Vanishing Point-based Line Sampling for Real-time People Localization

Kuo-Hua Lo and Jen-Hui Chuang

***Abstract*—In this paper, we propose a real-time multicamera people localization method based on line sampling of image foregrounds. For each view, these line samples are originated from the vanishing point of lines perpendicular to the ground plane. With these line samples, vertical line samples in the 3-D scene can be reconstructed for potential human locations. After some efficient geometric refinement and filtering procedures, the remaining qualified 3-D line samples are clustered and integrated for the identification of locations and heights of people in the scene. Both indoor and outdoor scenarios are examined to demonstrate the effectiveness of our approach in handling serious occlusion in crowed scenes. The average localization error of less than 15 cm for average viewing distance of 15 m suggests that our method can be applied to a broad range of surveillance applications that require the real-time computation of localization without using special hardware for acceleration.**

***Index Terms*—2-D/3-D line sampling, multicamera, people localization, real time, vanishing point.**

I. Introduction

In RECENT years, visual surveillance using multiple cameras has attracted much attention in the computer vision community. Moreover, vision-based localization and tracking have shifted from monocular approaches to multicamera approaches since the latter can often achieve better results. Especially when there are many people in the scene, serious occlusions may occur in multiple views, and real-time people tracking and localization become a challenging problem. Thus, the previous works on visual surveillance are reviewed in the following two categories: monocular approaches and multicamera approaches.

A. Monocular Approaches

In [1] and [2], location and intensity of image foreground are extracted to allow construction of a human model, which allows us to match a subject image for tracking in successive grayscale images. In [3], color information is used to construct human models, wherein a person is modeled by several parts of similar color, and a Bayesian framework is employed to handle occlusion in the tracking process. In [4], an extension of the particle filter using object contour is proposed to track the head of a person. In [5], a color-based tracking that integrates color distributions into particle filter is presented to describe people using ellipses and associated color histograms. The method is robust when dealing with partial occlusion, and is rotation and scale invariant. In [6], color, shape, and edge are integrated into the particle filter to create a robust tracking method. Additionally, it proposes adaptive scheme to choose the most effective cues in different situations. However, the performance of these methods might be seriously impaired when the human model of occluded persons is not updated in time that the appearance of a person may change significantly. To resolve such a problem, spatial/temporal features are used in [7] to train convolutional neural networks to achieve robust people tracking, wherein the appearances of a target object of different views are adopted in the training stage.

Since single-view tracking depends on inherently limited information from a single viewing angle, dealing with situations involving serious or full occlusions is quite difficult. Thus, many multiview tracking approaches have been proposed. Unlike single view, multiple views can provide more visual information to cope with occlusions in human localization. For example, a stereo camera with a small baseline can estimate depth information easily, whereas a set of wide-baseline cameras can decrease invisible regions. Finding feature correspondence is usually the most important step for many multicamera approaches since only the correct correspondences between multiple cameras can ensure the correctness of subsequent processes, e.g., localization and tracking.

B. Multicamera Approaches

There are several types of multiple camera approaches for tracking people. The first type of approach uses a stereo camera to obtain depth maps for tracking. The second type of approach can be divided further into two subcategories, region-based and point-based methods, both of which have to establish correspondence between different views for tacking. The third type of approach seeks to find locations of persons directly without the correspondences of people in different views.

For the first type of approach, such as in [8]–[10], a stereo camera is exploited to establish correspondence between two views to construct a depth map. By using such a map to avoid influences of moving shadows on foreground detection,

Manuscript received January 13, 2012; revised May 29, 2012; accepted October 15, 2012. Date of publication January 24, 2013; date of current version June 27, 2013. This work was supported in part by the NSC under Grant 100-2221-E-009-116-MY2 and Grant 101-2220-E-009-054, and the National Chiao Tung University and Ministry of Education, Taiwan, under the "Aim for the Top University Plan." This paper was recommended by Associate Editor H. Wang.

The authors are with the Department of Computer Science, National Chiao Tung University, Hsinchu 30065, Taiwan (e-mail: lokh@cs.nctu.edu.tw; jchuang@cs.nctu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2013.2242592

better segmentation results can be obtained and object tracking becomes more robust. However, using a pair of cameras with a small baseline may suffer from total occlusions frequently. Without information of occluded regions (e.g., behind of a person closer to a stereo camera), the tracking performance is impaired.

Region-based methods of the second type generally regard people as regions and use region features to match people in multiple views. Most of these methods use color as the main feature to find correspondences of regions in different views. For instance, color and 3-D position are utilized to match and track multiple objects by a tracking algorithm in [11]. Mittal and Davis [12] use Gaussian color models to segment foreground regions of people from each image. The results are then used to match regions from one view to another along epipolar lines to find correspondence across multiple views. After that, Kalman filters are used to track people on the ground plane. Chang and Gong [13] use Bayesian networks for object tracking in individual views independently. After that, both geometry-based (epipolar geometry, homographies, and landmarks) and recognition-based (height and color of target appearance) methods, are utilized to find correspondence across multiple views. However, one of the main disadvantages of these methods is that color information may degrade the performance of tracking since the appearance and color can change with scene illumination.

Point-based methods can be further divided into two additional subcategories: 3-D-based and 2-D-based methods. 3-D-based methods locate and find correspondence of target object in images based on 3-D geometric constraints. These 3-D-based methods often need a complete camera calibration. In [14], the location of a person is described by a Gaussian distribution of its center of gravity (COG) in the scene. The distribution, which denotes the probability of the existence of a COG point, is projected onto multiple views, respectively, and the correspondence of feature point can be found by maximizing the probability of the COG distribution in each view. In [15], people are modeled as vertical cylinders and tracked by optical flow. During the tracking process, the COG of human body in multiple views is used to estimate the people locations in the world coordinate. In [16], cameras are calibrated for the calculation of 3-D positions of feet points of target people, and the correspondences can be established from these feet points. In [17], feature points are extracted from a (vertical) major line of the upper part of a human body. The correspondence of the human body is found by matching intensity and location through epipolar constraints. However, the extracted feature points from each view may not always correspond to the same point in the 3-D space. In that case, the matching performance, the established correspondence, and tracking results may be impaired.

Unlike the above 3-D-based methods, some 2-D-based methods have been presented to establish correspondences between multiple cameras by matching locations of feature points on a reference plane. In [18]–[20], homography constraint is used to match the locations of feet points in different views. However, these feature points may be occluded between objects. In [20], a method, which can detect whether the feet points of a person are occluded, is proposed to select a best view for each person appearing in the scene. In contrast, Hu *et al.* [21] propose a method using the axes of people to estimate the feet points in images. They segment a group of people into individual persons and estimate an axis for each of them. Then, the location of the feet point of a person is estimated as intersection point of his/her axis and the bottom of his/her bounding box. In [22], foregrounds of a person are perspectively projected from each view to the ground plane, with the corresponding camera being the projection center. For each camera, a line passing through 1) the projected foreground and 2) the vertically projected camera center, both on the ground plane, is estimated. The person's location can then be estimated by calculating the intersection of these estimated lines on the ground plane based on the least-square criterion. For most of the aforementioned point-based approaches, accurate detection/estimation of point/line features, and their correspondences in different views, are required; otherwise the correctness of a person's location will be seriously impaired.

In recent years, approaches of the third type are proposed. These methods, which do not need a complete camera calibration, can locate people directly without finding the correspondences of the people between views. Eshel and Moses [23], [24] propose a method using cameras placed at high elevation to detect the heads of people. The method assumes the cameras are partially calibrated for homographic matrices for multiple planes with different heights. For each plane, intensity information of segmented foreground pixels is collected from all views, and head detection is achieved through intensity correlation. In [25], the authors propose an interesting method to track people by locating them on similar reference planes. The foreground likelihood information of all image pixels captured from different views is projected and integrated on each reference plane to form an occupancy probability. Such probabilities from several frames are then processed by a graph-cut algorithm to find trajectories of people. Although the correspondences of people between different views are not available,[1] such an approach performs quite well under serious occlusions in a crowed scene. Due to the high complexity of pixel-based processing, the approach is implemented with CUDA (Nvidia GeForce 7300 GPU) to achieve real-time performance.

In order to enhance the efficiency of the above approach, we propose a vanishing point-based line sampling technique in [26]. While the main idea of the approach presented in [25] is to project dense 2-D samples (image pixels) onto multiple (horizontal) planar surfaces in the 3-D space (before these data are fused into 3-D object distributions), it is simplified in [26] by projecting 1-D image samples,[2] i.e., lines passing through the vanishing point of vertical lines in the 3-D space, instead (before their intersections are grouped into 3-D line samples of the crowd through clustering). To further improve the efficiency of people localization, a novel approach is proposed

[1]For example, no additional image processing procedures are performed to identify each individual from a crowd, e.g., through connected component analysis and principal axis analysis as adopted in [21].

[2]In the rest of this paper, we will refer to these samples as 2-D line samples.

in this paper which projects the above line samples directly into the 3-D space, i.e., along a fan of vertical planes originated from the vertical axis containing the camera center, to generate possible 1-D (vertical line) samples of the 3-D object.[3] Since realistic constraints of a human body can be adopted to refine and to verify these object samples, localization results compatible with those in [25] can be achieved, but with more than an order of magnitude in processing speed.

The main improvements of this paper over [26] include: 1) new reconstruction and refinement procedures for possible 3-D (vertical) line samples of the human body, 2) addition of two new geometric rules (associated with the head level of a person) for the screening of these samples, and 3) a simple way of splitting a cluster of samples belonging to two persons very close to each other. While 1) reconstructs a sample directly (and efficiently) from a pair of foreground line samples from two views, as the intersection of two vertical triangles, the reconstruction is done in [26] in a much more complicated way as mentioned above. As for 2) and 3), both of them offer valuable improvements in the localization performance, in terms of precision and recall, with 2) also saving some computation time spent for invalid samples.

The rest of this paper is organized as follows. In Section II, a preliminary major axis-based method for locating nonoccluded persons is presented to convey the basic idea of the proposed approach. In Section III, a novel way of generating 3-D line samples for people under occlusion, via the vanishing point-based line sampling of image foreground, is proposed. Realistic constraints are also established to refine and verify these 3-D line samples, before they are clustered into 3-D major axes to represent individual persons in the scene. In Section IV, both indoor and outdoor video sequences are tested to show the efficiency and effectiveness of the proposed approach. Experimental results show that such an approach performs satisfactorily in terms of correctness and accuracy in people localization under serious occlusion. In Section V, some concluding remarks are given.

II. Construction of Major Axes for Nonoccluded Persons from a Pair of Views

For a better understanding of the basic ideas of the proposed localization, we begin by illustrating how to localize people using the major axes (MA) of the foreground regions in 2-D images. Assume the foreground of different persons do not overlap in a pair of views in which the major axis of each of them can be estimated correctly. By projecting these axes, instead of projecting all foreground pixels, as in [25], onto multiple reference planes parallel to the ground plane, a 3-D axis can be formed for each person by connecting corresponding intersection points of the projected 2-D axes on these reference planes. Furthermore, a more efficient scheme is introduced to find the above 3-D axis by calculating the intersection line segment of two triangles in the 3-D space if the cameras centers can be estimated in advance.

[3]In the rest of this paper, we will refer to these samples as 3-D line samples.

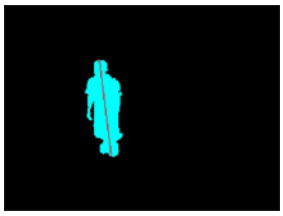


Fig. 1. Detected foreground regions and the estimated axis.

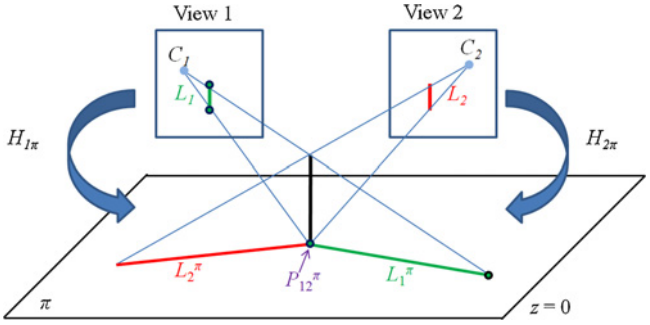


Fig. 2. Finding intersection points of two axes on a reference plane.

A. Major Axis Estimation for a Person in an Image

In order to segment foreground regions of a person from an image, the Gaussian mixture model (GMM) [27], [28] can be applied. Assume region R obtained from foreground segmentation contains a great percentage of a person, we can estimate the major axis for the person by PCA. An example of an axis thus estimated is shown in Fig. 1. One can see that the estimated major axis can represent the elongated shape of a person very well.

B. Finding a 3-D Major Axis of a Person—Two Approaches

As shown in Fig. 2, let L_1 and L_2 be the axes of a person obtained by PCA for View 1 and View 2, respectively. In addition, let P_{12}^{π} be the intersection point of the two lines containing the projections of L_1 and L_2, respectively, onto reference (ground) plane π from camera centers C_1 and C_2. Ideally, for reference planes of different heights, such intersection points will either 1) belong to both the projected axes, or 2) stay away from any of them if the corresponding heights are out of the range of the 3-D axis. Fig. 3 shows samples of the 3-D axis thus obtained for the person shown in Fig. 1. While intersection points satisfying 1) is colored in black, points not satisfying 1), including those contained in one but not both projected axes due to computation errors, are marked in red.[4]

The above results provide us an important cue to the estimation of a person's height. Additionally, one can see that the 2-D (horizontal) positions of these 3-D points are quite consistent that a roughly vertical major axis (MA) of the

[4]To find the above intersection points on reference planes of different heights, a method to produce multiple homographic matrices is introduced which can establish these matrices using only two marker points on each of the four calibrating pillars standing vertically on the ground plane. The detail can be found in Appendix A.

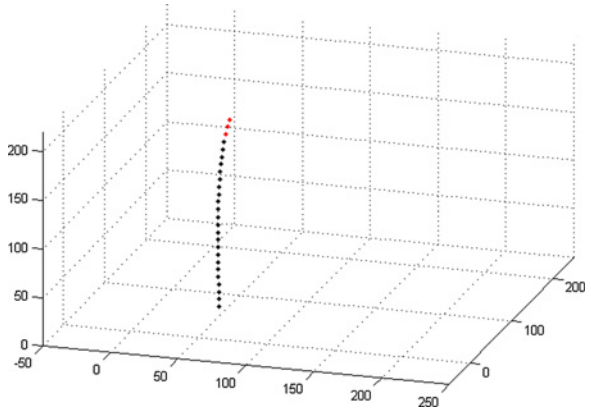


Fig. 3. Axis samples of the person shown in Fig. 1, which are reconstructed for reference (horizontal) planes with 4 cm spacing and up to 176 cm in height.

person can be constructed by connecting the black points, i.e.

$$\text{Axis\_}set_{1,2}{}^{h_b h_t} = \{P_{1,2}{}^{h_b}, ..., P_{1,2}{}^{h_t}\} \quad (1)$$

with h_b and b_t being the heights of bottom and top end points of the axis, respectively.

On the other hand, if the foreground object is a vertical axis standing on the ground plane, as shown in Fig. 2, a perfect reconstruction of the 3-D axis can be obtained by intersecting the two triangles formed by C_i and L_i^{π}, for $i = 1$ and 2, respectively.[5] By adopting this method, calculating a large number of intersection points, as shown in Fig. 3, for multiple reference planes is no longer needed and the computational time can be saved greatly. Axis points similar to that listed in (1) can then be estimated by simple interpolation along the 3-D axis if necessary.

C. Extension of Finding 3-D Major Axes for Nonoccluded Multiple Persons from a Pair of Views

The above method can be extended to estimate 3-D MAs for multiple people if an axis can be found for each of them in two different views. Without knowing the correspondence of the axes in the two views, candidate 3-D MAs can be constructed for all possible 2-D MA pairs. For example, for M persons in View 1 and N persons in View 2, a total of MN candidate MAs can be constructed (minus those associated with triangle pairs that do not intersect, like the two blue triangles shown in Fig. 4).

For a candidate 3-D MA obtained for person i in View 1 and person j in View 2, (1) can be rewritten as

$$Axis_{1i,2j}{}^{h_b h_t} = \{P_{1i,2j}{}^{h_b}, P_{1i,2j}{}^{h_t}\}. \quad (2)$$

Although we do not have correspondences of different people in these two views, it is possible to remove incorrect 3-D MAs by checking the consistency in the foreground coverage, as will be explained in Section III-B, with additional views. For example, while the two green axes in Fig. 4 are correct

[5]The camera centers can be found in advance by at least two of the aforementioned four pillars.

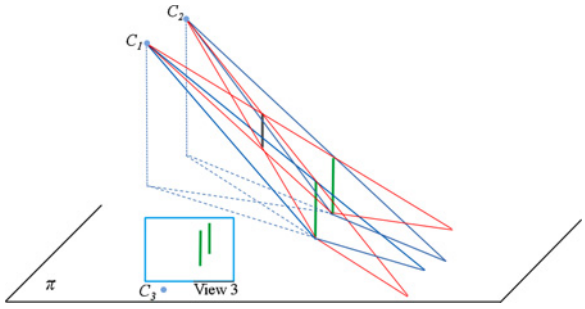


Fig. 4. Illustration of filtering out incorrect 3-D MAs by using an extra view.

Fig. 5. Example of overlap foreground and the estimated axis.

3-D MAs, the gray axis can be identified as an invalid axis from View 3.[6]

III. Construction of Major Axes for Multiple Persons with Occlusion

The above 2-D PCA-based axis estimation can only cope with situations under which the foreground of a person is separable from others in all views, and can be identified as one region by connected component analysis. However, in real applications, many people may appear in a monitored scene at the same time that each segmented foreground area may contain more than one person, as shown in Fig. 5, and the aforementioned axes detection approach will not work correctly. One possible solution proposed in [21] is to separate persons by projecting the foreground in the vertical direction to form a histogram, and then, determine the boundaries between persons based on the location of peaks and valleys in the histogram, before each person can be represented by one axis for localization and tracking. However, the above approach may not work well when there is a very dense group of people appearing in the scene, e.g., Fig. 6. For more complicated situations, instead of estimating a 2-D axis for each person, a 3-D sampling scheme is proposed in this section, wherein 2-D line samples of the foreground regions from multiple views are used to generate 3-D line samples of the foreground volume, based on the same idea described in Section II. Then, with noises filtered out, these 3-D line samples are refined and

[6]In general, incorrect MAs constructed from a pair of triangles can be removed by checking the consistency with an additional view point (in the 3-D space) except for those view points that are coplanar (in a 2-D subspace) with one of the two triangles mentioned above. Therefore, with the help of an additional camera, incorrect MAs will be removed completely, with zero probability for the above exceptions.

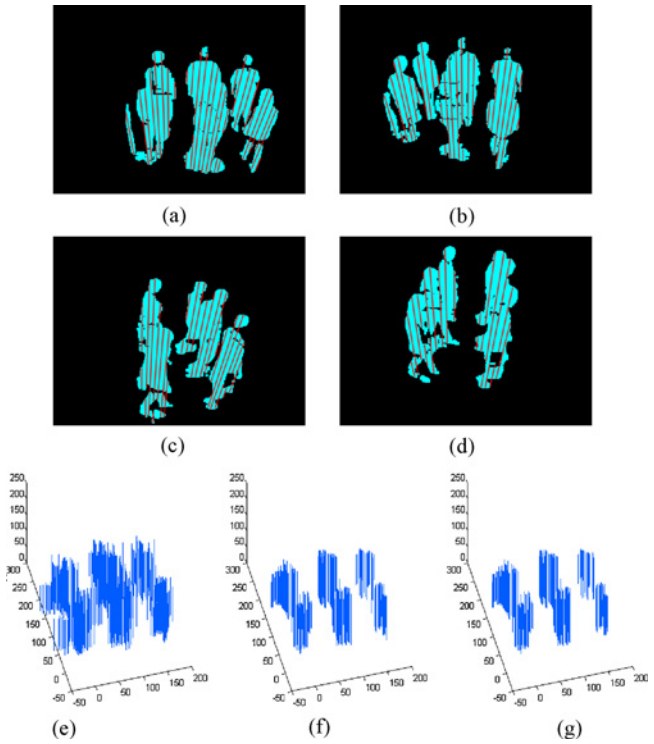


Fig. 6. (a)–(d) 2-D line samples in Views 1–4. (e) Unverified 3-D line samples that survive Rules 1–3. (f) Refined line samples that survive Rules 1–4. (g) Final line samples.

verified with respect to different views by a back projection procedure. Finally, a clustering algorithm is applied to the remaining samples in the scene, before members of each cluster are integrated into a 3-D MA.

A. Generating 3-D Line Samples Using Vanishing Points

Since the upper body of a person is almost always perpendicular to the ground plane when he/she is standing and walking in a monitored scene, we first generate 2-D line samples in each view, which are originated from the vanishing point of vertical lines in the 3-D scene [see Figs. 6(a)–(d)].[7] Thus, these 2-D line samples correspond to a fan of vertical sampling slices in the 3-D space originated from the vertical line containing the corresponding camera center. Note that generating 2-D line samples is much faster than the axis estimation discussed in Section II since no additional image processing is required. Very short 2-D sampling lines (less than a threshold T_p) will be discarded since they are expected to be far away from a major axis and will have little contribution to the estimation of a 3-D MA.

Next, for each pair of views, the remaining 2-D line samples are used to reconstruct 3-D line samples by the scheme described in Section II. Since there may still be incorrect 3-D line samples, such as the gray one shown in Fig. 4, three geometric rules can be used to filter out the 3-D line samples that will not correctly represent a person in the 3-D scene.

[7]The vanishing point in each view can be estimated by calculating the intersection points of the four lines extended from the four upright pillars mentioned in Section II-B.

1) The length of a 3-D line sample is shorter than T_{len}.
2) The height of its top P^{h_t} is lower than T_{tl}.
3) The height of its bottom P^{h_b} is higher than T_b.

Fig. 6(e) shows 3-D line samples that pass these rules, each adjusted slightly so that it is perpendicular to ground plane.

The main objective of the above three rules is to preserve two kinds of 3-D line samples that correspond to 1) the full length of a standing/walking person or 2) the head and torso of a person without his/her feet. By selecting appropriate thresholds, these three rules may also accommodate human activities such as jumping and squatting. In practice, these three rules can efficiently remove most inappropriate 3-D line samples, e.g., 84% of the originally reconstructed 3-D line samples for the above example. However, since each 3-D line sample is reconstructed by observations from two views only, the top and bottom ends of each 3-D line sample may not be very accurate in position. To deal with such a problem, a refinement procedure using information from additional views, as described next, is adopted to find more accurate positions of the two end points before further verification of the 3-D line samples are performed.

B. Refinement and Verification of Reconstructed 3-D Line Samples from Additional Views

In this section, a refinement and verification scheme is developed for further checking the validity of each 3-D line sample obtained in the previous section. The refinement is based on the fact that if a 3-D line sample corresponds to a real person in the scene, its image in all views should be covered by foreground regions. In other words, its top and bottom end points will be covered by some foreground regions in all views. If that is not the case, the 3-D line sample should be shortened until it falls within foreground regions in all views. Specifically, for each 3-D line sample, we can interpolate equally spaced sample points between P^{h_t} and P^{h_b} to form axis samples $\{P^{h_t},..., P^{h_b}\}$.[8] The refinement of a top end point corresponds to find the first sample point below P^{h_t} such that it is covered by some foreground regions in all views. Similarly, the refinement of the bottom end point can be done by searching in the upward direction from P^{h_b}.

After such a refinement (shrinking) procedure, Rules 1–3 can be applied again, as can the fourth rule, to filter out inappropriate 3-D line samples.

4) The height of top end point P^{h_t} is higher than T_{th}.

One can see from Fig. 6(f) that rough people locations can be distinguished visually from the remaining 3-D line samples. Finally, a threshold T_{fg} is used to filter out 3-D line samples that do not have sufficient average foreground coverage rate (AFCR), as shown in Fig. 6(g).[9]

[8]The interpolation spacing between two adjacent sample points corresponds to a total number of N_{plane} equally spaced reference planes between the ground plane and the plane with 250 cm in height.

[9]In our implementation, each sample point of a 3-D line sample is projected to all views to check if it is covered by foreground for the computation of AFCR. For example, AFCR for each of the green axes shown in Fig. 4 is equal to 100% with respect to all (three) views.

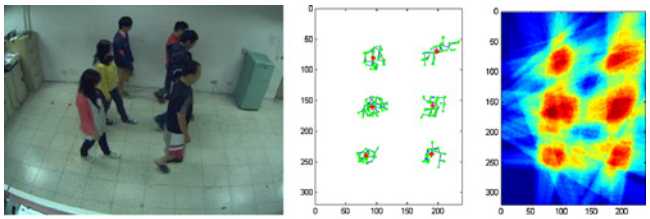


Fig. 7. Clustering and localization results. (a) Input frame 532. (b) Clustering sets. (c) Accumulated synergy map of all reference planes.

C. Integration of 3-D Line Samples to form 3-D Major Axes

After the above verification procedure, the major axis of a person can be estimated from the remaining 3-D line samples using the breadth-first search (BFS). Specifically, if the 2-D horizontal distance between two 3-D line samples is closer than a threshold T_c, an edge is established in an undirected graph. For example, Fig. 7(a) shows the input frame for Figs. 6(d) and 7(b) shows the top view of resultant graph obtained by the above BFS scheme, with green points representing the 3-D line samples. To avoid some false positives in the clustering, a cluster containing a total number of 3-D line samples less than threshold N_{line} will be removed.

To locate individual persons, the horizontal position of each of them can be estimated as the average, shown as red stars in Fig. 7(b), of the horizontal positions of the 3-D line samples in the corresponding cluster.[10] In Fig. 7(c), we show the synergy map obtained with a method modified from [25]. Instead of considering the foreground probability of all image pixels, only those inside of foreground regions are taken into account. One can see the above distribution of each cluster matches the corresponding occupied region (red color) in the map quite well, i.e., all red stars do fall inside of the occupied regions.

In the proposed localization scheme, each camera is treated separately only when 2-D line samples of foreground regions are generated in each image plane. The correspondences of people among different views, which are generally hard to determine for different views (especially with occlusion), are actually utilized implicitly in subsequent processes. First, vertical line samples of people are generated in the 3-D space for a pair of foreground line samples, with their correspondence yet to be determined, from two different views (cameras), as described in Section III-A. Second, the geometric constraints adopted in Sections III-A and III-B are necessary conditions for the correspondence of a person (in an upright posture) perceived from two views. Third, the refinement (shortening) and the AFCR check are developed in Section III-B to implicitly verify the correspondence for all views (image planes). Finally, the integration (clustering) is performed in Section III-C to spatially verify and establish the people correspondence in the 3-D space.

IV. Experimental Results

In this section, the proposed method is evaluated with several different videos taken from both indoor and outdoor scenes, with different degrees of occlusion. Comparisons with [25] and [26] are also included to show the proposed method can achieve comparable correctness/accuracy in localization but with much higher computation speed. Additionally, we investigate the performance of the proposed method with different numbers of cameras and densities of line samples in an image.

A. Experiments for Different Degrees of Occlusion with Indoor/Outdoor Sequences

To evaluate our methods under different degrees of occlusion, we captured several video sequences of indoor and outdoor scenes. For each scene, calibration pillars are placed vertically and then removed from the scene for the estimation of camera centers, vanishing points, and multiple homographic matrices (see Appendix A). These sequences are captured with different numbers and trajectories of people. The performance evaluation is implemented under Windows 7 with 4 GB RAM and a 2.4G Intel Core2 Duo CPU, without additional hardware.

Fig. 8 shows an instance of scenario S1 captured from four different viewing directions with a 360×240 image resolution. The average distance between the cameras and the monitored area is about 15 m. One can see that the lighting conditions are quite complicated. The sunlight may come through the windows directly and the reflections from the floor can be seen clearly. A total of 691 frames are captured for S1, wherein eight people are walking around; the ninth is standing near the center of the monitored area.

Fig. 9(a) and (b) shows 2-D line samples generated for Fig. 8(b) and the reconstructed 3-D MAs, viewing from a slightly higher elevation angle, respectively. In addition, for a closer examination of the correctness of the proposed people localization and height estimation scheme, bounding boxes with a fixed cross-section,[11] and with their height obtained from derived 3-D MAs, are back-projected to the captured images, as shown in Fig. 9(c) for the image shown in Fig. 8(b). One can see that these bounding boxes do overlay nicely with the corresponding individuals. The recall and precision rates for the whole sequence are evaluated as 96.5% and 95.6%, respectively.[12]

Fig. 10 shows similar localization results for scenario S2, which has the same people count as that for S1, but the nine people are walking randomly in the scene. While occlusion may become more serious in some instances, repeated occlusions caused by periodic walking pattern in S1 do not occur. As a result, both the average recall and precision rates are increased slightly. To further examine the robustness of our method under serious occlusion, scenario S3 is evaluated, which is similar to S2, except that it has 12 persons randomly

[10]The heights of the top and bottom ends of a 3-D major axis are assigned as the heights of the highest and lowest end points in the corresponding cluster, respectively.

[11]Since locations of ordinary persons are represented by these bounding boxes, a fixed 50×50 cm cross-section is adopted for each box. Thus, the width and length of these boxes will not be affected by the density of sampling, reducing possibly undesirable effects due to certain camera configurations, density of image line samples, and occlusion. For example, it is easy to see that the cross-section of the cluster of magenta color shown in Fig. 13(b) has an elongated shape, which does not represent the orientation of the corresponding human body.

[12]These values are generated by comparing with ground truth produced manually.

Fig. 8. Scenario S1, captured from four different viewing directions.

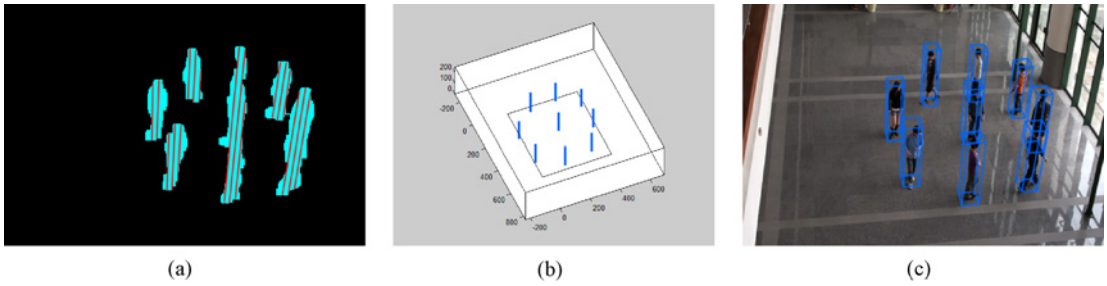


Fig. 9. Localization results for scenario S1. (a) Segmented foreground regions and 2-D line samples for Fig. 8(b). (b) 3-D major axes to represent different persons in the scene. (c) Localization results illustrated with bounding boxes.

TABLE I
LOCALIZATION RESULTS OF SEQUENCES S1−S3

Sequence	Number of frames/people	Method	Recall	Precision	Mean error (cm)	Frames per second
S1	691/9	This paper	**96.5**%	95.6%	**11.42**(5.89)	**33.41**(2.448)
		[26]	92.0%	**95.7**%	11.60(5.91)	11.62(1.008)
		[25]	93.8%	**95.7**%	11.78(6.12)	0.46(0.003)
S2	776/9	This paper	**96.8**%	97.0%	10.09(5.77)	**31.53**(3.089)
		[26]	94.9%	97.3%	**10.00**(5.66)	12.05(1.201)
		[25]	96.2%	**98.1**%	10.22(5.58)	0.46(0.003)
S3	271/12	This paper	**95.2**%	93.6%	10.55(6.01)	**21.61**(1.646)
		[26]	93.3%	**94.3**%	**10.28**(5.99)	8.34(1.025)
		[25]	93.3%	94.2%	10.93(5.87)	0.46(0.003)

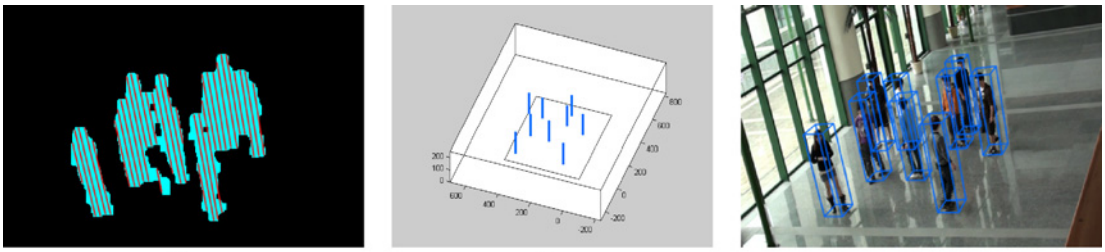
walking in the scene. Since the scene is becoming more crowded and serious occlusion may occur more frequently, foregrounds of different persons may easily merge into larger regions, as shown in Fig. 11(a). While satisfactory localization results are obtained in Fig. 11(b) and (c), the recall and precision rates for S3 are decreased to 95.2% and 93.6%, respectively.

Table I summarizes detailed localization results of the proposed method as well as two other methods. In addition to our previous work [26], a modified version[13] of the approach proposed in [25], is also implemented and tested. One can see the differences in recall and precision rates among the three methods are quite small (±2%) except for recall rates for S1. Specifically, the proposed approach achieves the highest recall rates for S1–S3 while the other two methods achieve the highest precision rates for two of the three video sequences. Similarly, very small difference (within 0.65 cm) among results obtained from these three methods can be found for the accuracy of derived people location. Overall, the mean value and standard deviation of (x–y) location errors of the proposed method for S1–S3, together, are equal to 10.70 cm and 5.90 cm, respectively, which can hopefully be regarded as sufficient for many surveillance applications.[14]

As for the computational speed, in frames per second (f/s), the values for different cases listed in Table I are evaluated without including the cost of foreground segmentation. One can see that speed-up of more than an order of magnitude from the method in [25] can be achieved by the proposed approach, with as much as 70 times acceleration (near 2.7 times in speed improvement from our previous approach in [26]) in the process speed of S1. While real-time performance can be achieved for S1 and S2, the computation speed is

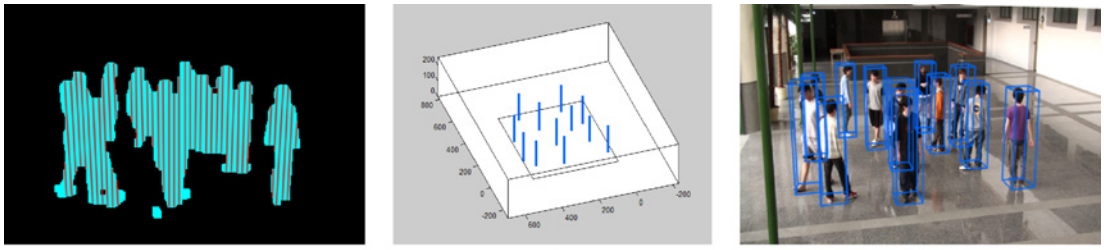
[13]In our implementation, which also does not perform people tracking, binary images of foregrounds are adopted as system input as the other two algorithms. A grid size of 100 × 100 is chosen for each of the 20 reference planes, with 10 cm grid spacing. A grid point on the ground is regarded as occupied if more than T_{acc} = 11 grid points with the same horizontal coordinates (but on reference planes of different heights) correspond to image foreground in all four views. Then, connected component analysis is applied to identify connecting occupancy regions. The connected occupancy regions with very small areas, i.e., smaller than 22% of average area of such regions, are regarded as noise and are removed.

[14]The errors are only calculated for correctly detected people locations, which contribute to the precision rates listed in Table I, i.e., with location errors less than 30 cm.



(a) (b) (c)

Fig. 10. Localization results, similar to those shown in Fig. 9, for scenario S2.



(a) (b) (c)

Fig. 11. Localization results, similar to those shown in Fig. 9, for scenario S3.

down to a near real time 21.61 f/s when the number of people increases to 12.[15] Note that for [25], the computation times (in f/s) are about the same for different cases. This is because the time complexity in the generation of synergy maps is mainly dependent on the size of image frame and the number of views.

Although the above evaluations show that the proposed method can often provide reasonably good localization results, there are extreme cases of poor foreground segmentation that cannot be well handled with the proposed method. Fig. 12(a)–(h) show localization results and foreground regions for the 51st frame of S1. In Figs. 12(a) and (e), one can see the foreground segmentation of a person (in red circle) is very poor because of reflections as well as a clustered background (see green arrows). Consequently, lesser 3-D line samples are retained after the screening process, as shown in Fig. 12(i), resulting in a failure. Since some more 3-D line samples can still be reconstructed correctly for that person as different time instances, erroneous results are generated for only three out of 20 frames (from frame 41 to frame 60), compared with 13 erroneous frames obtained from the method in [25].

On the other hand, problematic results may also be generated due to very serious occlusions. First, as shown in Fig. 13, there may be a ground region that is covered by foregrounds in all views. Whether a person exists or not, a 3-D MA will be generated. If such a 3-D MA cannot be filtered out by the aforementioned geometric rules, a false alarm will occur [see the yellow arrow in Fig. 13(c)].[16] Second, when the distances between people are too small [see the red arrow in Fig. 13(c)], their 3-D line samples will be clustered into the same group [see Fig. 13(b)], resulting in two missed detections and one false alarm. This is because, for localization efficiency, the BFS scheme only determines whether the distance between two line samples is smaller than a threshold when grouping 3-D line samples.[17] A more detailed discussion of the effect of the distance threshold can be found in Appendix C.

To further evaluate our method for an outdoor environment, S4 and S5 are captured from a real scenario with image resolution of 360 × 240. In general, working in such an environment may be challenging for visual surveillance systems since there are more time varying factors such as illumination for object, speed of wind, and shadows of various strength. For the real scene under consideration, groups of people of different sizes are walking quickly through the monitored area[18] (green polygons in Figs. 14 and 15). Thus, less image frames are captured for S4 and S5 than those in S1–S3. Figs. 14 and 15 show snapshots of localization results for S4 and S5, respectively, with more statistics summarized in Table II. One can see that the correctness/accuracy level similar to that shown in Table I can be achieved with the proposed approach, except for larger differences between: 1) recall and precision rates for S4, and 2) mean localization errors for S4 and S5. Such differences may result from a higher probability of the aforementioned occlusions for people walking together along a passage and/or complexities associated with an outdoor scene.

In practice, due to significant differences between the indoor and outdoor scenes where video sequences S1–S3 and S4–S5 are captured, respectively, different parameter values may need to be selected to achieve desirable localization results. In the

[15]This is because the computational time is dominated by the number of 2-D line samples, which will grow with the area of foregrounds.

[16]Such a problem may be eliminated by adopting additional temporal information, which is not discussed in this paper for brevity.

[17]To partially resolve this problem, a heuristic scheme is applied in our method. If a cluster contains a larger number of 3-D line samples, it will be divided into two clusters. Specifically, we calculate the average number of 3-D line samples, N_c, in all clusters, and divide a cluster into two groups if it contains more than more than $2N_c$ line samples.

[18]It is assumed that the evaluation of people localization is only preformed for the monitored area.

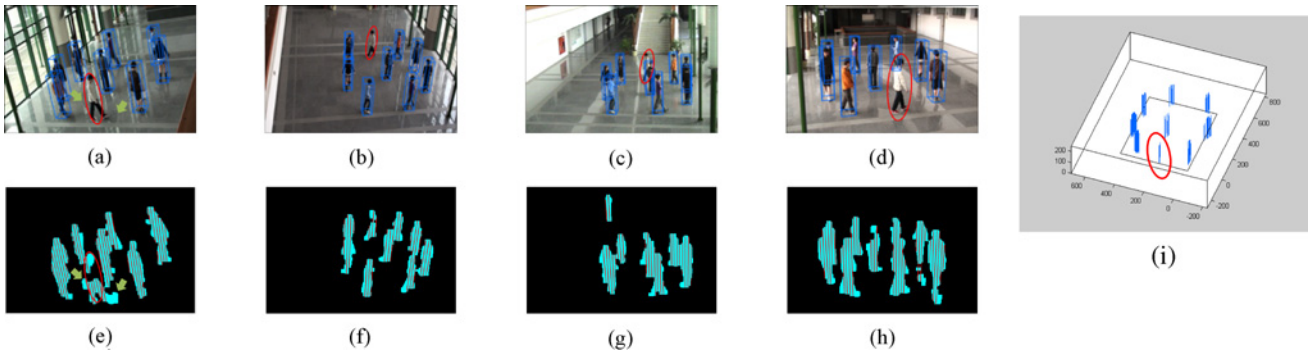


Fig. 12. Failure example of the proposed method. (a)–(d) Localization results (illustrated with bounding boxes) of four views. (e)–(h) Corresponding foreground regions and 2-D line samples. (i) 3-D line samples to represent different persons in the scene.

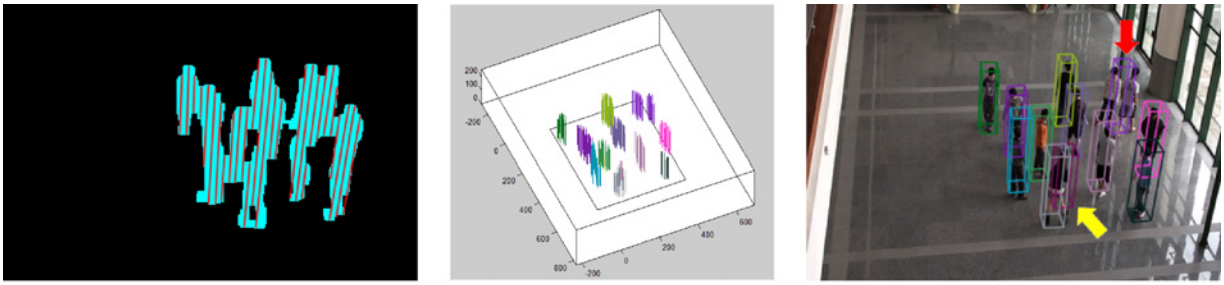


Fig. 13. Example of missed detections and false alarms of S3. (a) Segmented foreground regions and 2-D line samples. (b) 3-D line samples to represent different persons in the scene. (c) Localization results illustrated with bounding boxes. Note that corresponding colors are used in (b) and (c) for different clusters/bounding boxes after clustering.

TABLE II
LOCALIZATION RESULTS OF SEQUENCES S4 AND S5

Sequence	Number of frames/people	Method	Recall	Precision	Mean error (cm)	Frames per second
S4	70/6–7	This paper	97.5%	89.8%	**8.57**(5.05)	**28.61**(2.903)
		[26]	90.0%	75.4%	8.84(5.62)	9.20(1.153)
		[25]	**97.7**%	**91.1**%	9.08(5.30)	0.46(0.002)
S5	40/7	This paper	95.0%	96.0%	11.70(6.02)	**26.66**(1.684)
		[26]	**97.5**%	91.0%	**11.37**(6.52)	5.97(0.408)
		[25]	97.1%	**97.8**%	11.48(6.25)	0.46(0.001)

next subsection, effects of choosing different densities of 2-D line samples in each image, as well as incorporating different numbers of cameras in the proposed localization system, will be investigated (only for the indoor scene for brevity). While the two associated parameters will determine the initial amount of data to be processed by the proposed algorithm, other parameters will be used to tune the algorithm for better performance under different environmental conditions, as will be discussed in more detail in Appendix C.

TABLE III
RESULTS OF USING DIFFERENT NUMBERS OF CAMERAS

Number of cameras	3	4	5
Recall	95.4%	96.2%	98.3%
Precision	85.7%	95.0%	96.6%
Localization error (cm)	11.30	10.70	10.13
Frames per second	75.57	29.48	24.16

B. Experiments for Different Numbers of Cameras and Densities of Sampling

To investigate the relationship between performance of localization and the numbers of cameras, the indoor scenarios S1–S3 are examined with an additional view captured from a different camera, and the results are presented in Table III. One can see that while similar recall rates can be obtained by using different numbers of cameras, the precision rate of using three cameras is much lower than if four or five cameras are used. This implies that using only three cameras may not be sufficient when there are serious occlusions. In addition to the above performance indices, adding more cameras also improves the system performance in terms of the localization accuracy. However, if slight degradations in these performance indices are acceptable, a set of four cameras may be used if hardware (cameras) cost is of major concern.

In order to investigate the influence of densities of sample lines in an image on the localization performance, a very simple sampling scheme is adopted in our method. In par-

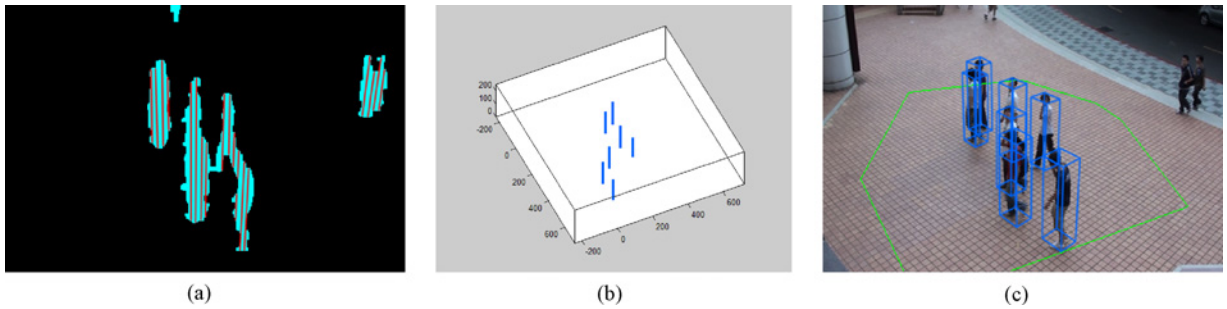


Fig. 14. Localization results, similar to those shown in Fig. 9, for scenario S4.

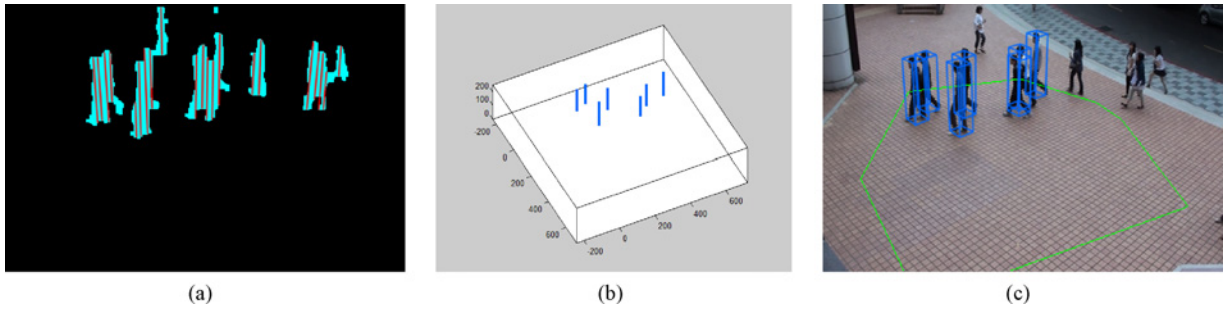


Fig. 15. Localization results, similar to those shown in Fig. 9, for scenario S5.

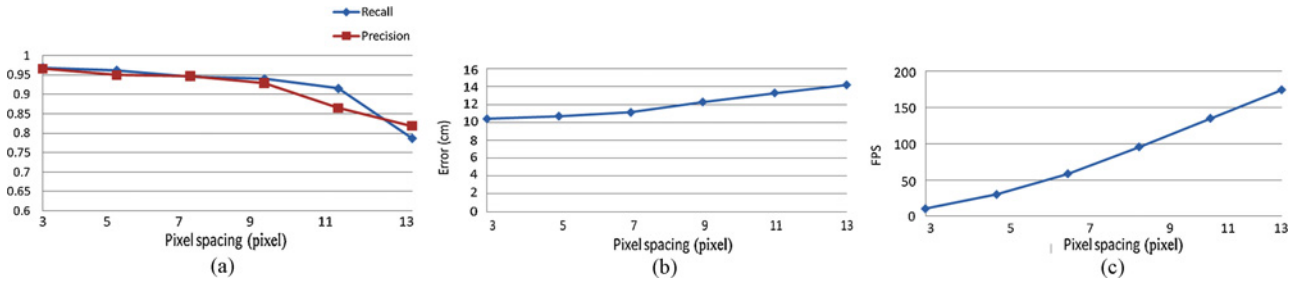


Fig. 16. Results of using different line densities (pixel spacings) with four cameras. (a) Recall and precision. (b) Localization error. (c) Computation speed.

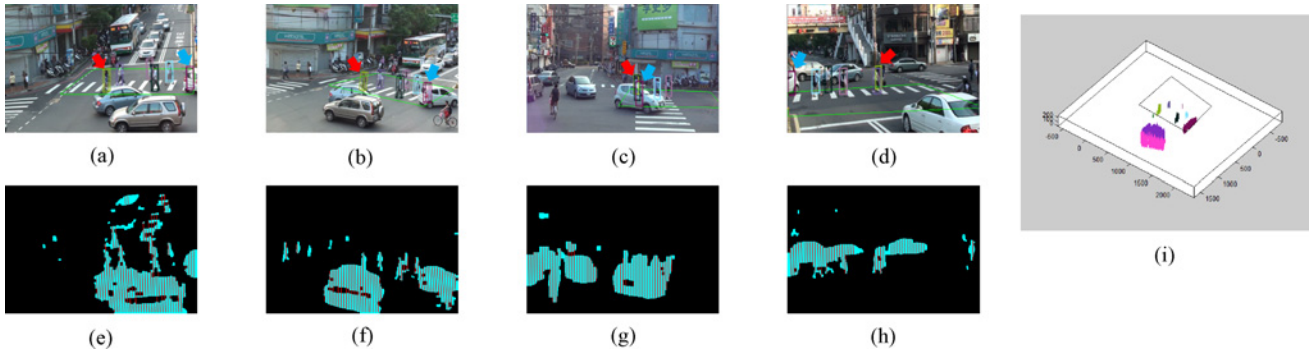


Fig. 17. More challenging localization example for a busy street scene. (a)–(d) Localization results (illustrated with bounding boxes) of four views. (e)–(h) Corresponding foreground regions and 2-D line samples. (i) 3-D line samples to represent different persons in the scene.

TABLE IV
RECOMMENDED VALUE RANGES OF PARAMETERS FOR S1−S3

Section used	Parameter	Function/description	Value range
Section III-A and B	T_{len}	Minimum length of a 3-D line sample (Geometric Rule 1)	[100 cm, 150 cm]
Section III-A and B	T_{tl}	Minimum height of a 3-D line sample (Geometric Rule 2)	[70 cm, 130 cm]
Section III-A and B	T_b	Minimum height of bottom of a 3-D line sample (Geometric Rule 3)	[70 cm, 105 cm]
Section III-B	T_{th}	Maximum height of a 3-D line sample (Geometric Rule 4)	[190 cm, 230 cm]
Section III-B	T_{fg}	Minimum AFCR of a 3-D line sample	[0.68, 0.97]
Section III-B	N_{plane}	Number of reference planes	[10, 45]
Section III-C	T_c	Maximum distance between 3-D line samples of a cluster	[15 cm, 40 cm]
Section III-C	N_{line}	Minimum number of 3-D line samples of a cluster	[1, 11]
Section III-A	T_p	Minimum number of foreground pixels of a 2-D line sample	see text

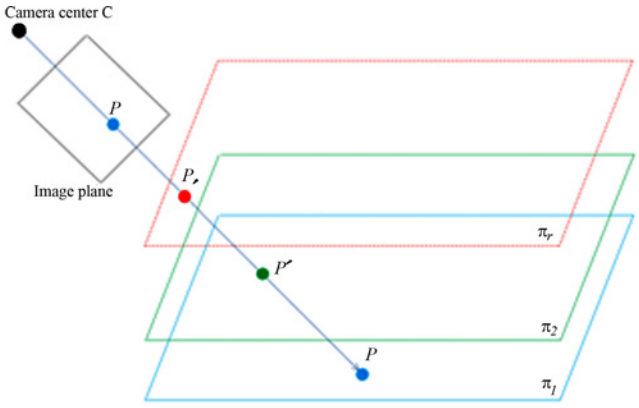


Fig. 18. Illustration of the calculation of a reference point on π_r.

ticular, the line samples are originated from the vanishing point to equally spaced image pixels at the bottom row of the captured image. Fig. 16(a) shows the decreases of both the recall and precision rates with such pixel spacing.[19] One can see that for spacing less than 10, similar recall and precision rates can be obtained, and a larger spacing seems to capture inadequate information for localization. Fig. 16(b) shows that the localization errors are growing slightly with pixel spacing. Whether the localization errors due to different pixel spacings are acceptable will depend on applications under consideration. Finally, Fig. 16(c) shows the growth of computation speed with pixel spacing. Again, the choice among different pixel spacing will depend on the requirement of system performance.

C. Exploring More Challenging Scenes

As a preliminary investigation of possible extensions needed for the proposed approach to work within more challenging scenes, a busy street scene is considered in this subsection. Fig. 17 shows people localization results obtained by directly applying our algorithm, for the monitored area marked in green,[20] for a time instance while six persons are crossing a street. Besides failure cases mentioned earlier (the red arrow indicates the merge of two persons, as in Fig. 13), additional interferences from nonhuman foreground objects (vehicles) include: 1) vehicle–people occlusion and 2) presence of vehicles in the monitored area. While 1) can be seen in all four views but does not result in a problem in this case, 2) does cause a false alarm [shown as a big (dark purple) cluster in Fig. 17(i)]. Overall, the recall and precision rates for this challenging scene are evaluated as 80.9% and 80.2%, respectively, for a total of 108 image frames.

[19]While a spacing of five pixels is selected for S1–S3, a spacing of 4.4 pixels is selected for S4–S5.

[20]Similar to the experiments conducted on S4 and S5, the evaluation of people localization is only preformed for the monitored area, and the image resolution is 360 × 240.

V. CONCLUSION

We proposed an efficient people localization method that is based on vanishing point-based line sampling. Thus, instead of using all foreground pixels, line samples from multiple views were used to find possible 3-D line samples of human bodies efficiently. While our earlier approach in [26] is a direct extension of the approach in [25] in that projection of pixels (lines in [26]) are computed for horizontal planes first, the algorithm presented in this paper reconstructs the above samples in the 3-D space directly. Additional efficiency of the proposed approach arises from effective screening of these 1D samples using geometric constraints of the body. Such efficiency is crucial for certain surveillance applications, which demand prompt attention (and high processing speed) with people localization being part of the complete process.[21] Experimental results demonstrate that the proposed method can handle serious occlusions in quite crowed scenes to provide localization results with correctness and accuracy, and localization accuracy, comparable to that attained with a modified version of [25], but with a much higher processing speed. Additionally, because the proposed localization approach is based on 3-D reconstruction/sampling, it is possible to extend the approach to locate people in the 3-D space.[22]

APPENDIX A
DERIVATION OF MULTIPLE HOMOGRAPHIC MATRICES FOR PLANES OF DIFFERENT HEIGHTS

Homographic matrices are required for projecting 2-D line samples onto the reference plane, as in Section II-B. Also, the

[21]For example, while localization-based people tracking is often needed in intruder detection and abnormal behavior detection, if such functions are to be implemented with no special hardware for acceleration, our approach will have a better chance of fulfilling the requirement of real-time performance than that presented in [25]. As another example, effective people tracking based on the localization results may need to be developed for similar applications, which may be more sophisticated than that presented in [25] and implemented without any special hardware.

[22]To that end, constraints for human standing on the ground plane should be removed, which include Rules 2–4.

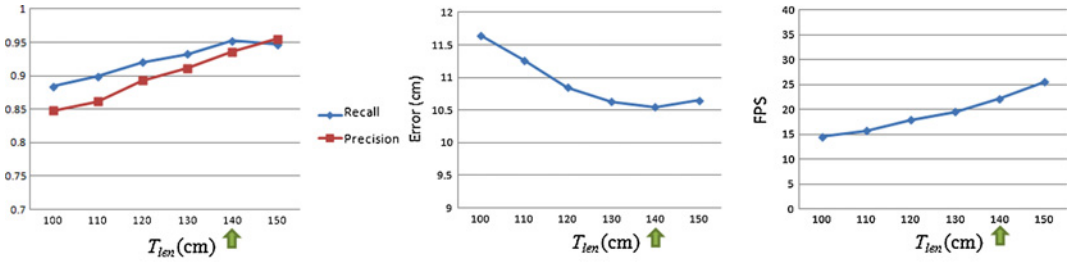


Fig. 19. Results of using different values of T_{len}. (a) Recall and precision. (b) Mean localization error. (c) Computation speed.

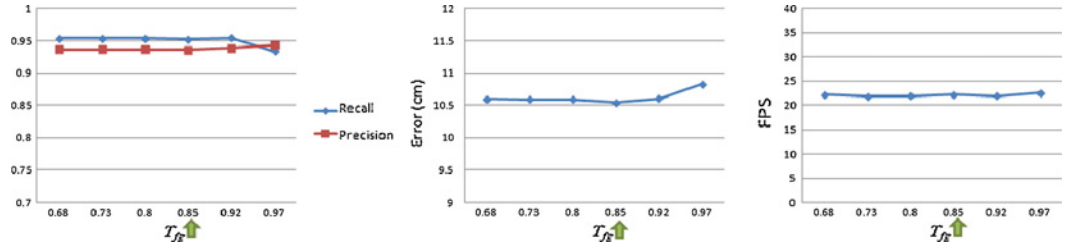


Fig. 20. Results of using different values of T_{fg}. (a) Recall and precision. (b) Mean localization error. (c) Computation speed.

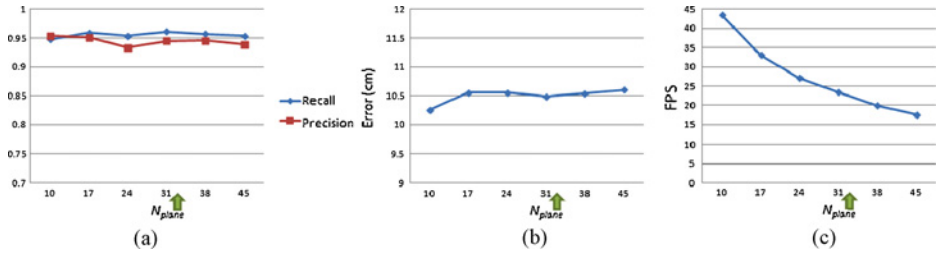


Fig. 21. Results of using different values of N_{plane}. (a) Recall and precision. (b) Mean localization error. (c) Computation speed.

homographic matrices of multiple reference planes at different heights can be used to back-project points on a reference plane to different views for the computation of AFCR, as in Section III-B.

Eshel and Moses [23], [24] use four vertical calibration pillars placed in the scene, with marker points at three known heights on each of them, to establish the homographies between image planes and reference planes at desired heights. Since a new reference point at any height along a pillar can be identified in the images of interest using the cross-ratio along that pillar, the above homographic relationship can actually be established for planes at arbitrary height. Thus, 12 (4×3) marker points are required for calculating all homographic matrices.

Instead of using 12 marker points, an approach for the derivation of multiple homographic matrices for planes of different heights, which only use eight (4×2) marker points on four pillars, is presented in the following. Assume each pillar has two marker points at planes π_1 and π_2 with heights h_1 and h_2, respectively. First, four marker points with height h_2 are used to calculate a homographic matrix H_{m2} between the image plane and the reference plane π_2 as shown in Fig. 18. Then, we will produce four reference points on π_2 by projecting the four marker points with height h_1, respectively. More specifically, the image point p corresponding to the marker point P can be projected to π_2 by H_{m2} to obtain the world coordinate of P' as shown in Fig. 18. After that, we can calculate a new reference point P_r on an arbitrary imaginary plane π_r with a specified height by calculating the intersection of $\overline{PP'}$ and π_r. Similarly, the other three marker points with height h_1 can be used to produce another three new reference points on π_r. Finally, a homographic matrix H_{mr} can be found by using the four new reference points. Thus, we can produce a set of homographic matrices for reference planes of various heights using only eight marker points.

APPENDIX B
SETTING THE PARAMETERS

In Section IV, satisfactory results of people localization are obtained with the proposed approach for selected values of some parameters. In this section, we will show that it is not too hard to set these parameters properly in practice for different scenes. Table IV shows a list of such parameters together with the section(s) in which each of them is used and a range of values tested for each of them. While the first three parameters are applied before and after the refinement process, in both Sections III-A and III-B, the rest are applied only in one subsection of Section III. As for their physical meanings, five of them are for measurements in the 3-D scene (in cm), one of them is based on percentage values, two of them are for number counts, and the last one is for measurements in 2-D image planes (in pixels).

In general, for satisfactory performance of the proposed localization approach, proper values should be assigned to the

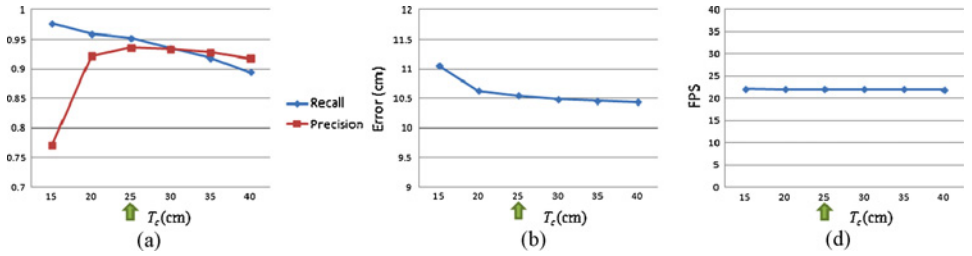


Fig. 22. Results of using different values of T_c. (a) Recall and precision. (b) Mean localization error. (c) Computation speed.

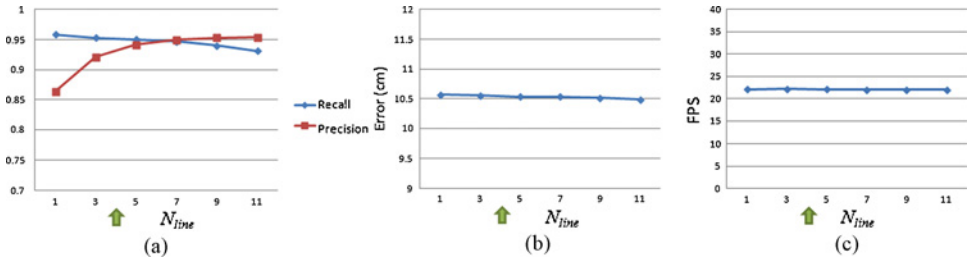


Fig. 23. Results of using different values of N_{line}. (a) Recall and precision. (b) Mean localization error. (c) Computation speed.

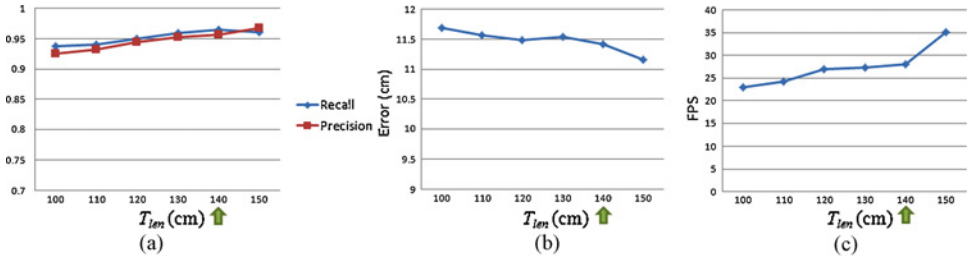


Fig. 24. Results of using different values of T_{len} for S1. (a) Recall and precision. (b) Mean localization error. (c) Computation speed.

above parameters for each scene, or camera configuration. In Table IV, appropriate value ranges, which yield reasonable localization results for S1–S3 taken from the indoor scene considered in Section IV, are listed for these parameters.[23] In particular, Figs. 19–23 show such results, only for the most complicated S3 (with four views) for brevity, for the most important five parameters.[24] For each of the five figures, only one parameter is adjusted for easy observation of the trend of localization performance, which has fairly low sensitivity to the adjustment, with the parameter value used in Table I indicated by an arrow.

For recall and precision rates shown in these figures, significant changes (still within ±2.4% of that in Table I) mainly exist at one end of each plot except for Figs. 19(a) and 22(a). Besides, the plots of recall and precision rates are intersected at one point in each figure. For example, threshold T_c in Fig. 22(a) specifies the maximum distance between two 3-D line samples that can be grouped into the same cluster. If T_c is too small, a cluster corresponding to a person may be split into several groups, resulting in poor precision rate due to a lot of false positives. In contrast, if T_c is too large, the recall rate tends to decrease due to missed detections, resulting from incorrectly merged clusters.

As for localization errors, variations caused by adjusting these parameters are fairly small, i.e., within ±0.50 cm, except for Fig. 19(b). Small variations in computation speed can also be found in these figures, except for Figs. 19(c) and 21(c). For Fig. 21(c), it is easy to see that the computation time is directly related to the number of sample points of a 3-D line sample, which need to be verified against image foregrounds.

Overall, threshold T_{len}, which specifies the minimum length of a 3-D line sample that should be covered by foreground regions in all views, seems to be most influential. While increasing its value to remove more (possibly incorrect) 3-D line samples will always reduce the computation time, the precision/recall rates and localization accuracy will increase monotonically, up to 9% and 1.1 cm in Fig. 19, respectively, as its value is increased from 100 to 140 cm.

In practice, different values of all these parameters may need to be selected for different scenes and camera configurations. Table V shows the two sets of (mostly different) parameter values selected for the indoor scene (for S1–S3) and the outdoor scene (for S4–S5) considered in Section IV. One can see that the values used for the latter are not far from the corresponding value ranges recommended in Table IV for the former. In general, once their values are determined,

[23]In all experiments in Section IV, T_p is arbitrarily chosen as 24 (pixels), i.e., 10% of the height of the input image.

[24]The other three parameters are associated with Geometric Rules 2 to 4, respectively. For their ranges of values listed in Table IV, recall and precision rates are basically the same as those listed in Table I. The screening with all three rules, on the other hand, does increase the computation speed by 17%.

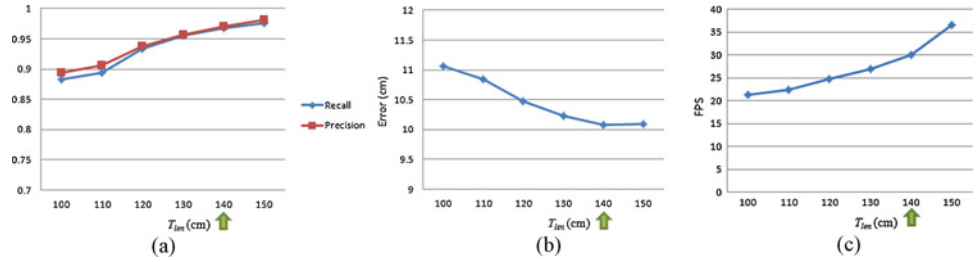


Fig. 25. Results of using different values of T_{len} for S2. (a) Recall and precision. (b) Mean localization error. (c) Computation speed.

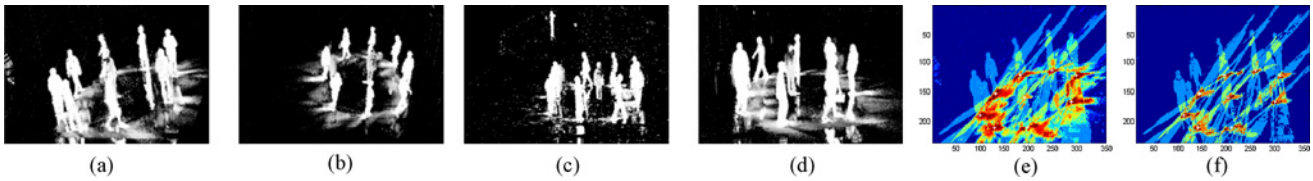


Fig. 26. (a)–(d) Foreground likelihood maps. (e) Synergy map used in [25]. (f) Synergy map obtained by using binary foreground images.

TABLE V
PARAMETER VALUES SELECTED FOR EXPERIMENTS PRESENTED IN SECTION IV

	T_{len}	T_{tl}	T_b	T_{th}	T_{fg}	N_{plane}	T_c	N_{line}
S1–S3	140	90	90	230	0.85	36	25	4
S4–S5	110	130	70	190	0.92	36	25	7

the algorithm will work consistently for the scene under consideration.[25] For example, Figs. 24 and 25 show testing results similar to Fig. 19, but for sequences S1 and S2, respectively. One can see that good localization results can also be obtained with $T_{\text{len}} = 140$ cm.

APPENDIX C
TWO TYPES OF SYNERGY MAPS

For better understanding of the effects of our implementation of [25], synergy maps created by 1) foreground likelihood maps used in [25] and 2) the binary version (foreground regions used in this paper) of 1) are both generated in this section. Fig. 26(a)–(d) shows foreground likelihood maps obtained for Fig. 8(a)–(d), respectively. Even with pixels of lower likelihood filtered out, these foreground maps are still influenced greatly by the cluttered background with strong shadows and reflections. Fig. 26(e) and (f) shows synergy maps generated by 1) and 2), respectively. One can see the positions with high occupancy likelihoods, which are also very close to the ground truth (marked as white crosses), are quite similar for these two types of synergy maps.

ACKNOWLEDGMENT

The authors would like to thank H.-H. Lin for providing many useful comments.

[25] As for automatic determination of appropriated parameter values, different approaches are currently under investigation. For example, by examining these three figures, it seems that it will not be necessary to consider larger values of T_{len} either: 1) when the recall rate drops or 2) when the mean localization error increases.

REFERENCES

[1] Q. Cai and J. K. Aggarwal, "Automatic tracking of human motion in indoor scenes across multiple synchronized video streams," in *Proc. Int. Conf. Comput. Vision*, Jan. 1998, pp. 356–362.
[2] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: A real time system for detecting and tracking people," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, Jun. 1998, p. 962.
[3] S. Khan and M. Shah, "Tracking people in presence of occlusion," in *Proc. Asian Conf. Comput. Vision*, Jan. 2000, pp. 1132–1137.
[4] M. Isard and A. Blake, "Condensation–conditional density propagation for visual tracking," *Int. J. Comput. Vision*, vol. 29, no. 1, pp. 5–28, Aug. 1998.
[5] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "An adaptive color-based particle filter," *Image Vision Comput.*, vol. 21, no. 1, pp. 99–110, Jan. 2003.
[6] H. Wang, D. Suter, K. Schindler, and C. Shen, "Adaptive object tracking based on an effective appearance filter," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1661–1667, Sep. 2007.
[7] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," *IEEE Trans. Neural Netw.*, vol. 21, no. 10, pp. 1610–1623, Oct. 2010.
[8] D. Beymer and K. Konolige, "Real-time tracking of multiple people using stereo," in *Proc. IEEE Frame Rate Workshop*, Sep. 1999.
[9] T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb, "Plan-view trajectory estimation with dense stereo background models," in *Proc. Int. Conf. Comput. Vision*, Jul. 2001, pp. 628–635.
[10] T. Darrell, G. Gordon, M. Harville, and J. Woodfill, "Integrated person tracking using stereo, color, and pattern detection," *Int. J. Comput. Vision*, vol. 37, no. 2, pp. 175–185, Jun. 2000.
[11] J. Orwell, S. Massey, P. Remagnino, D. Greenhill, and G. A. Jones, "A multi-agent framework for visual surveillance," in *Proc. Int. Conf. Image Anal. Process.*, Sep. 1999, pp. 1104–1107.
[12] A. Mittal and L. Davis, "M_2 Tracker: A multi-view approach to segmenting and tracking people in a cluttered scene," *Int. J. Comput. Vision*, vol. 51, no. 3, pp. 189–203, Feb./Mar. 2003.
[13] T.-H. Chang and S. Gong, "Tracking multiple people with a multi-camera system," in *Proc. IEEE Workshop Multi-Object Tracking*, Jul. 2001, pp. 19–26.
[14] A. Utsumi, H. Mori, J. Ohya, and M. Yachida, "Multiple-human tracking using multiple cameras," in *Proc. IEEE Int. Conf. Automatic Face Gesture Recognit.*, Apr. 1998, pp. 498–503.
[15] H. Tsutsui, J. Miura, and Y. Shirai, "Optical flow-based person tracking by multiple cameras," in *Proc. Int. Conf. Multisensor Fusion Integr. Intell. Syst.*, Aug. 2001, pp. 91–96.
[16] P. Kelly, A. Katkere, D. Kuramura, S. Moezzi, and S. Chatterjee, "An architecture for multiple perspective interactive video," in *Proc. ACM Int. Conf. Multimedia*, Nov. 1995, pp. 201–212.
[17] Q. Cai and J. K. Aggarwal, "Tracking human motion in structured environments using a distributed-camera system," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 11, pp. 1241–1247, Nov. 1999.

[18] S. Khan and M. Shah, "Consistent labeling of tracked objects in multiple cameras with overlapping fields of view," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1355–1360, Oct. 2003.

[19] S. Khan, O. Javed, and M. Shah, "Tracking in uncalibrated cameras with overlapping field of view," in *Proc. IEEE Workshop Performance Evaluat. Tracking Surveillance*, Dec. 2001, pp. 84–91.

[20] S. Sun, H. Lo, H. Lin, Y. Chen, F. Huang, and H. Liao, "A multi-camera tracking system that can always select a better view to perform tracking," in *Proc. APSIPA Annu. Summit Conf.*, Oct. 2009, pp. 373–379.

[21] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank, "Principal axis-based correspondence between multiple cameras for people tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 663–671, Apr. 2006.

[22] L. Sun, H. Di, L. Tao, and G. Xu, "A robust approach for person localization in multi-camera environment," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 4036–4039.

[23] R. Eshel and Y. Moses, "Homography based multiple camera detection and tracking of people in a dense crowd," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, Jun. 2008, pp. 1–8.

[24] R. Eshel and Y. Moses, "Tracking in a dense crowd using multiple cameras," *Int. J. Comput. Vision*, vol. 88, no. 1, pp. 129–143, May. 2010.

[25] S. M. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 505–519, Mar. 2009.

[26] K.-H. Lo and J.-H. Chuang, "Vanishing point-based line sampling for efficient axis-based people localization," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 529–532.

[27] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, vol. 2, Jun. 1999, pp. 246–252.

[28] H.-H. Lin, J.-H. Chuang, and T.-L. Liu, "Regularized background adaptation: A novel learning rate control scheme for Gaussian mixture modeling," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 822–836, Mar. 2010.

Kuo-Hua Lo received the B.S. degree in computer science and engineering from Tatung University, Taipei, Taiwan, in 2004, and the M.S. degree in computer science and information engineering from National Dong Hwa University, Hualien, Taiwan, in 2006. He is currently pursuing the Ph.D. degree in the Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan.

His current research interests include image processing, pattern recognition, computer vision, and computer graphics.

Jen-Hui Chuang received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1980, the M.S. degree in electrical and computer engineering from the University of California, Santa Barbara, in 1983, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, in 1991.

Since 1991, he has been with the Faculty of the Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, where he is currently a Professor. His current research interests include robotics, computer vision, 3-D modeling, and image processing.