t}(w), \tag{28}$$

where parameter $\lambda^v_{p,t} \in [0, 1]$ weights the contribution of the observed colour model to the updated one. In this work, this parameter is set to the mass of the *present* subset in the vth view

$$\lambda^v_{p,t} = m^{v,b}_{p,t}(present). \tag{29}$$

Please notice that $m^{v,b}_{p,t}(present)$ is not a fused evidence but the evidence calculated before the fusion step. Then, the colour model of each view is updated independently according to its circumstances. Parameter $\lambda^v_{p,t}$ is near 1 when the person is highly visible and the colour models $a^v_{p,t}$ and $\hat{a}^{v,b}_{p,t}$ are similar. If the occlusion is high or the colour models are different, $\lambda^v_{p,t}$ tends to 0. The goal is to prevent rapid colour changes that might be caused by occlusions or momentary tracking failures. Therefore, we are assuming that light changes occur smoothly.

Finally, the updated histogram is normalised so that their bins sum up to one.

5. Experimental results

This section explains the experiment conducted to test the proposed algorithm. Several video sequences have been recorded in two different scenarios. The first scenario is the PEIS room [41]; a robotised apartment employed for the development and research of mobile and embedded robotic systems. Four usb web cams were placed at a height of $\simeq 3$ m in slanting positions to cover an area of 3×4 m. The cameras were synchronised via software and set to record at 7 fps with a resolution of 320×240 pixels. In the second scenario, a total of five firewire cameras were placed at a height of 3 m to cover an area of approximately 3×3 m. The cameras were synchronised via software and set to record at 5 fps with a resolution of 320×240 pixels. The number of people in the recorded sequences varied from 2 to 6. The people were instructed to move about freely in the environment. Therefore, interactions and occlusions are frequent in the recorded videos.

The performance of the proposed algorithm depends on a set of parameters that needs to be estimated. These parameters are the number of particles N, the number of bins of the colour histograms n_h, n_s, n_v and the occupancy parameters θ_{occ} and γ_{occ} of Eq. (20). In order to determine the values for these parameters, the positions of the people in one of the sequences have been determined in each frame. For that purpose, a camera was mounted on the ceiling of the first scenario and synchronised with the usb cameras. The positions of the people being tracked were then extracted in a total of 2500 frames. In this way, quantitative measures of the tracking error can be obtained. The sequence was recorded in the first scenario and shows three people entering the environment and moving about while discussing a topic. Images of the sequence can be seen in Fig. 8 (explained later).

The best parameter configuration has been estimated in two phases in order to reduce the search space. In the first phase, the sequence was processed for $\theta_{occ} = \{0, 0.05, \ldots, 1\}$, $\gamma_{occ} = \{1, 2, 3, 4\}$ and $n_h = n_s = n_v = \{0, 2, 5, \ldots, 14\}$. The goal was to determine the best values for these parameters. It has been assumed that the quality of the colour acquisition is equal in all the colour channels so that $n_h = n_s = n_v$. In the first phase, the number of particles was set to a large enough value ($N = 300$) in order to prevent tracking failures due to the lack of particles. Because of the stochastic nature of the algorithm, the tests were repeated ten times with different seeds for the random number generator. The error measure employed is the root mean-square error (RMSE) of the manually extracted positions and the positions indicated by the trackers in the ten runs. The results can be seen in Fig. 5.

The graph labelled $n_h = n_s = n_v = 0$ represents the case when no colour information is employed. In that case, tracking is based exclusively on position information, i.e., people are tracked by intersecting the foreground information from all the cameras. As can be seen, the algorithm is very sensitive to the occupancy parameters in this case. In general, a value of $\theta_{occ} > 0.25$ is required in order to obtain good results. This means that at least one quarter of the points in the model projection are frequently background points. This occurs for two reasons: because the model is normally bigger than real person dimensions and because of errors in the foreground segmentation.

In the cases where $\theta_{occ} = 1$ and $\gamma_{occ} > 1$, background information is not employed, i.e., tracking is based mostly on colour information. For $\gamma_{occ} = 1$, the smooth curve transition means that foreground information is still to be considered. As γ_{occ} increases, the relevance of foreground information becomes null. It can be noticed that as θ_{occ} increases, a higher number of histogram bins is required in order to obtain good results, i.e., as foreground information becomes less relevant a more precise colour model is required. Nevertheless, the algorithm does not perform well when tracking is based exclusively on colour information. This might be explained by a drift in the colour models due to changes in illumination conditions during tracking. As can be seen, the best tracking performance is normally obtained for intermediate values of θ_{occ} and high values of the number of the histogram bins. As regards parameter γ_{occ}, we have observed that it is preferable to set low values for this parameter to obtain a smooth transition of the Butterworth filter. In light of the results obtained, we consider that good values for the parameters are $n_h = n_s = n_v = 8$, $\theta_{occ} = 0.65$ and $\gamma_{occ} = 1$. Although higher values for the number of bins also produce good results, we consider that 8 bins constitute an appropriate trade-off between performance and computational cost.

In a second phase, the impact of the number of particles on tracking error is analysed. The more particles employed, the higher the computational effort required, but also the higher the precision obtained. However, there must be a limit to the

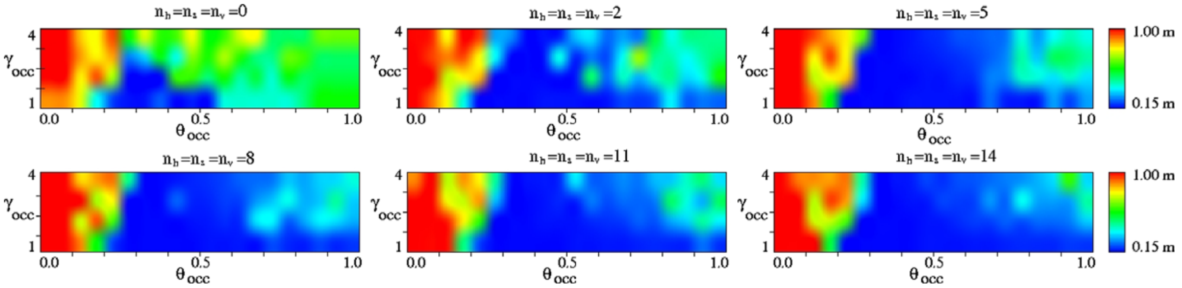

Fig. 5. Tracking errors for different values of the algorithm parameters.

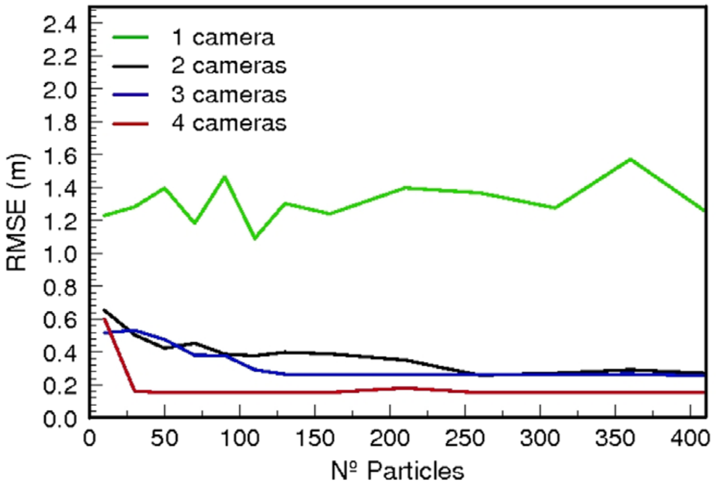

Fig. 6. Error evolution for several camera configurations as the number of particles grows.

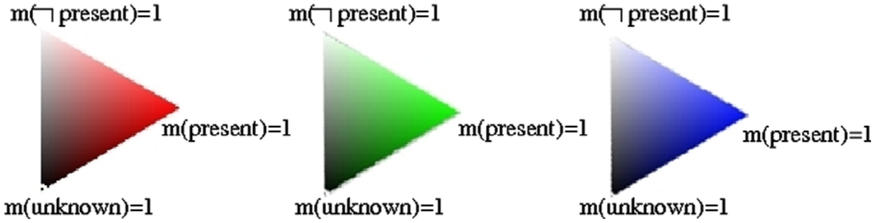

Fig. 7. Colour map used to represent the degrees of evidence in Fig. 8.

number of particles over which no significant improvement is obtained. Thus, it is desirable to determine the minimum number of particles required for an optimal trade-off between precision and computational cost. For that purpose, the sequence has been processed using the best parameter selection of the previous phase and a different number of particles. Furthermore, the sequence has been processed with an increasing number of cameras in order to determine the impact of the number of views on algorithm performance. The results obtained are shown in the graph of Fig. 6. The horizontal axis of the graph represents the number of particles employed for each tracker. The vertical axis represents the RMSE in determining the people's position. The RMSE for the different camera configurations are depicted with different coloured lines. As expected, error is reduced as the number of particles is increased. However, it can be observed that the reduction is greater from 15 to 30 particles. In fact, in the best camera configuration (four cameras) no error reduction is obtained for more than 30 particles. In that case, the mean error is 0.15 m.

The tests have been performed on an AMD Turion 3200 portable computer with 1 GB of RAM running Linux. In our tests, foreground extraction and colour conversion consumes 10 ms (for each image), while another 10 ms are required for the background update. Evaluation of the masses of 30 particles requires 11 ms for each view, while the final data fusion step requires 2 ms (for the four cameras).

Fig. 8 shows some scenes from the previously analysed sequence. In the sequence, three people enter the room and talk for approximately 3 min. The people move about the room causing frequent occlusions to each other in some of the cameras. The figure shows the tracking results in four different time instants. The odd rows show the camera images at a particular time instant. In the images, the models have been drawn at the position estimated by the tracker. Below the camera images, the figure shows a ground map of the monitored area where the location and orientation of the cameras have been superimposed. The particles have been drawn in the ground maps in the form of circles whose colour indicates their degree of evidence. The colour scheme employed is indicated in Fig. 7. Each target is represented by a different colour: red,[1] green and blue. The particles of the red target are drawn in pure red when $m(present) = 1$ and $m(unknown) = m(\neg present) = 0$. Black is used for particles with $m(unknown) = 1$ and $m(present) = m(\neg present) = 0$ and white for particles with $m(\neg present) = 1$ and $m(unknown) = m(present) = 0$. The rest of possible intermediate values for the masses of a particle are represented by the colours inside the corresponding triangle. The maps labelled *fusion* show the evidence resulting from the data fusion step.

[1] For interpretation of colour in figures, the reader is referred to the Web version of this article.

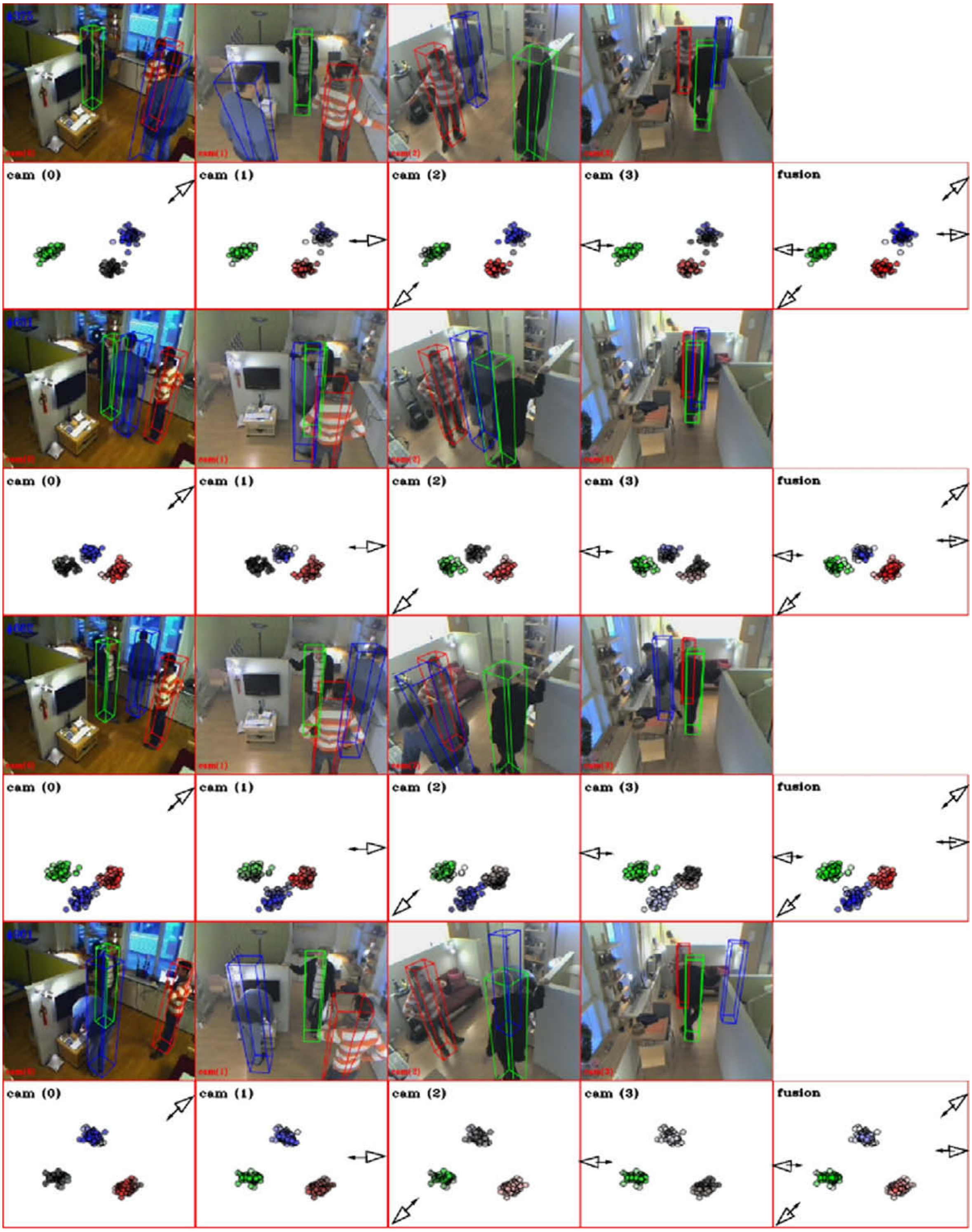

Fig. 8. Tracking results for three frames of a sequence. The upper rows show the camera images superimposing the positions estimated by the proposed tracking algorithm. In the bottom rows the evaluated particles are drawn in a ground map of the scenario. The colour of each particle represents the masses calculated according to the colour scheme shown at the bottom. See text for details.

In the first frame analysed, the red target is occluded by the blue one in the first camera. It can be noticed that the colour of the particles for the first camera are drawn in dark grey, thus indicating that the target is occluded in the camera. Nevertheless, since the target is properly seen in the rest of the cameras, the final location estimated by the tracker is very accu-

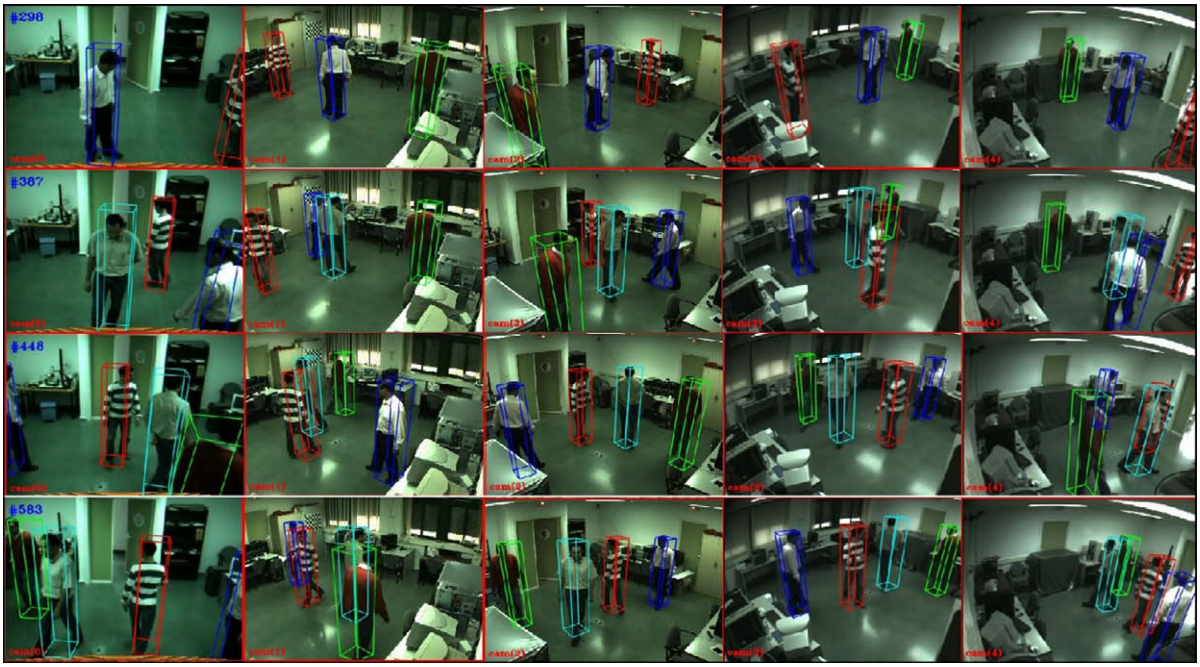

Fig. 9. Four people move randomly about a lab causing frequent occlusions and leave the field of view of some of the cameras.

rate. In the second frame, the blue target is passing between the other two targets. In this situation, the person is occluded simultaneously in at least two of the cameras. Nevertheless, as can be seen in the third frame, tracking can be done properly. Finally, the fourth frame shows another interesting situation: the blue target bends over to type on a keyboard. In this case, the portion of foreground points seen for the person is very low, thus causing the algorithm to assign a low relevance to the evaluated particles. However, the best hypothesis evaluated provides a good estimation of the person's location as can be seen in the figure.

Fig. 9 shows one of the test sequences recorded in the second scenario. In this sequence, four people enter the room and start to walk around it randomly. It is a difficult tracking situation since three of the people are wearing clothes with very similar colour distributions. The targets labelled in blue, light blue and red are wearing black and white clothes. Despite this, the algorithm is able to track them without confusing their identities and avoiding the coalescence problem. Another aspect that is worth mentioning about this video is that the people are not seen simultaneously in all the cameras. Certain positions of the environment are only covered by a subset of the cameras. This is especially evident for the first camera, which has a longer focal length. As previously explained, particles drawn in unreachable regions for a camera are set with high values of $m(unknown)$. Therefore, the rest of the cameras are used to determine the person's location.

From the experiments performed, we have seen that under-segmentation is the main source of error of the proposed algorithm. Under-segmentation occurs when the target's colour is similar to the colour of the foreground. In this case, the target cannot be distinguished from the environment since it produces no foreground points. This situation is shown in the first row of Fig. 10. The figure shows some frames of a sequence where six-people are tracked simultaneously. The figure shows both the camera images and the foreground maps immediately below. The foreground pixels are coloured with the colour of the person they belong to. The under-segmentation problem is particularly evident in the first scene (top row) for the person marked in red. He is wearing black clothes whose colour is very similar to the background in the cameras $cam(3)$ and $cam(4)$. The problem with under-segmentation is that the portion of background points becomes very high. Thus, the algorithm might consider that the region is empty. Of course, if the situation is repeated in most of the cameras, the person cannot be tracked. However, if there are more cameras without under-segmentation, the fusion method is able to determine the person's position.

6. Conclusions

In this paper we have proposed a novel evidential filtering approach that can be considered an extension of the Bayesian particle filters to the Demspter-Shafer theory of evidence. The proposed algorithm is specifically designed for tracking multiple targets by fusing information from multiple unreliable sensors. The management of uncertainty in our approach is particularly attractive due to its simplicity and because it does not require specifying priors nor conditionals that might be difficult to obtain in complex problems. To avoid the curse of dimensionality that arises when joint configurations are em-

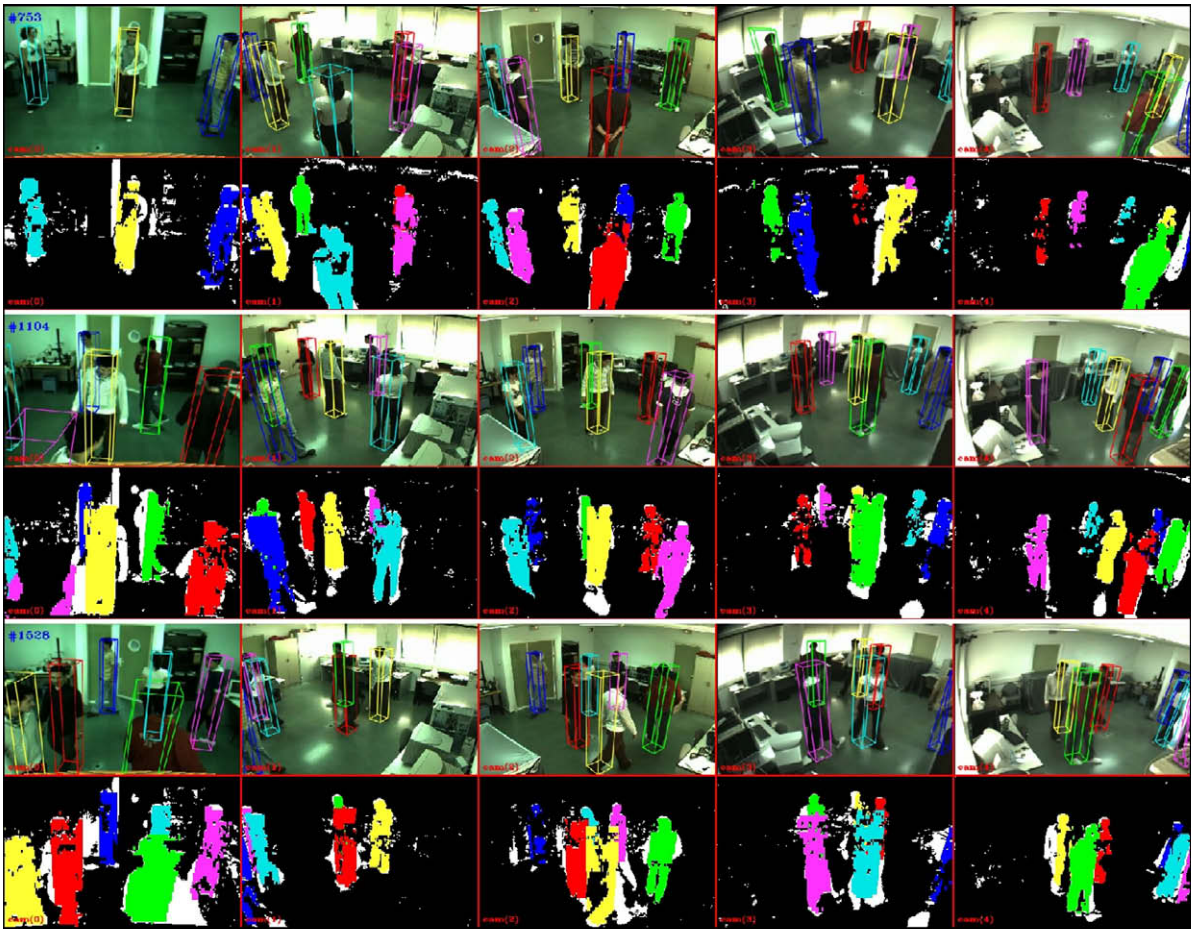

Fig. 10. Camera images and foreground extracted for a six-people sequence. Foreground segmentation errors cause problems in the tracking process (see text for details).

ployed, a separate tracker is used for each target. Interactions between targets are modelled in order to maintain multimodality.

The proposed algorithm is employed to provide a novel solution to the multi-camera people tracking problem. Targets are tracked combining foreground, colour and shape information. The proposed evidence particle filter is especially appropriate for modelling the frequent occlusions that occur in the multi-camera tracking problem. For that purpose, an occupancy map is used to detect target occlusions. The occupancy map is computed independently for each camera using a depth-ordered scheme. Therefore, the evidence can be estimated concurrently in each camera. When a particle is placed at a position hidden to a camera (due to occlusion or because the particle is out of the camera's field of view), the camera indicates that its knowledge about that location is unreliable. Therefore, information from unreliable cameras is weakly considered in the final data fusion step. The test performed shows that the proposed algorithm is able to estimate the locations of the people being tracked using a reduced number of particles and under severe occlusion conditions.

Acknowledgement

This work has been carried out with the support of the Research Projects DPI2006-02608 and TIN2007-66367 funded by the Spanish Ministry of Science and Technology and FEDER.

References

[1] F. Aherne, N. Thacker, P. Rockett, The Bhattacharyya metric as an absolute similarity measure for frequency coded data, Kybernetica 32 (1997) 1–7.

[2] A. Aregui, T. Denoeux, Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities, International Journal of Approximate Reasoning 49 (2008) 575–594.

[3] A.A. Argyrs, M.I. Lourakis, Three-dimensional tracking of multiple skin-colored regions by a moving stereoscopic system, Applied Optics 43 (2004) 366–378.

[4] J. Black, T. Ellis, P. Rosin, Multi view image surveillance and tracking, in: Workshop on Motion and Video Computing, 2002, pp. 169–174.

[5] Isabelle Bloch, Defining belief functions using mathematical morphology – application to image fusion under imprecision, International Journal of Approximate Reasoning 48 (2008) 437–465.

[6] H.-J. Böhme, T. Wilhelm, J. Key, C. Schauer, C. Schröter, H.-M. Gross, T. Hempel, An approach to multi-modal human–machine interaction for intelligent service robots, Robotics and Autonomous Systems 44 (2003) 83–96.

[7] A-O. Boudraa, A. Bentabet, F. Salzenstein, Dempster–Shafer's basic probability assignment based on fuzzy membership functions, Electronic Letters on Computer Vision and Image Analysis 4 (2004) 1–10.

[8] F. Caro, B. Ristic, E. Duflos, P. Vanheeghe, Least committed basic belief density induced by a multivariate Gaussian: formulation with applications, International Journal of Approximate Reasoning 48 (2008) 419–436.

[9] C. Hue, J.L. Cadre, P. Perez, Sequential Monte Carlo methods for multiple target tracking and data fusion, IEEE Transaction on Signal Processing 50 (2002) 309–325.

[10] C. Colombo, AD. Bimbo, A. Valli, Visual capture and understanding of hand pointing actions in a 3-D environment, IEEE Transactions on Systems, Man and Cybernetics – Part B 33 (2003) 677–686.

[11] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2000), vol. 2, 2000, pp. 142–151.

[12] T. Darrell, G. Gordon, M. Harville, J. Woodfill, Integrated person tracking using stereo, color, and pattern detection, International Journal of Computer Vision 37 (2000) 175–185.

[13] S. Démotier, W. Schön, T. Denoeux, Risk assessment based on weak information using belief functions: a case study in water treatment, IEEE Transactions on Systems, Man and Cybernetics C 36 (2006) 382–396.

[14] T. Denoeux, Conjunctive and disjunctive combination of belief functions induced by non-distinct bodies of evidence, Artificial Intelligence 172 (2008) 234–264.

[15] T. Denoeux, M. Masson, Evclus: evidential clustering of proximity data, IEEE Transactions on Systems, Man and Cybernetics B 34 (2004) 95–109.

[16] T. Denoeux, P. Smets, Classification using belief functions: the relationship between the case-based and model-based approaches, IEEE Transactions on Systems, Man and Cybernetics B 36 (2006) 1395–1406.

[17] Thierry Denoeux, Constructing belief functions from sample data using multinomial confidence regions, International Journal of Approximate Reasoning 42 (2006) 228–252.

[18] P.M. Djuric, L. Ting Lu, M.F. Bugallo, Multiple particle filtering, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2007, pp. 1181–1184.

[19] W. Du, J. Piater, Tracking by cluster analysis of feature points and multiple particle filters, in: IEEE Conference on Advanced Video and Signal Based Surveillance, 2005, pp. 165–170.

[20] A.M. Elgammal, D. Harwood, L.S. Davis, Non-parametric model for background subtraction, in: Lecture Notes in Computer Science, vol. 1843, 2000, pp. 751–767.

[21] Francois Fleuret, Jérôme Berclaz, Richard Lengagne, Pascal Fua, Multi-camera people tracking with a probabilistic occupancy map, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 267–282.

[22] D. Focken, R.Stiefelhagen, Towards vision-based 3-D people tracking in a smart room, in: IEEE International Conference on Multimodal Interfaces, 2002, pp. 400–405.

[23] J.D. Foley, A. van Dam, Fundamentals of Interactive Computer Graphics, Addison Wesley, 1982.

[24] N. Gordon, D. Salmand, Bayesian state estimation for tracking and guidance using the Bottstrap filter, Journal of Guidance, Control and Dynamics 18 (1995) 1434–1443.

[25] D. Grest, R. Koch, Realtime multi-camera person tracking for immersive environments, in: IEEE 6th Workshop on Multimedia Signal Processing, 2004, pp. 387–390.

[26] M.S. Grewal, A.P. Andrews, Kalman Filtering, Theory and Practice, Prentice Hall, 1993.

[27] M. Ha-Duong, Hierarchical fusion of expert opinions in the transferable belief model, application to climate sensitivity, International Journal of Approximate Reasoning 49 (2008) 555–574.

[28] Z. Hammal, L. Couvreur, A. Caplier, M. Rombaut, Facial expression classification: an approach based on the fusion of facial deformations using the transferable belief model, International Journal of Approximate Reasoning 46 (2007) 542–567.

[29] M. Harville, Stereo person tracking with adaptive plan-view templates of height and occupancy statistics, Image and Vision Computing 2 (2004) 127–142.

[30] M. Isard, A. Blake, Countour tracking by stochastic propagation of conditional density, in: European Conference on Computer Vision, 1996, pp. 343–356.

[31] M. Isard, A. Blake, Condensation – conditional density propagation for visual tracking, International Journal of Computer Vision 29 (1998) 5–28.

[32] H.H. Kagiwada, R.E. Kalaba, Fuzzy evidential filter for detection and tracking of dim objects, Applied Mathematics and Computation 69 (1995) 75–96.

[33] T. Kailath, The divergence and Bhattacharyya distance measures in signal selection, IEEE Transactions on Communication Technology 15 (1967) 52–60.

[34] J. Kang, I. Cohen, G. Medioni, Tracking objects from multiple and moving cameras. IEE Intelligent Distributed Surveilliance Systems, 2004, pp. 31–35.

[35] S.M. Khan, M. Shah, A multiview approach to tracking people in crowded scenes using a planar homography constraint, in: Lecture Notes in Computer Science, vol. 3954, 2006, 133–146.

[36] Z. Khan, T. Balch, F. Dellaert, MCMC-based particle filtering for tracking a variable number of interacting targets, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 1805–1819.

[37] K. Kim, L.S. Davis, Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering, in: ECCV, vol. 3, 2006, pp. 98–109.

[38] G. Kitagawa, Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, Journal of Computational and Graphical Statistics 5 (1996) 1–25.

[39] J. Klein, C. Lecomte, P. Miche, Preceding car tracking using belief functions and a particle filter, in: 19th International Conference on Pattern Recognition (ICPR 2008), 2008, pp. 864–871.

[40] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, S. Shafer, Multi-camera multi-person tracking for easyliving, in: Third IEEE International Workshop on Visual Surveillance, 2000, pp. 3–10.

[41] R. Lundh, L. Karlsson, A. Saffiotti, Dynamic self-configuration of an ecology of robots, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2007, pp. 3403–3409.

[42] R. Mahler, Can the Bayesian and Dempster–Shafer approaches be reconciled? yes, in: 8th International Conference on Information Fusion, 2005, pp. 864–871.

[43] E. Malis, F. Chaumette, S. Boudet, 21/2d visual servoing, IEEE Transactions on Robotics and Automation 15 (2) (1999) 238–250.

[44] M.-H. Masson, T. Denoeux, Ecm: an evidential version of the fuzzy c-means algorithm, Pattern Recognition 41 (2008) 1384–1397.

[45] B. Menser, M. Brunig, Face detection and tracking for video coding applications, in: Conference Record of the Thirty-Fourth Asilomar Conference on Signals, Systems and Computers, 2000, pp. 49–53.

[46] N. Milisavljevic, I. Bloch, S. Broek, M. Acheroy, Improving mine recognition through processing and Dempster–Shafer fusion of ground-penetrating radar data, Pattern Recognition 36 (2003) 1233–1250.

[47] A. Mittal, L.S. Davis, M2 tracker: a multi-view approach to segmenting and tracking people in a cluttered scene, International Journal of Computer Vision 53 (2001) 189–203.

[48] R. Muñoz-Salinas, A Bayesian plan-view map based approach for multiple-person detection and tracking, Pattern Recognition 41 (2008) 3665–3676.

[49] R. Muñoz-Salinas, E. Aguirre, M. García-Silvente, People detection and tracking using stereo vision and color, Image and Vision Computing (25) (2007) 995–1007.

[50] R. Muñoz-Salinas, E. Aguirre, M. García-Silvente, A. González, People detection and tracking through stereo vision for human–robot interaction, in: Lectures Notes on Artificial Intelligence, vol. 3789, 2005, pp. 337–346.

[51] R. Muñoz-Salinas, M. García-Silvente, R. Medina-Carnicer, Adaptive multi-modal stereo people tracking without background modelling, Journal of Visual Communication and Image Representation 19 (2008) 75–91.

[52] R. Muñoz-Salinas, R. Medina-Carnicer, F.J. Madrid-Cuevas, A. Carmona-Poyato, Depth silhouettes for gesture recognition, Pattern Recognition Letters 29 (2008) 319–329.

[53] K. Okuma, A. Taleghani, D. De Freitas, J.J. Little, D.G. Lowe, A boosted particle filter: multi target detection and tracking, in: Lectures Notes in Computer Science, vol. 3021, 2004, pp. 28–39.

[54] S. Panigrahi, A. Kundu, S. Sural, A.K. Majumdar, Use of Dempster–Shafer theory and Bayesian inferencing for fraud detection in mobile communication networks, in: Lecture Notes in Computer Science, 2007, pp. 446–460.

[55] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufman, 1989.

[56] W. Pieczynski, Multisensor triplet Markov chains and theory of evidence, International Journal of Approximate Reasoning 45 (2007) 1–16.

[57] G. Shafer, A Mathematical Theory of Evidence, Princeton Univ. Press, 1976.

[58] L. Sigal, S. Sclaroff, V. Athitsos, Skin color-based video segmentation under time-varying illumination, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2004) 862–877.

[59] P. Smets, The combination of evidence in the transferable belief model, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1990) 447–458.

[60] P. Smets, The combination of evidence in the transferable belief model, Pattern Analysis and Machine Intelligence 12 (1990) 447–458.

[61] P. Smets, Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem, International Journal of Approximate Reasoning 9 (1993) 1–35.

[62] P. Smets, R. Kennes, The transferable belief model, Artificial Intelligence 66 (1994) 191–243.

[63] P. Smets, B. Ristic, Kalman filter and joint tracking and classification based on belief functions in the tbm framework, Information Fusion 8 (2007) 16–27.

[64] Y. Sun, L. Bentabet, A sequential Monte-Carlo and dsmt based approach for conflict handling in case of multiple targets tracking, in: Lecture Notes in Computer Science, vol. 4633, 2007, pp. 526–537.

[65] J. Vermaak, A. Doucet, P. Perez, Maintaining multimodality through mixture tracking, in: Ninth IEEE International Conference on Computer Vision, 2003, pp. 1110–1116.

[66] J. Vermaak, S.J. Godsill, P. Perez, Monte Carlo filtering for multi-target tracking and data association, IEEE Transactions on Aerospace and Electronic Systems 41 (2005) 309–332.

[67] W.E. Vieux, K. Schwerdt, J.L. Crowley, Face-tracking and coding for video compression, in: International Conference on Computer Vision Systems, 1999, pp. 151–160.

[68] C.R. Wren, A. Azarbayejani, T. Darrell, A.P. Pentland, Pfinder: real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 780–785.

[69] H. Wu, M. Siegel, R. Stiefelhagen, J. Yang, Sensor fusion using Dempster–Shafer theory, in: IEEE Instrumentation and Measurement Technology Conference, 2002, pp. 21–23.

[70] D.B. Yang, H.H. Gonzalez-Banos, L.J. Guibas, Counting people in crowds with a real-time network of simple image sensors, in: IEEE International Conference on Computer Vision, 2003, pp. 122–129.

[71] J. Yen, Gertis: a Dempster–Shafer approach to diagnosing hierarchical hypotheses, Communications of the ACM Archive (1989) 573–585.

[72] Z. Yi, H.Y. Khing, C.C. Seng, Z.X. Wei, Multi-ultrasonic sensor fusion for autonomous mobile robots, in: SPIE Proceedings Series: Architectures, Algorithms and Applications IV, 2000, pp. 314–321.