Vanishing Point-Based Image Transforms for Enhancement of Probabilistic Occupancy Map-Based People Localization

Yen-Shuo Lin, Kuo-Hua Lo, *Student Member, IEEE*, Hua-Tsung Chen, *Member, IEEE*, and Jen-Hui Chuang, *Senior Member, IEEE*

***Abstract*— The widespread use of vision-based surveillance systems has inspired many research efforts on people localization. In this paper, a series of novel image transforms based on the vanishing point of vertical lines is proposed for enhancement of the probabilistic occupancy map (POM)-based people localization scheme. Utilizing the characteristic that the extensions of vertical lines intersect at a vanishing point, the proposed transforms, based on image or ground plane coordinate system, aims at producing transformed images wherein each standing/walking person will have an upright appearance. Thus, the degradation in localization accuracy due to the deviation of camera configuration constraint specified can be alleviated, while the computation efficiency resulted from the applicability of integral image can be retained. Experimental results show that significant improvement in POM-based people localization for more general camera configurations can indeed be achieved with the proposed image transforms.**

***Index Terms*— Image transform, people localization, video surveillance, multiple cameras, probabilistic occupancy map.**

I. Introduction

PUBLIC security is always a critical issue that ordinary people keenly concern. Recently, the installation of a huge number of security cameras has helped protect the people from dangerous or criminal activities, and significantly reduces security incidents [2]. However, for traditional security systems wherein personnel are employed to watch videos for abnormal incidents or inappropriate behaviors, misses may easily occur. Consequently, there has been an ever growing demand for computer vision techniques to assist the development of automatic/semi-automatic video surveillance system. Among them, people localization and tracking plays an important role in vision-based video surveillance systems, and most of the related works can be divided into two categories: monocular and multi-camera approaches, as reviewed in the following[1].

A. Monocular Approaches to People Localization

Until recent years, the bulk of research in the field of people localization and tracking has been concentrated on single camera approaches since it is the most common way to monitor an area by just one camera. Even for most multi-camera surveillance systems, single camera-based analysis still acts as a fundamental and essential processing component.

Many research works regard each person as a blob (a single object) and establish descriptive models for each detected (located) person in an image. For instance, Wang et al. [4] propose an appearance model of spatial-color mixture of Gaussians for particle filters in people tracking, wherein an adaptive scheme is adopted for the model to choose effective cues in different situations. Fan et al. [5] apply a convolutional neural network (CNN) model based on spatial visual features and temporal motion information to track people. However, being updated by features from just two adjacent frames, the CNN model does not handle full and long-term occlusions very well. For better people segmentation and tracking, Zhao et al. [6] propose a unified framework which utilizes models of 3D human bodies, backgrounds, and the camera. In addition to a person's appearance, Yamaguchi [7] et al. also consider a behavior model, which takes into account personal, social, and environment factors in people tracking.

Since the above methods often rely on the assumption of the availability of complete object shape, their performances are limited for crowded scenes. To better deal with such a problem, some research works [8]–[10] regard each person as an assembly of body parts and build models for each part. Felzenszwalb et al. [8] utilize mixtures of models for multi-scale deformable parts to detect people, wherein methods of (i) discriminative training of classifiers using latent information and (ii) matching deformable models to images are developed. To extend the above approach, Shu et al. [9] develop an occlusion handling scheme to indicate occluded parts of target person and ignore their effects, which is based on an SVM

Manuscript received December 4, 2013; revised June 6, 2014; accepted August 28, 2014. Date of publication September 18, 2014; date of current version November 18, 2014. This work was supported in part by the National Science Council of Taiwan under Grant NSC-102-2221-E-009-120 and Grant NSC-102-2218-E-009-003 and in part by the Aim for the Top University Plan through the National Chiao Tung University, Hsinchu, Taiwan, and Ministry of Education, Taiwan. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jianfei Cai.

Y.-S. Lin, K.-H. Lo, and J.-H. Chuang are with the Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan (e-mail: linys.cs00g@nctu.edu.tw; lokh@cs.nctu.edu.tw; jchuang@cs.nctu.edu.tw).

H.-T. Chen is with the Information and Communications Technology Laboratory, National Chiao Tung University, Hsinchu 300, Taiwan (e-mail: huatsung@cs.nctu.edu.tw).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes a supplementary video for Vanishing Point-Based Image Transforms for Enhancement of Probabilistic Occupancy Map-Based People Localization. The video sequence corresponds to the video sequences for Fig. 12(a) and (c) (and localization results, as shown in Fig. 14(c) for Fig. 12(c)). The total size of the video is 22 MB. Contact jchuang@cs.nctu.edu.tw for further questions about this work.

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/TIP.2014.2359152

[1]For a more thorough discussion of tracking techniques, please refer to the comprehensive survey by Yilmaz et al. [3].

classifier trained with the articulations of human bodies. For further enhancement of the performance in crowded scenes, Rodriguez et al. [10] utilize a score map of the part-based model [8] and a crowd density map [11] to detect people.

Although most of the single camera approaches can cope with partial occlusions in people localization, they are inherently limited due to insufficient information captured by only one camera.

B. Multi-Camera Approaches to People Localization

To address the above problem of monocular approaches, multi-camera approaches are proposed so that more visual information can be captured and fused to obtain more precise people locations in crowded scenes. Such approaches can roughly be divided into three categories. The first type of approaches is based on blob information obtained for image foreground. In [12], Otsuka and Mukawa use elliptic cylinder blobs to represent human torso at different 3D locations, and match people in different views. The tangency combination between the objects and the edges of the visual angles is considered as an occlusion problem, and solved with recursive Bayesian estimation. In [13], Chang and Gong use a geometry model and blob features, i.e., people's heights and appearances, to match people in different views. Bayesian networks are then utilized to track people in each view independently. In [14], people locations are obtained by finding correspondence of foreground blobs along epipolar lines for each pair of camera views. Then, Kalman filter is used to track people's trajectories. Usually, complete object shapes are assumed in these methods and the localization may be impaired in crowded scenes.

For the second type of approaches such as [15]–[17], the depth information obtained by a stereo camera is used to achieve better foreground segmentation, and to reduce the influences of shadows. In [15], depth and occlusion information recovered from stereo views of the scene is integrated with color and motion cues to realize the segmentation of individual persons from a group of people. An appearance-based tracking approach is then employed wherein the motion occlusion is handled as layer transition. In [16], Nedevschi et al. exploit the depth information to deal with different scales of a person at different positions for enhancing the performance of Kalman filtering-based people tracking. Motion fields of the tracked people are then analyzed to eliminate false positives. By utilizing depth information to reconstruct 3D object representations, Mitzel and Leibe [17] propose a tracking-before-detection scheme to track people as well as the carried objects in street scenes. However, the use of a pair of cameras with a small baseline may suffer from total occlusions frequently, resulting in impaired tracking performance.

The third type of approaches estimate people locations based on the concept of the occupancy map established with image intensity/foreground information from all camera views, which is firstly proposed in the area of robotic navigation by Elfes [18] and has become one of the main research trends recently. Similar maps are used in a series of people localization and tracking approaches [19]–[26], in which the foreground information from each view are projected onto a reference plane and fused. While these approaches locate and track people without finding correspondences of the people between different views, additional information may be used to improve the performance of localization. For instance, Muñoz-Salinas et al. [19] utilize the occupancy map as well as a height map based on the projected foreground and a distance-based confidence map to infer people locations. In [20], information from all views is fused for ground occupancies with Dempster-Shafer theory to reduce the sensitivity to the foreground quality.

Instead of considering one single occupancy map, approaches in [21]–[24] generate multiple occupancy maps by projecting foreground information from all views onto multiple planes at different heights from the reference (ground) plane. Similarly, intensity variances on multiple reference planes are evaluated in [25] and [26] to detect/locate people heads by finding 2D patches with lower intensity variance. To enhance the efficiency of occupancy map approaches, Lo et al. propose a line sampling scheme in [27], based on the vanishing point of vertical lines (VPVL), to avoid projecting all foreground information from different camera views. To further improve the efficiency, the above line samples are projected directly into the 3D space to generate possible 1D (vertical line) samples of the 3D object in [28].

Rather than analyzing projections of image pixels on reference plane(s), another sub-class of occupancy map-based approaches rely on probabilistically validating in images human models defined for a grid of locations on the reference (ground) plane. In particular, Fleuret et al. [1] propose a probabilistic occupancy map (POM) framework, which uses rectangles (both in the scene and in the images) as human models to locate and track people over time with dynamic programming. Such an approach seems to work fairly well and several enhancements are proposed subsequently. In [29], the localization and tracking problem is first converted to a bipartite graph to improve the performance when people are standing close together. In [30], k-shortest paths algorithm is adopted so that the dynamic programming used in [1] will be less likely to be trapped in a local optimum. In [31], the tracking problem is reformulated by taking appearance model into account so that the problem of identity switch in [30] can be alleviated.

Although the above POM-based approaches perform very well, their applicability is somehow limited by the major assumption that all videos are taken at head or eye level (so that rectangular human models, and the computationally efficient *integral image*, can be used). One possible way of extending the POM framework in [1] to cope with more general camera configurations, whereas rectangular human models are no longer valid from different perspectives, is to rectify all images via image transform. Several image rectification/restoration schemes have been proposed for document images [32]–[35], satellite images [36], [37], and face images [38], [39]. However, since such schemes are mainly concerned with images of a *single* surface in the scene, images of rectangular models at different locations cannot be rectified simultaneously. In [21], VPVL is mapped to infinity via homographic transformation so that all people in the scene will have *upright* appearances. Nonetheless, through a plane-to-plane

mapping, transformed images of the above human models will not be rectangular in general since *horizontal* lines of each rectangular model may have a different vanishing point[2].

In this paper, several VPVL-based image transforms are proposed, as extensions of our previous work proposed in [40] wherein the transformed human models are rectangular and the computation efficiency resulted from the use of *integral image* can be retained. By redefining the *orientation* of each vertically standing human model, the transformed models can be closely approximated by rectangles having the following desirable attributes: (i) each rectangle is symmetric with respect to the major axis of the human model and (ii) models being equidistance from the camera will have identical shape and size. Moreover, by closely (linearly) relating the transformed coordinate system to that of the real world, i.e., the ground plane, superior performance of people localization, compared with [40], can be obtained with different versions of the proposed image transforms.

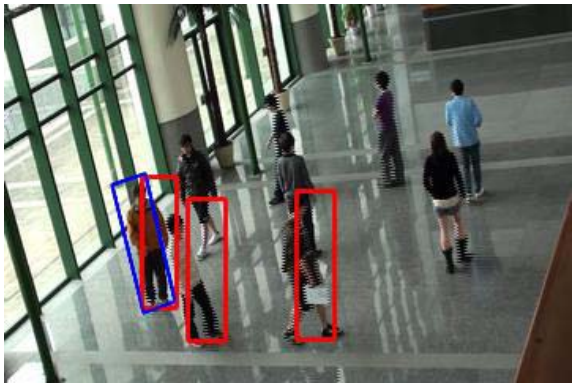
The rest of this paper is organized as follows. In Section II, the basic idea of generalizing POM-based people localization in [1] will be presented. In Section III, the development of a series of image transforms, together with the redefinition of human model's orientation, is elaborated. The experimental results, including a comparison with the original POM-based approach, will be presented in Section IV. Finally, some concluding remarks will be given in Section V.

II. Generalizing POM-Based People Localization

Fleuret et al. [1] propose a robust approach that can reliably track multiple persons in a complex environment and provide accurate location estimation. They compute the occupancy probabilities of the ground plane at each time step, and then combine these probabilities with a color and a motion model to accurately follow individuals by Viterbi algorithm [41][3]. The area, which is visible from multi-camera at the same time, is divided into a regular grid of locations. Then, the binary images obtained from the input images by background subtraction are used to estimate occupancy probabilities of each grid location. To determine whether people present at given locations, and assuming images are taken at head/eye level, a simple rectangle is used for each grid location to represent a person, allowing the use of integral image to speed up the algorithm. Then, the occupancy probabilities at each grid location are approximated as the margins of a product law minimizing the Kullback-Leibler divergence from the "true" conditional posterior distribution. Subsequently, they re-estimate the marginal probability until a stable solution is obtained.

In [1], it is assumed that videos are taken near head or eye level so that the rectangles of human sizes at the occupied ground locations will contain enough foreground pixels to support a POM well, as shown in Fig. 1(a). However,

(a)



(b)

Fig. 1. (a) Image captured by a camera located slightly higher than head level, wherein rectangular regions of human sizes can cover the persons appropriately. (b) Image captured by an imprecisely aligned camera located much higher than a person, wherein upright rectangular regions (in red) may not give satisfactory occupancy measures.

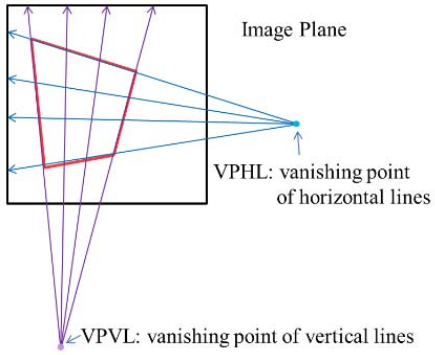


Fig. 2. Illustration of VPVL and VPHL (vanishing point of horizontal lines).

surveillance videos are often taken from security cameras located at a higher-up location with an oblique viewing angle, and may not be precisely aligned, as shown in Fig. 1(b). Such more general camera configurations often lead to the phenomenon that the extensions of vertical and horizontal lines in video frames will intersect at VPVL and a vanishing point of horizontal lines (VPHL), respectively, as illustrated in Fig. 2. Consequently, people standing vertically in the real world may look slanted in an image, and an upright rectangle may not give a satisfactory occupancy measure, as shown in Fig. 1(b). In fact, the higher the cameras are placed, the more slanted people may pose, which will significantly degrade the correctness of the POM-based people localization methods.

[2]Note that the transformed models can only be rectangles (with same height but possibly different widths due to perspective projection) for coplanar models. For other out-of-plane models, trapezoids of vertical edges will be obtained (unless all these horizontal lines are parallel to the intersection of the image plane and ground plane in the 3D space).

[3]Fleuret et al. do provide an on-line POM program for the location estimation part in http://cvlab.epfl.ch/software/pom/index.php, which will be used to validate the effectiveness of our approach.

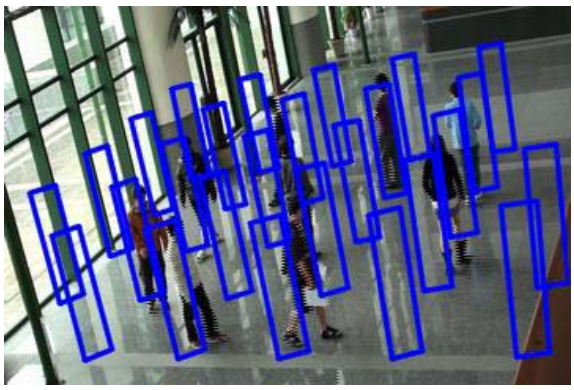


Fig. 3. Equally-spaced virtual billboards (VBs) of human size. The grid resolution is 160cm × 160cm.

To resolve the above problem, vanishing point-based image transforms, entitled *VP-transform/GVP-transform*, are developed to produce rectified images, wherein the people are no longer slanted, regardless of the installation heights or orientations of cameras. The generality and practicability of the POM-based people localization method in [1] can thus be substantially enhanced.

III. The Proposed Image Transforms

In this section, we firstly present an image plane-based VP-transform[4] to improve POM-based people localization results by ensuring each standing/walking person in the transformed image will have an upright appearance. A physically more general, ground plane-based transform, entitled *GVP-transform*, is then proposed for further enhancement of the people localization results.

A. The Image Plane-Based VP-Transform

To extend the POM-based approach for more general situations wherein each standing/walking person in the captured image may not have an upright appearance, an image transform is introduced in this subsection for performance enhancement of people localization. Only simple camera calibrations are needed for such a transform, to provide (i) VPVL in the scene, (ii) camera position, and (iii) image-ground plane homographic matrices[5].

While rectangular regions are adopted in [1] as bounding box of human bodies for efficient evaluation of POM using integral images, such regions, as shown in Fig. 1, are not suitable for more general situations mentioned above. Fig. 3 shows, similar to that described in [1] for people localization, virtual billboards (VBs) of human size (172cm × 40cm) on a grid of 160cm × 160cm for the image shown in Fig. 1(b). Images of such VBs will in general have quadrilateral shapes, as that shown in Fig. 2, due to perspective projection. Although the original POM may be modified to incorporate non-rectangular shapes, the computational advantages associated with integral images will be lost.

[4]Part of the work has been presented in [40].

[5]In fact, information similar to (ii) and (iii) is also needed to generate images of rectangles of human size for the original POM-based localization method presented in [1].

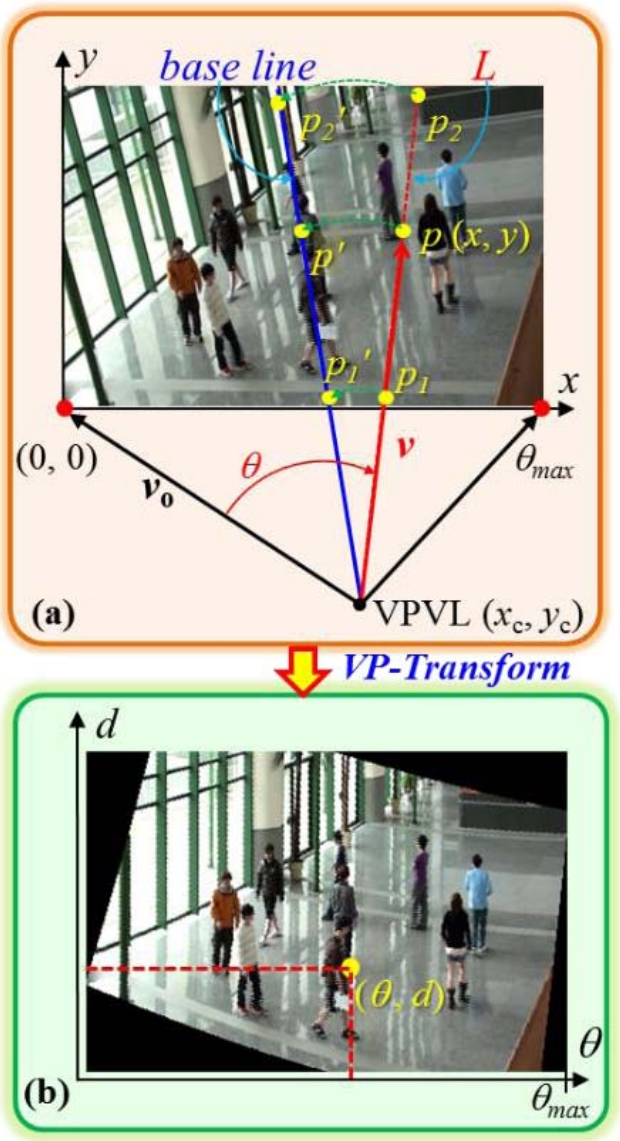


Fig. 4. The basic concept of VP-transform. (a) Original image for View 1 in the ordinary (u, v) plane. (b) Transformed image for View 1 in the (θ, d) plane, wherein the people are no longer slanted.

One possible way of resolving the above problem is to rectify an image, i.e., with the VP-transform proposed in the following, so that rectangular regions can be used to reasonably represent the VBs. First of all, although extensions of the left and right sides of each VB will intersect the VPVL (by definition), these extensions will be *parallel* to one another if they are arranged along *columns* of a (one-dimensionally) rectified image. Moreover, by a simple assumption of orientations of vertically standing VBs, two ends of the two sides of each VB can be connected to form a rectangular region by properly scaling and aligning the above image columns in the corresponding (two-dimensionally) rectified image, as discussed next.

Fig. 4 shows the basic concept of transforming the ordinary $x-y$ image system (Fig. 4(a)) to the rectified $\theta-d$ coordinate system (Fig. 4(b)). In Fig. 4(a), let (x_c, y_c) denote the coordinates of VPVL, $\boldsymbol{v_0}$ represent the vector pointing from (x_c, y_c) to the image origin (0, 0), and $\boldsymbol{v}$ represent the vector

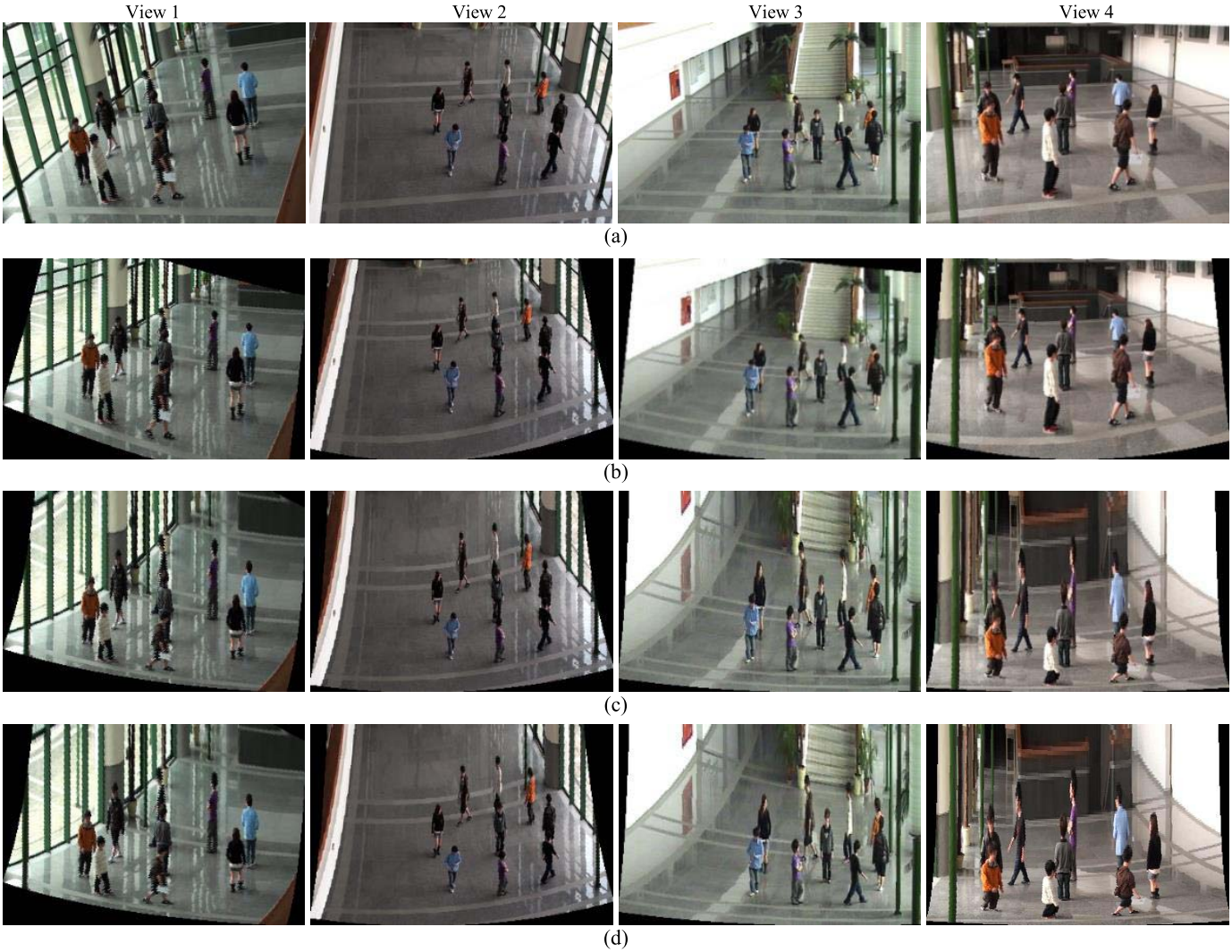


Fig. 5. (a) Original images of four different views. (b) Images rectified by the VP-transform. (c) Images rectified with an additional modification for dimension d of the VP-transform. (d) Rectified images obtained by the proposed GVP-transform.

pointing from $(x_c,\ y_c)$ to an arbitrary image pixel (x,y). The column variable θ in the rectified image can simply be defined as the angular displacement between $\boldsymbol{v}$ and $\boldsymbol{v_0}$, or more formally,[6]

$$\theta = \cos^{-1}\left(\frac{v \cdot v_0}{|v|\,|v_0|}\right) = \cos^{-1}\left(\frac{x_c^2 + y_c^2 - x \cdot x_c - y \cdot y_c}{\sqrt{(x-x_c)^2 + (y-y_c)^2} \times \sqrt{x_c^2 + y_c^2}}\right) \quad (1)$$

As for the other dimension d, it is assumed that all VBs are facing the vertical line containing the camera so that each of them can be well represented by a rectangular region in the rectified $\theta - d$ image plane established in this paper. Considering a VB under such an assumption, it is not hard to show that images of the VB's two upper corners will be transformed, via image-ground plane homography obtained in camera calibration, to two locations with the same distance to the origin C_0 (projection of the camera) of the ground plane. This is because the two corners have the same height, and also have the same distance to the above vertical line. Thus, if variable d corresponds to the distance (D) to C_0 on the ground plane, the two corners will stay in the *same row* in the rectified $\theta - d$ image. Since similar arguments also apply to the lower corners of a VB, its four corners can be connected into a rectangle.

The establishment of dimension d of the rectified image can be accomplished by (i) obtaining a mapping between variables d and D for a one-dimensional base line (BL) image[7] and (ii) scaling and, for each row, aligning the rectified image columns with that of BL for a constant d. For (i), in order to minimize the changes of object shape/size in the rectified image, the line going through VPVL and the center point of the frame (see Fig. 4(a)) is selected as the BL. For a point on BL, its d is computed as the distance from $(x_c,\ y_c)$ to (x,y), i.e.,

$$d = \sqrt{(x-x_c)^2 + (y-y_c)^2} \quad (2)$$

[6]A scaling factor will then be applied to ensure that the monitored area is well covered by the transformed image (columns), as shown in Fig. 4(b).

[7]A scaling factor will then be applied to ensure that BL is well covered by the transformed image (rows), as shown in Fig. 4(b).

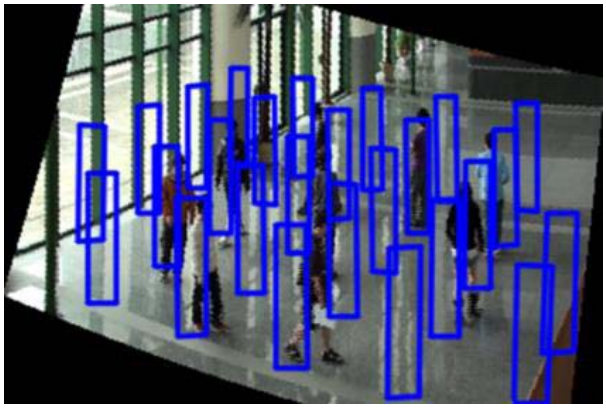


Fig. 6. The VBs in the image rectified from that shown in Fig. 3.

while the corresponding D value can be found via the homographic relation mentioned above.

As for (ii), e.g., for line L shown in Fig. 4(a), we can arbitrarily select two reference points, p_1 and p_2, and compute their distances to C_0 on the ground plane, termed $D(p_1)$ and $D(p_2)$ respectively, by using the homographic matrices obtained from the camera calibration. For each point on BL, we also compute its distance to C_0, and find the corresponding points p'_1 and p'_2 such that $D(p_1) = D(p'_1)$ and $D(p_2) = D(p'_2)$, respectively. Then, p_1 and p_2 are assigned with the same d-coordinates (rows) as those of p'_1 and p'_2, respectively, in the (θ, d) space. For a point p other than p_1 or p_2 on L, its d value can be found by first computing the cross-ratio on the ground plane as:

$$R = \frac{D(p)(D(p_2) - D(p_1))}{(D(p) - D(p_1))D(p_2)} \tag{3}$$

Since (3) is view-invariant, we also have

$$R = \frac{d(d_2 - d_1)}{(d - d_1)d_2} \tag{4}$$

where d_1 and d_2 are the d-coordinates of p_1 and p_2 in the (θ, d) space, respectively. Thus, the row number of p in the rectified image is equal to the rounded value of

$$d = \frac{Rd_1d_2}{d_1 + (R-1)d_2} \tag{5}$$

Fig. 5(a) shows original images obtained from four viewing angles and Fig. 5(b) shows the corresponding images obtained with VP-transform[8]. One can see that all people in the rectified images are indeed having up-right appearances. Fig. 6 shows the VBs in the image rectified from that shown in Fig. 3. While the two top (and bottom) ends of left and right sides of each transformed VB (TVB) are now staying in the same row, the top (and bottom) edge of the TVB, obtained with (5), is not straight in general. Nonetheless, because such an effect is quantitatively minimal, the rectangular approximation of TVB, with the two ends of its two sides being the vertices, will be used in the rest of this paper.

[8]The VP-transform is actually implemented with an efficient table-lookup process.

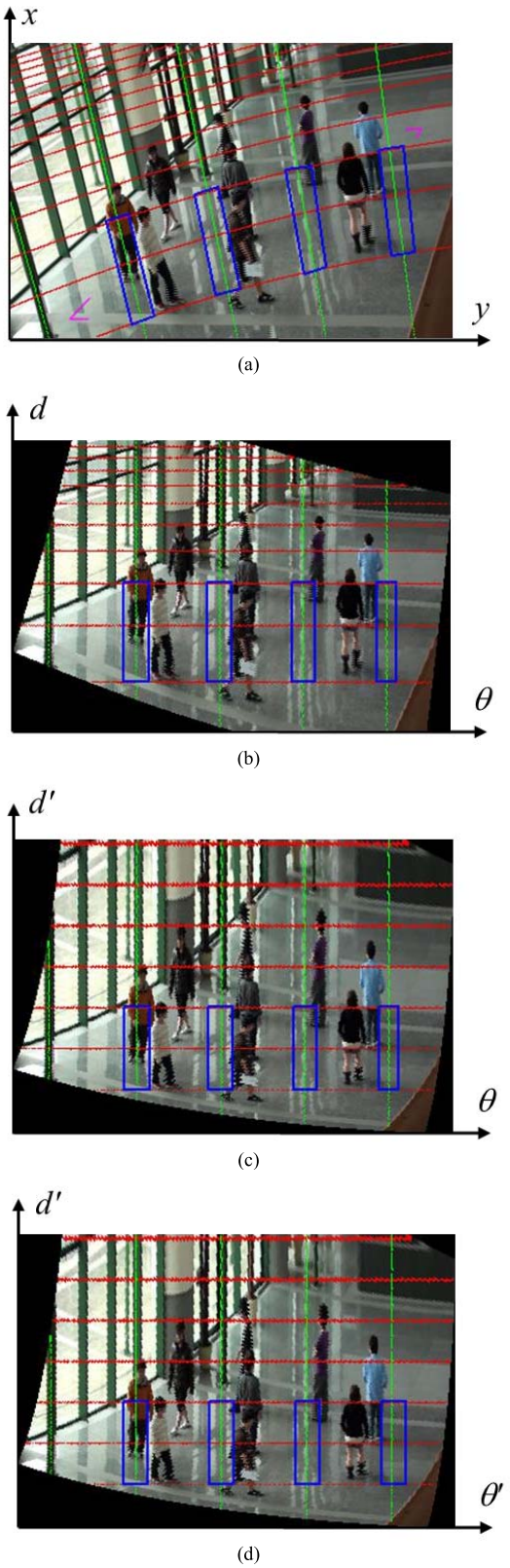


Fig. 7. (a) Fig. 1 (b) overlaid with equidistant (red) contours and equally-spaced angular positions (green lines) on the ground plane, with C_0 being the origin. (b) (a) rectified by the VP-transform. (c) (a) rectified by an intermediate transform. (d) (a) rectified by the GVP-transform.

B. The Ground Plane-Based GVP-Transform

While the above VP-transform can indeed provide rectified TVBs for POM-based people localization, further improve-

ment of the image rectification scheme is possible and will be investigated in this subsection. In particular, a ground plane-based transform, the GVP-transform, will be considered, which is only based on the coordinate system on the ground plane, instead of considering different polar coordinate systems (each corresponds to a VP-transform for a particular camera view). Furthermore, two undesirable properties of VP-transform can also be avoided: (i) the relationship between real-world and transformed angular coordinates is non-linear and (ii) the distance mapping function is not unique (depending on the selection of BL).

To show the effect of (i) more clearly, four VBs are generated in the scene, as shown in Fig. 7(a) with blue quadrilaterals[9]. The four VBs are placed at a distance of 12m, with equally-spaced angular positions (green lines), from C_0. Although the green lines (and the two sides of the rectified TVBs) become *vertical* in the rectified image through VP-transform, as shown in Fig. 7(b), the widths of the TVBs are different. For example, the leftmost TVB has a width of 22 pixels while the rightmost one is only 17-pixel wide. It is easy to see that such a phenomenon also exists for other distances from C_0. As for (ii), the non-uniqueness is due to the fact that such a mapping, as shown in Fig. 4, is determined by the selection of BL. In order to address the above two issues, which are owing to the fact that the establishment of VP-transform is mainly based on the polar coordinate system of the image plane, the enhanced GVP-transform will be developed with respect to the ground plane ($\theta' - d'$plane), as presented in the following.

The establishment of dimensions d' and θ' of the GVP-transform is based on independent adjustments of dimensions d and θ of the VP-transform, respectively. Instead of using (3)-(5), the adjusted dimension d' is defined by establishing a *linear* relationship, which is unique up to a scaling factor, between d' and D. Consider the red lines (rows) shown in Fig. 7 (b), which correspond to equidistant contours on the ground plane with a 2.5-meter spacing. Fig. 7 (c) shows the rectified image obtained from Fig. 7 (b) based on

$$d' = (D - D_{min})/\Delta D \tag{6}$$

where $D_{min} = 10.19$m is the minimum distance (the bottom of Fig. 7 (c)) between the monitored area and C_0 on the ground plane, $\Delta D = 0.07$m is determined such that the person in the middle has about the same size as in Fig. 7 (b), and $d' \in [0, 240)$ corresponds to image row. Note that the red lines are now equally spaced with the adjusted dimension d'. Figs. 8(a) and (b) show close-up views of the third person from the left in Figs. 7(b) and (c), respectively. It is easy to see that the relative length of the upper body of the person shown in Fig. 8(b) is longer than that shown in Fig. 8(a). Fig. 5(c) shows more examples of such an intermediate transform (from d to d').

Similarly, physical information of the ground plane is also adopted directly for the establishment of dimension θ'. Specifically, the adjusted dimension θ' is defined by establishing a linear relationship between θ' and Θ which represent the angular displacement between the projection of $\boldsymbol{v}$ and $\boldsymbol{v_0}$ on the ground plane, as shown in Fig. 4. Consider the green lines (columns) shown in Fig. 7 (c), which correspond to equally-spaced angular positions on the ground plane with a spacing of 7.2° (but with non-uniform 72/70/68/69-pixel spacing in the image plane). Fig. 7 (d) shows the rectified image obtained from Fig. 7 (c) based on[10]

$$\theta' = (\Theta - \Theta_{min})/\Delta\Theta \tag{7}$$

where $\Theta_{min} = 0^\circ$ is defined for the angular position of the projection of $\boldsymbol{v_0}$ in Fig. 4 onto the ground plane, $\Delta\Theta = 0.1^\circ$ is determined such that the person in the middle has about the same size as in Fig. 7 (b), and $\theta' \in [0, 360)$ corresponds to image column. Note that the green lines are equally spaced (with a constant spacing of 70 pixels) with the adjusted dimension θ'. Similarly, while the TVBs in Fig. 7(b), and Fig. 7(c), have different widths (ranging from 17 to 22 pixels),

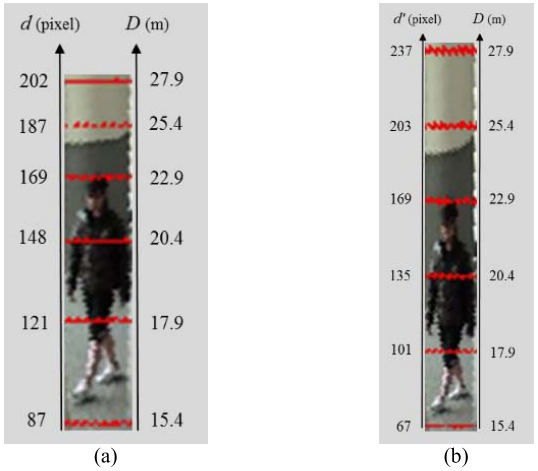


Fig. 8. Close-up views of the third person from the left in (a) Fig. 7(b) and (b) Fig. 7(c).

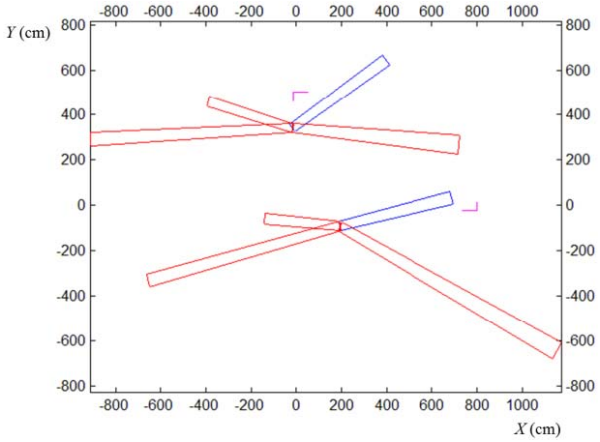


Fig. 9. Transforming TVBs in GVP-transformed images onto the reference (ground) plane. Simple (linear) transformation between θ'-d' coordinates and the $X - Y$ coordinates of the ground plane can be established (see text) for blue TVBs from View 1 (Fig. 7 (d)). Similar transformations can also be found for other views (red TVBs).

[9]See the caption of Fig. 9 for a description of the two L-shaped (purple) markers on the ground.

[10]Two L-shaped (purple) marks are added in this figure, and in Fig. 7(a), to better illustrate the geometric relationship between the two coordinate systems.

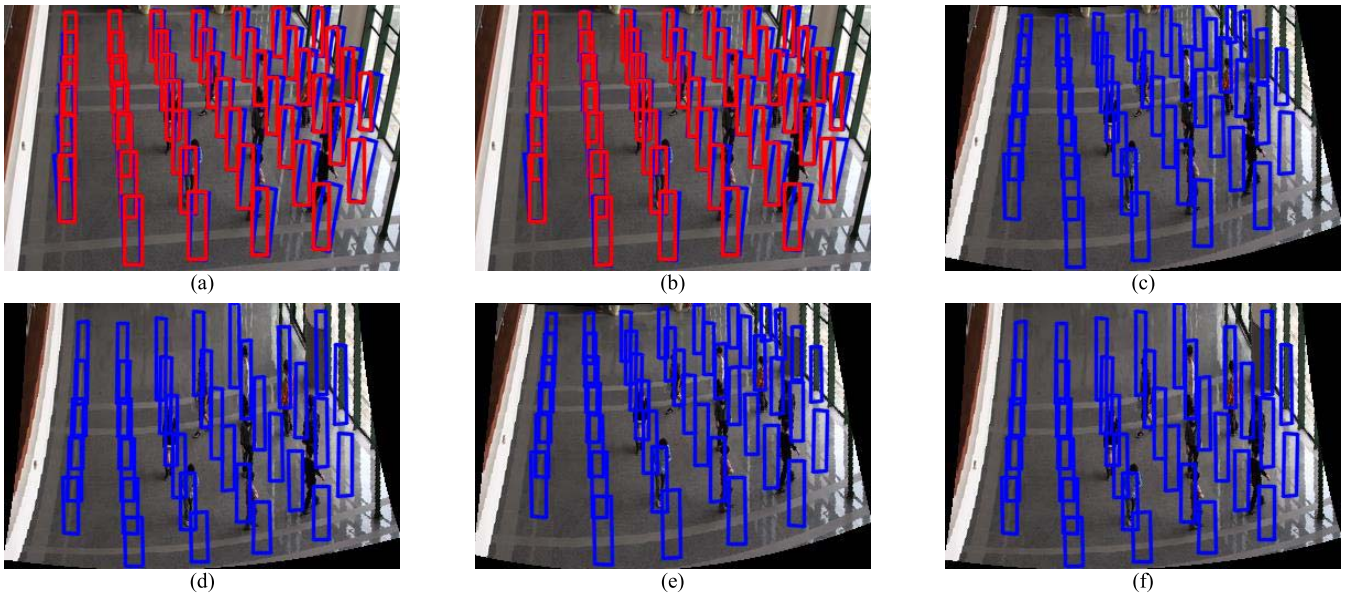


Fig. 10. (a)–(b) The original images of View 2 overlaid with (blue) VBs and type I and type II (red) rectangles, respectively. (c) The VP-transformed image and associated (rectangular) TVBs. (d) The $VP_{d'}$-transformed image and TVBs. (e) The $VP_{\theta'}$-transformed image and TVBs. (f) The GVP-transformed image and TVBs. A sparse 160cm × 160cm grid is adopted in (a)–(f) for better visualization.

TABLE I
LOCALIZATION RESULTS OF VIDEO SEQUENCES S1–S3

Sequence	Number of frames/persons	Method	Recall (%)	Precision (%)	Mean error (cm)
S1	691/9	ORG_I	96.83	92.83	12.18
		ORG_{II}	97.10	92.60	12.20
		VP-transform	99.14	96.38	10.44
		VP_d-transform	99.39 [A]	97.92 [A]	10.19
		VP_θ-transform	99.02	96.16	10.51
		GVP-transform	99.36 [B]	97.85 [B]	10.47
S2	776/9	ORG_I	96.72	94.39	11.78
		ORG_{II}	97.15	94.56	12.05
		VP-transform	99.16	97.12	9.87
		VP_d-transform	98.91	96.82	9.71
		VP_θ-transform	99.41 [A]	97.51 [B]	9.71
		GVP-transform	99.33 [B]	97.69 [A]	9.53
S3	271/12	ORG_I	90.71	90.02	13.39
		ORG_{II}	92.00	90.22	13.18
		VP-transform	96.55 [B]	95.09	9.97
		VP_d-transform	96.25	95.98 [B]	9.82
		VP_θ-transform	96.49	95.12	9.73
		GVP-transform	96.74 [A]	96.77 [A]	9.61

A: the best result, B: the second best

they all have the same width of 20 pixels in Fig. 7(d). Such a property will be desirable for an image rectification scheme, which corresponds to *unbiased* localization for people staying at various angular positions in the scene. Fig. 5(d) shows more examples of the GVP-transform.

For better understanding of the difference between GVP-transform and VP-transform, consider the transforms of the TVBs in Fig. 7 (for View 1) onto the reference (ground) plane, as shown in Fig. 9 wherein only the rightmost and leftmost TVBs (*blue* isosceles trapezoids[11]) are shown for better demonstration. (Additional *red* isosceles trapezoids are also illustrated for other views shown in Fig. 5.) While the trapezoids in Fig. 9 can be obtained easily from Fig. 7 (d) via

[11]It is easy to show that the projection of a TVB on the ground plane, with the corresponding camera being the projection center, is an isosceles trapezoid.

two independent linear mappings, i.e., the inverse mappings of (6) and (7), and

$$X = C_0 + D\cos(-\Theta + \Theta_0) \quad (8)$$

$$Y = C_0 + D\sin(-\Theta + \Theta_0) \quad (9)$$

with $C_0 = (-1056.30, -415.14)$ and $\Theta_0 = 45.89^\circ$ for View 1, more complex (non-linear) mappings are required for the VP-transformed image (Fig. 7 (b)) due to the effect of perspective projection.

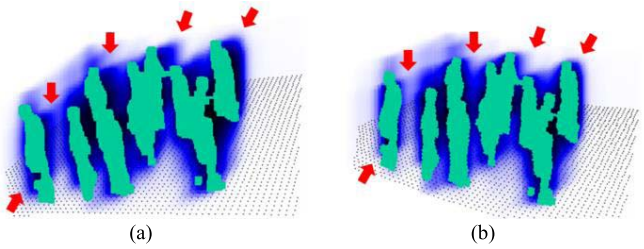
IV. Experimental Results

To validate the effectiveness of our approaches, we compare results of the people localization using the proposed VP-transform/GVP-transform with those from the original POM approach presented in [1]. Two types of input data, i.e., foreground images and rectangles for all grid locations, are required for the online POM program mentioned earlier. For the approaches proposed in this paper, the transformed foreground images and the corresponding rectangles (TVBs) will be provided. For the approach presented in [1], the original foreground images and the same number of rectangles will be used. Two ways of converting the non-rectangular VBs to (type I and type II) rectangular ones will be considered, as discussed next.

For a type I rectangle, its width is simply set to the length of the bottom of the VB at the same grid point, while its height is set to the mean value of left and right sides of the VB. On the other hand, each of the type II rectangles is generated in a simple way, from a VB (an isosceles trapezoid) on BL[12] which has the same distance to C_0 on the ground plane: only the lengths of the upper and lower sides of the VB are changed to their mean value. Therefore, the rectangle will have the same area/height of the VB. Fig. 10 shows different sets of rectangles used in our experiments. One can see that definitions of these two types of rectangles, as shown in Figs. 10(a) and (b), can also be regarded as a reasonable way to rotate each of the original VBs and then modify its height and width.

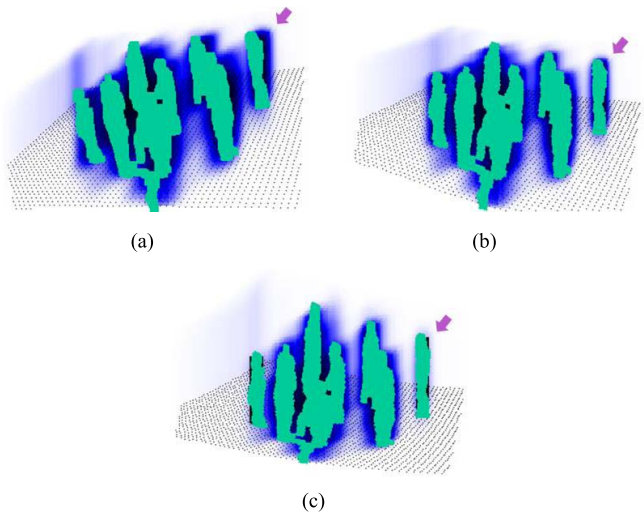
To show the effectiveness of the proposed approaches in improving the accuracy of people localization under general camera configurations, we first use three video sequences (S1, S2, and S3) of an indoor environment, each with four camera views. The three videos contain different numbers of people walking in different trajectories, as described in the following. A total of 691 frames were captured for S1 wherein eight people are walking around the person standing in the middle, as shown in Fig. 5(a). For S2, which has 776 image frames, the same group of people are walking randomly. Although the occlusion among them may increase at times, repeated occlusion instances caused by the periodic walking pattern in S1 do not occur. The walking pattern of S3 is similar to that of S2, except for the increased people count (twelve people). Apparently, the occlusion in S3 is most serious compared to S1 and S2. In the following location estimation experiments, the actual grid resolution is 15cm × 15cm and the size of a VB is 172cm × 40cm.

[12]Note that due to perspective projection, VBs on the BL are also isosceles trapezoids.



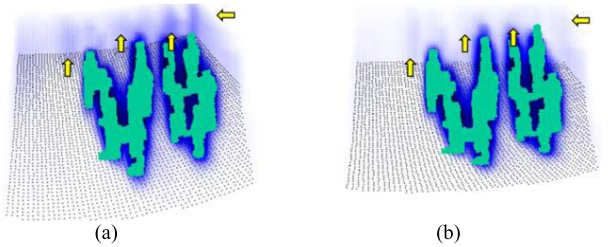
(a) (b)

Fig. 11. Images combining the foreground and the synthetic average image (see [1]) for evaluating the occupancy probabilities (for S3, frame 35). (a) ORG_I. (b) The VP-transform.



(a) (b) (c)

Fig. 12. Images combining the foreground and the synthetic average image for evaluating the occupancy probabilities (for S3, frame 194). (a) ORG_I. (b) The VP-transform. (c) The GVP-transform.



(a) (b)

Fig. 13. Images combining the foreground and the synthetic average image for evaluating the occupancy probabilities (for S3, frame 24). (a) The VP-transform. (b) The GVP-transform.

Table I shows detailed localization results of adopting various image transforms (and the associated rectangular TVBs) as well as those obtained by using type I/II rectangles in the original images (ORG_I/ORG_{II}). As mentioned in [40], adopting the image plane-based VP-transform together with the associated rectangular TVBs results in much better localization than those obtained with ORG_I or ORG_{II} for all performance indices, with more than 5% improvements in recall/precision rates and 3cm reduction in mean localization error for the scenario with the most serious occlusion (S3). While ORG_I and ORG_{II} lead to very similar results for precision rate and mean localization error, within 0.23% and

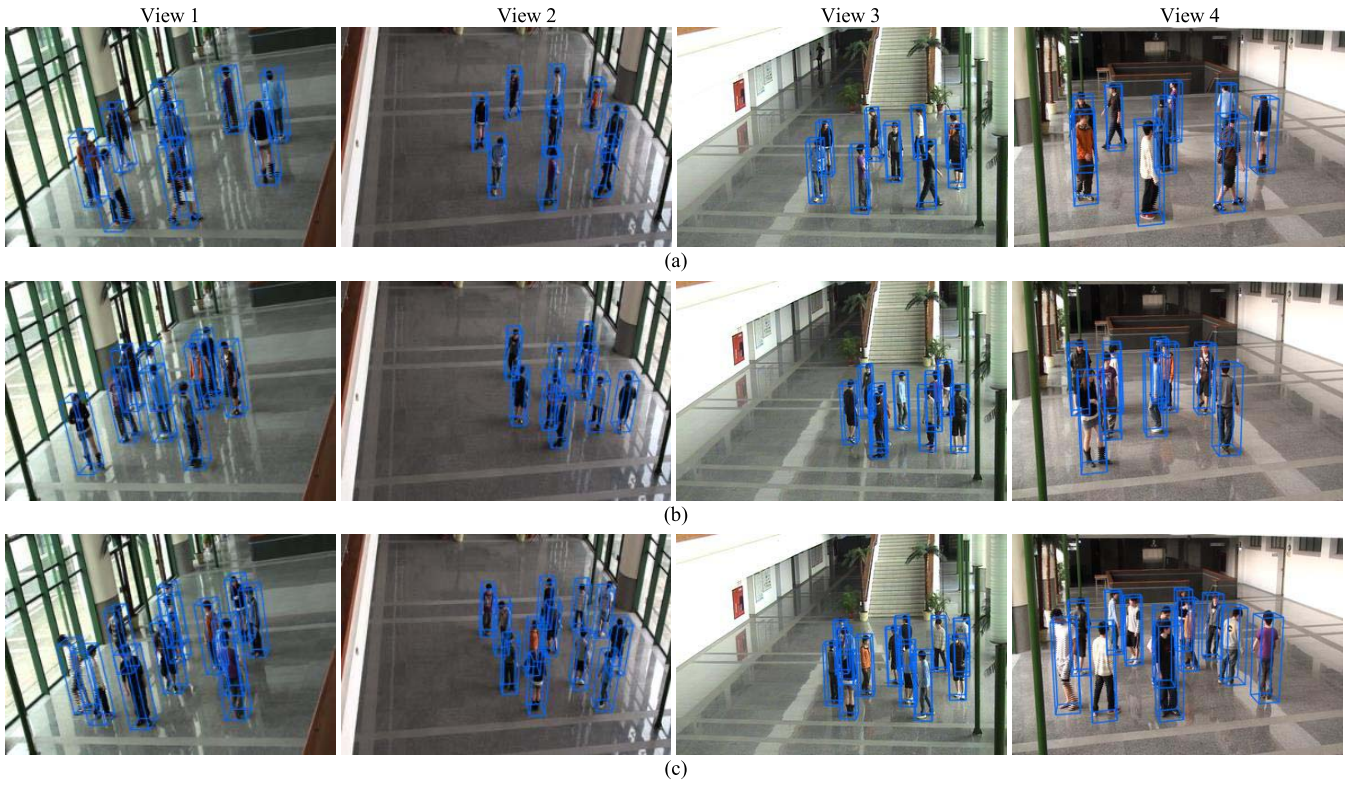


Fig. 14. Some localization results of the GVP-transform for (a) S1, (b) S2, and (c) S3.

0.27cm, respectively, the recall rate obtained from ORG_{II} is better than that from ORG_{I} by 1.28%.[13]

As for the effect of using the transforms based on the ground plane coordinate system, we also consider $\text{VP}_{d'}$-transform (VP-transform with dimension d changed to d'), $\text{VP}_{\theta'}$-transform (VP-transform with dimension θ changed to θ'), in addition to the GVP-transform, as shown in Figs. 10 (d) to (f)[14]. While the overall performance are quite similar for all transformed images, more clear comparison can be carried out by considering the 1st and the 2nd best recall and precision rates for each scene, as marked with A and B respectively in Table I. One can see that while the VP-transform gets only one B, the $\text{VP}_{\theta'}$-transform obtains one A and one B, and the $\text{VP}_{d'}$-transform obtains two As and one B. Moreover, the GVP-transform achieves the best or the second best performance for every video sequence, i.e., three As and three Bs. Besides, the GVP-transform also has smaller mean localization errors than the VP-transform for all scenes except for S1.

Two desirable properties of the ground plane-based GVP-transform may contribute to the extraneous performance improvement over already reasonable localization results obtained with the image-based VP-transform. First of all, while there is linear relationship between real-world and transformed angular coordinates for GVP-transform, such a relationship is nonlinear, and view dependent, for VP-transform. Thus, the rectangular TVBs for grid locations at the same distance from the camera will have same width, and are symmetric to the center line, in GVP-transformed image but not for the VP-transformed one (see Fig.7). Secondly, although such TVBs have a constant height for each transform, it is view independent (up to a scaling factor) for GVP-transform but view (BL) dependent for VP-transform. Additionally, for cameras mounted at higher up positions, a person in a GVP-transformed image will have a relatively longer upper body, which is usually less occluded and has less variations in its appearance, and may also result in more accurate people localization.

For better understanding of the underlying measures of occupancy probability, Figs. 11 and 12 show two images (each combining the foreground and the synthetic average images) for a group of people at two time instances, respectively.[15] In these combination images, blue areas represent the probability of occupancy, with lower (higher) probability represented by lighter (darker) blue, while green areas correspond to image foregrounds. In Fig. 11, the blue areas indicated by red arrows show where the VP-transform achieves better results in locating a person (separating individuals) than the original approach presented in [1]. In addition, one can also see that the blue areas indicated by purple arrows in Fig. 12(c) are

[13]One possible explanation of the difference in recall rate is that type II rectangles are defined with area/height constraint, from VBs (isosceles trapezoids) along the BL, and may represent human size more closely.

[14]In order to perform fair comparisons, the height (in pixels) of a person in the two transformed images based on d' coordinate, i.e., $\text{VP}_{d'}$-transform and GVP-transform, is further adjusted (normalized): similar to that done for VP-transform, the normalizations are based on the height of the person standing in the center area of each view.

[15]They are intermediate results generated by the on-line POM program mentioned earlier. Results obtained from using ORG_{I} are used here for better comparison (with larger differences) between location estimation performed with or without image rectification.

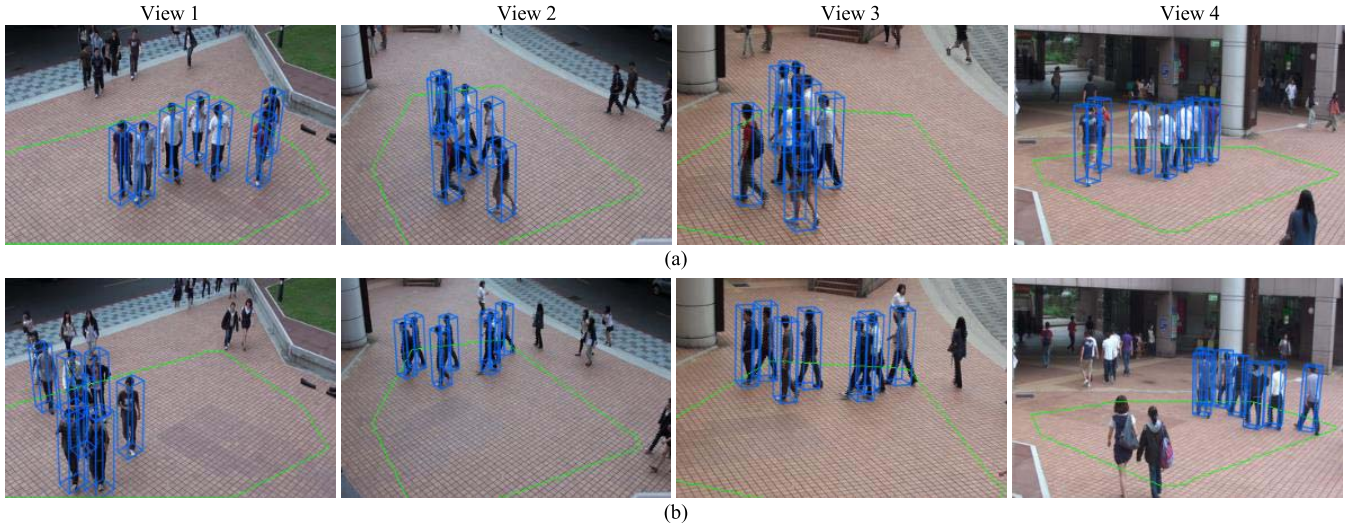


Fig. 15. The monitored area and localization results of the GVP-transform for (a) S4 and (b) S5.

TABLE II
LOCALIZATION RESULTS OF VIDEO SEQUENCES S4 AND S5

Sequence	Number of frames/persons	Method	Recall (%)	Precision (%)	Mean error (cm)
S4	70/6-7	ORG_{II}	97.45%	90%	**9.17**
		VP-transform	98.73%	92.63%	9.43
		GVP-transform	**99.58%**	**96.9%**	9.64
S5	40/5-7	ORG_{II}	88.93%	96.89%	11.09
		VP-transform	**92.86%**	95.94%	10.08
		GVP-transform	92.14%	**98.1%**	**9.47**

more compact than those in Figs. 12(a) and (b), showing that the above observations are particularly clear for more isolated individuals. On the other hand, Figs. 13 (a) and (b) show that GVP-transform tends to generate less ghost/phantom regions, as indicated by yellow arrows, compared with the VP-transform. Finally, Fig. 14 show some localization results using the GVP-transform for S1, S2, and S3 by overlaying bounding boxes of localized people in each camera view (for the first frame of each sequence).

For more investigation into the localization performance of the proposed image transforms, two outdoor sequences captured from some real scenarios, i.e., S4 and S5 shown in Fig. 15, are considered. In general, working with outdoor videos are much challenging due to several time-varying factors such as ambient lighting, wind speed, and shadows of various strengths. In addition, more people may show up outside the monitored area (shown with green border) from time to time. Finally, since a group of people may quickly walk through the monitored area, less image frames are captured for S4 and S5 compared with those in S1–S3. One can see that people location can also be found satisfactorily with the GVP-transform. (Note that people groups of two different formations can be found in Figs. 15(a) and (b), respectively.)

Table II shows localization results for ORG_{II}, the VP-transform, and the GVP-transform to S4 and S5. One can see that adopting the two transforms results in better localization than those obtained with ORG_{II} and achieves recall rates and mean localization errors similar to those shown in Table I. However, the GVP-transform substantially outperforms the VP-transform in precision rates, with 4.27% and 2.16% improvements for S4 and S5, respectively.

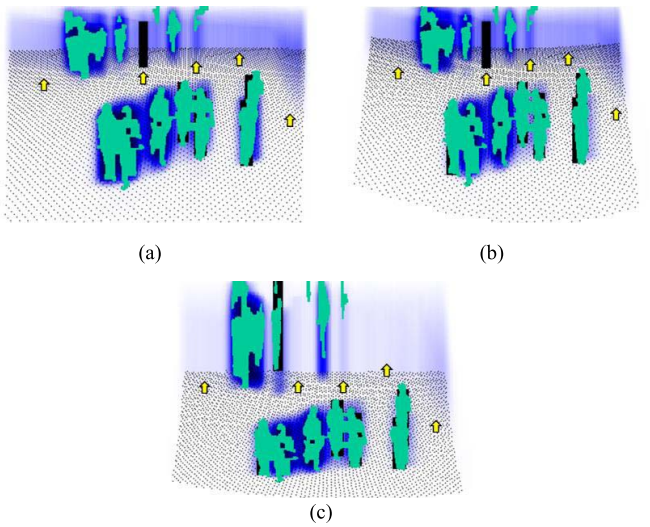


Fig. 16. Images combining the foreground and the synthetic average image for evaluating the occupancy probabilities (for S4, frame 69). (a) ORG_{II}. (b) The VP-transform. (c) The GVP-transform.

Figs. 16(a), (b) and (c) show synthetic images, similar to that shown in Fig. 13, for ORG_{II}, the VP-transform and the GVP-transform, respectively. One can see that the blue areas indicated by yellow arrows in Fig. 16(c) contain much less phantom regions than those in Figs. 16(a) and (b). In summary, the newly proposed GVP-transform, which are developed based on the ground plane coordinate system, does achieve better results for both indoor and outdoor scenes.

V. Conclusion

A series of novel image transforms, based on image or ground plane coordinate system, using the vanishing point of vertical lines are proposed in this paper to generalize the POM-based people localization presented in [1] so that more general camera configurations, e.g., those located higher than head level, can be used. The main feature of the transformed images is that each standing/walking person will have an upright appearance; therefore, rectangular billboard, similar to that adopted in [1], can still be used to model the human body for efficient establishment of the probabilistic occupancy map using integral image. Experimental results, for video sequences obtained from four camera views in both indoor and outdoor environments, show that the proposed image transforms can significantly improve the performance of the people localization for more general camera configurations, with the best localization results achieved by the ground plane-based GVP-transform.

References

[1] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 267–282, Feb. 2008.

[2] M. Nieto, *Public Video Surveillance: Is It an Effective Crime Prevention Tool?* Sacramento, CA, USA: California State Library, 1997.

[3] A. Yilmaz, O. Javed, and S. Shah, "Object tracking: A survey," *ACM J. Comput. Surv.*, vol. 38, no. 4, pp. 1–45, Dec. 2006, Art. ID 13.

[4] 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.

[5] 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.

[6] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1198–1211, Jul. 2008.

[7] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1345–1352.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[9] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1815–1821.

[10] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert, "Density-aware person detection and tracking in crowds," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2423–2430.

[11] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Neural Inf. Process. Syst.*, 2010, pp. 1324–1332.

[12] K. Otsuka and N. Mukawa, "Multiview occlusion analysis for tracking densely populated objects based on 2D visual angles," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun./Jul. 2004, pp. 90–97.

[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. Mittal and L. S. Davis, "M_2Tracker: A multi-view approach to segmenting and tracking people in a cluttered scene," *Int. J. Comput. Vis.*, vol. 51, no. 3, pp. 189–203, Feb. 2003.

[15] Q. Zhang and K. N. Ngan, "Segmentation and tracking multiple objects under occlusion from multiview video," *IEEE Trans. Image Process.*, vol. 20, no. 11, pp. 3308–3313, Nov. 2011.

[16] S. Nedevschi, S. Bota, and C. Tomiuc, "Stereo-based pedestrian detection for collision-avoidance applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 380–391, Sep. 2009.

[17] D. Mitzel and B. Leibe, "Taking mobile multi-object tracking to the next level: People, unknown objects, and carried items," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 566–579.

[18] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, Jun. 1989.

[19] R. Muñoz-Salinas, R. Medina-Carnicer, F. J. Madrid-Cuevas, and A. Carmona-Poyato, "People detection and tracking with multiple stereo cameras using particle filters," *J. Vis. Commun. Image Represent.*, vol. 20, no. 5, pp. 339–350, Jul. 2009.

[20] M. Morbee, L. Tessens, H. Aghajan, and W. Philips, "Dempster-Shafer based multi-view occupancy maps," *Electron. Lett.*, vol. 46, no. 5, pp. 341–343, Mar. 2010.

[21] D. Delannay, N. Danhier, and C. De Vleeschouwer, "Detection and recognition of sports(wo)men from multiple views," in *Proc. 3rd ACM/IEEE Int. Conf. Distrib. Smart Cameras*, Aug./Sep. 2009, pp. 1–7.

[22] 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.

[23] A. Utasi and C. Benedek, "A 3D marked point process model for multi-view people detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3385–3392.

[24] A. Utasi and C. Benedek, "A Bayesian approach on people localization in multicamera systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 1, pp. 105–115, Jan. 2013.

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

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

[27] 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. 521–524.

[28] K.-H. Lo and J.-H. Chuang, "Vanishing point-based line sampling for real-time people localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 7, pp. 1209–1223, Jul. 2013.

[29] M. Liem and D. M. Gavrila, "Multi-person localization and track assignment in overlapping camera views," in *Proc. 33rd Annu. Symp. German Assoc. Pattern Recognit.*, 2011, pp. 173–183.

[30] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1806–1819, Sep. 2011.

[31] H. Ben Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Tracking multiple people under global appearance constraints," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 137–144.

[32] M. S. Brown and Y.-C. Tsoi, "Geometric and shading correction for images of printed materials using boundary," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1544–1554, Jun. 2006.

[33] L. Zhang, Y. Zhang, and C. L. Tan, "An improved physically-based method for geometric restoration of distorted document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 728–734, Apr. 2008.

[34] J. Liang, D. DeMenthon, and D. Doermann, "Geometric rectification of camera-captured document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 591–605, Apr. 2008.

[35] N. Stamatopoulos, B. Gatos, I. Pratikakis, and S. J. Perantonis, "Goal-oriented rectification of camera-based document images," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 910–920, Apr. 2011.

[36] S. Vassilopoulou *et al.*, "Orthophoto generation using IKONOS imagery and high-resolution DEM: A case study on volcanic hazard monitoring of Nisyros Island (Greece)," *J. Photogramm. Remote Sens.*, vol. 57, nos. 1–2, pp. 24–38, Nov. 2002.

[37] M. O. Karslioglu and J. Friedrich, "A new differential geometric method to rectify digital images of the Earth's surface using isothermal coordinates," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 666–672, Apr. 2005.

[38] D. Gonzalez-Jimenez and J. L. Alba-Castro, "Toward pose-invariant 2D face recognition through point distribution models and facial symmetry," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 413–429, Sep. 2007.
[39] X. Chai, S. Shan, X. Chen, and W. Gao, "Locally linear regression for pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 7, pp. 1716–1725, Jul. 2007.
[40] Y.-S. Lin, K.-H. Lo, H.-T. Chen, and J.-H. Chuang, "VP-transform: A novel vanishing point-based image transform for enhancement of people localization," in *Proc. Int. Conf. Multimedia Expo*, Jul. 2013, pp. 1–6.
[41] J. B. Francois, J. Berclaz, F. Fleuret, and P. Fua, "Robust people tracking with global trajectory optimization," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. 2006, pp. 744–750.

Yen-Shuo Lin received the B.S. degree in electronic engineering from Chang Gung University, Taoyuan, Taiwan, in 2009, and the M.S. degree in space science from National Central University, Taoyuan, in 2011.

He is currently pursuing the Ph.D. degree with the Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan. His research interests include remote sensing, image processing, pattern recognition, and computer vision.

Kuo-Hua Lo (S'11) received the B.S. degree in computer science and engineering from Tatung University, Taipei, Taiwan, in 2004, the M.S. degree in computer science and information engineering from National Dong Hwa University, Hualien, Taiwan, in 2006, and the Ph.D. degree from the Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, in 2013.

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

Hua-Tsung Chen (M'07) received the B.S., M.S., and Ph.D. degrees in computer science and information engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2001, 2003, and 2009, respectively.

He is currently an Assistant Research Fellow with the Information and Communications Technology Laboratory, National Chiao Tung University. His research interests include computer vision, video signal processing, content-based video indexing and retrieval, multimedia information system, and music signal processing.

Jen-Hui Chuang (SM'06) 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 at Santa Barbara, Santa Barbara, CA, USA, in 1983, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1991.

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