3D Pedestrian Tracking Based on Overhead Cameras

Zhognchuan Zhang and Fernand Cohen
Electrical and Computer Engineering Department
Drexel University
Philadelphia, USA
zz57@drexel.edu and fscohen@coe.drexel.edu

***Abstract*—This paper proposes a method to track pedestrians in crowded scenes based on the detection of the 3D head position of a person using two overhead cameras. A possible head area in one frame acquired from one of the overhead cameras is determined by evaluating a head area existence probability based on the integral polar mapped image, where a foreground pixel is assigned a probability to belong to the head area. A segment passing through the head top is estimated for each clustered head area. The disparities along each segment are calculated using the synchronized frame from the other overhead camera. The center of the points with the largest disparity on the segment is determined as the head point and its 3D position is computed using triangulation. It is then tracked using common assumptions on motion direction and velocity. This is efficiently done notwithstanding the fact that several segments may exist in a single foreground blob with each segment corresponding to a different person. The approach is tested using a publicly available visual surveillance simulation test bed. The experiments show that the 3D tracking errors are around 5 cm. The method allows for the capture of high quality close-up facial images.**

Keywords—3D head position detection; pedestrian tracking; overhead camera; crowded scene; facial image capture

I. Introduction

With the prevalence of video surveillance, face recognition and tracking is drawing more attention and is more rigorously pursued. Accurately tracking the 3D position of a person is fundamental in capturing close-up facial images that are required by most face recognition systems. In this paper, we focus on pedestrian tracking in indoor environments, such as train stations, airports, shopping malls and hotel lobbies where scenes can be crowded.

Many existing tracking methods use a single side view camera, which cannot handle occlusions between people well, a phenomenon bound to happen in crowded scenes. To resolve the occlusion problem, multiple side view cameras are used. The more cameras are used, the more accurate the targets are localized. This, in turn, increases the computation and data transmission load as each view of the scene should be transmitted to the computer and processed.

Overhead cameras, which are usually deployed in indoor environments, offer advantages over side view cameras. As shown in Fig. 1, occlusion is much less likely to happen in an overhead view compared to a side view where almost no person is viewed by him/herself. However, when the scene becomes more crowded, inter-object occlusion can occur especially for those near the boundary of field of view (FOV). In this paper, we use two identical cameras that look straight down and are installed at the same height to track the pedestrians in crowded scenes based on the detection of their 3D head points. The existence probability of a head area is calculated for each foreground pixel in a left image. The regions consisting of the pixels whose existence probabilities are higher than a threshold are regions of interest (ROIs) and are clustered according to the distances between them if there is more than one ROI in a foreground blob. A short segment passing through the head top is then established based on the centroid of the clustered ROIs. The disparity of each pixel on the segment is calculated using the synchronized left and right images. With disparity distribution along the segment, the center of the highest points on the segment is determined as the 3D head point. And then people are tracked across frames under the assumption of non-erratic and smooth walking.

The major contributions of this paper are twofold: 1) using an image from just one camera, a potential head top segment is determined based on the existence probability of a head area for each person disregarding the number of people in a foreground blob; 2) combining the image from the other camera, the 3D head point is localized efficiently without using depth image. Our approach has the following advantages: 1) lower computation and data transmission load when compared to using several side view cameras; 2) no full disparity map of the scene is needed unlike other methods using stereo vision, resulting in a large saving on the computational load; 3) better scalability since the common FOV of two overhead cameras is rectangular and easy to be calculated and measured.

The paper is organized as follows. Section II presents a short review of pervious work. Section III and IV respectively focus on the 3D head point detection and tracking. Experiment results are presented in section V. In section VI, conclusions are drawn.

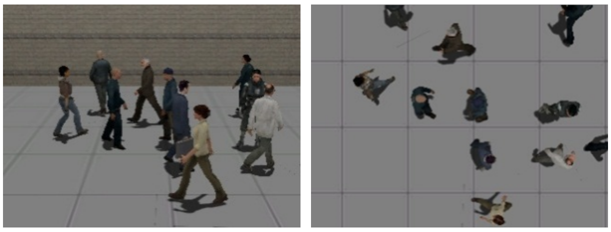

(a) Side view (b) Overhead view

Fig. 1. Occlusions under side and overhead views of the same scene

II. RELATED WORK

Side view cameras are extensively used to track people due to their applicability in both indoor and outdoor environment. To solve the occlusion problems, multiple side view cameras are deployed. In [1], the connected blobs obtained from background subtraction are modeled using color histograms and then used to match and track objects. Krumm et al. [2] combine the information from multiple stereo cameras to detect human-shaped blobs in 3D space. Color histograms are created for each person and are used to identify and track people. Mittal and Larry [3] match object region based on the color characteristics in each camera pair. For each pair of matched region the back projection in 3D space is performed in a manner that yields 3D points that guaranteed to be inside the object. Although these methods attempt to solve the occlusion problem, they may fall short due to either near total occlusion or when people are dressed in similar colors. Instead of using color or shape cues of a person, a planar homography constraint that combines foreground likelihood information from different views is applied to resolve occlusion and to localize people on the ground plane [4]. A homography constraint and a multi-view geometric constraint (regarding the perpendicular projection of a camera's optical center) are used in [5]. Multiple planes parallel to the ground are used in [6] to increase localization robustness and in [7] to solve the optimal height of a person and track his/her head. Similar to [8] and [9] our method integrates the information of all parallel planes by considering all foreground pixels along a vertical segment.

For indoor environments, overhead cameras are also used. In [10] and [11] one overhead camera is used to localize a person, and the centroid of the foreground blob is taken as the ground position. The method is not accurate especially when people are close to the camera or walking around the boundaries of the FOV, and it fails when more than one person exists in a foreground blob. Boltes et al. [12] detect people's heads from an overhead view by placing pasteboards with markers on the heads. To reduce the perspective distortion error, the height of a person is needed and color coded as a marker on the pasteboard. This method, however, is not applicable in general scenarios because the assumption of known heights and requiring people to have markers on their heads are not practical. To obtain the 3D position of a person without these constraints, stereo overhead cameras are applied and robust background subtraction is done in 3D space. Beymer [13] reprojects 3D points to a top-down orthographic view to track the people's ground position. Oosterhout et al. [14] detect 3D head positions in highly crowded situations by matching a sphere crust template on the foreground regions of the depth map and then track the head using Kalman filters. In another paper [15], they localize people in the scene by maximizing the similarity between the depth map obtained from a stereo camera and that reconstructed by projecting a certain number of templates at certain locations. By using stereo images, Boltes and Seyfried [16] build the perspective height field of pedestrians which are represented by a pyramid of ellipses. A person is then tracked using the center of the second ellipse from the head downward. This paper, like our previous work [19], detects and tracks people's 3D position without using depth images. It differs from [19] in that it can handle the case where more than one person is detected inside a foreground blob, and hence is geared towards dealing with crowded scenes.

III. PEDESTRIAN LOCALIZATION

In our work, we assume that people in the scenes are upright and the head tops are their highest parts. This implies that the head top of a person is generally visible from an overhead view even in crowded scenes. Note that the 3D head point is defined as the highest point of a person and its projection is roughly the center of the head top in an image. We also assume the scenes are not extremely crowded (e.g., subway station in Tokyo), i.e., people's bodies are not touching each other, although they are in very close proximity, with the distance between two people being greater than a shoulder width.

A. Background Subtraction

Unlike the method using multiple side view cameras, the foreground segmentation is only implemented for frames captured from one overhead camera (the left one in this paper), which makes our approach more efficient. The foreground blobs of pedestrians are extracted using background subtraction done in HSV rather than in the RGB color space to remove the shadow caused by the lighting. This shadow removal technique is sufficient for indoor environment where the shadows are small and diffused. The small 'holes' inside each foreground blob are removed using binary area openings [17], leaving the big ones to be the background.

B. Potential Head Top Segment Detection

Any segment, which partly contains the head top points, is called potential head top segment. A 3D head point is detected from a potential head top segment.

1) Head area existence probability

To detect a potential head top segment, we estimate the probability of a pixel being inside a head area. This probability is referred as the head area existence probability. We roughly model a person as a cylinder as illustrated in Fig. 2. P, with the height h, is a point on a person and the cylinder is the part of the person from the ground plane π to the height h. A$'$ is the area corresponding to the projection of the cylinder on the image plane π'. Fig. 3 shows a part of the image captured by the left camera, where O$'$ is the image center. From Fig. 2, O$'$ is also the vanishing point of all vertical segments on the image plane. 3D points located on the same vertical segment are projected on the same segment along the vanishing line. In Fig. 3, P$'$ is the image of P and the segment along the vanishing line $\overline{P'Q'}$ denotes the height h in the image plane. Usually the longer $\overline{P'Q'}$ and the more number of foreground pixels (highlighted in green) inside A$'$ indicate a larger probability that P$'$ is within a head area.

A polar mapping is performed to each left image to efficiently calculate both l (the length of $\overline{P'Q'}$) and the foreground area within A$'$. Fig. 4 shows a part of the image obtained by performing a polar mapping on the left image mentioned in Fig. 3. As a result, the image center O$'$ becomes a line (the solid blue shown in the first row). The segment along the vanishing line $\overline{P'Q'}$ becomes $\overline{P''Q''}$ that is in the same

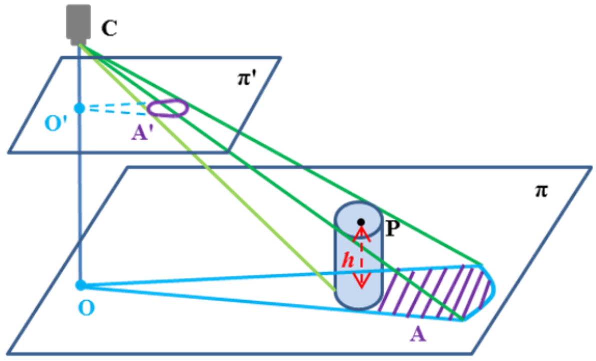


Fig. 2. The projection of a person on the ground and image plane

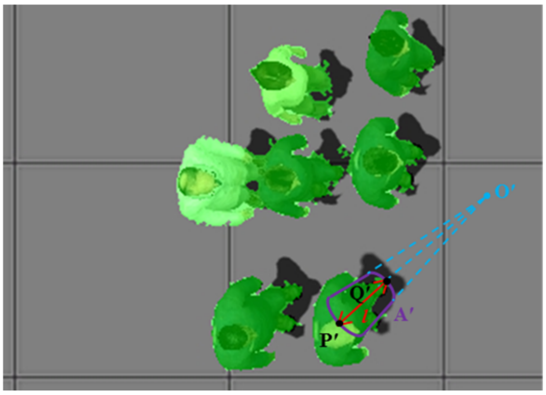


Fig. 3. Part of the image captured by the left camera

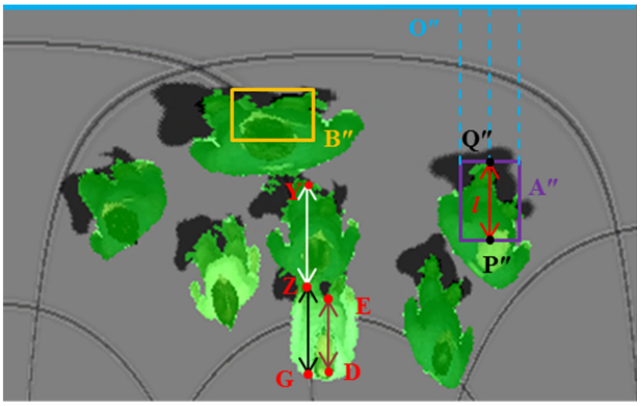
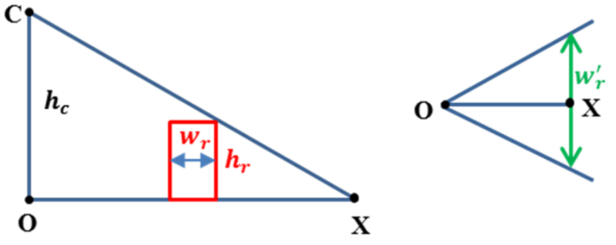


Fig. 4. Part of the image obtained by performing polar mapping to the left image in Fig. 3

column and has the same height l, and A′ is approximated by a rectangle A″ with the length being the height l and the width being the person's width (diameter of the cylinder model). Thus l can be evaluated by the number of foreground pixels right above P″, and the foreground area of A″ can be computed using integral image of binary foreground image.

Due to the perspective projection and polar mapping, the size of A″ associated with a fixed point on a person varies with his/her location. In Fig. 4, the rectangles A″ and B″ correspond to the same size cylinder but at different locations. So the size of A″ must be normalized to accurately evaluate the head area existence probability of each foreground pixel. Using a person



(a) A plane perpendicular to the ground

(b) The ground plane

Fig. 5. Pre-computing a normalization factor of the height (a) and width (b) for each foreground pixel

with the average height h_r and width w_r as reference, normalization factors of the height and width is pre-computed respectively for all pixels in the polar mapped image. The normalization factors of a pixel are the height and width in pixels associated with a head pixel. Since the widths of a person from the front and side views are different, and a person is modeled as a cylinder, w_r is obtained by fitting an average chest size into a circle. In Fig. 5, C is a camera with height h_c, O is its projection on the ground plane, and X is a projected point on the ground corresponding to a pixel X″ in the polar mapped image. Fig. 5(a) shows a plane perpendicular to the ground plane depicted in Fig. 5(b). If X″ is a head pixel, then the red rectangle in Fig. 5(a) is the cross section of the reference person with height h_r and width w_r, and $w'_r = h_c \cdot w_r/(h_c - h_r)$ in Fig. 5(b) is the projected width of w_r. The height and width normalization factors for pixel X″, $\eta_h(X")$ and $\eta_w(X")$ are computed as

$$\eta_h(X") = h_r \frac{d(\overline{O"X"})}{h_c} + d(w_r) \tag{1}$$

$$\eta_w(X") = 2tan^{-1}(\frac{d(w'_r)}{2d(\overline{O"X"})})/\Delta\alpha \tag{2}$$

where d(·) denotes the distance in pixels and $\Delta\alpha$ is the sampling angle for polar mapping. $d(\overline{O"X"})$ is the row number of X″ in the polar mapped image and O″ is image center after polar mapping, shown as the blue line in Fig. 4. For example, $d(\overline{O"P"})$ is the row number of pixel P″. The number of pixels corresponding to a segment on the ground plane can be obtained by simple camera calibration.

The height of a foreground pixel X″, H(X″), is obtained by counting the number of consecutive pixels right above it in the same foreground blob. The height is calculated column by column from up to down. If H(X″) is greater than $\eta_h(X")$ or the pixel one row above X″ belongs to the background, the pixel counting restarts from 0 (this also infers that H(X″) ≤ $\eta_h(X")$). In Fig. 4, $\overline{ED}$ denotes H(D) and the pixel counting starts from E because of the background pixels above E. H(G) is the length of $\overline{ZG}$ instead of $\overline{YG}$ since H(Z) = $\eta_h(Z)$. Thus occluded people can be roughly separated along a vanishing line (a column in the polar mapped image). The head area existence probability of a foreground pixel X″ in a polar mapped image is defined as

$$p_e(X") = \frac{N(X")}{\eta_h(X")\eta_w(X")} \tag{3}$$

where N(X″) is the number of foreground pixels inside the rectangle with its bottom side centering at X″, width being

$\eta_w(X'')$ and height being H(X″). Compared to H(X″), N(X″) can determine a possible head pixel better, since a head pixel needs to be not only 'high' enough but also around the vertical central axis of a person. N(X″) embeds the both features of a head point. In Fig. 4, H(G) is greater than H(D), but G does not necessarily have higher odds over D to be a head pixel as D is closer to the vertical central axis of the person.

2) *Potential head top segment*

We consider all pixels with the existence probability greater than a threshold t_p and not just those with the largest probability, as possible head area pixels. Since people are not perfect cylinders and a person's height and width can be somewhat different from the averages, the probabilities of the head area pixels of one person may be lower than those of another person. Some head area pixels will not be detected if just those with maximal probability are considered. The threshold t_p (set as 0.5 here) cannot be too high, since the computed existence probability of a true head area pixel is relatively low if part of a person is blocked by another. The region formed by those detected possible head pixels is the region of interest (ROI). Thus the ROIs mainly lie on the top central part of a person.

It's possible that more than an ROI exists in a foreground blob especially when it contains more than one person, thus ROI clustering is needed. To cluster the ROIs corresponding to a person, the polar mapped image is converted back to the original one. In Fig. 6, ROIs are shown as colored masks in a part of an original image (O′ is the image center), with the reddest indicating existence probability of 1 and the greenest a probability t_p. The pixels with high probabilities (those marked in red) are located not just in the head area but also on the shoulder and around the neck. ROIs with very small areas are considered not robust and removed before clustering. The clustering starts by merging the largest ROI with any other ROI whose centroid is within an average shoulder width to the centroid of the largest ROI. The clustering continues with the largest ROI of the remaining ROIs until all the ROIs are clustered. The number of persons in a foreground blob can be estimated from the number of clusters. The three persons around the middle of Fig. 6 are very close to each other and thus detected in the same foreground blob with the various ROIs clustered into 3 groups, G_1, G_2 and G_3, corresponding to the three persons. The centroids of the clustered ROIs, considered as the points of interest (POIs), are shown as blue dots. When a foreground blob contains the image center, it spreads after polar mapping over the entire rows, and the POI cannot be obtained using the aforementioned method. This happens when a person is very close to the FOV center. In this case, if the scene is not extremely crowded, s/he will not be occluded and the head area roughly locates at the center of the foreground blob, thus the ROI can be considered as the center part of the blob and the POI the blob centroid. Since an ROI (in either case) appears not just in the head area, the POI is not necessarily located at the head top center. To get an accurate 3D head point, a potential head top segment is established by extending the POI to both sides by a short length (e.g., the diameter of the head top) along the vanishing line and only the part lying on the foreground is considered. In Fig. 6, the POI P_1 is almost outside the head top but the potential head top segment (the white one) passes through the whole head top.

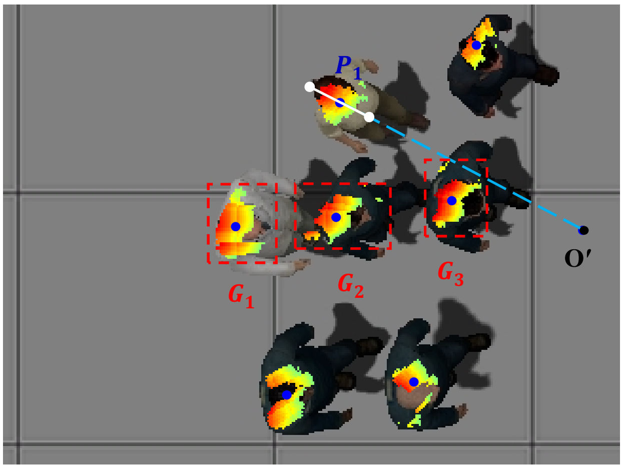


Fig. 6. Establishing a potential head top segment

C. *3D Head Position Estimation*

With the detected potential head top segment, locating the 3D head point reduces to finding points on the segment which are closest to the cameras and thus have the largest disparity.

To calculate the disparity of each pixel on the potential head top segment, its corresponding pixel needs to be found on the synchronized right image. For each pixel on the segment from the left image, we compare the RGB values of an N*N region about the pixel (the template) with a series of regions of the same size extracted from the right image (the samples). The center of each sample, the candidate matching pixel, has the same row number as the pixel in the left image, since the left and right camera are aligned horizontally. The search area on the row can be largely narrowed down with the disparity range of a head point. A pixel on the potential head top segment is described as an N*N*3 vector $\boldsymbol{L}$, containing the RGB values of all pixels in the template. The k^{th} candidate matching pixel is described by the same size vector $\boldsymbol{R_k}$. The similarity of the two vectors is evaluated by

$$S_k = ||\boldsymbol{L} - \boldsymbol{R}_k|| \tag{4}$$

A corresponding point of the point on the potential head top segment is established if

$$S_a < \gamma \cdot S_b \tag{5}$$

where S_a and S_b are the minimum and second minimum of S_k and γ is the similarity ratio (typically $\gamma = 0.8$). The disparity of the point on the segment is computed from the difference of the two matching pixels.

To get a more accurate 3D position, we estimate the sub-pixel disparity by considering S_a that satisfies (5) and its two neighboring values instead of just taking the point with minimum S_k as the matching point. A parabola is fitted to the three values and the minimum is analytically solved for to get the sub-pixel correction. The disparities on the potential head top segment may have outliers caused by mismatching of the pixels from the left and right image. The 3D head point detection will be highly affected if the outlier has the maximum disparity on the segment. A correct disparity distribution along the segment has only one peak because of the ‘Ω’ shape of the upper body. To remove the outlier robustly, the disparities are first rounded. If there are multiple local maxima in a rounded disparity distribution, those looking like spikes are considered as the outliers.

After outlier removal, the center of the pixels with the largest rounded disparity instead of only the pixel with largest disparity on the potential head top segment is determined as the head point. Both the disparity and the position of the head point have sub-pixel resolution, making the localization of the 3D head point more robust and accurate. With the head point in the image and its disparity (not rounded), the 3D head position can be computed by triangulation.

IV. Pedestrian Tracking

Once the 3D positions of pedestrians, denoted as the 3D head points, are obtained in each frame, they are tracked by assuming a constant moving direction and velocity within two consecutive frames. The position of a person is predicted at the next time interval and a search is implemented in a neighborhood around the predicted point. The position of the person is then updated by the estimated 3D head point that is nearest to the predicted point. If no head point is found in the search area, the person’s location is updated using the predicted one. A person is deleted if not found over certain extend periods of time. Similarly, if an object is not associated with any object in the previous frame over some frame intervals, it is regarded as a new object.

V. EXPERIMENTS

We test our approach using a publicly available visual surveillance simulation test bed, ObjectVideo Virtual Video (OVVV) [18]. OVVV is based on a commercial game engine, which can make human models behave like people in real world. It allows placing and configuring static and pan-tilt-zoom (PTZ) cameras freely and can generate the true 3D position of an object.

A virtual scene of a train station concourse is created, where two group people walk towards two different directions. 14 people walk in an area of about 4*4.5 m, which is a crowded scene. The foreground blobs of people merge from time to time even in the overhead view. The ceiling is 8.84 m high from the ground. Two identical horizontally aligned cameras are installed on the ceiling with the image planes parallel to the ground plane. The frame rate is 15 frames per second and the frame size is 640*480 pixels. Fig. 7 shows two frames captured by the left camera. The white dots are the projections of the detected 3D head points on the image plane. In both frames, people are close to each other and occlusions happen. Most of the detected head pixels are very close to head top centers.

To better show the tracking results, the estimated 3D tracks are projected onto the X-Y plane (ground) and Z plane (height) separately. Fig. 8 depicts the ground plane tracking results before anyone in Fig. 7 walks off the FOV. Thus the scene is generally crowded. The solid and dashed lines of the same color represent the ground truth and estimated trajectory of the same person, respectively. The brown dot is the FOV center. The number at one end of each trajectory denotes the object ID. The estimated trajectories are oscillating around but very close to the ground truth.

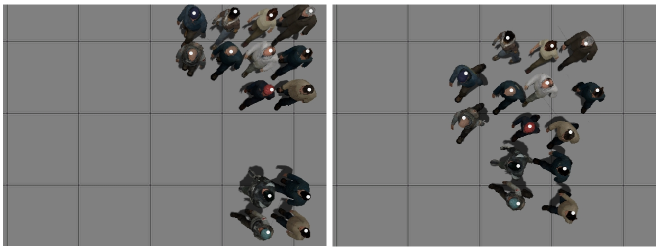


Fig. 7. Two frames captured by the left camera

The tracking errors are reported as the average distances between the true and estimated positions across all frames for each object. The smallest and largest four average ground plane tracking errors are tabulated in TABLE I together with the corresponding height and 3D tracking errors. TABLE II shows the overall tracking errors. To the best of our knowledge, the smallest error of the ground plane trajectories in others’ work is reported around 5 cm as in paper [5] where more than 2 side view cameras are used and the scene is sparse. Our ground tracking error (4 cm) is even smaller than that. From the two tables, we can see that the ground plane, height and 3D tracking errors are dependent. When estimating the 3D head position, the height of person is first calculated, and based on which the ground plane coordinates are obtained. The 3D head position errors mainly result from the fact that the estimated potential head top segment is slightly off the head top center. Pedestrians walking in the scene are not perfect cylinders; hence the foreground blobs are not completely symmetric about the vanishing line through the head top center. Although the detected head point misses the center, it still lies inside the head top. This results in an error on the X-Y plane that is usually smaller than the head top radius (about 8 cm on average for adults). Nevertheless, the estimated height is very close to the true value since the head top is relatively flat. This is also validated in TABLE I and TABLE II where the height errors are usually smaller than the corresponding ground plane errors.

To demonstrate that our accurately estimated 3D head positions can be helpful in face tracking and recognition, a PTZ camera with a 4m height, ground plane coordinate (-1, -9) (the same coordinate system shown in Fig. 8) and resolution 320*240 pixels is installed on the wall. The FOV of the PTZ camera is set small to 1.27*0.95 m to capture high quality close-up facial images. The pan and tilt angles of the PTZ cameras are determined by the ground plane position and the height of the target, respectively. In Fig. 9 the three close-up facial images from left to right are captured when person 3 arrives at the locations marked by the circles in Fig. 8. The arrow denotes the walking direction. Our method is very effective in capturing close-up facial image, with almost all the captured faces around the image center. Even for the second capture location where both X-Y and Z plane errors are relatively big, the whole face is still captured (the middle image in Fig. 9).

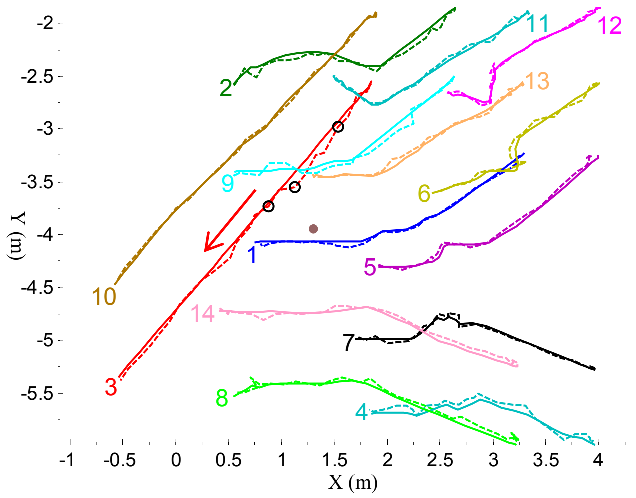


Fig. 8. The ground plane tracking results

TABLE I. THE SMALLEST AND LARGEST FOUR AVERAGE ERRORS OF THE ESTIMATED TRACKS

Object ID	Ground Plane Errors (cm)	Height Errors (cm)	3D Errors (cm)
11	2.40 ± 1.28	2.07 ± 0.62	3.31 ± 1.06
13	2.48 ± 1.59	2.21 ± 1.31	3.42 ± 1.90
12	2.83 ± 1.83	2.50 ± 1.96	3.89 ± 2.51
1	3.15 ± 2.32	4.07 ± 2.62	5.41 ± 3.07
8	4.90 ± 2.50	3.87 ± 2.54	6.36 ± 3.35
3	5.06 ± 2.47	3.56 ± 3.94	6.59 ± 4.03
4	5.30 ± 2.05	3.87 ± 2.74	6.83 ± 2.82
6	5.67 ± 1.72	3.95 ± 3.90	7.18 ± 3.76

TABLE II. THE OVERALL TRACKING ERRORS

Ground Plane Errors (cm)	Height Errors (cm)	3D Errors (cm)
4.02 ± 2.38	3.08 ± 2.80	5.34 ± 3.25

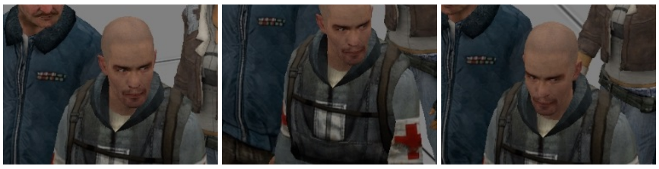


Fig. 9. Close-up facial images of person 3 captured by the PTZ camera

VI. CONCLUSIONS

We present an approach based on 3D head point detection to track pedestrians in a crowded indoor environment using two overhead cameras. We use the clustered possible head areas to establish the potential head top segment for each person inside a foreground blob and then detect the highest point(s) on the segment to estimate the 3D head position efficiently. The possible head area is determined by evaluating the head area existence probability based on the integral polar mapped image. Our method works well for accurate 3D tracking without using full disparity map of a scene which is computationally expensive. The experiments show that the average errors of the estimated ground plane positions and heights of the pedestrians are around 4 and 3 cm, respectively and that our method is well suited for capturing high quality close-up facial images.

REFERENCES

[1] J. Orwell, S. Massey, P. Remagnino, D. Greenhill and G. Jones, “A multi-agent framework for visual surveillance,” In IEEE International lst Conference on Image Processing, 1999.

[2] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale and S. Shafer, “Multi-camera multi-person tracking for easy living,” In Third IEEE International Workshop on Visual Surveillance, 2000.

[3] A. Mittal and S. Larry, “M2tracker: a multi-view approach to segmenting and tracking people in a cluttered scene,” International Journal of Computer Vision, 51(3), pp. 189-203.

[4] S.M. Khan and M. Shah, “A multi-view approach to tracking people in crowded scenes using a planar homography constraint,” In ECCV, 2006.

[5] L. Sun, H. Di, L. Tao, G. Xu, “A robust approach for person localization in multi-camera environment,” In International Conference on Pattern Recognition, 2010, pp. 4036-4039.

[6] S.M. Khan and M. Shah, “Tracking multiple occluding people by localizing on multiple scene planes,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 31 (3), pp. 505–519.

[7] R. Eshel and Y. Moses, “Homography based multiple camera detection and tracking of people in a dense crowd,” In IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-8.

[8] D. Delannay, N. Danhier, and C. D. Vleeschouwer, "Detection and recognition of sports(wo)man from multiple views," in Third ACM/IEEE International Conference on Distributed Smart Cameras, 2009.

[9] T.T. Santos, C.H. Morimoto, “Multiple camera people detection and tracking using support integration,” Pattern Recognition Letters, 32 (2011), pp. 47–55.

[10] O. Ozturk, T. Yamasaki and K. Aizawa, “Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis,” In International Conference on Computer Vision. Kyoto, Japan, 2009, pp. 1020-1027.

[11] N. Bellotto, E. Sommerlade, B. Benfold, C. Bibby, I. Reid, D. Roth, C. Fernandez, L. V. Gool and J. Gonzalez, “A distributed camera system for multi-resolution surveillance,” In ACM/IEEE International Conference on Distributed Smart Cameras, 2009.

[12] M. Boltes, A. Seyfried, B. Steffen and A. Schadschneider, “Automatic extraction of pedestrian trajectories from video recordings,” Pedestrian and Evacuation Dynamics 2008, Berlin, Germany: Springer.

[13] D. Beymer, “Person counting using stereo,” In Workshop on Human Motion, 2000, pp. 127-133.

[14] T. V. Oosterhout, S. Bakkes, B. J. A. Kröse, “Head detection in stereo data for people counting and segmentation,” In International Conference on Computer Vision Theory and Applications, 2011, pp. 620-625.

[15] T. V. Oosterhout, B. J. A. Kröse, G. Englebienne, “People Counting with Stereo Cameras - Two Template-based Solutions,” In International Conference on Computer Vision Theory and Applications (2) 2012, pp. 404-408.

[16] M. Boltes, and A. Seyfried, “Collecting pedestrian trajectories,” Neurocomputing 100(2013), pp. 127–133.

[17] L. Vincent, “Gray scale area openings and closings, their efficient implementation and applications", In Workshop on Mathematical Morphology Applications Signal Processing, 1993, pp. 22 -27.

[18] G. R. Taylor, A. J. Chosak and P. C. Brewer, “OVVV: using virtual worlds to design and evaluate surveillance systems,” In IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-8.

[19] Zhongchuan Zhang, Fernand Cohen, “Pedestrian tracking based on 3D head point detection,” In International Conference on Computer Vision Theory and Applications (2), 2013, pp. 382-385.