Available online at www.sciencedirect.com

ScienceDirect

Image and Vision Computing 25 (2007) 995–1007

www.elsevier.com/locate/imavis

People detection and tracking using stereo vision and color ☆

Rafael Muñoz-Salinas [a,*], Eugenio Aguirre [b], Miguel García-Silvente [b]

[a] *Department of Computing and Numerical Analysis, University of Córdoba, 14071 Córdoba, Spain*
[b] *Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain*

Received 22 August 2005; received in revised form 10 June 2006; accepted 12 July 2006

Abstract

People detection and tracking are important capabilities for applications that desire to achieve a natural human–machine interaction. Although the topic has been extensively explored using a single camera, the availability and low price of new commercial stereo cameras makes them an attractive sensor to develop more sophisticated applications that take advantage of depth information. This work presents a system able to visually detect and track multiple people using a stereo camera placed at an under-head position. This camera position is especially appropriated for human–machine applications that require interacting with people or to analyze human facial gestures. The system models the background as height map that is employed to easily extract foreground objects among which people are found using a face detector. Once a person has been spotted, the system is capable of tracking him while is still looking for more people. Our system tracks people combining color and position information (using the Kalman filter). Tracking based exclusively on position information is unreliable when people establish close interactions. Thus, we also include color information about the people clothes in order to increase the tracking robustness. The system has been extensively tested and the results show that the use of color greatly reduces the errors of the tracking system. Besides, the people detection technique employed, based on combining plan-view map information and a face detector, has proved in our experimentation to avoid false detections in the tests performed. Finally, the low computing time required for the detection and tracking process makes it suitable to be employed in real time applications.

© 2006 Elsevier B.V. All rights reserved.

Keywords: People detection; People tracking; Stereo vision; Human–machine interaction; Color processing

1. Introduction

The topic *human–machine interaction* (HMI) has drawn a lot of attention in the last decade. The objective is to create intelligent systems capable of extracting information about the context or about the actions to perform through a natural interaction with the user, for example, through their gestures or voice. One fundamental aspect in that sense is people detection and tracking, existing an extensive literature about the topic [14,25,36,40].

Although people detection and tracking with a single camera is a well explored topic, the use of stereo technology for this purpose concentrates now an important interest. The availability of commercial hardware to solve the low-level problems of stereo processing, as well as the lower prices for these types of devices, turn them into an appealing sensor to develop intelligent systems. Stereo vision provides a type of information that brings several advantages when developing human–machine applications. On one hand, the disparities information is more invariable to illumination changes than the information provided by a single camera. It is a very advantageous factor for the development of background estimation techniques [6,9,19]. Furthermore, the possibility to know the distance from the camera to the person is of great assistance for tracking as well as for a better analysis of his gestures.

This paper presents a system able to detect and track multiple people with a stereo camera placed at an

☆ This work has been partially supported by the MEC project TIC2003-04900.

* Corresponding author. Tel.: +34 658101520.
E-mail addresses: salinas@decsai.ugr.es, rmsalinas@uco.es (R. Muñoz-Salinas), eaguirre@decsai.ugr.es (E. Aguirre), M.Garcia-Silvente@decsai.ugr.es (M. García-Silvente).

0262-8856/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2006.07.012

under-head position. In a first phase, the system creates a background model of the environment using a height map. It can be constructed even in the presence of moving objects in the scene (like people passing by) and is an appropriate tool for using on mobile devices (such as mobile robots). Using this structural map, foreground objects are easily detected and those that are potential candidates to people are analyzed to detect whether they are people. Our system detects people by combining plan-view map information with a face detector on the raw images. Once an object is identified as a person, the system keeps track of him as well as of the rest of detected people while still looking for more candidates. The Kalman filter has been employed to estimate the position of each person in the next image captured. Nonetheless, when people become close to each other, the estimation of the position becomes unreliable. Thus, information about the color of the person is combined with the predicted position to achieve a more robust tracking.

The remainder of this paper is structured as follows. Section 2 explains some of the more relevant related works and exposes the main differences with our approach. Section 3 explains the basis of the background modelling and foreground extraction techniques. In Section 4 it is shown how people detection and tracking are performed. Section 5 presents the experimentation carried out and Section 6 draws the conclusions and possible future work.

2. Related works

Among the first projects related to people detection and tracking using stereo vision we find the one by Darrel et al. [7]. They present an interactive display system capable of detecting and tracking several people. Person detection is based on the integration of the information provided by three modules: a skin detector, a face detector and the disparity map provided by a stereo camera. First, independent objects (*blobs*) detected in the disparity map are treated as candidate to people. Then, the color of the image is analyzed to identify those areas that could be related to skin. Finally, a face detector is applied on selected regions of the camera image. These three items are merged in order to detect and track several people. However, a main drawback to their approach can be pointed out as the system relies on a predefined color model to detect skin, a degradation on the tracking performance can be expected when the illumination conditions differ significantly from the training ones [27]. A dynamic skin color model [35] could have solved this problem.

Grest and Koch [15] developed a system able to detect and track a single-person using stereo vision. It allows the user to navigate in a virtual environment by walking through a room using virtual reality glasses. The user is detected using the face detector proposed by Viola and Jones [39]. Once the user is located, both a color histogram of the face region and a color histogram of the chest region are created and employed by a particle filter to estimate his position in the next image. Then, stereo information assists in determining the real position of the person into the room. The real 3D position is employed to calculate the corresponding position in the virtual environment. In their work, the stereo processing is performed using the information gathered by different cameras located at different positions of the room. The main limitation of this contribution is that their system requires the face of the user to be visible in the image to perform the tracking.

A very interesting method to locate and track multiple people in stereo images (using plan-view maps) is presented by Harville [18]. Before the detection process takes place, a model of the environment is created through a sophisticated image analysis method [19]. Once the background image is created, objects that do not belong to it are easily isolated. Then, both an occupancy map and a height map are created. The information from both maps is merged to detect people through the use of simple heuristics. Person tracking is performed using the Kalman filter combined with deformable templates. The stereoscopic system used in his work is located three meters above the ground in a fixed slanting position. The main draw back of his approach is that the simple detection heuristics employed may lead the system to incorrectly detect as people new objects in the scene whose dimensions are similar to human beings (as indicated by the author). That is the case of coat placed on a hanger or a big box into the room.

Hayashi et al. [20] present a people detection and tracking system especially designed for video surveillance. The environment is modelled using an occupancy map an the stereo points are projected as “*variable voxels*” to deal with stereo errors. People detection is performed in the occupancy map based on a simple heuristic: a person is a peak in the map whose height is into a normal range. As in the previous case, many false positives can be expected because of the simplicity of the detection scheme.

Tanawongsuwan presents in [37] the initial steps for designing a robotic agent able to follow a person and recognizing his gestures. His system employs a basic technique to find the arms and head of a person combining the information provided by a skin color filter, a movement detector and depth. Once a person has been located, Hidden Markov Models are used to recognize his gestures among a set of previously learned ones. Nevertheless, the detection problem is not dealt in depth assuming that a person if found when three skin colored blobs (corresponding to head and hands) are located in the camera image.

2.1. Our approach

This paper presents a system able to detect and track multiple people. It is specially designed for situations in which the camera must be placed at an under-head position. The position of the camera is a problem-dependent issue related with the purpose of the system. Several authors have mounted their cameras in the ceiling to perform people detection in Ambient Intelligence domains

[17,36]. Others have used slating cameras in overhead positions [3,18,20] to achieve the same functionality. Nevertheless, in most of the works that seek interacting with people, the position of the camera is usually lower than them [7,30,37]. This approach is mainly supported by two facts. First, this camera configuration allows to see the faces and arms of the people and thus, be able to analyze their facial expressions and gestures. Second, studies in the human–robot field reveal that people tend to feel threaten by big robots [12]. However, the main drawback of low camera positions is that occlusions between people are more frequent that in elevated positions.

Most of people detection and tracking systems have an initial phase where a background model is created. It is employed to easily detect the moving objects in the environment (foreground). Techniques usually employed for that purpose consist in creating a background image (pixel by pixel) using several images of the scene, if possible, without motion elements in them [14,18,25,36]. The simplest approach consists in using the average of the sequence in each pixel as the background value. Other authors have used the median value and even Kalman filters to perform the update process. In this paper, a height map of the environment (built using stereo information) is employed as background model. Height maps bring several advantages over traditional techniques to create background models. First, the background model created is more invariant to sudden illumination changes because of stereo information is used instead of intensity values. Second, the background model created contains structural information of the environment so they seem to be specially appropriated for mobile devices like autonomous robots [1,29] or mobile stereoscopic systems [3]. In fact, height maps have been widely used in mobile robotics to describe the environment and calculate trajectories on them [4,8,34,38]. Once the height map of the environment is created, the foreground is modelled as an occupancy map that registers the position of the moving objects in the environment.

Can be found mainly, in the related literature, two approaches for people detection when using stereo vision. The first one consists in considering a person as an object in an occupancy map with sufficient weight [17,18,20]. This approach is commonly used when the camera is placed at elevated positions. As we have previously commented, the main problem of that approach is that objects with dimensions similar to human beings that enters in the scene can be incorrectly detected as people. The second approach consists in looking for faces in the camera image [7,15,30]. This approach seems to be more appropriate when low camera positions are employed. However, if no additional information is employed, this approach is sensible to the false positives of the face detector. The system proposed in this work combines these two approaches to avoid the drawbacks of each one them. An object detected in the foreground occupancy map is considered as person if it has appropriate dimensions (human being dimensions) and if a face is detected on it. To speed up computation, the face detector is only applied on selected regions of the image where it seems possible to find faces. It is important to remark that the face detector is only employed for people detection. Once a person is detected, the tracking process does not employ the face detector so the person does not need to look at the camera to be tracked.

When a foreground object is identified as a person, the system starts to track him in the occupancy map. The system is able to simultaneously keep track of the detected people and to look for new people. People is tracked by combining position information (using the Kalman filter [16]) with information about the color of their clothes. Most of the works that perform people tracking using stereo vision rely uniquely on position information. However, when several people come close to others, the position prediction is not reliable. In close distances people tend to change their trajectories to avoid a collision or even stop walking to begin an interaction. Our system combines color and position information to achieve a robust tracking event when people interact at close distances.

3. Environment map building

This section presents the basis of the stereo processing, background modelling and foreground extraction. It is structured in three parts. Subsection 3.1 explains the basis of stereo calculation and how the 3D points captured by the stereo camera are translated to another reference system more appropriate for our purposes. Then, Subsection 3.2 explains the technique employed to create the background height map using the translated 3D points. Finally, Subsection 3.3 explains how the height map is used to extract the foreground objects.

3.1. Stereo processing

A commercial stereo camera [32] has been employed in this work. It can capture two images from slightly different positions (stereo pair) that are transferred to the computer to calculate a *disparity image* I_d containing the points matched in both images. Knowing the extrinsic and intrinsic parameters of the stereo camera it is possible to reconstruct the three-dimensional position p_cam of a pixel (u, v) in I_d. Let us denote by $P_\mathrm{cam} = \{p^0_\mathrm{cam}, \ldots, p^{np-1}_\mathrm{cam} | p^i_\mathrm{cam} = (X^i_\mathrm{cam}, Y^i{}_\mathrm{cam}, Z^i_\mathrm{cam}, 1)^T\}$ the set of three dimensional points (in homogeneus coordinates) captured by the camera that are calculated using Eq. 1. Where, f is the focal length of the cameras, b is the baseline distance between the cameras and d the disparity value of the pixel (u, v) in the image I_d.

$$\begin{aligned} Z_\mathrm{cam} &= \frac{fb}{d} \\ X_\mathrm{cam} &= \frac{uZ_\mathrm{cam}}{f} \\ Y_\mathrm{cam} &= \frac{vZ_\mathrm{cam}}{f} \end{aligned} \tag{1}$$

The three-dimensional positions p^i_{cam} calculated by Eq. 1 (affected by typical stereo errors [28,33]) are referred to the stereo camera reference system. In our case it is centered at the right camera (also known as *reference camera image*). However, this reference system may be changed from one application to another, i.e., the stereo camera can be placed at different positions and with different orientations. Hence, it is preferable for our purposes to translate the position of the points captured to a "*world*" reference system placed at ground level and parallel to it. Knowing the position and orientation of the camera in relation to the floor plane, it is possible to calculate the linear transformation matrix T that translates the points p^i_{cam} into $p^i_{\mathrm{w}} = (X^i_{\mathrm{w}}, Y^i_{\mathrm{w}}, Z^i_{\mathrm{w}}, 1)^T$ using Eq. 2. For more information about three-dimensional transformations the interested reader is referred to [10].

$$p_{\mathrm{w}} = Tp_{\mathrm{cam}}. \tag{2}$$

Fig. 1(a) shows an example of an scene captured with our stereo camera (the image corresponds to the right camera). Fig. 1(b) shows the three-dimensional reconstruction of the scene captured using the points detected by the stereo camera. The "*world*" and camera reference systems have been superimposed in the Fig. 1(b). As it can be seen in Fig. 1(b), the number of points acquired by an stereo camera can be very high (they are usually referred to as point cloud). For that reason, many authors perform a reduction of the amount of information by orthogonally projecting them into a 2D plan-view map [17,18,20]. This decision is also supported by the fact that people do not tend to be overlapped in the floor plane as much as they are in the original captured images. Therefore, the detection and tracking process is more reliable in the 2D projection.

A plan-view map divides a region of the floor plane into a set of *nxm* cells of fixed size δ. In this work, the cell $(x, y) = (0,0)$ coincides with the "*world*" positions (0, 0, Z_{w}, 1)^T (see Fig. 1(b)). Hence, the cell (x^i, y^i) in which a three-dimensional point p^i_{w} is projected can be calculated as:

$$x^i = (X^i_{\mathrm{w}}/\delta);\; y^i = (Y^i_{\mathrm{w}}/\delta) \tag{3}$$

Every time a stereo pair is captured, the set of 3D points that are projected on each cell is calculated as:

$$P_{(x,y)} = \{i | x^i = x \wedge y^i = y \wedge Z^i_{\mathrm{w}} \in [h_{\min}, h_{\max}]\}.$$

Where [$h_{\min}$, $h_{\max}$] is a height range that has two purposes. On one hand, the superior limit $h_{\max}$ avoids using points from the ceiling or from objects hanging from it (e.g. lamps). On the other hand, the inferior limit $h_{\min}$ excludes from the process low points that could not be relevant for a particular application (floor points or even the legs of people) and thus helps to reduce the computing time. The height range [$h_{\min}$, $h_{\max}$] should be such that, at least, the head and shoulders of the people to detect should fit in it. The rest of points p^i_{w} whose projection is outside the limits of the plan-view map are not considered.

The selection of δ must be made taking into account several aspects. A high value helps to decrease the computational effort and the memory used. The side effect is that it causes a loss of precision in the estimation of the position. On the other hand, a low value increases the precision (up to the limit imposed by the errors of the stereo computation [28,33]) but also the computational requirements. Ismail et al. [17] propose the use of $\delta = 0.2$ cm while Harville [18] uses $\delta \in [2,4]$ cm. We have selected a value of $\delta = 1$ cm which according to our experimentation is an

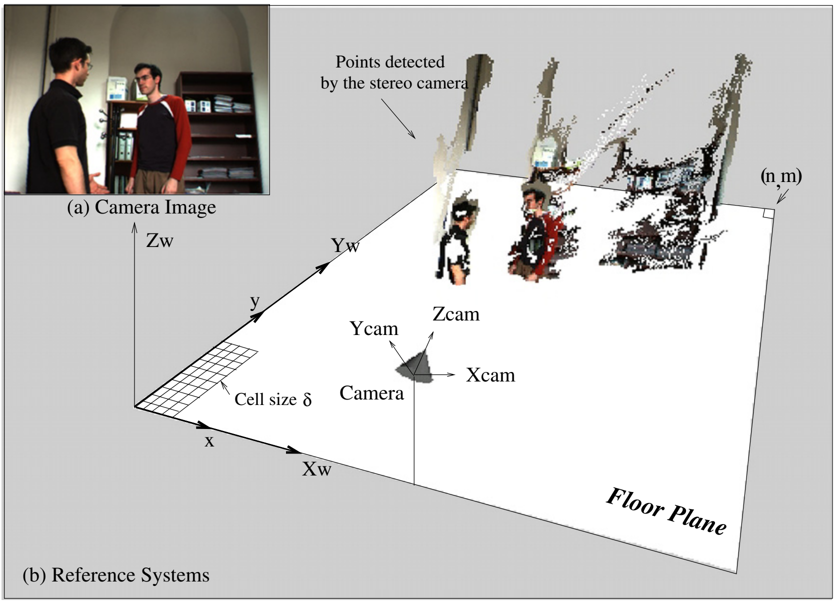


Fig. 1. (a) Image of the right camera captured with the stereo system. (b) Three-dimensional reconstruction of the scene showing the reference systems employed.

adequate balance between both requirements (precision and memory).

3.2. Background modelling

People can be considered movable elements in the environment. Thus, it is very helpful to separate the points that belong to the environment (background) from those that do not (foreground). Our background modelling approach consists in creating a geometrical height map of the environment $\hat{\mathscr{H}}$ [6] that indicates in each cell $\hat{\mathscr{H}}_{(x,y)}$ the maximum height of the points projected in it. We might think of $\hat{\mathscr{H}}$ as a representation of the environment over which foreground objects move. To avoid including objects momentarily passing by in the background map, $\hat{\mathscr{H}}$ is created aggregating several instantaneous height maps $\mathscr{H}^t$ with a robust estimator as the median:

$$\hat{\mathscr{H}}_{(x,y)} = \text{median}(\mathscr{H}^{t=t_0}_{(x,y)}, \ldots, \mathscr{H}^{t=t_0+\Delta t}_{(x,y)}). \qquad (4)$$

Each cell of an instantaneous $\mathscr{H}^t$ is calculated as:

$$\mathscr{H}^t_{(x,y)} = \begin{cases} \max(Z^j_w \mid j \in P_{(x,y)}) & if \ P_{(x,y)} \neq \emptyset \\ h_{\min} & if \ P_{(x,y)} = \emptyset \end{cases} \qquad (5)$$

Fig. 2 shows the evolution of $\hat{\mathscr{H}}$ from a set of 13 instantaneous height maps $\mathscr{H}^t$ captured at time intervals of $\Delta t = 400$ ms. Due to space reasons, the Fig. 2 shows only the status of the map at the time instants $t = \{0,\ 1600,\ 4000,\ 5200\}$ ms. Fig. 2(a–d) (upper row) show the images captured by the stereo system. Fig. 2(e–h) (middle row) are the corresponding instantaneous height maps $\mathscr{H}^t$. Dark areas represent the highest zones and white areas represent the lowest ones $h_{\min}$. Finally, Fig. 2(i–l) (lower row) show the evolution of the height map $\hat{\mathscr{H}}$ as more instantaneous height maps are employed to calculate it. Notice that $\hat{\mathscr{H}}$ has been created in the presence of people moving in the environment (their positions have been marked in the instantaneous height maps). As it can be seen, at the beginning $\mathscr{H}^{t=0} = \hat{\mathscr{H}}$ and thus the moving person appears in the height map. However, $\hat{\mathscr{H}}$ tends to truly represent the motionless characteristics of the environment as more instantaneous height maps are employed. The height maps shown Fig. 2 have been created using $h_{\min} = 0.5$ m and $h_{\max} = 2.1$ m.

$\hat{\mathscr{H}}$ can be periodically updated in order to dynamically adapt to the changes in the environment. Nonetheless, the number of instantaneous height maps $\mathscr{H}^t$ employed to create $\hat{\mathscr{H}}$ must be limited to the more recent ones in

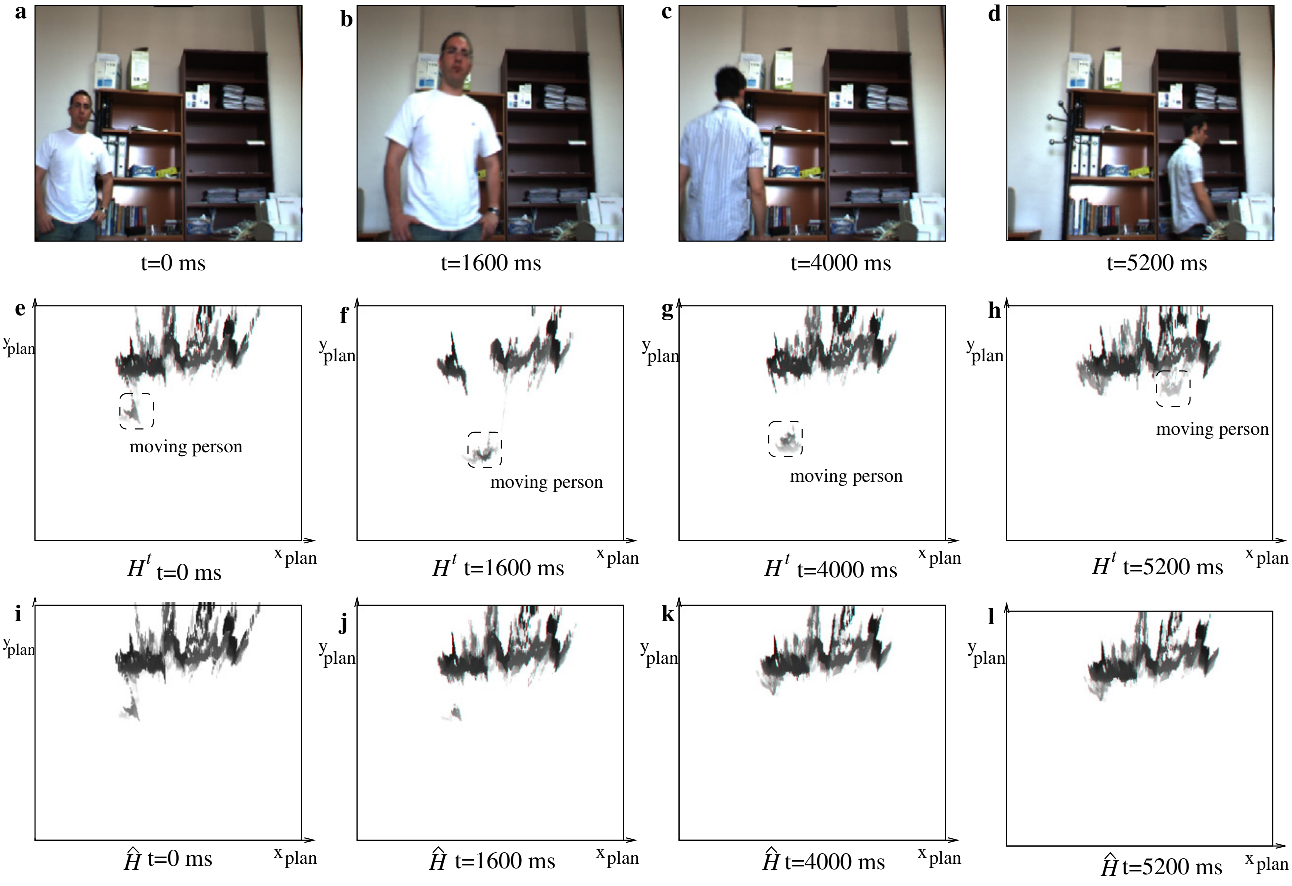


Fig. 2. Creation of the height map. Upper row (a–d) shows the images in instants {0, 1600, 4000, 5200} ms. Central row (e–h) shows the instantaneous height maps $\mathscr{H}^t$ for each one of the upper images. Lower row (i–l) shows the evolution of the height map $\hat{\mathscr{H}}$ created as the median of the height maps $\mathscr{H}^t$ created until that moment.

order to avoid old data influence the updating process. In our experiments, $\hat{\mathscr{H}}$ was appropriately updated using the last 10 instantaneous height maps. The frequency employed to update the height map should be smaller than the employed to detect and track people. In this way, it is possible to avoid including as part of the background people momentarily standing by or other moving objects. The update frequency employed in our experiments has been set to 0.1 Hz. Thus, the complete height map used was updated every 10 s. However, the update frequency should be set for each particular application.

3.3. Foreground extraction

Foreground points detected in each captured stereo pair can be easily isolated using the height map $\hat{\mathscr{H}}$. We have employed a plan view-map $\mathscr{O}$ (called *occupancy map*) that registers in each cell $\mathscr{O}_{(x,y)}$ the amount of foreground points projected in it. Lets denote by

$$F_{(x,y)} = \{i | i \in P_{(x,y)} \wedge Z^i_{\mathrm{w}} > \hat{\mathscr{H}}_{(x,y)}\},$$

to the set of points that projects in cell (x,y) and are above the height indicated in $\hat{\mathscr{H}}_{(x,y)}$, i.e., the foreground points detected over the background surface that represents $\hat{\mathscr{H}}$. Then, each cell of the occupancy map is calculated as:

$$O_{(x,y)} = \sum_{j \in F_{(x,y)}} \frac{(Z^j_{\mathrm{cam}})^2}{f^2} \tag{6}$$

The idea is that each foreground point increments the cell in which it is projected by a value proportional to the surface that it occupies in the real scene [18]. Thus, points close to the camera correspond to small surfaces and vice versa. If the same increment is employed for every cell, the same object would have a lower sum of the areas the farther it is located from the camera. This scale in the increment value will compensate the difference in size of the objects observed according to their distance to the camera. Fig. 3(b) shows the occupancy map $\mathscr{O}$ of the scene in the Fig. 3(a) using the height map $\hat{\mathscr{H}}$ from Fig. 2(l). Darker values represent areas with high occupancy density. The image has been manually retouched to make the occupied areas visible. As it can be seen, there are small dark dots in the upper area of Fig. 3(b) that are caused by errors of the stereo correlation process. However, the person that stands in the scene is clearly projected in the occupancy map as a connected group of cell with high occupancy level.

The next step in our processing, is to identify the different objects present in $\mathscr{O}$ that could correspond to human beings. For that purpose, $\mathscr{O}$ is processed with a closing operator in order to link possible discontinuities in the objects caused by the errors in the stereo calculation. Then, objects are detected as groups of connected cells. Those objects whose area is similar to the area of a human being and whose sum of cells (occupancy level of the object) is above a threshold θ_{occ} are employed in the next phase for people detection and tracking. This test is performed in a flexible way so that it is possible to deal with the stereo errors and partial occlusions. Fig. 3(c) shows the unique object detected in the occupancy map of Fig. 3(b) after the above mentioned process. This approach is very fast but, the presence of several people together is a problematic case since a single face is being looked for into the detected area. However, it is not a big problem because the system will catch each person as soon as they separate or, a face detector over the whole foreground can be performed for that special case (foreground detection and no people detection).

4. People detection and tracking

People detection and tracking are treated as separate processes in this work. Every time a new scene is captured, the system must decide first, which one of the objects detected in $\mathscr{O}$ corresponds to each one of the people that are being tracked (an assignment problem). Then, the system applies a face detector on the remaining objects in order to detect new people. Notice that the face detector is only applied on those objects that have not been detected as people yet. It allows to manage false negatives detections of the face detector, i.e., although the face detector might fail in detecting a person once, it could succeed in the next image. Our people tracking approach consist in: (i) predicting the

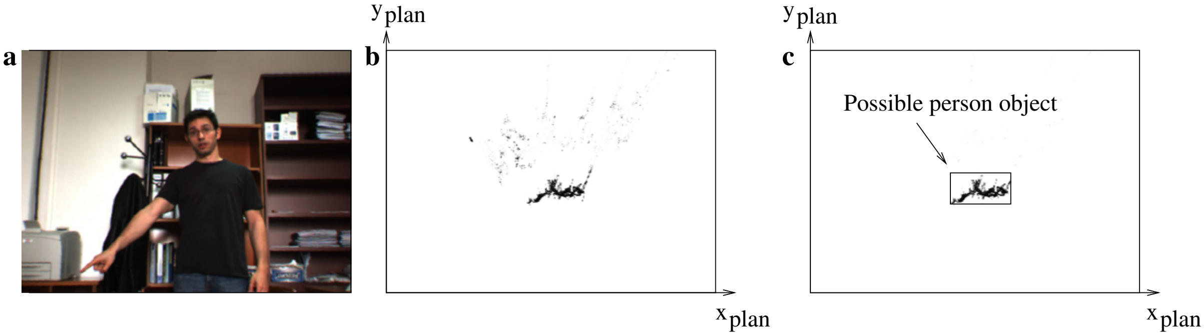


Fig. 3. (a) Right image of the pair in an instant, the environment with an object not in background. (b) Occupancy map $\mathscr{O}$ corresponding to the environment. (c) Framed information corresponding to the object, detected using $\mathscr{O}$.

future positions of each person according to its past movement (using the Kalman filter) and, (ii) solving the assignment problem combining position information with color information about the person's clothes. Thus, a color model of each object in $\mathcal{O}$ is created and compared with the color model of each person being tracked.

The system divides each object in two different parts. The fist one is the top region of the object that should correspond to the person's head (*head region*) and is employed for face detection purposes. The second one encloses the torso (and part of the legs) of the person (*body region*) and is employed to create the person's color model. These regions are determined in the following way. The 3D top head point is determined as the highest point in the mass center of the object. Then, the head region is considered to be a 30 × 15 cm sized rectangle centered below the top head point. The body region is considered to be a 45 × 100 cm sized rectangular region placed below the head region. The 2D projection of these regions in the reference camera image are calculated projecting them back by inverting the stereo calculation explained in Subsection 3.1. Thus, the size of the two regions in the reference camera image are appropriately fitted to the varying size of the person projection at different distances. Fig. 4 shows a reference camera image where the head and body regions have been superimposed.

Next subsections explain in detail the detection and tracking phases. Subsection 4.1 explains how the color model of each object is created. Later, in Subsection 4.2 it is shown how people detection is performed. And finally, Subsection 4.3 explains how these pieces of information are fused to perform the tracking.

4.1. Color modelling

A color model $\hat{q}$ of each object is created to assist the tracking process. The color model aims to capture information about the color of the clothes of the people in the scene. The color model $\hat{q}$ of each object is modelled as an histogram using technique described by Comaniciu et al. [5]. The *HSV* space [11] has been selected to represent color information. The histogram $\hat{q}$ is comprised by $n_{\mathrm{h}}n_{\mathrm{s}}$ bins for the hue and saturation. However, as chromatic information is not reliable when the value component is too small or too big, pixels on this situation are not used to describe the chromaticity. Because these "color-free" pixels might have important information, the histogram is also populated with n_{v} bins to capture its illuminance information. The resulting histogram is composed by $m = n_{\mathrm{h}}n_{\mathrm{s}} + n_{\mathrm{v}}$ bins.

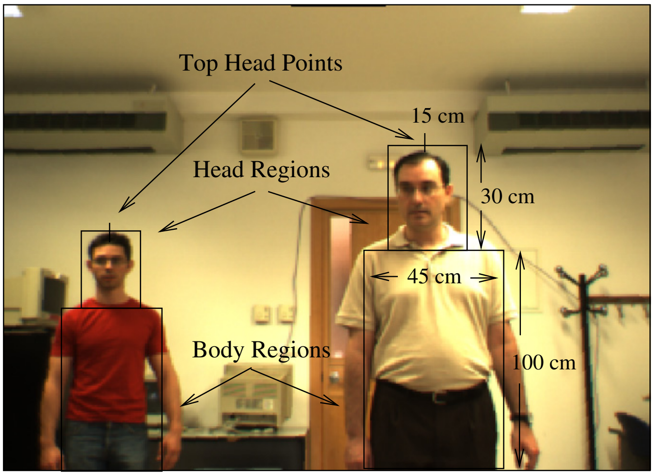


Fig. 4. Example of scene showing the head and body regions of two people.

Let $\{x_i^*\}_{i=1\ldots n}$ be the locations of the pixels employed to create the color model. We define a function $b : \Re^2 \rightarrow \{1\ldots m\}$ which associates to the pixel at location x_i^* the index $b(x_i^*)$ of the histogram bin corresponding to the color of that pixel. The color density distribution for each bin $\hat{q}_{\mathrm{u}}$ of the region x^* is calculated in the following way:

$$\hat{q}_{\mathrm{u}} = K\sum_{i=1}^{n} w(x_i^*)\kappa[b(x_i^*) - u]. \tag{7}$$

The weighting function w gives more relevance to pixels near the central point of the region x^* (p_{chest}) and thus reduces the influence of background pixels that might be incorrectly included. Function κ represents the Kronecker delta function. Finally, K is a normalization constant calculated by imposing the condition $\sum_{u=1}^{m}\hat{q}_u = 1.$, from where

$$K = \frac{1}{\sum_{i=1}^{n} w(x_i^*)} \tag{8}$$

since the summation of the Kronecker delta functions is equal to 1.

Once the color model $\hat{q}$ of an object is created, it can be compared with other color model $\hat{p}$ using the Bhattacharyya coefficient [2,22]. In the case of a discrete distribution of our color models it can be expressed as:

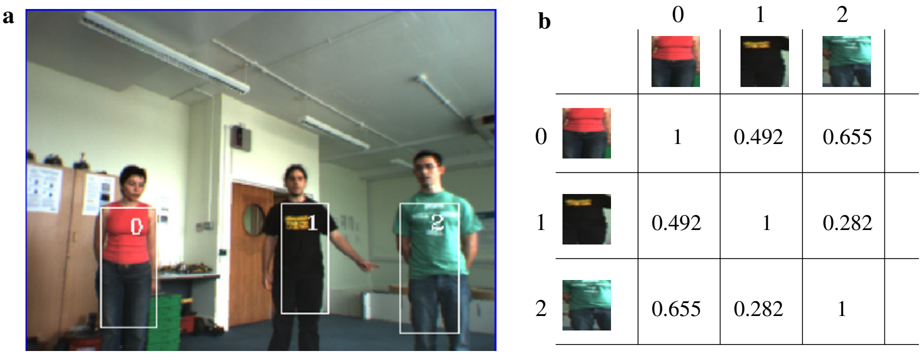
$$\rho(\hat{q},\hat{p}) = \sum_{u=1}^{m}\sqrt{\hat{q}_u\hat{p}_u}. \tag{9}$$

The value $\rho(\hat{q},\hat{p})$ gives a similarity measure in the range [0,1] between two color models. Values near 1 indicate that both color models are very similar and vice versa. Fig. 5(b) shows as a table the value of this metric for the body regions selected in Fig. 5(a). There have been employed histograms with a total of $m = 30$ bins ($n_{\mathrm{h}} = n_{\mathrm{s}} = n_{\mathrm{v}} = 5$).

Changes in the illumination conditions might change the observed color distribution of the person's clothes. Therefore, it is necessary to update the color model to achieve a robust tracking. If $\hat{q}$ is the person's color model and $\hat{q}'$ is the observed color model in the next frame, an updated model $\hat{q}''$ can be efficiently computed as [31]:

$$\hat{q}'' = (1-\alpha)\hat{q} + \alpha\hat{q}', \tag{10}$$

where α ponderates the contribution of the observed color model to the updated one. High values for α makes the



	0	1	2
0	1	0.492	0.655
1	0.492	1	0.282
2	0.655	0.282	1

Fig. 5. (a) Scene with three people in it where the regions employed to create their color models have been marked. (b) Table showing the similitude $\rho(\hat{q},\hat{p})$ between the color distribution of the three people's clothes.

updated color model forget old values thus adapting rapidly to changes, and low values produce a slow adaptation to illumination changes.

4.2. People detection

Our approach for people detection consists in analyzing if an object detected in $\mathcal{O}$ shows a face in the camera image. As previously indicated, the detection process is performed only on the remaining objects after tracking of known people has been completed. Face detection is a process that can be time consuming if applied on the entire image. Thus, it is only applied on these regions of the camera image where the head of each object should be. As the human head has an average dimensions, the system analyzes whether the 3D points in the top part of the object reveals a width similar to the width of the human head. The test is performed in a flexible manner so that it can handle stereo errors and people with different morphological characteristics can pass it. If the object passes the test, the head region of the reference camera image is analyzed using a face detector. The reduction of the search region where the face detector is applied brings two main advantages. First, it reduces the computational time as smaller regions are analyzed. Second, it reduces the number of false positives as stated in [23].

The system employs the face detector provided by the OpenCv's Library [21]. It is not in the scope of this paper to develop face detection techniques since there is plenty literature about it [41]. The face detector is based on the of Viola and Jones [39] method which was later improved by Lienhart [26]. The implementation is trained to detect both frontal and lateral views of human faces and works on gray level images. Fig. 6(a) shows a scene where there is a person that has entered and has hanged his coat. Fig. 6(d) shows the occupancy map $\mathcal{O}$ of that scene. As it can be noticed, two objects are detected, the person and the coat. Using our people detection procedure explained, it is detected that the size of the upper part of the two objects are similar to human's heads. Hence, the regions of the image that should contain their faces (Fig. 6(b) and (c)) are processed by the face detector. In that case, the face detector only detects a face in the left object (Fig. 6(b)).

4.3. People tracking

Once an object has been identified as a person, it is necessary to keep track of him in the following images. The tracking problem can be seen as an assignment problem, i.e., relate a person that is being tracked with an object currently detected in $\mathcal{O}$. The problem has been solved using the

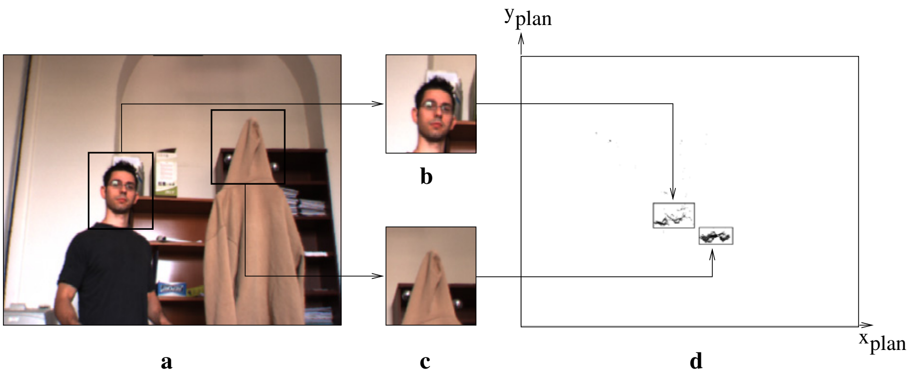


Fig. 6. Example of people detection (a) Image captured by the camera (b and c) Images of the upper part of the objects detected. (d) Occupancy map of the scene.

Kuhn's well-known Hungarian Method for solving optimal assignment problems [24].

Let us denote by $\Upsilon = (\Phi, \Pi)$ the set of n objects detected in $\mathcal{O}$. The set $\Phi = \{\text{obj}^1,\ldots,\text{obj}^n\}$; $\text{obj}^i = (x^i_{\text{obj}}, y^i_{\text{obj}})$ denotes the location of their center of masses in the plan-view map, and $\Pi = \{\hat{p}^1,\ldots,\hat{p}^n\}$ their corresponding color models, being $\hat{p}^i$ the color model of the object obj^i.

Let also denote by $\Psi = (\varphi, \Omega)$ the set of m people detected and being tracked. The set $\varphi = (s^1\ldots s^m)$; $s^j = (x^j_p, y^j_p, v^j_x, v^j_y)$, denotes their positions and velocities, and $\Omega = (\hat{q}^1,\ldots,\hat{q}^m)$, their color models created when their faces were detected.

The assignment problem consists in determining the optimal assignment of currently detected objects Υ, to the people being tracked Ψ. In order to use the Hungarian method, it is necessary to calculate a cost value of assigning the object obj^i to the person s^j. In this work, this cost is calculated accordingly to two features. The first one is the difference between the position of the object and the predicted position for the person. The second one is the similitude between the color models of the object and the person by Eq. 9.

Kalman filter is employed to predict the new position $s^j_{\text{pred}} = (x^j_p, y^j_p)$ of each person in the plan-view maps using a linear model of his movement:

$$x(t+1) = x(t) + v_x t; \;\; y(t+1) = y(t) + v_y t,$$

where v_x and v_y are the velocities of the person in the plan-view map. Although it is a basic movement model, it is able to successfully predict the position of people when images are captured at short time intervals. Fig. 7 shows the tracking results of a real sequence of 9 s captured at 7 Hz where a person walks 6.93 m at steady speed. The solid line represents the observed path of a person moving in the environment while the dashed line represents the predictions of the Kalman filter. As it can be observed, the prediction model is able to estimate the trajectory of the person with a low error rate. Fig. 8 shows the estimation errors. Notice that the maximum error does not exceed 15 cm, and it occurs when the person is turning.

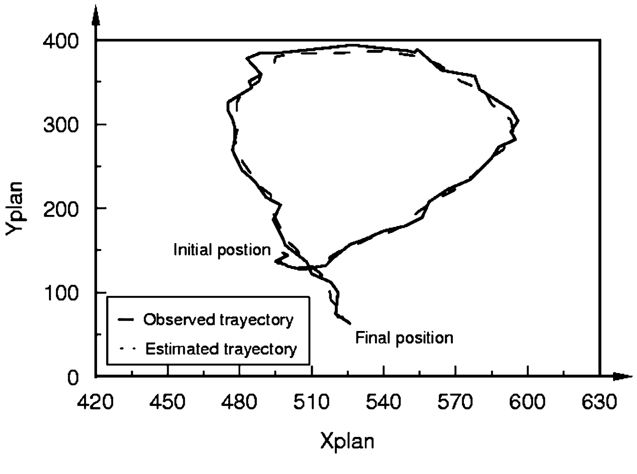


Fig. 7. Example of trajectory of a person. Lines show both the observed positions and the positions estimated by the Kalman filter.

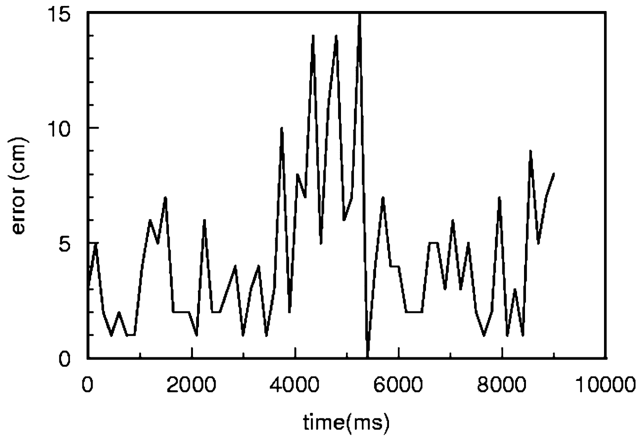


Fig. 8. Errors in the predictions while tracking a person.

In order to combine both pieces of information (color and position) into a single cost value, we have employed the cost function expressed in Eq. 11. The exponential term of the equation measures the distance between the position of the object $\text{obj}^i(x^i_o, y^i_o)$ and the prediction for the person $s^j(x^j_p, y^j_p)$. Values close to 1 indicate that the object is near the predicted position of the person. The parameters σ^j_x and σ^j_y are a measure of the uncertainty associated to the position predicted for s^j and are given by the Kalman filter prior uncertainty matrix. They decrease when the person does not move and increase when the person moves (in proportion to the speed) or when the person cannot be tracked in a image (indicating that his position has a higher uncertainty). The second term of the Eq. 11 compares the color models of the object and the person via Eq. 9. Eq. 11 provides values in the range $[0, 2]$. Values close to 0 indicate that an object in far from the expected position of a person and they have different color distribution. Values close to 2 are achieved for objects placed at the position predicted by the Kalman filter and whose color distribution are very similar to the person being tracked.

$$S(o^i, \hat{p}^i, s^j_{\text{pred}}, \hat{q}^j) = \mathrm{e}^{-\left(\frac{\left(x^i_o - x^j_p\right)^2}{2\left(\sigma^j_x\right)^2} + \frac{\left(y^i_o - y^j_p\right)^2}{2\left(\sigma^j_y\right)^2}\right)} + \rho(\hat{q}^i, \hat{p}^j) \qquad (11)$$

The combined use of color and position information is specially useful in case of close interactions. When two or more people come close to each other, position information is not reliable. In these situations, people tend to alter their trajectory and a tracking scheme based exclusively on position might fail. Our approach of combining also color information allows to overcome the previous problem. If two or more people wearing clothes of different colors become close to each other, color information helps to distinguish between them. Of course, if the people wear clothes of very similar colors, the system has not enough information to solve the problem. In these cases, other sources of information should be added to enhance tracking.

A robust tracking system must deal with occlusions. If the person is partially visible, it will be detected as an object if the sum of its cells in $\mathcal{O}$ is higher than the threshold θ_{occ} (previously explained in Section 3.2). The value θ_{occ} is selected so that a person could be projected as an object in case of partial occlusion and dealing with stereo errors. However, if a person s^j is very occluded, it is not extracted any object from $\mathcal{O}$ for that person. In that situation the system must take care of not to assign an incorrect object obj^i to that person. For that reason, the system only considers valid an assignment when the probability value, given by Eq. 11, overcomes a threshold value θ_{assig}. If it does not happen for the person s^j, the system keeps predicting his position and still looking for it for a maximum number of times. If the person s^j remain unseen for too much time, the system assumes that the person has definitively left the scene and deletes him from Ψ.

When the assignment problem has been solved, both the Kalman filter and the color model of each "*assigned*" person are updated using the information of its corresponding objects. The color model $\hat{q}^j$ of the person s^j is updated using the color model $\hat{p}^i$ of its corresponding object obj^i as indicated in Eq. 10.

5. Experimentation

During the explanation of the model we have shown examples of its performance. A broader experimentation has been done to test detection and tracking of different people under different illumination conditions and different distances from the stereo vision system. We have employed 320 × 240 sized images and sub-pixel interpolation to enhance the precision in the stereo calculation. The operation frequency of our system is near 10 Hz on a 3.2 Ghz Pentium IV laptop computer running with Linux. More than the fifty percent of the computing time is dedicated to image capturing and stereo computation (about 50 ms) and the rest to detection and tracking (about 40 ms). It indicates that the proposed system is fast enough to be used in real time applications.

The face detector used has been configured to detect people at distances ranging from 0.5 to 2.5 m. Although it could be configured for detecting people at larger distances, we have noticed that it substantially increments the time required to analyze an image. However, once a person has been located, it can be tracked up to distances of 5 m. At higher distances, the errors of the stereo system employed are so high that people cannot be correctly tracked.

A set of 18 color-with-depth video sequence, captured at 7 Hz, has been recorded in order to test the performance of the tracking process and the influence of color in it. The total time of all the sequences sum 10′23″ and they were recorded for camera heights ranging between 0.5 and 1.2 m. Some of them were recorded using an stereo camera of $f = 6$ mm and the others using an stereo camera of $f = 4$ mm. The number of people in each sequence is different and it varies from 2 to 4. In the sequences, people perform several types of interactions: walk at different distances, shake hands, cross their paths, jump, run, embrace each other and even quickly swap their positions trying to confuse the system.

To evaluate the success of the system in tracking people, we have manually count the number of potentially "*conflicting*" situations that take place in each video sequence. A conflicting situation is considered when: (i) most of the body of a person is out of the camera image, (ii) a person is almost totally occluded by another person and (iii) two or more people collide, embrace or touch each other. Fig. 9 shows some images of one of the video sequences. An example of the conflicting situation (i) can be seen in Fig. 9(a). Examples of the conflicting situation (ii) can be observed in Fig. 9(b, c, g), and examples of the conflicting situation (iii) can be seen in Fig. 9(e and h).

Each video sequence has been processed twice: one time using both color and position information and a second using only position information. The processed video sequences can be downloaded in *avi* format at http://decsai.ugr.es/~salinas/humanrobot.htm. After processing them, the results have been examined and we have apunted the number of times that the system was able to successfully detect the object associated to each person when a conflicting situation had ended. Table 1 summarizes the results obtained. Column #*people* indicates the number of people in the test, f the focal length of the camera used and #*conflicts* the number of all the conflicting situations counted in these video sequences. Column #*nc* indicates the success of the system in tracking the people in the conflicting situations without using color information. Finally, column #*c* indicates the success of the system when color information is employed.

At the light of the results showed in Table 1, it can be pointed out that the use of color information is a powerful clue when tracking people. We have also observed that for the $f = 6$ mm stereo camera, more than three people make the tracking process unreliable because of the excessive occlusions that occurs when people is near the camera. Similarly, the maximum number of people for an appropriate tracking using the $f = 4$ mm camera is 4. It can be observed that as the number of people increases, the system tends to fail more (specially when color information is not employed). It must also be mentioned, that no false positives where detected on the video sequences. Thus, the combination of face detection and object detection in the occupancy map seems to be a very appropriate method for an accurate people detection.

Additionally, we have observed that an important advantage when using color: despite the system can incorrectly assign a person to the object of another person in a image, the error can be corrected in the next image if both individuals are wearing clothes of different colors. This is because the color comparison is continuously done and the correct assignment can be done in the next image (while

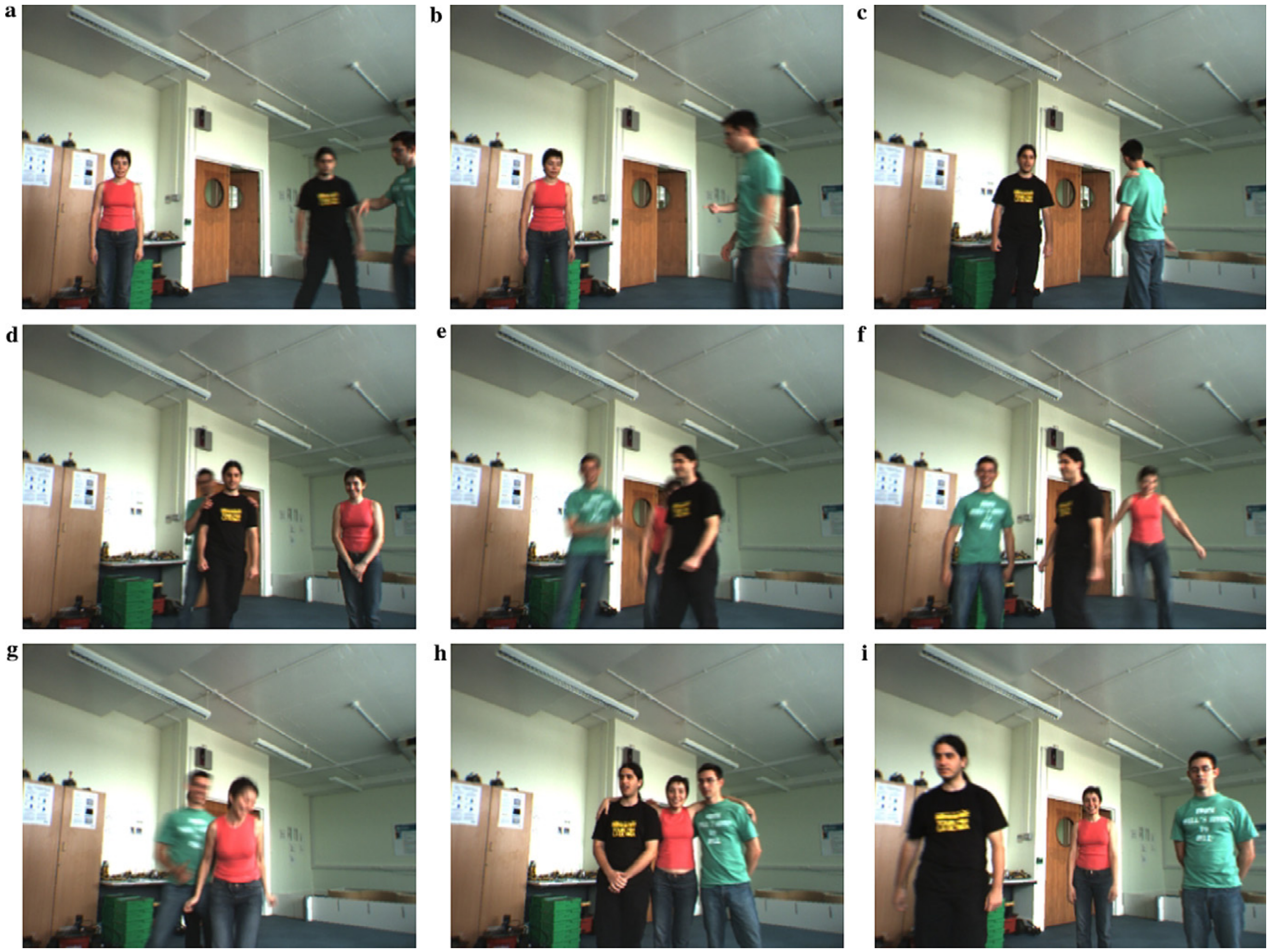


Fig. 9. Images of a sequence employed to test the system.

Table 1
Success of the tracking system when using and not using color

#People	f (mm)	#Conflicts	#Nc (%)	#c (%)
2	6	30	86	100
3	6	17	58	82
3	4	52	69	100
4	4	13	50	100

the people incorrectly classified still close). This correction is not possible when tracking is only based on position information.

6. Conclusions and future work

We have presented a system able to detect and track multiple people using an stereo camera placed at underhead positions. The method proposed is especially indicated for applications that require analyzing the user gestures and facial expression because of the position of the camera. The system uses a height map built using depth information to model the environment. The background model obtained is more robust to sudden illumination changes than approaches based on intensity values because of the use of depth information [6]. Besides, it is specially appropriated for mobile devices.

Each time a new stereo pair is captured, an occupancy map that registers the position of the foreground objects is created. Foreground objects with dimensions similar to human beings are considered as potential candidates to people and a color model of each one of them is created. Then, the system performs the tracking of the already known people. Tracking is considered as an assignment problem, i.e., assign known people to the objects detected in the occupancy map. To calculate the best assignment scheme, the Hungarian Method [24] has been employed. For that purpose it is necessary to calculate a cost value of each object detected to be a known person. This cost is calculated combining information about a color model of each person and its predicted position using the Kalman filter. The combined use of color and position information allows a robust tracking avoiding tracking failures that occur when using only position information. Once the assignment problem is solved, both the Kalman filter and the color models of the tracked people are updated. The remaining objects not belonging to any known person are examined using a face detector in order to detect new people in the scene. Our system does not apply the face detector on the entire image but only in the image region where the head should be placed. The people detection technique employed by our system allows to greatly reduce the computing time required and helps to avoid false positives of the face detector [23].

The system has been extensively tested on 18 color-with-depth video sequences (summing a total time of $10'23''$) where several people move freely in the environment. The video sequences have been processed twice, with and without color, in order to analyze the influence of color in the tracking process. The results show that the use of color allows to greatly reduce the number of errors of the tracking system. The proposed system is able to track multiple-people up to distances of 5 m with very high success rate without using complex three-dimensional models to describe the scene. Besides, the time required to perform the detection and tracking is only 40 ms, which induces to think that it could be suitable for real time applications. It is also remarkable that no false detections were registered in the tests performed.

As future work, we consider of interest the use of multi-scale techniques for creating the plan-view maps [13] and avoiding the use of a unique cell size. This could allow to manage the errors of the stereoscopic system in a more flexible way. We also consider that it could be interesting the use of additional features that could discriminate between people wearing similar colored clothes. In particular, the use of face identification techniques seem interesting for that purpose. Finally, we find that the system is specially appropriated for mobile devices. Therefore, as future work we also plan to include the system into a bigger architecture that controls and autonomous mobile robot [1,29] and to test it suitability for human–machine applications that require to operate in real time and in moving conditions.

References

[1] E. Aguirre, A. González, Fuzzy behaviors for mobile robot navigation: design, coordination and fusion, International Journal of Approximate Reasoning 25 (2000) 255–289.

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

[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] W. Burgard, A.B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, S. Thurn, Experiences with an interactive museum tour-guide robot, Artificial Intelligence 144 (1999) 3–55.

[5] 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, vol. 2, 2000, pp. 142–149.

[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, International Journal of Computer Vision 37 (2000) 175–185.

[8] C. Eldershaw, M. Yim, Motion planning of legged vehicles in an unstructured environment, in: IEEE International Conference on Robotics and Automation (ICRA'2001), vol. 4, 2001, pp. 3383–3389.

[9] C. Eveland, K. Konolige, R.C. Bolles, Background modelling for segmentation of vide-rate stereo sequences, in: IEEE Conf. on Computer Vision and Pattern Recognition, 1998, pp. 266–271.

[10] J. Foley, A. Van Dam, S. Feiner, J. Hughes, Computer Graphics: Principles and Practice, Addison Wesley, 1990.

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

[12] T. Fong, I. Nourbakhsh, K. Dautenhahn, A survey of socially interactive robots, Robotics and Autonomous Systems 42 (2003) 143–166.

[13] M. García-Silvente, J.A. García, J. Fdez-Valdivia, The novel scale-spectrum space for representing gray-level shape, Pattern Recognition 30 (3) (1997) 367–382.

[14] D.M. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding: CVIU 73 (1) (1999) 82–98.

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

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

[17] I. Haritaoglu, D. Beymer, M. Flickner, Ghost 3d: detecting body posture and parts using stereo, in Workshop on Motion and Video Computing, 2002, pp. 175–180.

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

[19] 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.

[20] K. Hayashi, M. Hashimoto, K. Sumi, K. Sasakawa, Multiple-person tracker with a fixed slanting stereo camera, in: 6th IEEE International Conference on Automatic Face and Gesture Recognition, 2004, pp. 681–686.

[21] Intel. OpenCV: Open source Computer Vision library. <http://www.intel.com/research/mrl/opencv/>.

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

[23] H. Kruppa, M. Castrillon-Santana, B. Schiele. Fast and robust face finding via local context, in: Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2003.

[24] H.W. Kuhn, The hungarian method for the assignment problem, Naval Research Logistics Quarterly 2 (1955) 83–97.

[25] W. Liang, H. Weiming, L. Tieniu, Recent developments in human motion analysis, Pattern Recognition 36 (2003) 585–601.

[26] R. Lienhart, J. Maydt. An extended set of haar-like features for rapid object detection, in: IEEE Conf. on Image Processing, 2002, pp. 900–903.

[27] B. Martinkauppi, M. Soriano, M. Pietikainen, Detection of skin color under changing illumination: a comparative study, in: 12th International Conference on Image Analysis and Processing, 2003, pp. 652–657.

[28] R. Mohan, G. Medioni, R. Nevatia, Stereo error detection, correction, and evaluation, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (1989) 113–120.

[29] R. Muñoz-Salinas, E. Aguirre, M. García-Silvente, M. Gómez, A multi-agent system architecture for mobile robot navigation based on fuzzy and visual behaviours, Robotica 23 (2005) 689–699.

[30] K. Nickel, E. Seemann, R. Stiefelhagen. 3D-tracking of head and hands for pointing gesture recognition in a human–robot interaction scenario, in: Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'04), 2004, pp. 565–570.

[31] K. Nummiaro, E. Koller-Meier, L. Van Gool, An adaptive color-based particle filter, Image and Vision Computing 21 (2003) 99–110.

[32] PtGrey. Bumblebee. <http://www.ptgrey.com/>.

[33] J.J. Rodriguez, J.K. Aggarwal, Stochastic analysis of stereo quantization error, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1990) 467–470.

[34] K. Sabe, M. Fukuchi, J.-S. Gutmann, T. Ohashi, K. Kawamoto, T. Yoshigahara,Obstacle avoidance and path planning for humanoid robots using stereo vision, in: IEEE International Conference on Robotics and Automation (ICRA'04), vol. 1, 2004, pp. 592–597.

[35] 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.

[36] L. Snidaro, C. Micheloni, C. Chiavedale, Video security for ambient intelligence, IEEE Transactions on Systems, Man and Cybernetics, Part A 35 (2005) 133–144.

[37] R. Tanawongsuwan, Robust Tracking of People by a Mobile Robotic Agent. Technical Report GIT-GVU-99-19, Georgia Tech University, 1999.

[38] S. Thompson, S. Kagami, Stereo vision terrain modeling for non-planar mobile robot mapping and navigation, in: IEEE International Conference on Systems, Man and Cybernetics, vol. 6, 2004, pp. 5392–5397.

[39] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2001, pp. 511–518.

[40] 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.

[41] M.H. Yang, D. Kriegman, N. Ahuja, Detecting faces in images: a survey, IEEE Trans. on Pattern Analysis and Machine Intelligence 24 (1) (2002) 34–58.