ROBUST 3D PEOPLE TRACKING AND POSITIONING SYSTEM IN A SEMI-OVERLAPPED MULTI-CAMERA ENVIRONMENT

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

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

ABSTRACT

People positioning and tracking in 3D indoor environments are challenging tasks due to background clutter and occlusions. Current works are focused on solving people occlusions in low-cluttered backgrounds, but fail in high-cluttered scenarios, specially when foreground objects occlude people. In this paper, a novel 3D people positioning and tracking system is presented, which shows itself robust to both possible occlusion sources: static scene objects and other people. The system holds on a set of multiple cameras with partially overlapped fields of view. Moving regions are segmented independently in each camera stream by means of a new background modeling strategy based on Gabor filters. People detection is carried out on these segmentations through a template-based correlation strategy. Detected people are tracked independently in each camera view by means of a graph-based matching strategy, which estimates the best correspondences between consecutive people segmentations. Finally, 3D tracking and positioning of people is achieved by geometrical consistency analysis over the tracked 2D candidates, using head position (instead of object centroids) to increase robustness to foreground occlusions.

***Index Terms*—** Background modeling, human template, 3D tracking, geometrical consistency, occlusion robustness.

1. INTRODUCTION

Visual positioning and tracking of multiple people in indoor environments is an active research topic due to its applicability to surveillance systems, security, and restricted access area control, intelligent rooms, etc. Many works have addressed this task using one single camera (as in [1]), and have satisfactorily solved it in controlled environments: i.e. low cluttered background. However this single-camera approaches fail in presence of foreground occlusions. Recently, moving foreground occlusions have been partly overcome in [2], by using several views (obtained by different uncalibrated cameras) warped to find the ground areas where a moving object may stand with high probability. However, this approach fails when the static foreground clutter occludes the lower part of the moving objects (which is a usual situation in real environments), avoiding a correct projection of foreground objects on the ground plane.

To overcome these problems, a novel system for 3D positioning and tracking of multiple people in highly cluttered environments is proposed, solving both static and moving foreground occlusions using a multi-camera strategy. The overview of the system is depicted in Fig. 1. Moving regions are segmented in each frame by means of a novel motion detection technique based on background subtraction, that performs a multidimensional description of the background using a bank of Gabor filters. The resulting segmentation is correlated with a set of human-shape templates with different scales and orientations according to both the size of the room and the information from the camera calibration, to find the most probable people segmentations, discarding erroneously segmented regions mainly due to shadows. People tracking is performed by means of a graph-based matching strategy that finds the best correspondences between people segmentations in consecutive frames. Resulting correspondences representing people tracks are smoothed by means of a Kalman filtering. Obtained 2D tracking information from different cameras is then fed into the 3D positioning and tracking module. This final stage decides which 2D objects correspond to the same 3D moving object according to its geometrical coherence, and then models these 3D detected objects as cylinders to track them using a Kalman filter. The system assumes fully-calibrated cameras and works on the topmost points of moving objects (heads), as this feature shows greater robustness to static foreground occlusions than others such as, e.g. the centroid.

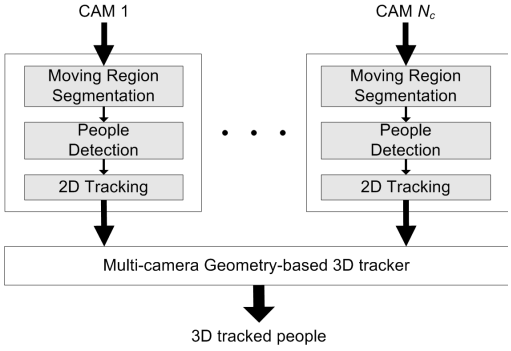


Fig. 1. Block diagram of the 3D positioning and tracking system

This paper is organized as follows: Section 2 describes the moving region segmentation strategy, which is used to detect people as it is explain in the Sec. 3. Tracking process within each camera is described in Sec. 4, while Sec. 5 presents the details of the 3D people tracking and positioning module. Experimental results are presented in Sec. 6, followed by the conclusions in Sec. 7.

2. MOVING REGION SEGMENTATION

Moving regions are segmented in the current image I^t, where t is the time instant, by means of a background substraction strategy

This work has been partially supported by the Ministerio de Ciencia e Innovación of the Spanish Government under project TEC2007-67764 (SmartVision), and by the Comunidad de Madrid under project S-0505/TIC-0223 (Pro-Multidis).

performed in a Gabor-based feature space. The background B^t is modeled by a multidimensional image where each pixel is a feature vector $\vec{F}$ that characterizes its neighborhood, computed using a bank of 16 Gabor-based filters according to (1).

$$\vec{F} = [G_Y(\theta_m), R_R(\theta_m), R_G(\theta_m), R_B(\theta_m)], m \in \{1, ..., 4\} \quad (1)$$

where $G_Y(\theta_m)$ is the Gabor filter [3] response applied over the luminance channel of the image, which describes the spatial luminance variations; $R_R(\theta_m)$, $R_G(\theta_m)$ and $R_B(\theta_m)$ are the half-wave rectified Gabor filter responses applied respectively over the red, green and blue image channels of the image, which characterize the spatial color distribution; and θ_m is the filter orientation, whose range is $\{0, 45, 90, 135\}$ degrees.

Each pixel of B^t is updated between the instant $t-1$ and t through (2) to make the model robust to slow illumination variations and changes due to image noise

$$B^t(x, y) = (1 - \alpha)B^{t-1}(x, y) + \alpha M^t(x, y) \quad (2)$$

where M^t is a multidimensional image obtained by means of the application of the aforementioned bank of Gabor-based filters over I^t; and α is a variable that controls the weight of the current image in the background model. Its value is a trade-off between adapting to slow illumination variations (a low value of α) and fast ones (a high value of α). Satisfactory results have been obtained using values in the range $(0.01, 0.1)$, depending on the variation of the background.

Background substraction is carried out by computing the Euclidean distance between the feature vectors of B^t and M^t. The resulting image is thresholded by means of the Median Absolute Deviation technique [4], obtaining the binary moving region segmentation S^t.

3. PEOPLE DETECTION

Shadows and static foreground occlusions can produce situations where a person is divided into several independent regions or where static regions are segmented in S^t. These problems have been solved correlating a set of human silhouettes, used as templates, over S^t to estimate the most probable people segmentation. Since the camera is calibrated, the range of the scale and the 3D orientation of the human silhouettes can be restricted according to the camera position and size of the room. In the current implementation 16 normalized human silhouettes have been used, four of them are shown in Fig. 2 as example. The correlation of the n^{th} template, T_n, is computed using the Mean Squared Error (MSE) as shown (3)

$$MSE_n^t(x, y) = \sum_{i=1}^{H_n} \sum_{j=1}^{W_n} \Big(T_n(x + i, y + j) - S^t(x, y) \Big)^2 \quad (3)$$

where H_n and W_n are respectively the height and the width of the n^{th} template. All MSE_n^t are combined by selecting the minimum value for each pixel coordinate, and the result is used to perform a non-minimal suppression, obtaining a combined matching error image ,P^t, that represents the most probable people locations. P^t must be properly thresholded to remove false detections due to segmented shadows (a high threshold value), but allowing a certain mismatching produced by static foreground occlusions (a low threshold value). According to the performed experiments, a threshold value in the range $(0.6, 0.75)$ allows to obtain an accurate people detection. Figure 3 shows four images depicting the people detection process.

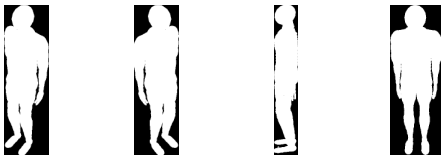


Fig. 2. Four of the 16 human templates used in the correlation process to detect people.

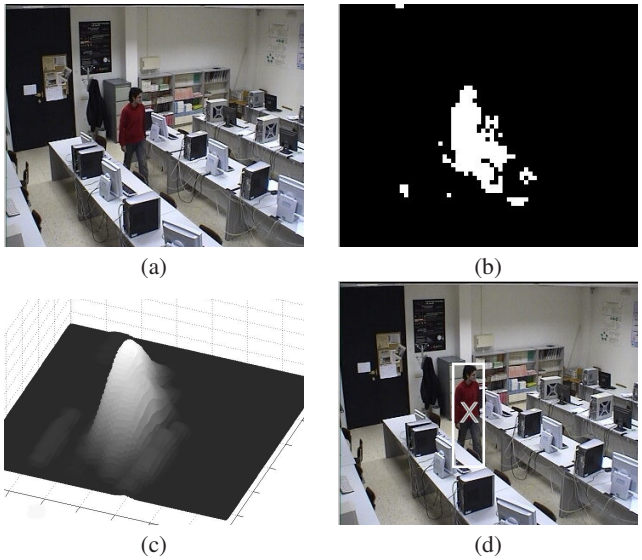


Fig. 3. (a) Frame showing a single person. (b) Moving region segmentation corresponding to (a), affected by shadows. (c) The combined matching error image P^t. (d) Person detection where the person has been bounded by a rectangle and his centroid has been marked according to the human silhouette dimensions. Note that shadows have been satisfactory removed.

4. 2D TRACKING

People tracking is carried out by means of a graph matching strategy [5], that estimates the best correspondence between people segmentations in consecutive frames based on the corresponding circumscribed rectangles. In absence of occlusions, the people matching is straightforward, giving correct correspondences. As explained in Sec. 5 the 3D tracking uses the head coordinates instead of the correspondence coordinates (which represent the centroid of the human template). This is accomplished by vertically translating the correspondence coordinates according to the dimensions of related human template. Head coordinates along with the dimensions of the corresponding template are smoothed using a Kalman filter to minimize the impact of low accuracy people detections in situations where static foreground objects occlude a significant part of a person.

In the presence of occlusions, the graph matching strategy may produce multiple correspondence for each occluded person. This situation is addressed by the 3D tracker, which solves the uncertainty by means of the tracking information of multiple cameras.

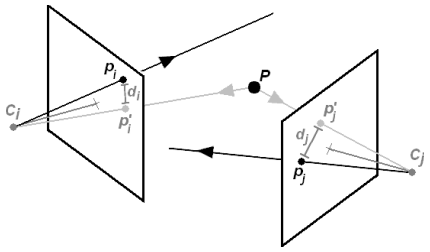


Fig. 4. Pixel error measurement of a correspondence between two points (final mean error would be $(d_i + d_j)/2$, as in this case only 2 cameras are involved).

5. MULTI-CAMERA GEOMETRY-BASED 3D TRACKER

Once detection and tracking of people have been performed for each of the system cameras separately, valuable 3D information can be inferred through fusion. Particularly, 3D position, height and width can be robustly estimated.

The proposed 3D tracking module takes, as input, 2D objects tracked separately in each of the cameras composing the system. The 3D tracker is able to infer which 2D segmentations correspond to different views of a real 3D object, and also to decide whether a 2D segmentation must be seen as an error, and thus discarded. Once correspondences between objects from different views are established, position of the 3D objects can be estimated by triangulation. Correspondences between 2D objects can be performed using appearance modeling of objects [6], or geometrical consistency between 2D objects [7]. The proposed system follows the geometrical consistency approach, allowing greater robustness against differences in acquisition conditions in different cameras (due to possible diverse sensor responses, direct or varying illumination, etc.).

Although the centroid is the preferred feature for performing geometrical consistency calculations [7], it depends directly on the shape and size of the whole 2D blob. This results, for a 3D object, in significant variations in 2D centroid positions due to static foreground occlusions of the lower part of objects (mainly because of furniture), and thus 3D object centroid projection and 2D blob centroids may differ dramatically. Using the topmost central point of 2D blobs (2D head positions) increases the invariance to usual static foreground occlusions. Estimation of the 3D head position is performed by applying the linear triangulation method [8] to the 2D head coordinates in the different cameras. At this point, 3D object position can be inferred projecting it onto the (known) ground plane.

Correspondences between tracked 2D objects from different cameras are established through exhaustive search across the whole set of posible combinations, including those in which the 3D object is not being seen by some of the cameras. Each possible combination gives an error measurement E_m expressing its geometrical coherence, and combinations with E_m above a certain pixel threshold p_{Th} are immediately discarded. Computation of E_m is depicted in Fig. 4, and follows the expression

$$E_m = \frac{1}{N_c} \sum_{\forall c} d_{L2}\Big(\vec{h}_c^{2D}, P_c\big(\vec{h}^{3D}\big)\Big) \tag{4}$$

where N_c represents the number of cameras supporting the 3D object, $d_{L2}(\cdot)$ is the Euclidean distance, $\vec{h}_c^{2D}$ is the head position of the supporting blob in camera c, and $P_c(\vec{h}^{3D})$ is the projection of the 3D reconstructed head position $\vec{h}^{3D}$ onto camera c image plane. This mean pixel error measurement conveys a clear geometrical meaning, as opposed to the utilization of the residual r of the equation system of the linear triangulation method used in [7]. An additional corrected error measurement E_c is also needed to promote those correspondences involving a higher number of cameras, as E_m tends to penalize these desirable sets of 2D objects. E_c has been calculated using Eq. (5). E_c has proved a good trade-off between different numbers of cameras (see Sec. 6).

$$E_c = E_m / 10^{N_c} \tag{5}$$

The 3D tracking module deals essentially with 3D objects. A particular combination of 2D object correspondences between cameras is promoted to "3D object" whether has appeared in the last N_{app} frames, and has simultaneously showed the lesser corrected error E_c among all combinations. Results analysis shows that $N_{app} = 5$ is a good trade-off between 3D object detection latency and false alarm rate (see Sec. 6). New 3D objects are looked for amongst those 2D objects that do not support any existing 3D object. The system models each 3D object as a cylinder characterized by $\vec{x}_i^{gp}$ (center of the cylinder base on the ground plane), h_i (height) and R_i (radius). Both $\vec{x}_i^{gp}$ and h_i are immediately extracted from the linear triangulation method output. The radius R of the cylinder can be estimated from the 3D head position $\vec{h}_{3D}$ and the 2D corresponding objects using triangle similarity, through the expression

$$R = \frac{1}{N_c} \sum_{\forall c} r_c = \frac{1}{N_c} \sum_{\forall c} \left[\frac{1}{2} \frac{w_c}{\beta_c} \big(\cos^2 \alpha_c\big)\ d_{L2}\big(\vec{C_c}, \vec{h^{3D}}\big) \right], \tag{6}$$

where

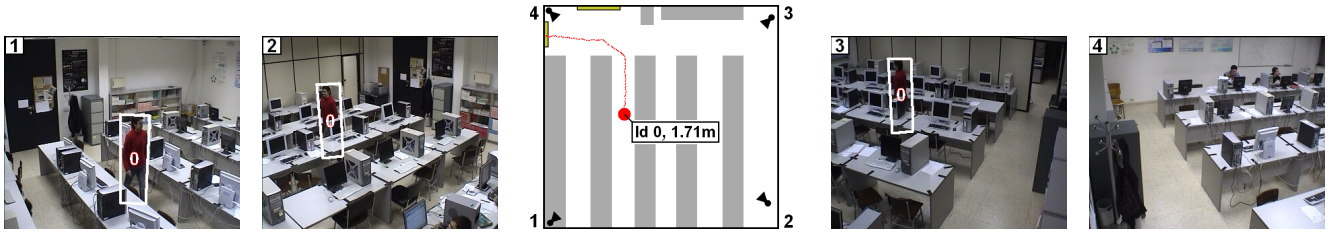
$$\alpha_c = \arctan(d_c), \quad \text{with } d_c = \frac{1}{\beta_c} d_{L2}\big(\vec{h}_c^{2D}, \vec{p}_c\big), \tag{7}$$

and where w_c is the width of the bounding box of the corresponding 2D object, β_c represents the focal length of the camera in terms of pixel dimensions, $\vec{C_c}$ is the camera optical center, and $\vec{p}_c$ is the principal point position (in pixels) of camera c. Estimated cylinder parameters are smoothed by a Kalman filter to ensure consistency when the number of supporting 2D views changes.

In addition, every 3D object stores its supporting 2D object identifiers in the last time step. In each time step all existing 3D objects are updated, estimating the best 2D object combination among those that differ, at most, in one of the objects regarding to the previous correspondence. A 3D object is discarded whether, during N_{dis} consecutive frames, E_m is above a certain pixel threshold p_{Th}, or no set of two or more 2D objects can support it. The analysis of the results shows that $N_{dis} = N_{app}$ is a reasonable selection.

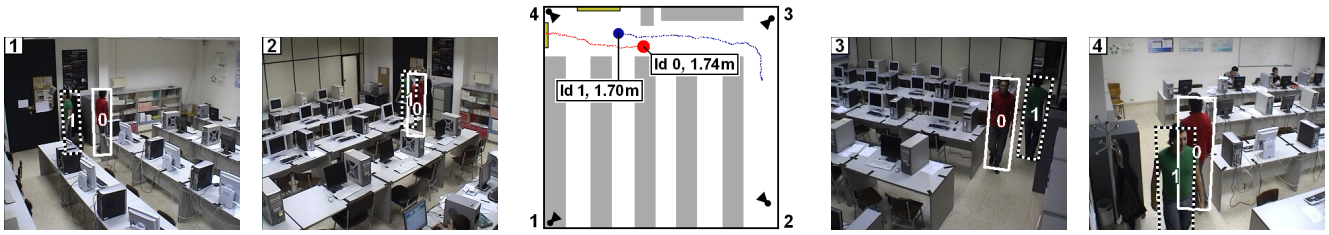
6. RESULTS

The proposed multi-camera 3D tracking system has been evaluated in a rectangular test-room (8×8.5 meters) with two entrance doors reproducing a typical office environment (see Fig. 5 and 6). Four identical and overlapping field-of-view cameras have been placed in the four topmost corners of the room, although the implemented system could handle an arbitrary number of different cameras (greater than 2). All cameras have been manually calibrated, and referenced with respect to a 3D coordinate system $(\vec{x}, \vec{y}, \vec{z})$ in which, for convenience, ground-plane equation is $z = 0$. Processed video streams have a frame rate of 25 fps, and a resolution of 352×288 pixels. Radial distortion has been previously compensated, as it prevents correct 3D reconstructions and geometrical correspondences. A simple,



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

Fig. 5. Single person undergoing static foreground occlusion. (a)(b)(d)(e) Four different camera views of the test room, with modeling cylinder projection superimposed. (c) Bird's-eye view of the test room containing 3D position and trajectory of the person.



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

Fig. 6. Two people with crossing trajectories, undergoing severe people-to-people occlusion. (a)(b)(d)(e) Four different camera views of the test room, with modeling cylinder projections superimposed. (c) Bird's-eye view of the test room containing 3D positions and trajectories.

bird's-eye view model of the test room has been generated to show the object evolution (Fig. 5 and 6).

Different situations have been evaluated in the test environment, where multiple people undergo both inter-people and static foreground occlusions, and enter and leave the room. System parameters (listed across this paper) have been tuned according to an exhaustive analysis of 2D and 3D tracking results. Some of them are dependent on the working conditions (resolution and frame rate of the video streams, extent of the monitored area, camera position, etc.).

Figure 5 shows a single man walking between two office desk rows. The moving target is within the field of view of three of the four cameras composing the system, undergoing severe static foreground occlusions. However, the system is able to accurately locate him in the room and even to estimate its height with great precision (Fig. 5 (c)). Figures 5 (a)(b)(d)(e) show the projection of the modeling cylinder onto the camera planes, demonstrating the accuracy of the 3D positioning.

Figure 6 shows a more complex situation where two people follow crossing trajectories, undergoing severe people-to-people occlusion in two of the four system cameras (Fig. 6 (b)(e)). The proposed system is able to correctly track both people from non-occluded views (Fig. 6 (a)(d)), as demonstrated in Fig. 6 (c).

7. CONCLUSIONS

The presented system positions and tracks multiple people in 3D in complex scenarios, addressing not only inter-people occlusion but also static foreground occlusion. A novel multidimensional background substraction technique along with a human template correlation process allow to accurately detect people even when they are significatively occluded by static foreground objects. However, the accuracy of people locations may be decreased. On the other hand, situations of strong occlusion between people can yield incorrect 2D tracking information. These problems are overcome by means of a 3D geometrical approach with multiple cameras. This approach uses the field of view of each camera to solve detection and 2D tracking uncertainties by means of the efficient selection of the 3D people location among the multiple 3D possible locations related to each person. This allows to achieve an accurate 3D tracking and positioning of people. Excellent results have been obtained in high-cluttered scenarios with multiple people.

8. REFERENCES

[1] M. Isard and J. MacCormick, "Bramble: A bayesian multiple-blob tracker," in *Proceedings of the 8th IEEE Int. Conf. on Computer Vision (ICCV)*, 2001, pp. 34–41.

[2] S. M. Khan and M. Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *Proceedings of the 9th IEEE Euroean. Conf. on Computer Vision (ECCV)*, 2006, pp. IV: 133–146.

[3] Tianding Chen, "Biologically-inspired model for multi-order coloring texture boundary detection," in *Proc. ICIA*. IEEE, 2006, pp. 183–188.

[4] P. Rosin, "Edges: Saliency measures and automatic thresholding," *Machine Vision and Applications*, vol. 9(4), pp. 139–159, 1999.

[5] C. Gomila and F. Meyer, "Graph-based object tracking," in *Proc. ICIP*. IEEE, 2003, vol. II, pp. 41–44.

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

[7] D. Focken and R. Stiefelhagen, "Towards vision-based 3-d people tracking in a smart room," in *Proceedings of the 4th IEEE Int. Conf. on Multimodal Interfaces (ICMI)*, 2002.

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