



People detection and tracking with multiple stereo cameras using particle 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 28 August 2008

Accepted 17 March 2009

Available online 24 March 2009

Keywords:

Plan-view maps

Stereo vision

Person tracking

Particle filtering

Colour processing

Condensation algorithm

Person detection

Gesture recognition

ABSTRACT

In this work, a novel approach for people detection and tracking using multiple stereo cameras is proposed. Our proposal consists in combining information from all the available cameras using three different plan-view maps. Occupancy and height maps register the volume and height of the objects that are visible in the stereo cameras, respectively. We also propose the use of a novel map, named *confidence map*, which registers the confidence of the information projected in each cell. The proposed confidence map is employed to fuse the information captured by each camera so that the most reliable information is kept in each cell. We then propose a particle filter algorithm for tracking people in the fused plan-view maps. The observation model employed considers height, occupancy and confidence information so that information from the most reliable camera is employed at each time instant. The experiments conducted show the validity of our proposal.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

People detection and tracking have been active areas of research in the computer vision community. Although most research has focused on single-camera approaches, this configuration is not the ideal solution for multi-people tracking due to occlusions 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.

Detection and tracking in monocular images are widely explored topics in the related literature. However, using stereo vision for these purposes is an emerging research area. Stereo vision has several advantages over monocular images. First, all of the methods designed for tracking in monocular images can be applied but with much richer per-pixel information (colour or luminance plus depth). Depth information can be employed to track people in a more precise manner and gain a better understanding of their gestures [30]. Moreover, depth is an important piece of information for developing robust background estimation techniques [6,15,28]. Second, disparity information (from which depth is obtained) is relatively invariable to illumination changes. Therefore, systems that employ stereo vision are expected to be more robust in real scenarios where sudden illumination changes might occur.

Several approaches using stereo vision have been proposed to solve people detection and tracking problems. The work of Cipolla and Yamamoto [5] is among the first in this area. They present a stereo-based method for tracking objects using the *visualized locus* method. In [9], Eklundh et al. study the uses of active vision and conclude that the integration of multiple cues is necessary to accomplish real-life tracking tasks. In [7], Darrell et al. present an interactive display system that is capable of detecting and tracking several people by combining a skin detector, a face detector and a disparity map. Perseus [21] constitutes a remarkable example of using stereo vision for interacting with human users. Perseus is able to detect and track a single user and to determine where the subject is pointing. The Perseus architecture is also employed by Franklin et al. [10] to develop a robotic waiter controlled by simple hand gestures. In [12], Haritaoglu et al. present a complete system that integrates grey-level and stereo information for tracking people.

A common aspect of these works is that tracking is performed directly in the camera image. However, many authors have shown that a ground-based representation of stereo data (named *plan-view map*) has several advantages. First, an orthogonal projection allows the enormous amount of information provided by stereo systems to be reduced, while preserving meaningful spatial information that is useful for tracking purposes. Second, people do not tend to be overlapped in floor plane representations as much as they are in original camera images. Therefore, tracking is more reliable in the projection and it is not necessary to use a segmentation method that determines which person each pixel of the camera image belongs to. Third, plan-view maps allow the information provided by several stereo sensors to be easily integrated into a single representation model.

[☆] 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.

^{*} Corresponding author.

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

Most of the work on plan-view maps has focused on single stereo cameras. In Ref. [1], Beymer introduces the concept of *occupancy map* to count people entering a shop using a ceiling stereo camera. The occupancy map is a 2D map of the environment in which each cell registers the amount of 3D points projected in it. The work of Harville [13] constitutes an excellent example of using plan-view maps for people detection and tracking in video surveillance tasks combining occupancy and *height* maps. The latter is employed to register the height of the detected objects. Later, in Ref. [14], they show that height maps can be effectively employed to learn human body poses. In [3], Checka et al. propose the use of stereo plan-view maps with sound plan-view maps. Sound and stereo information are combined to detect a speaker in a conference room. Hayashi et al. [16] present a people detection and tracking system whose main contribution is the use of a *variable voxel* voting scheme in which stereo points are projected in an area proportional to their error. A serious problem of their approach is that the complexity of the map calculation grows exponentially according to the distance of the stereo points. Later, in Ref. [17], the same authors propose the use of a variable window size to represent people according to their distance to the camera. Thus, when people are far from the camera, the window size is enlarged to account for stereo errors. However, their proposal does not allow the integration of multiple stereo cameras.

The authors of this work have proposed several contributions for people detection and tracking using a single stereo camera. While Ref. [29] proposes a tracking approach combining colour and stereo extracted directly from the camera image, the works in [28] propose the use of plan-view maps to represent stereo information more efficiently. Later, in Ref. [27] the *colour map* is proposed in order to manage colour information.

Despite the fact that stereo people detection and tracking problems have been addressed using a single view, few works have addressed the problem using multiple views. In [26], Mittal and Davis present a probabilistic approach for tracking people using multiple monocular cameras. They perform a separate tracking procedure in each of the monocular cameras and then extract stereo information which is projected in a ground map. A great disadvantage of their method is that they repeat the segmentation process several times for each frame until convergence is achieved. In [24], Krumm et al. show a people tracking system for a smart room using a pair of stereo cameras. Rather than employing plan-view maps, people are detected by grouping 3D blobs extracted from the stereo information. A plan-view map projection could have speed up computation by reducing the amount of information. Finally, in Ref. [34], Zhao et al. present an approach for people detection and tracking using multiple short base line stereo cameras. A separate tracker determines people's positions in each camera and then the local positions are fused by a centralized tracker.

1.1. Proposed contribution

Stereo people detection and tracking using multiple stereo views remains an open research field. This work proposes a novel solution based on detecting and tracking people directly in plan-view maps (unlike Refs. [24,26,34]). For each stereo camera, we create three different plan-view maps. The first is an occupancy map indicating in each cell the amount of stereo information projected in the corresponding space region. The second is a height map indicating in each cell the maximum height of the objects in the corresponding space region. Finally, we define a novel map called the *confidence map* to register the reliability of the information projected in the previous maps. The maps of each camera are then fused in order to obtain a global occupancy map, a global height map and a global confidence map. Unlike the above-mentioned works on multiple stereo camera tracking which do not con-

sider stereo errors, we propose a stereo error model in order to combine information from multiple unreliable stereo cameras.

The global maps we create are employed for detecting and tracking people. To do so, we employ a multiple particle filtering approach, i.e., a separate particle filter is used for tracking each person. This work proposes an observation model that combines information from the occupancy, height and confidence maps to perform robust tracking even in the presence of severe occlusions and large stereo errors. As we show in this paper, the use of confidence information significantly reduces tracking errors. To avoid the coalescence problem when people approach each other, the weight of particles is modified by an interaction factor [22]. The algorithm also avoids the coalescence problem in the case of total occlusion by means of an occlusion detection and recovering mechanism. The numerous experiments conducted in several real-life environments validate our proposal.

The remainder of the paper is structured as follows. The basis of stereo computation and how the three plan-view maps are generated and fused are explained in Section 2. In Section 3 our proposal for people detection and tracking is presented. In Section 4, the experiment is presented, while some conclusions are drawn in Section 5.

2. Stereo plan-view maps

In this section the basis for stereo plan-view map calculation is explained. First, a brief introduction to stereo processing is given in Section 2.1 (for an updated review of the state of the art, the interested reader may refer to [2]). In Section 2.2 details are given on how the plan-view maps are created. Finally, in Section 2.3 we explain how the plan-view maps generated for each stereo camera are merged into global plan-view maps.

2.1. Stereo processing

The minimal possible stereo system is composed of a pair of cameras whose optical centres (O_l and O_r) are separated by a distance b . For purposes of simplicity, let us assume that both cameras have identical optical characteristics and that their optical sensors are coplanar (as shown in Fig. 1). A stereo system is able to capture two images (I_l and I_r) at the same instant. In our case, the centre of the reference system employed is the centre of the right image I_r (called *reference image*).

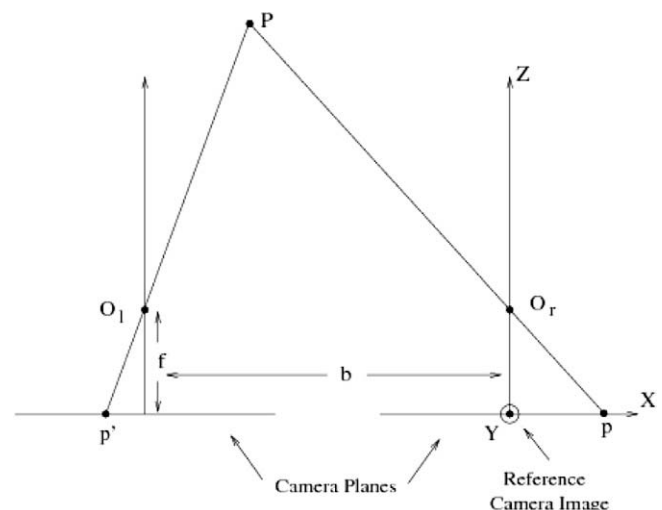


Fig. 1. Minimal stereo system composed by two cameras.

A point $P = (X, Y, Z)$ in the space is projected to two locations ($p = (x, y)$ and $p' = (x', y')$) at the same scan line in both camera images. The displacement of the projection in one image with respect to the other is called *disparity*. By knowing the intrinsic parameters of the stereo system it is possible to reconstruct the three-dimensional structure of a pixel whose disparity is known. Depth is calculated by triangulation as:

$$Z = \frac{fb}{d}, \quad (1)$$

where f is the focal length of the cameras and d is the disparity calculated as $d = |x - x'|$ ($||$ stands for the absolute value). It is also possible to compute the other two components (X and Y) of the three-dimensional position of P as:

$$X = \frac{xZ}{f}; \quad Y = -\frac{yZ}{f}. \quad (2)$$

Disparity is obtained by stereo matching methods. However, it is not always possible to obtain disparity due to the lack of texture. Points for which disparity can be obtained are called *matched points*. By

$$P_{cam} = \left\{ p_{cam}^i = (X_{cam}^i, Y_{cam}^i, Z_{cam}^i, 1)^T \mid i = 1 \dots n \right\},$$

let us denote the set of matched points (expressed in homogeneous coordinates). The three-dimensional positions p_{cam}^i are referred to the stereo camera reference system. However, for our purposes it is preferable to translate the position of the points to a “world” reference system, which is placed at ground level and parallel to it.

The chessboard panel shown in Fig. 2(a) is employed to determine the extrinsic camera parameters using the OpenCv library [18]. The extrinsics of each camera are used to determine the matrix that transform the 3D points from the local reference system of each camera to the global coordinate system centered in the chessboard panel. Therefore, all cameras share a common reference system. Let us denote

$$P_w = \left\{ p_w^i = (X_w^i, Y_w^i, Z_w^i, 1)^T \mid i = 1 \dots n \right\},$$

to the 3D points P_{cam} expressed in relation to the “world” reference system.

Fig. 2 helps to clarify the above explanation. Fig. 2(a) shows the reference camera image captured by a stereo system, while Fig. 2(b) shows the three-dimensional reconstruction of the scene.

The above-mentioned reference systems have been superimposed in Fig. 2(b).

2.2. Occupancy, height and confidence maps

As can be observed from Fig. 2(b), a large amount of 3D points are obtained from a single stereo pair. Nevertheless, since people move along the floor plane, determining their 2D positions in the plane is sufficient for many tracking applications. Thus, it is preferable to translate the 3D information into a simpler representation model that reduces the amount of information while preserving meaningful spatial properties that are useful for tracking. Plan-view maps are employed for this purpose.

A plan-view map divides a region of the floor plane into a set of $n \times m$ cells of fixed size δ . In this work, we set the origin of the map, cell (0,0), as the origin of the world reference system (see Fig. 2(b)). Hence, the cell coordinates (x^i, y^i) in which a three-dimensional point p_w^i is projected are calculated as:

$$x^i = \left\lfloor \frac{X_w^i}{\delta} \right\rfloor; \quad y^i = \left\lfloor \frac{Y_w^i}{\delta} \right\rfloor \quad (3)$$

The selection of δ must be made taking into account several aspects. On the one hand, a high value helps to decrease the computational effort and the memory used. The side effect is that it leads to a loss in precision. On the other hand, a low value increases the precision (up to the limit imposed by the errors of the stereo computation) but also the computational requirements. In this work, we have opted for setting $\delta = 3$ cm which is an adequate balance between both requirements (precision and memory), according to our experimentation.

Every time a stereo pair is captured, the set of points that project in each cell is computed as:

$$P_{(x,y)} = \left\{ i \mid x^i = x \wedge y^i = y \wedge Z_w^i \in [h_{min}, h_{max}] \right\},$$

where $[h_{min}, h_{max}]$ is a height range that has two purposes. First, the upper limit h_{max} avoids using points from the ceiling or from objects hanging from it (e.g., lamps). Second, the lower limit h_{min} excludes low points from the process that are irrelevant for our purposes (e.g., floor points).

In this work, three different plan-view maps are employed. These are called occupancy map $\mathcal{O}$, height map $\mathcal{H}$ and confidence map $\mathcal{C}$. Each cell $\mathcal{O}_{(x,y)}$ of the occupancy map represents the amount of points projected into it and is calculated as indicated in Eq. (4).

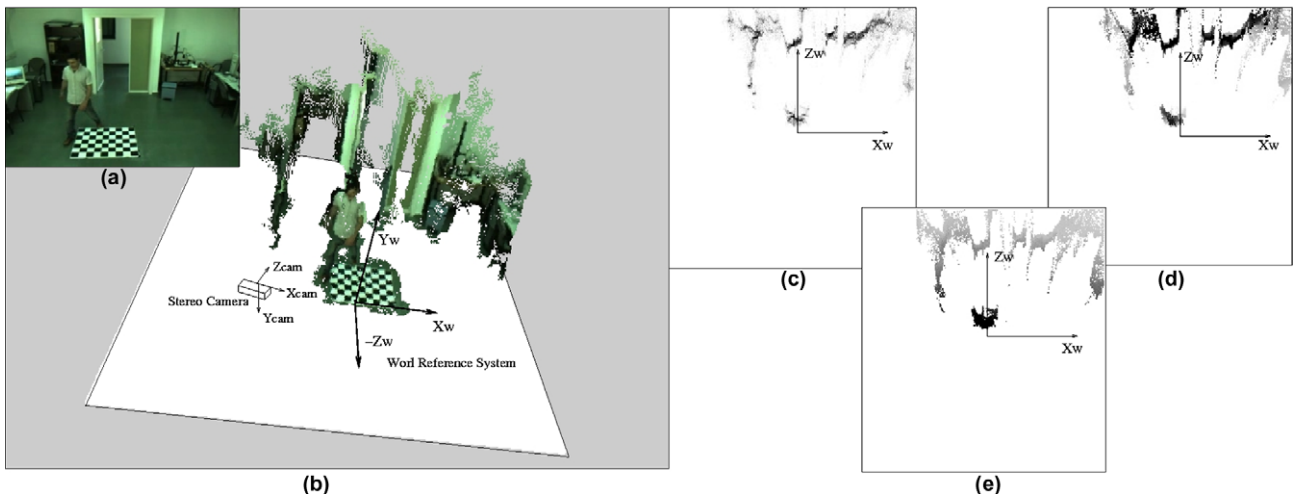


Fig. 2. (a) Reference camera image captured with an stereo system. (b) Three-dimensional reconstruction of the scene showing the reference systems employed in this work. (c) Occupancy map of the scene shown in (a). (d) Height map of (a). (e) Confidence map of (a).

$$\mathcal{O}_{(x,y)} = \sum_{j \in P_{(x,y)}} \frac{(Z_{cam}^j)^2}{f^2} \quad (4)$$

The idea underlying the equation is that each point increments its corresponding cell by a value proportional to the surface occupied in the real scene. Thus, points close to the camera correspond to small surfaces and vice versa. If the same increment is employed for every cell, an object would have a lower sum of the areas the farther it is located from the camera. This scale in the increment value compensates for the difference in the size of the objects observed according to their distance to the camera. The occupancy map of the scene shown in Fig. 2(a) can be seen in Fig. 2(c). In Fig. 2(c), the white cells represent empty regions (i.e., cells with no points projected in them), while the grey cells represent occupied areas. The darker the colour of a cell, the higher the number of points projected in it.

The cells of the height map, $\mathcal{H}_{(x,y)}$, register the height of the highest point projected in them. The height map is calculated as:

$$\mathcal{H}_{(x,y)} = \begin{cases} \max(V_w^j | j \in P_{(x,y)}) & \text{if } P_{(x,y)} \neq \emptyset \\ h_{\min} & \text{if } P_{(x,y)} = \emptyset \end{cases} \quad (5)$$

Fig. 2(d) shows the height map of the scene in Fig. 2(a). In Fig. 2(d), the darker the colour of a cell, the higher its height value.

Finally, this work defines the *confidence map*, $\mathcal{C}$, registering the confidence of the stereo information projected in each cell. As explained in the previous section, the three-dimensional coordinate of a matched point is obtained from the inverse of its disparity Eq. (1). Consequently, the reconstruction error of a point becomes larger as it moves away from the camera. The proposed confidence map registers a certainty value at each cell which indicates the reliability of the information contained in the corresponding cell of the occupancy and height maps. The confidence map is calculated as

$$\mathcal{C}_{(x,y)} = \begin{cases} 0 & \text{if } |P_{(x,y)}| = 0 \\ \frac{1}{|P_{(x,y)}|} \sum_{j \in P_{(x,y)}} C(Z_{cam}^j) & \text{otherwise} \end{cases} \quad (6)$$

where $||$ denotes the cardinal. Each cell $\mathcal{C}_{(x,y)}$ indicates the average confidence of the points projected in it. The confidence of a point is estimated using a confidence function $C(Z_{cam})$ that assigns values in the range (0, 1) according to the distance from the camera. Confidence tends to 1 for points near the camera and to 0 as points are far from the camera. Thus, low values for a cell $\mathcal{C}_{(x,y)}$ indicate whether there is no information in it or that it is unreliable. However, if a cell value is close to 1, we can be confident that the error of the information projected in it is small.

The confidence function is defined in this work as a Gaussian function:

$$C(Z_{cam}) = \exp\left(-\frac{Z_{cam}^2}{2\sigma_c^2}\right), \quad (7)$$

so that $C(Z_{cam})$ decreases as Z_{cam} increases and vice versa. Parameter σ_c is defined as:

$$\sigma_c = \frac{Z_{\max}}{3}. \quad (8)$$

Therefore, when $Z_{cam} = Z_{\max}$, the value of $C(Z_{cam})$ is very small, indicating a lack of confidence. In fact, parameter $Z_{\max}$ can be seen as a upper limit indicating the maximum distance at which stereo errors can be considered too high. From Eq. (1), we can derive:

$$Z_{\max} = \frac{fb}{d_{\min}}. \quad (9)$$

Then, $Z_{\max}$ is expressed in terms of a disparity limit $d_{\min}$. The disparity limit $d_{\min}$ must be selected according to the precision of the

stereo matching algorithm employed. Experimentally, we have determined that $d_{\min} = 3$ provides appropriate results. Expressing $Z_{\max}$ in terms of $d_{\min}$ allows a different $Z_{\max}$ to be calculated for each camera according to its optical properties. In consequence, a camera with a long focal length can be far from an object and obtain a higher confidence than a closer camera with smaller focal length.

The confidence map of the scene shown in Fig. 2(a) can be seen in Fig. 2(e). The white cells represent low confidence regions, while the grey cells represent high confidence regions.

Colour information can be managed by the use of colour plan-view maps as proposed in Ref. [27]. The use of colour helps to reduce the likelihood of confusing people identities in cases of extreme occlusion but only if they wear clothes of different colours. However, we have focused this work only on the minimum configuration that allows us to achieve a reliable people detection and tracking, i.e., using only occupancy and height plan-view maps. Nevertheless, the extension of the methods proposed to account for colour maps is a straightforward procedure. To do so, a separate colour map should be created by each camera and then they should be fused into a global one using the philosophy explained in the next section.

2.3. Plan-view map merging

The process explained above is employed to create an occupancy, height and confidence map for each stereo camera of our system. However, our aim is to create a single map of each type by fusing the information from all of them, i.e., we merge the maps in order to obtain a single occupancy map $\mathbb{O}$, a single height map $\mathbb{H}$ and a single confidence map $\mathbb{C}$. However, a direct combination of the information might lead to problems because of stereo errors. Thus, our combination approach is based on the confidence map. The underlying idea is that the merged maps contain the most reliable information at each cell. By

$$\mathbf{O} = \{\mathcal{O}^v | v = 1 \dots V\}$$

let us denote the set of occupancy maps created for the V stereo systems available. Let us also denote by

$$\mathbf{H} = \{\mathcal{H}^v | v = 1 \dots V\},$$

and

$$\mathbf{C} = \{\mathcal{C}^v | v = 1 \dots V\}$$

to the set of height and confidence maps respectively. By

$$\mathbf{v}_{(x,y)} = \underset{v}{\operatorname{argmax}} \mathcal{C}_{(x,y)}^v, \quad (10)$$

let us define the index of the map with greater confidence in cell (x, y) . Then, the merged maps are calculated setting in each cell the information from the most reliable map, i.e.,

$$\mathbb{O}_{(x,y)} = \mathcal{O}_{(x,y)}^{\mathbf{v}_{(x,y)}}, \quad (11)$$

$$\mathbb{H}_{(x,y)} = \mathcal{H}_{(x,y)}^{\mathbf{v}_{(x,y)}}, \quad (12)$$

and

$$\mathbb{C}_{(x,y)} = \mathcal{C}_{(x,y)}^{\mathbf{v}_{(x,y)}}. \quad (13)$$

Fig. 3 shows the results obtained when maps generated by two stereo cameras are merged. Fig. 3(a and b) are the camera images showing a scene with two people in opposite parts of a corridor. The occupancy maps created for each image are shown in Fig. 3(c) and (d); the corresponding confidence maps are shown in Fig. 3(e) and (f). Fig. 3(e) shows the fused occupancy map, while Fig. 3(h) shows the fused confidence map. Height maps have been omitted for reasons of space.

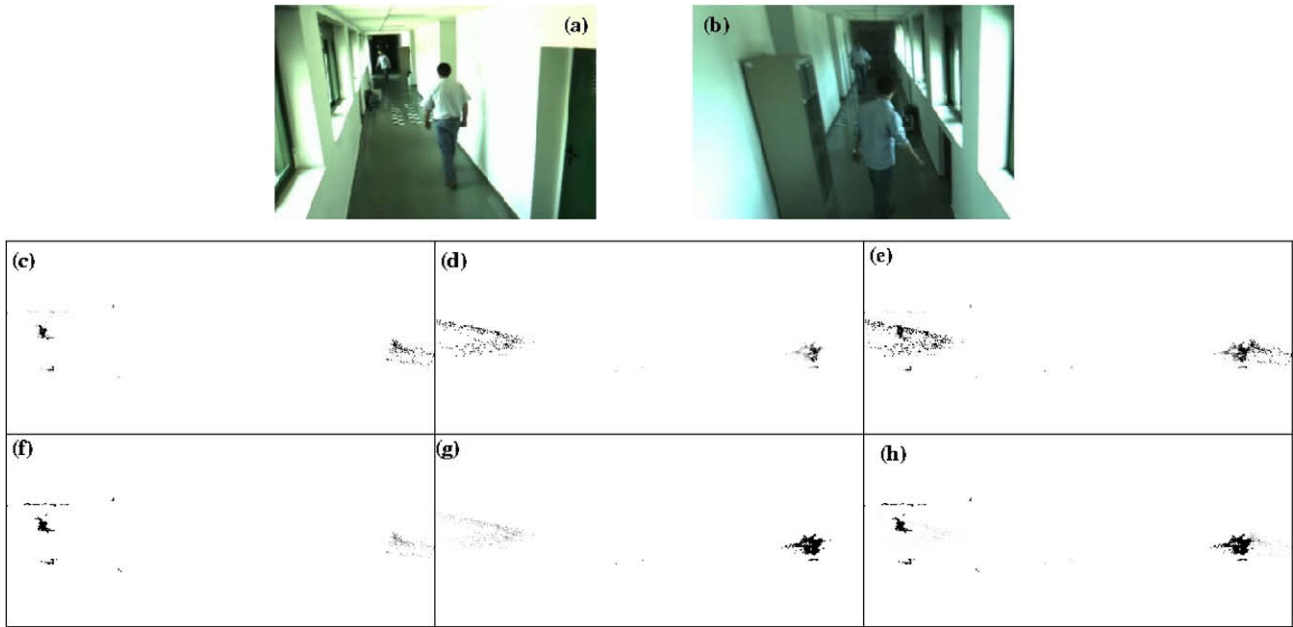


Fig. 3. (a and b) Camera images. (c) Occupancy map of image (a). (d) Occupancy map of image (b). (e) Fused occupancy map of (c) and (d). (f) Confidence map of (a). (g) Confidence map of (b). (h) Fused confidence map of (f) and (g). See text for discussion.

Fig. 3(c–h) shows the impact of stereo errors in plan-view maps at large distances. As can be seen, people are projected as a spread cloud of points when they are far from the cameras. Ideally, information from both cameras should coincide in the fused map. However, projections do not coincide due to stereo errors. Nevertheless, confidence maps register the reliability of the projected information. As explained below, this information will be employed in the tracking process to reduce tracking errors.

3. People detection and tracking

In this section we explain our proposal for people detection and tracking. We have opted for employing a multiple particle filtering approach, i.e., an independent tracker is employed for each target.

Our approach relies on a background model [15] that defines a mixture of Gaussian combining colour and depth information at each pixel. When stereo pairs are captured, foreground pixels are extracted and only these pixels are employed to create the plan-view maps. Afterwards, the known people are tracked and the new people in the scene are looked for. Finally, the background models are updated. However, the pixels belonging to the people being tracked are not considered in the update process. Thus, people do not become part of the background even if they remain static for long periods of time.

The rest of this section is structured into three parts. Section 3.1 explains the basis of the particle filter employed, while Section 3.2 explains the proposed observation model. Section 3.3 briefly explains how people entering and leaving the monitored area are detected.

3.1. Multiple particle filtering

Particle filtering techniques are widely employed for estimating the state of a dynamic system $X(t)$ from sequential observations $Z(t)$ of the system [11,19,20,23]. To do so, a weighted set of particles $\{(s_j(t), \pi_j(t))\}$ is used where $s_j(t)$ represents a possible state of the system, and $\pi_j(t)$ is a non-negative numerical factor called *importance weight* which represents an estimation of the observation density $p(Z(t)|s_j(t))$.

Particle filters are especially interesting because they can deal naturally with systems where both the posterior density and the observation density are non-Gaussian and they are able to manage multiple hypotheses simultaneously. Nevertheless, the straightforward application to multi-target tracking presents several problems. First, the standard particle filter does not define a method for identifying individual modes (or targets). Second, the particles of a standard particle filter quickly collapse to a single mode discarding the rest of them (usually referred to as *coalescence* or the particle “hijacking” problem) [4,32]. Third, a joint particle filter suffers from exponential complexity as the number of targets grows [33]. Recent approaches overcome these problems through the use of multiple particle filters (MPFs) [8,22,31,32]. A MPF consists of employing an independent PF for each target and an *interaction factor* which modifies the weight of particles in order to avoid the coalescence problem.

The typical structure of each independent tracker (based on the Condensation algorithm [20]) is outlined in Fig. 4. At the beginning, the algorithm is provided by an initial sample set $S(0)$ of N particles, where all of them are equally weighted. In this work, the state vector for each target is defined by

$$X(t) = [x(t), y(t)]^T$$

the plan-view map authors' previous works. Would you provide a comparison with these previous works to show the improvement of this work. For example, [27] employed a color map to in location of the person.

At each iteration, the algorithm uses the sample set $S(t-1)$ to create a new sample set $S(t)$. The new samples $s_j(t)$ are obtained by selecting, with replacement, N samples $s'_j(t-1)$ from $S(t-1)$. For that purpose, the cumulative probabilities $c_j(t-1)$ of the particles $s_j(t-1)$ are first calculated. Then, the new particles $s'_j(t-1)$ are selected by finding the particle whose cumulative probability is closer to a selected random number r using binary subdivision. Afterwards, for each selected sample $s'_j(t-1)$, the algorithm computes its next state $s_j(t)$ according to the dynamics of the model plus some random noise.

In this work, we have employed a random walk movement model due to the stochastic nature of people's movements, partic-

Condensation $(S(0) = \{(s_j(0), \pi_j(0)) | j = 1, \dots, N\} | \pi_j(0) = \frac{1}{N})$

From the sample set $S(t-1)$ at time step $t-1$ construct a new sample set $S(t)$ as:

1. **Select** N samples from the set $S(t-1)$ as follows:

- Calculate the normalised cumulative probability of each sample

$$c_j(t-1) = \sum_{i=1}^j \pi_i(t-1)$$

- Generate a random number $r, \mathcal{U}(0, 1)$ and find by binary subdivision the smallest k for which $c_k(t-1) \geq r$.
- Set $s'_j(t-1) = s_k(t-1)$.

2. **Propagate** each selected sample $s'_j(t-1)$ according to a movement model of the target in order to obtain the new sample $s_j(t)$.

$$s_j(t) = A s'_j(t-1) + w(t-1)$$

3. **Measure** and weight each sample in terms of the new observations $Z(t)$:

$$\pi_j(t) = \mathcal{I}_j(t) \cdot p(Z(t) | s_j(t)). \quad (14)$$

Then normalise so that $\sum_{j=1}^N \pi_j(t) = 1$.

4. **Estimate** the mean of the state $S(t)$ and calculate its weight:

$$\mathcal{E}[S(t)] = \sum_{j=1}^N \pi_j(t) s_j(t). \quad (15)$$

$$\pi_{\mathcal{E}[S(t)]} = p(Z(t) | \mathcal{E}[S(t)]). \quad (16)$$

5. If $\pi_{\mathcal{E}[S(t)]} < \pi_{min}$ then, person is **occluded**:

- If the person is occluded for a maximum number of times, then exit.
- Else
 - Increment the number of particles N and the standard deviation of $w(t-1)$

6. Else

- Set N and $w(t-1)$ to their initial values.

Fig. 4. Structure of an independent particle filter tracker.

ularly when people interact with each other. The random noise applied is assumed to follow a Gaussian distribution $N(0, \sigma_m^2)$ whose deviation 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 σ_m^2 is calculated as:

$$\sigma_m = \frac{1}{fps}. \quad (17)$$

The weight $\pi_j(t)$ of each particle is computed based on the new observations obtained $p(Z(t) | s_j(t))$ and the interaction factor $\mathcal{I}_j(t)$

(see Eq. (14)). The interaction factor takes a value in the range $[0, 1]$ and is used to modify the final weight of a particle according to its distance to the rest of the targets. This factor has values near 0 for particles drawn close to the position of targets from other trackers, and values near 1 in the opposite case. Thus, the interaction factor can be defined as a function that is inversely proportional to the distance between a particle and a target. In this work, the interaction factor is defined as a Gaussian function

$$\mathcal{I}_j(t) = 1 - e^{-\frac{dm^2}{2\sigma_{dm}^2}}, \quad (18)$$

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 is assumed to be an upper bound for the width of adult people.

After computing the particle weights, the mean estimated position $\mathcal{E}[S(t)]$ is assumed to be the current person's location (see Eq. (15)). However, a person frequently becomes occluded by another person or by structural elements in the environment. Although the standard particle filter is able to work in the case of partial occlusion, total occlusion represents a major problem. Occlusion occurs when the view of a person is blocked by another person or object in the surroundings. In that case, no 3D points are detected for the blocked person, and consequently, the occupancy, height and confidence maps are empty in the region where the person is. Then, tracking can not be done since there is information to track. Since several cameras are employed in this work, occlusion in one camera can be solved by another camera. Therefore, the problem scarcely occurs. However, in crowded scenarios, a person might be occluded in all cameras. If occlusion is not detected and handled, the tracker of the occluded target will end up tracking one of the remaining visible targets. To avoid that problem, the likelihood of the mean estimated position $\pi_{\mathcal{E}[S(t)]}$ is calculated (Eq. (16)). The value $\pi_{\mathcal{E}[S(t)]}$ reflects the degree to which the estimated position is really occupied by a person in the map. Therefore, an occlusion can be detected when $\pi_{\mathcal{E}[S(t)]}$ is below a given threshold π_{min} .

When occlusion is detected, the tracker incrementally increases the number of particles N as well as σ_m (five percent at each iteration). Thus, while the person remains lost, there are more particles and they are more spread out in order to cover the region where the lost person might reappear. If after several iterations the person is not found, the tracker stops. However, if the person is found, both N and σ_m are set to their default values.

3.2. Observation model

When using stereo plan-view maps, the stereo information of the visible people is projected in an area of the maps that is proportional to their real dimensions. Usually, a person whose arms are partially extended fits in a rectangular region whose size, ζ_r , varies from 0.4 m to 0.6 m [7,28]. Thus, our people detection and tracking approach is based on the analysis of rectangular regions (with a size of $\zeta_r = 0.5$ m in this work). Let

$$\mathcal{R}(x, y) = \left\{ i | \max(|x^i - x|, |y^i - y|) < \frac{\zeta_M}{2} \right\}$$

represent the set of cells from a squared region with a size of ζ_M centred at (x, y) . The parameter ζ_M represents the dimension of a person's projection, ζ_r , expressed in cells. By $\mathcal{R}(s_j(t))$, let us represent the region centred at the particle $s_j(t)$.

The observation model proposed in this work combines occupancy, height and confidence information from the generated plan-view maps. Occupancy information is managed by defining the variable $O_{\mathcal{R}(x, y)}$ representing the occupancy level of the region $\mathcal{R}(x, y)$. It is calculated as:

$$O_{\mathcal{R}(x, y)} = \sum_{i \in \mathcal{R}(x, y)} \mathbb{1}_{(x^i, y^i)}. \quad (19)$$

Assuming that occupancy

$$O_{\mathcal{R}(s_j(t))} = O_j(t) \sim N(\mu_o, \sigma_o^2),$$

follows a Gaussian distribution, we shall denote its probability distribution function (pdf) as:

$$p_o(O_j(t)|s_j(t)) = \frac{1}{\sigma_o \sqrt{2\pi}} \exp\left(-\frac{(O_j(t) - \mu_o)^2}{2\sigma_o^2}\right). \quad (20)$$

The parameters μ_o and σ_o represent the mean and standard deviation of the variable $O_{\mathcal{R}(x,y)}$. These parameters must be estimated by experimentation. In our experiments, the values $\mu_o = 2500$ and $\sigma_o = 1000$ provided good results.

Height information is managed in our model by defining the variable $H_{\mathcal{R}(x,y)}$ representing the maximum height in the region $\mathcal{R}(x,y)$. It is defined as:

$$H_{\mathcal{R}(x,y)} = \max(\mathbb{H}_{(x^i, y^i)}) \forall i \in \mathcal{R}(x,y). \quad (21)$$

We assume that the maximum height of a region $H_{\mathcal{R}(s_j(t))}$ follows a Gaussian distribution

$$H_{\mathcal{R}(s_j(t))} = H_j(t) \sim N(\mu_h, \sigma_h^2).$$

Its pdf is then defined as:

$$p_h(H_j(t)|s_j(t)) = \frac{1}{\sigma_h \sqrt{2\pi}} \exp\left(-\frac{(H_j(t) - \mu_h)^2}{2\sigma_h^2}\right). \quad (22)$$

Parameters μ_h and σ_h represent the mean and standard deviation for the height of a person. In our case, we have employed the values $\mu_h = 1.5$ m and $\sigma_h = 0.8$ m because they provided good results for tracking people in both standing and sitting positions.

Finally, confidence information is managed by defining $C_{\mathcal{R}(x,y)}$ representing the confidence level of the region $\mathcal{R}(x,y)$. It is calculated as

$$C_{\mathcal{R}(x,y)} = \frac{1}{|\mathcal{R}(x,y)|} \sum_{i \in \mathcal{R}(x,y)} C_{(x^i, y^i)}. \quad (23)$$

where $|\cdot|$ denotes the cardinal. As can be observed, $C_{\mathcal{R}(x,y)}$ is in the range $[0, 1]$ where 0 either means that there is no information in the region or that the information is not reliable. However, it tends to 1 as the confidence of the information in $\mathcal{R}(x,y)$ increases. We define the pdf

$$C_{\mathcal{R}(s_j(t))} = C_j(t) \sim iB(f_c, n),$$

to follow what we have denominated the *inverse Butterworth* distribution that is defined as:

$$\varphi(x, f_c, n) = \begin{cases} \frac{1}{N_{cte} \left(1 + \left(\frac{1-x}{1-f_c}\right)^{2n}\right)} & \text{if } x \in [0, 1]; \\ 0 & \text{otherwise.} \end{cases} \quad (24)$$

In Eq. (24), N_{cte} is a normalisation constant calculated as

$$N_{cte} = \int_0^1 \frac{1}{1 + \left(\frac{1-x}{1-f_c}\right)^{2n}} dx,$$

so that

$$\int_{-\infty}^{\infty} \varphi(x, f_c, n) dx = 1.$$

Fig. 5 shows the inverse Butterworth distribution for different values of its parameters f_c and n . The parameter f_c represents a cut-off point at which the confidence of a region is considered too low. Thus, when $C_{\mathcal{R}(s_j(t))}$ is below f_c , $\varphi(x, f_c, n)$ has low values. However, as the confidence value calculated in Eq. (23) is greater than f_c , $\varphi(x, f_c, n)$ tends to its maximum value. The parameter n is the or-

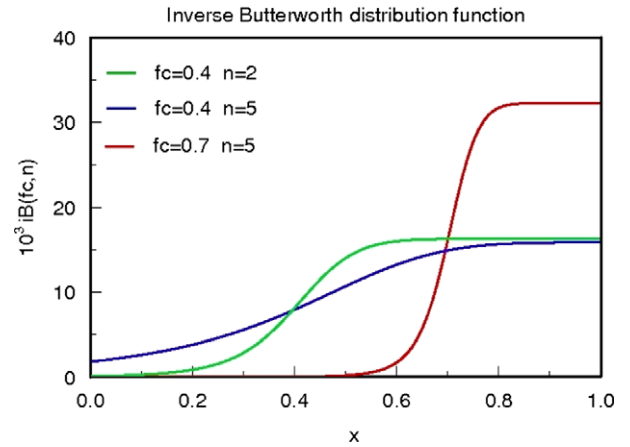


Fig. 5. Inverse Butterworth distribution for different values of its parameters.

der of the filter and determines the smoothness of the curve transition. Then, we define

$$p_c(C_j(t)|s_j(t)) = \varphi(x, f_c, n). \quad (25)$$

The idea is that high confidence values are always accepted. However, as the confidence of a region decreases below f_c , so does the probability assigned to particles. These parameters must be determined. In our case, $f_c = 0.15$ and $n = 2$ provided good results.

Using the three probabilities explained above, the overall likelihood of a state (see Eq. (14)) is calculated as:

$$p(Z(t)|s_j(t)) = p_o(O_j(t)|s_j(t))p_h(H_j(t)|s_j(t))p_c(C_j(t)|s_j(t)). \quad (26)$$

In the final implementation, computing time can be reduced by taking into account that in order to achieve a high likelihood, the three likelihoods must be high. Hence, if the calculation of $p_c(C_j(t)|s_j(t))$ reveals a low value, then the calculation of $p_h(H_j(t)|s_j(t))$ and $p_o(O_j(t)|s_j(t))$ is not required since the final likelihood will remain low.

Fig. 6(a) shows the results of applying Eq. (26) to every cell in the corridor example of Fig. 3(a and b). As previously explained, the figure shows the plan-view maps obtained in a test where two cameras placed at opposite ends of a corridor record two people. Notice that the proposed observation model generates two clear modes (one for each person). However, we must remember that stereo errors obtained in the plan-view map construction are high. To better show the impact of the confidence map, Fig. 6(b) shows the result of applying Eq. (26) without confidence information, i.e., setting $p_c(C_j(t)|s_j(t)) = 1$ for every cell. In that case, four different modes appear. When confidence information is not employed, stereo errors cause "phantoms" that might confuse the tracking algorithm. However, the use of confidence information allows the problem to be solved.

3.3. Detection of people entering and leaving the scene

Before using the above method for tracking, people detection is required. Once a person is detected, an independent tracker is em-

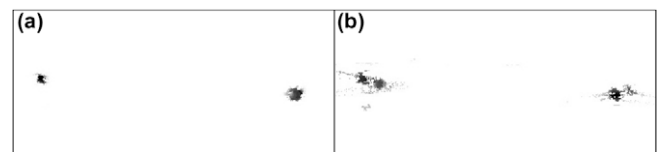


Fig. 6. (a) Likelihood map of the scene shown in Fig. 3 using confidence information. (b) Likelihood map of the scene shown in Fig. 3 without using confidence information.

ployed for that person. The people detection method employed in this work is similar to the one employed in [27]. However, in this work confidence information is also considered.

Detection is performed after the tracking step to avoid already detected people from being redetected. In controlled environments, people enter and leave the scene in known regions (e.g., doors) that can be located in the map. People detection is performed by applying Eq. (26) to each cell of these regions. Detection consists in iteratively seeking peaks above a threshold θ_e in these processed regions. Once a peak is detected, we consider that a person might have entered the scene and the cells in the region $\mathcal{R}(x, y)$ around the peak are excluded in the following iteration. The search process stops when no peaks above θ_e are found.

The peaks detected are not directly considered for tracking. Instead, a separate list of candidate detections is maintained and updated at each time step. When the same candidate is detected in a number of consecutive frames, it is considered that a new person has entered in the scene and a new tracker is added to the system. This strategy has proved to reduce the number of false positives.

On the other hand, if a tracker detects that a person has come back into one of these regions and then becomes occluded for a consecutive number of frames, we consider that the person has left the scene and that person's tracker is eliminated.

4. Experiments and discussions

This section explains the experiments that were conducted to test the proposed algorithm. Several video sequences were recorded in three different scenarios. The first scenario is the PEIS room [25]; a robotised apartment employed for the development and research of mobile and embedded robotic systems aimed to help elderly people. Two stereo systems were employed for the scenario. The first one was placed on the ceiling at a height of ≈ 4.5 m and pointing downwards. The second one was placed at a height of ≈ 3 m in a slanting position. The area covered by the cameras was 3×4 m. The two 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 three stereo cameras were placed at a height of 3 m to cover an area of approximately 4×4 m. The cameras were synchronised via software and set to record at 15 fps with a resolution of 320×240 pixels. Finally, the third scenario is the corridor shown in Fig. 3. Two stereo systems were placed at a height of 2 m at opposite ends of a corridor measuring 15 m in length. The stereo correspondence algorithm employed in all cases is an improved version of SAD that performs subpixel interpolation. The number of people in the recorded sequences varied from 2 to 4. The people were instructed to move freely about the environment. Therefore, interactions and occlusions are frequent in the recorded videos. A total of 10 different people participated in the experiments.

This section is divided into two parts. In the first one (Section 4.1), we evaluate the performance of our proposal, while in the second one (Section 4.2) we analyse the influence of using the proposed confidence map.

4.1. Performance evaluation

In this section we show the experiments conducted in order to evaluate two crucial aspects of the proposed method. First, the error in determining the positions of the people being tracked (tracking error), and second, the error evolution as the number of particles grows. Regarding the first objective, people's positions in four different sequences have been determined in each frame in order to obtain quantitative measures of the tracking error. The positions have been manually extracted to indicate the locations of the people's heads in the plan-view maps. A total of

4182 positions have been manually extracted from the 1103 frames employed. The tracking error is calculated for each person as the distance from his/her position (manually determined) to the position estimated by the proposed tracking method.

Regarding the second objective, the performance of particle filter based methods increases as the number of particles grows. Nevertheless, the higher the number of particles employed, the higher the computational effort required for the algorithm. Moreover, there is usually a limit to the number of particles over which no significant improvement is obtained. Our aim is to examine the error evolution in order to determine the minimum number of particles required for the proposed method to perform adequately. Thus, the tracking error of our algorithm has been analysed in the sequences recorded for an increasing number of particles. Furthermore, due to the stochastic nature of the algorithm, each test has been repeated 10 times with different seeds for the random number generator.

The results of our experimentation are shown in Fig. 7. The horizontal axis in the graph in Fig. 7 represents the number of particles employed for each tracker. The vertical axis represents the RMSE (root mean-square error) in determining the people's position. The RMSE has been obtained from the errors of the 10 runs for all the people in the sequence. As expected, the tracking error is reduced as the number of particles increases. However, it can be observed that the reduction is greater from 15 to 40 particles. In light of these results, we consider that the proposed method performs well for 40 particles, obtaining an error of 0.15 m. In fact, no substantial error reduction is obtained using more particles.

The tests were performed in an AMD Turion 3200 portable computer with 1 GB of RAM running GNU+Linux. In our tests, background removal consumes 30 ms, while plan-view map creation and fusion consumes 9 ms. In fact, background removal is the most expensive part of our algorithm; a shortcoming that future works will aim to improve. Finally, the time each tracker requires for estimating the person's position using 40 particles is 2.5 ms.

In what follows, we show images from some of the sequences employed to test our proposal. Fig. 8 shows images from one of the tests performed in the first scenario: the PEIS room. As previously indicated, two stereo cameras were placed in the monitored area: one on the ceiling and another one in a slanting position. The second one is comprised of two low-cost webcams employing the rectification procedure provided by the OpenCv library [18]. The ceiling camera is a stereo system manufactured by Videre Design. The camera images are shown in the two upper rows of Fig. 8. The likelihood obtained at each pixel is shown in the bottom row. Additionally, the people's locations estimated by our trackers

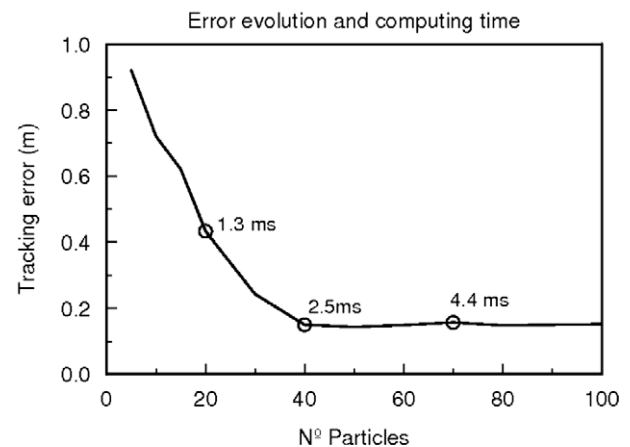


Fig. 7. Tracking error as the number of particles grow.

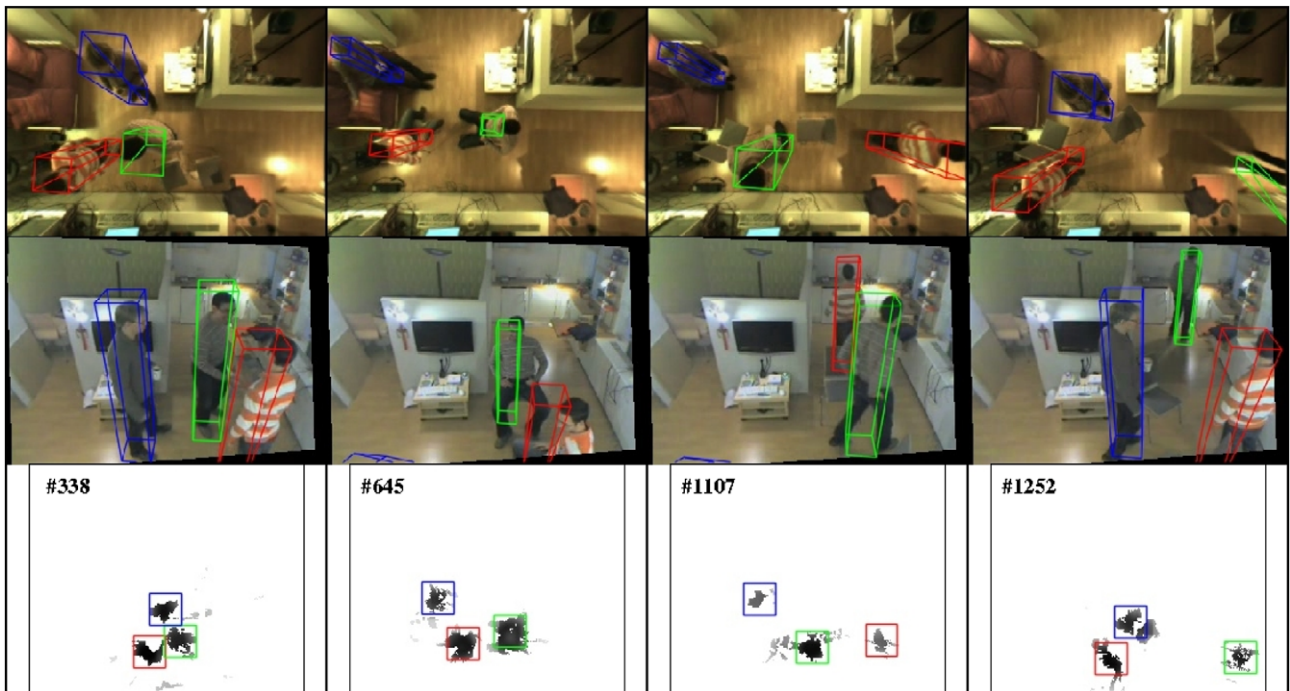


Fig. 8. Three people in the PEIS room. People interact, sit and move around the room.



Fig. 9. Four people moving randomly and sitting causing frequent occlusions.

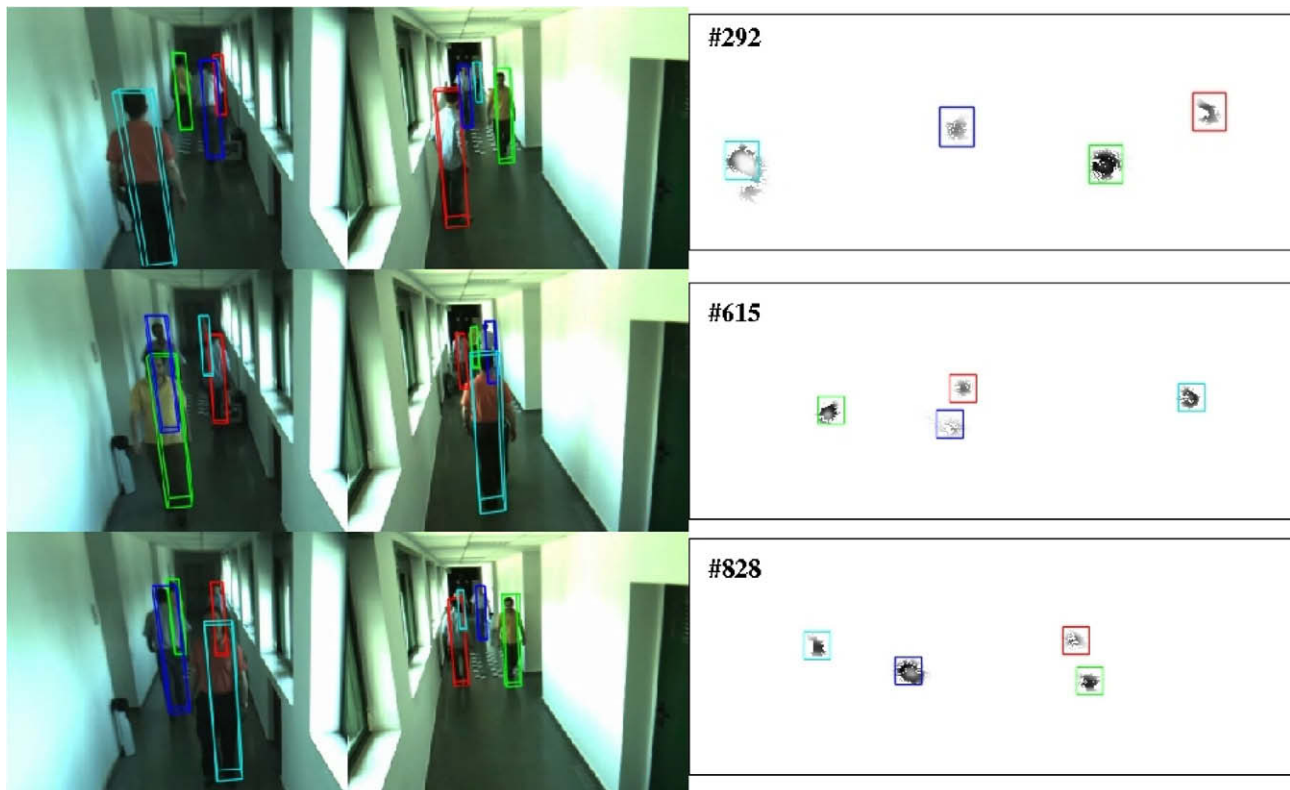


Fig. 10. Four people moving in a corridor causing severe occlusions.

have been superimposed in the map using a different colour for each person. The scene in Fig. 8 shows three people that enter the room and introduce themselves to each other. They then take a pair of chairs and sit down to talk. Finally, they move around inspecting the room. In this scenario, the distance of the people to both cameras is small (≈ 3 m). Therefore, no significant stereo errors are obtained. It is interesting to note that a ceiling camera reduced occlusions significantly in this particular configuration. Nevertheless, such a configuration requires very wide angle lenses and a high ceiling (which is not always available). On the other hand, slanting cameras monitor wider areas than cameras placed on ceilings but occlusions are frequent.

Fig. 9 shows one of the sequences employed to test our system in the second scenario. In this scenario, three *Videre* stereo cameras were placed in slanting positions at a height of 3 m. The sequence shows four people entering the room, moving about randomly and sometimes sitting on chairs. In this case, the cameras are also near the people resulting in small stereo errors. However, this is a challenging tracking scene since occlusions are very frequent. People are sometimes seen by three, two or only one camera. The tests performed show that the proposed approach permits people to be tracked in a reliable manner despite frequent occlusions. Note that if a person is occluded in one camera, another camera might observe him/her and thus the person will be projected in the map. The proposed fusion approach set information from the most reliable camera in each cell of the global map. However, if only one camera sees the target, the information from that camera is employed for tracking. Thus, occlusions in one camera are effectively solved by the rest of the cameras.

Finally, Fig. 10 shows one of the tests performed in the corridor scenario. In the figure, we can see four people moving from one end of the corridor to the other. This is an especially difficult tracking situation for two reasons. First, because the distance between the cameras causes high stereo errors as previously shown. But

there are also very marked and frequent occlusions. In this scenario, people are severely occluded in both cameras simultaneously. However, tracking can be recovered thanks to the occlusion handling procedure employed.

4.2. Influence of the confidence map

This section aims to analyse the impact of using confidence information in the tracking process. To test the influence of the proposed confidence map, we have performed a test in the corridor. We have chosen this scenario because it is large enough to see the impact of stereo errors in the tracking task.

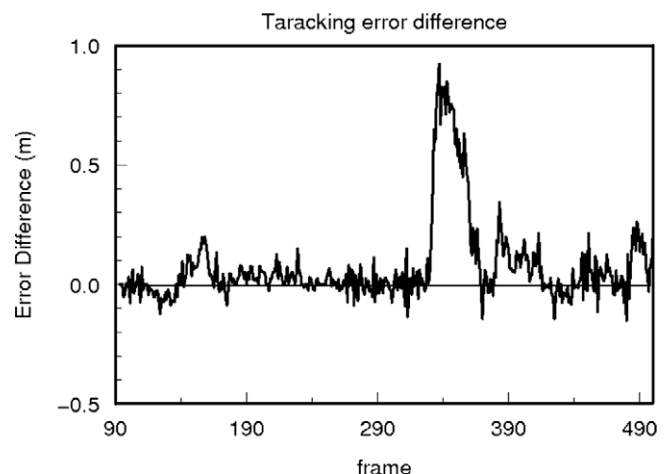


Fig. 11. Tracking error difference for the sequence shown in Fig. 12. Positive values indicate higher errors for the tracker that do not employ confidence information.

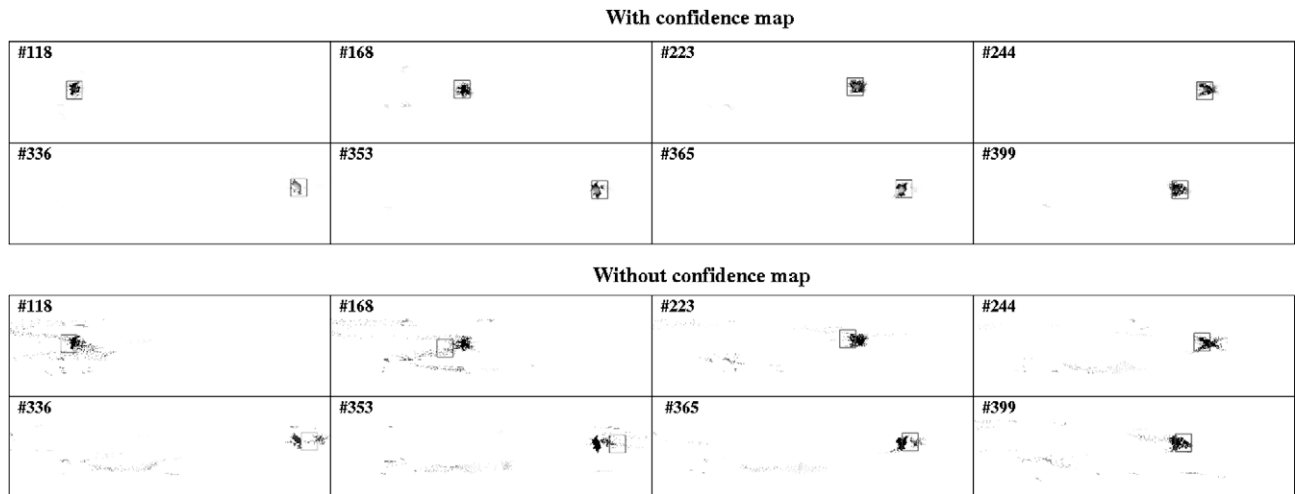


Fig. 12. A person moves from one extreme of a corridor to the other. Top rows shows the results when confidence map are employed. Bottom rows correspond to the results when no confidence maps are employed. Without confidence information, stereo errors causes tracking errors. Confidence information alleviates the problem.

The test consists of a single person moving from one end of the corridor to the other. The sequence has been recorded at 15 fps and processed with and without the confidence map using a total of 40 particles. To avoid using confidence information, we have assumed that $p_c(C_j(t)|s_j(t)) = 1$ (see Eq. (26)) for all the particles. The sequence was also manually processed to determine the location of the person in each frame so that tracking errors can be determined.

Fig. 12 shows the tracking results. The top rows show the likelihood Eq. (26) of all the cells map when confidence information is employed. The location estimated by the tracker has been superimposed on the maps as a red box. The bottom rows of Fig. 12 shows the likelihood of the cells map/cell maps when confidence information is not employed and the tracking result of the algorithms is also superimposed. The tracking errors of the sequence are shown in Fig. 11 as the difference between the error of the tracker that does not employ confidence information and the one that uses it. Therefore, positive values indicate that the version without confidence has a higher error than the version using confidence information.

Notice in Fig. 12 that when confidence information is not employed, the likelihood obtained is “noisy”. As previously explained, this is because stereo errors create false modes (or phantoms). The results clearly show that the use of confidence information outperforms the tracking results when no confidence information is employed. This case is not so evident when the person is in the middle of the corridor. In that case, the confidence of both stereo systems is similar. However, the tracking error obtained for the version without confidence information is very high when the person is near the ends of the corridor. In particular, in frame #360, the tracking error difference is nearly 1 m (see Fig. 11).

5. Conclusions

This paper has proposed a novel approach for people detection and tracking using multiple stereo cameras. Our proposal consists in fusing information from available cameras using plan-view maps. In this work, we have employed occupancy and height maps in combination with a novel map called the confidence map. The cells of the confidence map register the reliability of the information projected in the corresponding cells of the occupancy and height maps. Since stereo errors increase as the distance to the objects is greater, distant objects produce a low confidence in the map.

A separate occupancy, height and confidence map is created for each camera. Information from the maps is then merged into a

global occupancy, height and confidence map. Fusion is carried out by keeping the most reliable information in each cell, i.e., the information from the map with higher confidence. However, if only one camera sees the target, the information from that camera is the one set in the global maps. Thus, occlusions in one camera can be effectively solved by the rest of cameras.

The fused maps are employed for detecting and tracking people. To do so, we use a multiple particle filtering strategy (i.e., a separate particle filter is employed for each person). In order to maintain multimodality, particles employ an interaction factor that avoids the coalescence problem. The proposed observation model combines occupancy, height and confidence information to achieve robust tracking.

The proposed approach has been tested in numerous experiments in three different scenarios. The results show that tracking can be done reliably using only 40 particles and with an error of approximately 0.15 m. Moreover, the tests show that the use of the proposed confidence map allows the tracking error to be reduced by employing the information from the most reliable camera.

References

- [1] D. Beymer, Person counting using stereo, in: Intl. Workshop on Human Motion, 2000, pp. 127–133.
- [2] M.Z. Brown, D. Burschka, G.D. Hager, Advances in computational stereo, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (2003) 993–1008.
- [3] N. Checka, K. Wilson, V. Rangarajan, T. Darrell, A probabilistic framework for multi-modal multi-person tracking, in: Conference on Computer Vision and Pattern Recognition Workshop, 2003, pp. 100–107.
- [4] 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.
- [5] R. Cipolla, M. Yamamoto, Stereoscopic tracking of bodies in motion, Image and Vision Computing 8 (1990) 85–90.
- [6] T. Darrell, D. Demirdjian, N. Checka, P. Felzenszwalb, Plan-view trajectory estimation with dense stereo background models, in: Eighth IEEE International Conference on Computer Vision (ICCV 2001), vol. 2, 2001, pp. 628–635.
- [7] T. Darrell, G. Gordon, M. Harville, J. Woodfill, Integrated person tracking using stereo, color, and pattern detection, Int. J. Comput. Vis. 37 (2000) 175–185.
- [8] 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.
- [9] J.-O. Eklundh, P. Nordlund, T. Uhlin, Issues in Active Vision: Attention and Cue Integration/selection, in: British Machine Vision Conference, 1996, pp. 1–12.
- [10] D. Franklin, R.E. Kahn, M.J. Swain, R.J. Firby, Happy patrons make better tippers: Creating a robot waiter using Perseus and the animate agent architecture, In Int. Conf. Automat. Face Gesture Recogn. 1996, pp. 14–16.
- [11] N. Gordon, D. Salmand, Bayesian state estimation for tracking and guidance using the bootstrap filter, Journal of Guidance, Control Dynamics 18 (1995) 1434–1443.

- [12] I. Haritaoglu, D. Harwood, L.S. Davis, W4: real-time surveillance of people and their activities, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (2000) 809–830.
- [13] M. Harville, Stereo person tracking with adaptive plan-view templates of height and occupancy statistics, *Image and Vision Computing* 2 (2004) 127–142.
- [14] M. Harville, Li. Dalong, Fast, integrated person tracking and activity recognition with plan-view templates from a single stereo camera, in: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004, pp. 398–405.
- [15] M. Harville, G. Gordon, J. Woodfill, Foreground segmentation using adaptive mixture models in color and depth, in: *IEEE Workshop on Detection and Recognition of Events in Video*, 2001, pp. 3–11.
- [16] K. Hayashi, M. Hashimoto, K. Sumi, K. Sasakawa, Multiple-person tracker with a fixed slanting stereo camera. in: *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 681–686.
- [17] K. Hayashi, T. Hirai, K. Sumi, K. Sasakawa, Multiple-person tracking using a plan-view map with error estimation, *Lecturer Notes on Computer Sciences* 3851 (2006) 359–368.
- [18] Intel. OpenCV: Open source Computer Vision library. Available from: <<http://www.intel.com/research/mrl/opencv/>>.
- [19] M. Isard, A. Blake. Countour tracking by stochastic propagation of conditional density, in *European Conference on Computer Vision*, 1996, pp. 343–356.
- [20] M. Isard, A. Blake, CONDENSATION—conditional density propagation for visual tracking, *Int. J. Comput. Vis.* 29 (1998) 5–28.
- [21] R.E. Kahn, M.J. Swain, P.N. Prokopowicz, R.J. Firby, Gesture recognition using the Perseus architecture, in: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'96)*, 1996, pp. 734–741.
- [22] 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.
- [23] G. Kitagawa, Monte Carlo filter and smoother for non-gaussian nonlinear state space models, *Journal of Computational and Graphical Statistics* 5 (1996) 1–25.
- [24] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, S. Shafer, Multi-camera multi-person tracking for easy living, in: *Third IEEE International Workshop on Visual Surveillance*, 2000, pp. 3–10.
- [25] R. Lundh, L. Karlsson, A. Saffiotti, Dynamic self-configuration of an ecology of robots, in: *Proc of the IEEE/RSJ Int Conf on Intelligent Robots and Systems (IROS)*, 2007, pp. 3403–3409.
- [26] 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.
- [27] R. Muñoz Salinas, A bayesian plan-view map based approach for multiple-person detection and tracking, *Pattern Recognition* 41 (2008) 3665–3676.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] K. Okuma, A. Taleghani, D. De Freitas, J.J. Little, D.G. Lowe, A Boosted Particle Filter: Multi Target Detection and Tracking, *Lectures Notes in Computer Science* 3021 (2004) 28–39.
- [32] J. Vermaak, A. Doucet, P. Perez, Maintaining multimodality through mixture tracking, in: *Ninth IEEE International Conference on Computer Vision*, 2003, pp. 1110–1116.
- [33] 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.
- [34] T. Zhao, M. Aggarwal, R. Kumar, H. Sawhney, Real-time wide area multi-camera stereo tracking, in: *Computer Vision and Pattern Recognition (CVPR 2005)* (2005) 976–983.