Human Area Refinement for Human Detection

Rong Xu(✉), Satoshi Ueno, Tatsuya Kobayashi, Naoya Makibuchi, and Sei Naito

KDDI R&D Laboratories Inc., Fujimino-shi, Saitama 356-8502, Japan
ro-xu@kddilabs.jp

Abstract. Human detection technologies are very useful tools to understand human activity for various purposes, such as surveillance. Recently, tracking-by-detection methods have also become popular for analyzing human activity, but their performance is greatly affected by the accuracy of detected human areas because they use online learning based on the detected results. In order to improve the performance of such tracking methods, the inclination of human bodies in the image is considered as a way to refine the detected human bounding boxes. Based on background subtraction and a novel scheme of estimating human foot position, a refinement scheme is proposed to estimate a bounding box more accurately, which can better fit the contours of inclined human bodies than the conventional method. Experimental results illustrated that the bounding boxes refined by the proposed algorithm achieved a higher cover rate of 92.7 % and a smaller mean angle error of 0.7° compared with the cover rate of 83.7 % and mean angle error of 3.8° obtained using the conventional method, as determined by comparison with the ground truth, and a real-time detection speed of 32.3 fps on a 640 × 480 video has been realized. Thus, tracking performance is significantly enhanced by refining the human areas, with a mean improvement of 42.4 % in the F-measure when compared with the conventional method.

Keywords: Human detection · Background subtraction · Foot position estimation · Refinement scheme · Human tracking

1 Introduction

In computer vision, human detection in still images and videos has become a very hot research topic in the last few years. It is critical in applications such as surveillance systems, assisted driving, robotics, and smart homes. It can also be used in shops, supermarkets and stores to count the number of people present and to analyze customer behavior and interactions with clerks for business optimization.

In recent decades, several methods have been proposed for human detection, in which some typical descriptors include a Histogram of Oriented Gradients (HOG) feature [1], an Integral Channel Feature (ICF) [2], Local Binary Patterns (LBP) [3], and the CENTRIST feature [4]. The effectiveness of these methods has been proven in practice for the detection of upright complete humans. With the development of human detection technologies, an approach called tracking-by-detection [5] has become popular recently. This approach treats the tracking problem as a detection task applied over time. Such a method learns classifiers for tracking online using detected

V. Murino and E. Puppo (Eds.): ICIAP 2015, Part II, LNCS 9280, pp. 130–141, 2015.
DOI: 10.1007/978-3-319-23234-8_13

human bounding boxes (b-boxes) instead of using offline labeled data for training, and thus the quality of the classifiers is greatly affected by the accuracy of the detected human areas, which contributes to the final tracking performance.

Although most detection methods can provide a high detection rate, accurate depiction of human postures and regions still cannot be achieved, i.e., all existing methods can only detect approximate human locations denoted by upright b-boxes, and cannot deal with the contour of an inclined human body very well. In order to improve the accuracy of the detected human areas, in this paper we propose a refinement algorithm for the detected human bounding box (b-box) to fit the contour of the inclined human body based on background subtraction, human detection, and a novel scheme of estimating human head and foot position using a predefined human height.

The rest of this paper is organized as follows. Section 2 briefly introduces related work. Section 3 describes the details of the proposed approach. Section 4 presents the experimental results and discussion, and Section 5 concludes the paper.

2 Related Work

Certain features are commonly used for human detection, such as, Haar features [7], edgelet [8], Integral Channel Feature (ICF) [2], HOG feature [1], LBP [3], and the CENTRIST feature [4]. Papageorgiou et al. [7] proposed a sliding window-based target detector combined with multi-scale Harr features, which identifies the object by the SVM classifier in [9]. Wu et al. [8] treated the human body as several body parts and proposed part-detectors learned by boosting a number of weak classifiers based on edgelet features, which can detect multiple and partially occluded humans. Dollar et al. [2] studied integral channel features coupled with a standard boosting algorithm for pedestrian detection, which can efficiently extract gradient and color channels from a transformed image to represent image features by computing the cumulative integral value of special channel areas. Dalal and Triggs et al. [1] provided a feature using a histogram of oriented gradients (HOG) for pedestrian detection. This method is effective, and has reduced the missed detection ratio by at least one order of magnitude, relative to the Harr-based detector. Mu et al [3] improved the original LBP descriptor by proposing two variants of LBP: Semantic-LBP and Fourier-LBP for human detection, and achieved performance comparable to other descriptors based on the INRIA human database. In addition, real-time detection has attracted more and more attention, e.g., the CENTRIST feature [4] achieves a much higher speed than existing human detectors, and can detect humans on a 640×480 video at 20 fps using an ordinary CPU. In pursuit of a better detection rate, combining multiple information sources has become a trend, e.g., HOG-LBP [11] achieves the best detection rate in the literature (about 86%), while multiple information increases the cost in terms of detection time.

However, all current methods focus on detection rate and efficiency, and none of them can detect human body areas with sufficient accuracy to fit the contour of an inclined human body, which results in poor tracking performance for the tracking-by-detection method [5]. In order to resolve this problem, we propose a refinement

system to generate human body areas more accurately using the detected b-boxes. Since detection efficiency is also a critical factor in tracking performance, the fastest existing algorithm for human detection is employed in this paper, i.e., the CENTRIST feature [4] with a detection rate of 83.5%, which is comparable to the state-of-the-art [2,11]. The refined human areas will be produced in real-time by the proposed method and applied as learning and tracking targets to improve the tracking performance of the tracking-by-detection method [5].

3 Proposed Method

3.1 Basic Ideas

There are two advantages of the proposed method compared with other methods:

(1) A novel refinement scheme is proposed to estimate human areas more accurately to fit the contour of an inclined human body;

(2) Real-time detection and refinement of human areas can be realized due to the high efficiency of the CENTRIST feature [4], and by detecting humans and computing an integral image only on the foreground extracted by the Radial Reach Filter (RRF) [6].

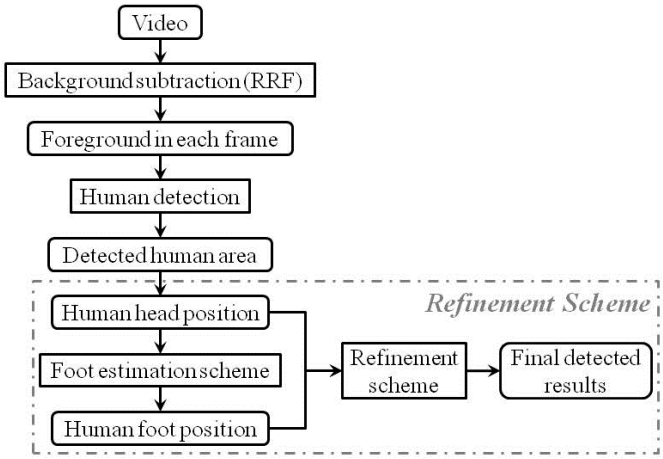


Fig. 1. Flow chart of the proposed method

The flow chart of the proposed method is shown in Fig. 1. First, RRF [6] is utilized to extract the foreground of each frame from an indoor video, which can detect new objects in a time-series image even if they stop moving after they enter the scene. Next, the CENTRIST feature [4] is applied to detect humans only in the foreground regions of each frame, to improve the computational efficiency. After human areas are detected, the corresponding human head position is estimated from the detected b-box, which is more reliable than the foot position extracted from the b-box, since feet are easily occluded by tables or other objects. Subsequently, the human foot position

is calculated based on the estimated human head position and the predefined human height by the proposed foot estimation scheme. Finally, the refinement scheme will create a refined b-box to fit the contour of the inclined human body.

3.2 Coarse Foot Position Estimation

In order to estimate foot position, a coarse foot position estimation scheme is proposed based on background subtraction by RRF [6], human detection by the CENTRIST feature [4], and projection and back-projection by direct linear transformation (DLT) [13], as depicted in Fig. 2. In Step-1, the human head position in the image is extracted from the detected b-box, i.e., the top and central point of the b-box (P_h in Fig. 4 (c)). In Step-2, the human head position in a 3D world coordinate system is calculated by back-projection based on the projection matrix estimated by DLT [13]. In Step-3, a coarse foot position is estimated from the human head position in the 3D world coordinate system, which will be projected onto the image to get the coarse foot position in the image in Step-4.

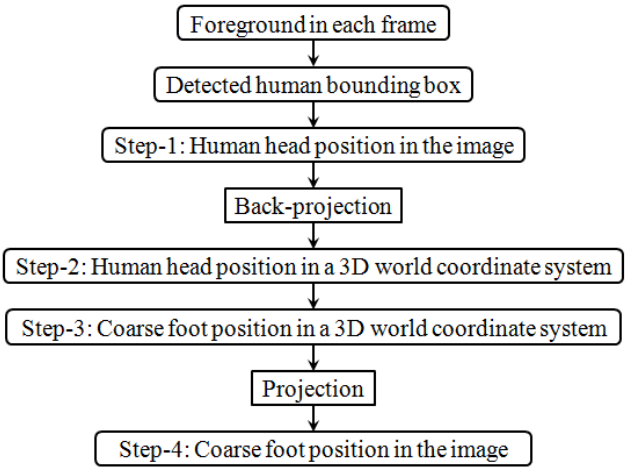


Fig. 2. Flow chart of the coarse foot position estimation

Background Subtraction. Radial Reach filter (RRF) [6] is an effective statistical measure for robust object detection. It can detect new objects in a time-series image even if they stop moving after they enter the scene.

For usual background subtraction methods, to detect new objects they simply subtract the current image from a background image. However, simple background subtraction is easily affected by illumination changes such as shadows. Furthermore, when the brightness difference between objects and a background is small, it cannot detect the difference. In order to solve such problems, the RRF method evaluates a local texture, i.e., measurement of the brightness difference between one pixel and its adjacent neighbors, and realizes robust object detection.

Fig. 3 illustrates one example of background subtraction by RRF, where (a) is the background image, (b) is one scene, and (c) is the detected foreground. Although there is some additional noise as shown in red ellipses in Fig. 3 (c), there is no negative impact on human detection since all of the persons in the scene have been successfully extracted from the background.

Human Detection. CENTRIST is short for CENsus TRansform hISTogram, and has been used for human detection by a cascade classifier called C^4 [4]. The CENTRIST visual descriptor can succinctly encode the crucial sign information (signs of local comparisons) and implicitly encodes the global human contour, and thus it is a suitable representation for detecting human contours. For CENTRIST, the histogram intersection kernel [12] is used to compute similarity scores, which will be used in the refinement scheme for selecting the best detection result described later in Section 3.3.

In this paper, the CENTRIST feature [4] is applied only to regions of interest (ROIs) extracted from the foreground in each frame by RRF [6]. Fig. 4 shows an example of detection, where (a) is the original image, and (b) is the detected results.

(a) (b) (c)

Fig. 3. An example of results from RRF

(a) (b) (c)

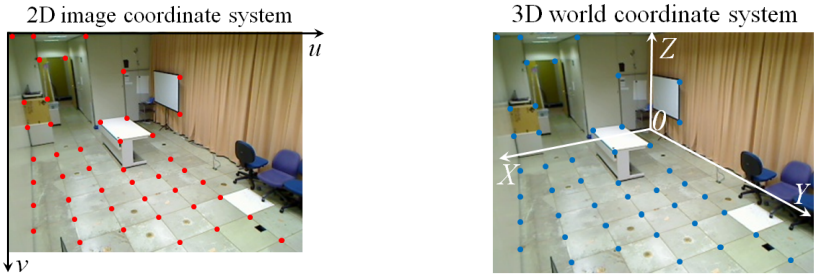
Fig. 4. Results of human detection

Projective Transformation. In order to perform projective transformation from a 3D world coordinate system (Fig. 5 (b), where the units are meters) to a 2D image coordinate system (Fig. 5 (a), where the units are pixels), a projection matrix is estimated in offline processing. In Eq. (1), the projection matrix [P] between the 3D

world coordinate system and the 2D image coordinate system is estimated by the Direct Linear Transformation (DLT) method [13],

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = [P]\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \end{bmatrix}\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (1)$$

where, s is the scale factor, (u, v) gives the coordinates of one pixel in the image, and (X, Y, Z) the corresponding coordinates in the world coordinate system. About 40-50 points (red points in Fig. 5 (a)) are selected from the image, and the corresponding points in the 3D world coordinate system (blue points in Fig. 5 (b)) are measured manually. In order to simplify the estimation process, we choose a regularly shaped room, and its floor is composed of regularly shaped floor tiles each of which is 0.6m by 0.6m, as shown in Fig. 8. Then the corners of the floor tiles and other stationary objects (e.g., table, whiteboard, wall, etc.) are extracted from the image for projective transformation, which can be easily distinguished from the image and the real room. On the other hand, the 3D world coordinate system of the room is constructed as shown in Fig. 5 (b), where the origin is set at one corner of the room, the X-Y plane is the floor, the X and Y axes are parallel to the respective sides of the rectangular floor, and the Z axis is vertical to the floor. Accordingly, we measure actual distances in meters between each corresponding point in the real room and X, Y, and Z axes to get their 3D coordinates.



(a) (b)

Fig. 5. 2D image coordinate system and 3D world coordinate system

Coarse Foot Position Estimation. Based on the human head position in the age($P_h(u_h, v_h)$ in Fig. 4 (c) and Fig. 6) and the estimated projection matrix [P], a human head position (V_h in Fig. 6) is calculated in the 3D world coordinate system by back-projection, using a predefined human height (1.7 meters in experiments), to define the head's z position in the 3D world coordinate system. Considering only upright humans in the scene, a coarse foot position (V_f in Fig. 6) in the world coordinate system can be estimated from the head position $V_h(\tilde{X}, \tilde{Y}, \tilde{Z})$ by setting $\tilde{Z} = 0$, i.e., the coarse foot position is $V_f(\tilde{X}, \tilde{Y}, 0)$. Finally, the coarse foot position ($P_f(u_f, v_f)$ in Fig. 6) in the image is computed from V_f by projection using the estimated projection matrix [P].

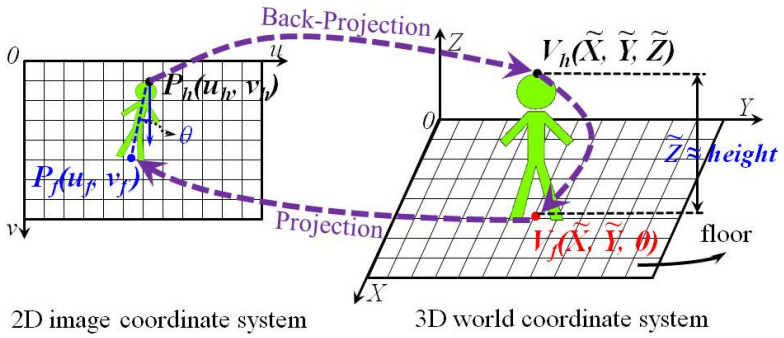


Fig. 6. Foot Estimation Scheme

Note that a predefined human height of 1.7 meters will lead to some error in estimated foot positions since actual human heights will differ. However, such errors can be removed by the refinement scheme because a human re-detection mechanism will be utilized to estimate foot positions more accurately.

3.3 Refinement Scheme

Based on the b-box initially detected by the CENTRIST feature [4], the extracted head position (P_h) and the estimated foot position (P_f), a new ROI is created the uppermost and lowermost centers of which are P_h and P_f, respectively, with a width equal to the width of the initial b-box(ROI1 in the leftmost image of Fig. 7). The steps of the refinement scheme are shown in Fig. 7 and are as follows:

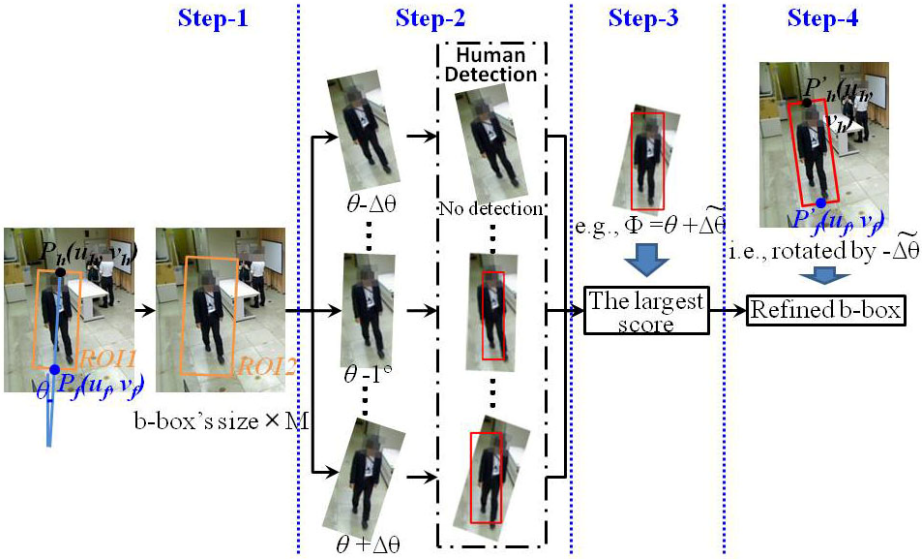


Fig. 7. Example of the refinement scheme

Step 1. ROI1 is enlarged M times to obtain a new region of ROI2 for human detection, the center of which is the same as that of ROI1, and the width and height are M times larger than those of ROI1. Meanwhile, angle θ of ROI1 is calculated from the lines from P_h to P_f and the v axis in the 2D image coordinate system. Here, M = 1.2 is selected based on experimental experience.

Step 2. The area of ROI2 in the image is extracted, and rotated in the range of $[\theta - \Delta\theta,\ \theta + \Delta\theta]$ in increments of τ in a clockwise direction. Thus a number of cases corresponding to different rotation angles are examined by the CENTRIST descriptor, and their similarity scores calculated by the histogram intersection kernel [12] will be recorded if a human is detected as shown in the dashed rectangle (Human Detection) in Fig. 7. If no human is detected, the corresponding similarity score is set to 0. Here, $\Delta\theta = 15°$ and $\tau = 2°$ are selected based on experiments.

Step 3. All scores from those cases are compared, and the detected result with the largest score is selected as the best. Also, the rotation angle of the image corresponding to the best detected result is recorded as $\Phi = \theta + \widetilde{\Delta\theta}$, as shown in Fig. 7.

Step 4. The refined result is achieved by rotating the best detected result (b-box) counter-clockwise around its center by an angle of $\widetilde{\Delta\theta}$. Then a fine head position (P_h') and foot position (P_f') are extracted from the refined b-box using its uppermost and lowermost central points, as shown in the right most image of Fig. 7.

4 Experiments

In the experiments, a camera was set in each of four corners of an indoor laboratory to cover all areas of the room, the layout of which is shown in the upper right image in Fig. 8. Fig. 8 also shows a set of four captured images. To simulate the recording of customer behavior in a shop, two groups of videos were recorded for human detection and tracking, with each group containing four videos captured by the four cameras. The first group shows a simple case with four people in the scene. The second group is a more complex case with ten people in the scene. Each 9-minute video was shot at 10 fps, and thus contains about 5400 frames. For one group, all of the 2D pixels in each video can be back-projected into the same 3D world coordinate system by the estimated projection matrices. Thus, humans are detected and tracked in each video, and integrated in the 3D world coordinate system.

4.1 Evaluation of Projection Transformation

As the basis of the proposed method, the accuracy of the projective transformation is very critical for estimating coarse foot position, which contributes to the accuracy of the refinement scheme and the final results. For this reason, we use the reconstruction error to measure transformation errors, specifically the RMS distance between the reconstructed coordinates and the measured ones (i.e., the ground truth). In experiments, the estimated fine foot positions in Section 3.3 from the four cameras are back-projected

into the 3D world coordinate system, and compared with the ground truth measured manually to calculate the reconstruction error, where the mid-point between the feet is considered to be the foot position. The mean and standard deviation of the reconstruction error of the cameras was $0.28 \pm 0.19\ meters$, which is sufficiently acceptable for practical applications.

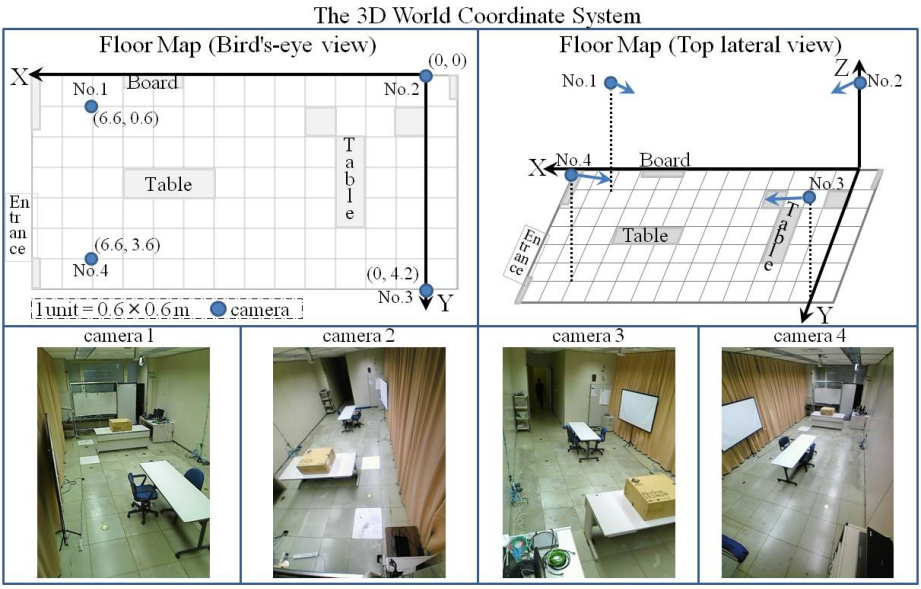


Fig. 8. Experimental environment

4.2 Evaluation of Refined B-Boxes

To evaluate the accuracy of the results, the ground truth was extracted manually for each person, and the following criteria were calculated to measure the similarity between the result and the ground truth for all videos. The *cover rate* equals the size of the overlap area of the result and the ground truth divided by the size of their union, i.e., $\frac{G \cap R}{G \cup R}$ in Fig. 9, where the blue (G) and red (R) b-boxes represent the ground truth and the result. The *mean angle error* is the mean of angle ε in Fig. 9.

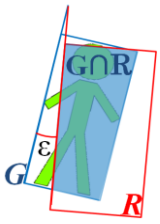


Fig. 9. Similarity criteria

In Table 1, the proposed method achieves a much better accuracy than the conventional method [4] for each criterion. Meanwhile, the standard deviations of each criterion show that the proposed method is more robust than the conventional method [4]. In addition, the proposed method is highly efficient at human detection, due to the combination of the conventional method [4] and ROIs of the foreground extracted by RRF [6], which yields a mean processing time of $31 \pm 14\,ms$ (32.3 fps) on a 640×480 video using 4 processing cores of a 3.5GHz CPU. This shows that real-time detection can be achieved by running the proposed method on a common desktop PC. Moreover, it is faster than the conventional algorithm [4] (i.e., 25.5 fps on a 640×480 video by the same PC). The higher efficiency of the proposed method is attributed to detecting humans and computing an integral image, two of the most time-consuming parts of the algorithm, only on the foreground.

Table 1. Accuracy of the detected results

	Cover rate	Mean angle error
Ref. [4]	$83.7 \pm 16.6\%$	$3.8 \pm 3.5°$
The proposed method	$\mathbf{92.7 \pm 9.0\%}$	$\mathbf{0.7 \pm 1.8°}$

4.3 Evaluation of Tracking Performance

The performance of the tracking-by-detection [5] method is strongly affected by the detected human areas. To evaluate how the proposed method affects tracking performance, the tracking quality of the conventional method [4] and the proposed method are compared. These are frequently estimated from three fundamental measurements: precision, recall, and F-measure [14]. High values of precision, recall, and F-measure indicate good tracking quality, where the F-measure that is calculated from the harmonic mean of precision and recall is the best measurement. The ground truth of human tracking is manually generated.

Since four cameras are used for human detection and tracking, integration of detection results from different cameras is critical for human tracking. In our tracking system, we calculate physical distances between detected humans from different cameras and tracked persons in a tracking list to judge whether they are the same person. Herein, we use local and global to separate detected humans from the cameras and tracked persons in the 3D world coordinate system. For example, we suppose that local person A, B, C, and D are detected by camera 1, 2, 3, and 4, respectively. First, the local person A's foot position is estimated in the 3D world coordinate system based on the projection and back-projection transformation, as described in Section 3.2 and 3.3. Then the distances between the local person A and the global tracked persons registered in the tracking list will be computed, and the minimal distance between the local person A and some global tracked person will be selected. If such a minimal distance is smaller than one threshold (in the experiments we use 0.5 meters), then the local person A will be considered to be the same person who has been registered in the tracking list. Meanwhile, tracking information of the corresponding tracked person will be updated based on the local person A detected by camera 1. If such a minimal distance is larger than the threshold, then a new person will be added to the tracking list.

Subsequently, the local person B, C, and D will be integrated into the tracking list in the same way, and the corresponding tracking information will be updated. However, the position of the global tracked person will be updated after checking all detected results from four cameras. For example, if the local person A, B, C, and D are assumed to be the same person, then the position of the corresponding global tracked person is updated by calculating the central point of four persons' positions in the 3D world coordinate system. Therefore, as minimum conditions, our tracking system requires the synchronization of four fixed cameras, projection transformations estimated between each camera and a global 3D world coordinate system, and background images captured by each camera for object extraction.

The tracking results are listed in Table 2, and we find that compared with the tracking results of the conventional method [4], improvements of about 45.9%, 18.6%, and 32.5% for precision, recall, and F-measure, respectively, are achieved by the proposed method using in the videos of the first group, and improvements of about 63.2%, 41.4%, and 52.2% for precision, recall, and F-measure, respectively, are achieved in the videos of the second group. A mean 42.4% improvement in F-measure was achieved by the proposed method for tracking performance, compared with the conventional method [4], showing that the refined human b-boxes of the proposed method can contribute to a significant improvement in tracking performance, especially for cases involving more persons.

Table 2. Comparison of tracking accuracy

	Videos of the first group		Videos of the second group	
	Ref. [4]	Proposed method	Ref. [4]	Proposed method
Precision	0.37	**0.54**	0.19	**0.31**
Recall	0.43	**0.51**	0.29	**0.41**
F-measure	0.40	**0.53**	0.23	**0.35**

Although the proposed method has been compared with only one conventional method [4] by b-box similarity and tracking performance, this was sufficient to verify its superiority because all other detection methods are like the conventional method [4] in only providing similar upright b-boxes for human detection.

5 Conclusions

In this paper, we have proposed a novel approach to refine b-boxes to better fit the contours of inclined human bodies based on background subtraction technology, human detection technology, and a novel scheme for estimating human foot position. The results showed that the proposed approach performs well at extracting human areas accurately, i.e., a cover rate of 92.7% and a mean angle error of 0.7° compared with the ground truth, and it contributes to a roughly 42.4% improvement in tracking performance. Moreover, a real-time detection speed of $31 \pm 14\,ms$ on a 640×480 video has been achieved. In the future, accuracy of head position estimation will be further improved by introducing human head detection technology. Also, some open datasets such as PETS 2009 will be used for evaluating the proposed approach.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
2. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: BMVC, vol. 3, p. 5 (2009)
3. Mu, Y., Yan, S., Liu, Y., Huang, T., Zhou, B.: Discriminative local binary patterns for human detection in personal album. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
4. Wu, J., Geyer, C., Rehg, J.M.: Real-time human detection using contour cues. In: 2011 IEEE International Conference on Robotics and Automation (ICRA), pp. 860–867 (2011)
5. Hare, S., Saffari, A., Torr, P.H.: Struck: structured output tracking with kernels. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 263–270 (2011)
6. Satoh, Y., Tanahashi, H., Wang, C., Kaneko, S.I., Niwa, Y., Yamamoto, K.: Robust event detection by radial reach filter (RRF). In: 16th International Conference on Pattern Recognition, pp. 623–626 (2002)
7. Papageorgiou, C., Poggio, T.: A trainable system for object detection. International Journal of Computer Vision **38**(1), 15–33 (2000)
8. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. International Journal of Computer Vision **75**(2), 247–266 (2007)
9. Maji, S., Berg, A.C., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
10. Zhu, Q., Yeh, M.-C., Cheng, K.-T., Avidan, S.: Fast human detection using a cascade of histograms of oriented gradients. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1491–1498 (2006)
11. Wang, X., Han, T.X., Yan, S.: An HOG-LBP human detector with partial occlusion handling. In: IEEE 12th International Conference on Computer Vision, pp. 32–39 (2009)
12. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision **7**(1), 11–32 (1991)
13. Shapiro, R.: Direct linear transformation method for three-dimensional cinematography. Research Quarterly American Alliance for Health, Physical Education and Recreation **49**(2), 197–205 (1978)
14. Smith, K., Gatica-Perez, D., Odobez, J.-M., Ba, S.: Evaluating multi-object tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 36–43 (2005)