J. Vis. Commun. Image R. 35 (2016) 36–54

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

People tracking in an environment with multiple depth cameras: A skeleton-based pairwise trajectory matching scheme ☆

Shih-Wei Sun [a,b,*], Chien-Hao Kuo [c], Pao-Chi Chang [c]

[a] Department of New Media Art, Taipei National University of the Arts, Taipei, Taiwan
[b] Center for Art and Technology, Taipei National University of the Arts, Taipei, Taiwan
[c] Communication Engineering, National Central University, Jong-Li, Taiwan

ARTICLE INFO

Article history:
Received 10 December 2014
Accepted 18 November 2015
Available online 2 December 2015

Keywords:
People tracking
Multiple depth cameras
Trajectory matching
Fusion
Skeleton
Pairwise
Occlusion
Hand gesture

ABSTRACT

This paper proposes a pairwise trajectory matching scheme from multiple cameras for people tracking, handling the mistracking situations caused by occlusion events occurred in one of the cameras. In a multiple cameras environment, a geometric calibration process is necessary for the co-plane of the overlapping field of views from different cameras as the initial step. Once the geometry is calibrated, according to the 2D positions of the analyzed foot joints from the depth cameras. Homography transformation is applied to project the detected foot points from different views into a synergistic virtual bird's eye view for people tracking. At the virtual bird's eye view, the people tracking results from each of the cameras based on Kalman filter are fused according to the proposed pairwise trajectory matching scheme. The contribution of this paper is trifold: (1) The proposed hand-gesture-triggered calibration process with temporally synchronization capability can effectively build and calibrate the geometry in a region of interest. (2) The proposed interleaving-based skeleton obtaining and moving average based valid skeleton determination can extend the skeleton tracking capability to track more people. (3) The proposed pairwise trajectory matching scheme effectively manages occlusion situations happened in one of the depth cameras. In addition, in the extensive experimental results, the proposed method can track up to six simultaneously freely moving persons in the field of view, with affordable complexity for real-time applications. Furthermore, the infrared-based depth cameras track people satisfactorily from bright to extremely dark environments.

1. Introduction

People tracking in a camera monitoring environment attracts intensive attention from researchers and engineers in the field of video surveillance, video understanding, and behavior analysis. When more cameras are adopted to monitor a region of interest, such as the surrounding area of an ATM, the area of a checked baggage inspection, or the entrance of a shopping mall, people tracking becomes a challenging task since the people severely occluded by others causes the mistracking issue and incorrectly assign the correspondence in the tracking process. On the other hand, when the above mentioned areas with different lighting conditions need to be monitored, people tracking becomes another challenging task, because the color distribution of the target people and the background color model are quite different in bright environment and in dark environment.

To achieve people tracking, researchers started to utilize a single camera [1–6] based on the color information. However, unstable lighting conditions and complex background cause imperfect people detection and tracking results. When considering multiple cameras facing to the same region of interest for people tracking, spatial geometries among the cameras [7–9] are built as the initial process and combine with proper tracking algorithms. Nevertheless, imperfect built spatial geometries and occlusion situations make the tracking task challenging.

To achieve people tracking in complex environments and cluttered backgrounds, an RGB-D camera, Kinect [10] of Microsoft, is proposed for people tracking, pose recognition, and behavior analysis. Kinect has one color (RGB) camera, one infrared emitter with structured light, and one infrared receiver (so-called depth camera), to monitor and analyze the people in front of it. In this paper, the Kinect cameras with the official Kinect software development kit (sdk) [11] are adopted for reliable people detection. As shown

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author at: Department of New Media Art, Taipei National University of the Arts, Taipei, Taiwan.
E-mail address: swsun@newmedia.tnua.edu.tw (S.-W. Sun).

http://dx.doi.org/10.1016/j.jvcir.2015.11.012
1047-3203/

in Fig. 1(a), based on the Kinect sdk, the Kinect cameras mounted in different locations (c_1, c_2, and c_3) can detect and track the people reliably from the depth channel when no occlusion event happened. However, as shown in Fig. 1(b), when one person was occluded by another, the occluded person could not be seen in c_2. Meanwhile, the skeletons and joints of the occluded person cannot be detected for further tracking.

In this paper, Kinect cameras with infra-red emitter and receiver modules are adopted for reliable people detection both in bright environment and in dark environment. The gesture recognition results based on the joints and skeletons revealed from Kinect sdk are utilized for building the spatial geometry in a multiple cameras environment. Furthermore, a pairwise trajectory matching scheme is proposed for managing occlusion events by compensating from the occlusion-free cameras. Therefore, the contribution of this paper is trifold: (1) a hand-gesture-triggered spatial calibration process, (2) a pairwise trajectory matching to manage occlusion events, and (3) accurate people detection and tracking capabilities, compared to the state-of-the-art methods, with affordable for realtime applications.

The remainder of this paper is organized as follows. Related work is discussed in Section 2. The proposed process for calibrating the geometries of the multiple depth cameras is explained in Section 3. In Section 4, the proposed people tracking system, including occlusion detection and pairwise trajectory matching among cameras, is described in detail. The experimental results are reported in Section 5, and finally, a conclusion is provided in Section 6.

2. Related work

According to the relevant literature, people tracking started from using a single camera [1–6], based on obtaining the correspondences from successive frames. The statistics of geometric distribution [1–3] and color distribution [4–6] of the detected blobs with further analysis and filtering process are utilized to determine the detected people area. However, the above-mentioned color-based approaches are sensitive to color changing, especially for the scenarios with unstable lighting conditions. When noticing the people detection accuracy degrading situations caused by unstable visible lighting conditions, Shotton et al. [12] applied a Kinect camera with invisible light (infrared) emitter and receiver to measure the depth in different planes, i.e., a 3D space, for people tracking. The detection and tracking tasks are not only achieved to the person level, but also to the joints and skeletons level. Even so, when using a Kinect camera, the occlusion situations still cause the mistracking issues, due to the front view mounting criteria. To suppress the occlusion effects, Ozturk et al. [13] mounted a single camera at a bird's eye view, using the motion vectors, optical flows of scale-invariant feature transform (SIFT) [14] points, color features, and edge histograms for people tracking. Furthermore, Baum et al. [15] mounted a Kinect camera from a bird's eye view to detect objects according to a conventional frame differencing technique [1]. Despite that, in many environments with a high ceiling or no ceiling, mounting a camera to obtain a bird's eye view is difficult, limiting the capability of this kind of approaches.

It is feasible to mount multiple cameras at the side views for people tracking, when single view approaches meet the limitations. Hu et al. [7] proposed a principal axis-based correspondence examination among multiple cameras, based on homography transformation [16,17] for cross camera matching. To handle the occluded people, Khan and Shah [8] proposed to localize on multiple scene planes, with a planar homography occupancy constraint and the foreground likelihood information extracted from different views. Berclaz et al. [9] proposed a probabilistic occupancy map across consecutive frames to estimate the most likely trajectories of an unknown number of targets, including entrances and departures. Though, increasing the unit square size of the grid defined in their paper reduces precision, and reducing the unit square size increases precision as well as computation time.

In addition, the conventional probability occupancy maps [18] is not applied in color cameras [9], but also extended to deal with depth map [19] from a single depth camera. The authors propose a generative model to predict the distribution of depth images to deal with the potentially occluding objects in a scene. However, as mentioned by the authors, to process a single depth frame needs several seconds on a 2.3 GHz Intel CPU would limit the range of the real-time applications. For real-time applications, Jafari et al. [20] proposed a depth-based upper body detector to be mounted on mobile robots and head-worn cameras. The object detect and tracking can be achieved to 24 fps, using a GPU-based ground HOG detector, the system still provide 18 fps results. On the other hand, people tracking from the RGB-D data captured from mobile service robots, methods by Munaro and Menegatti [21] provided 26 fps detector and tracker results. As a result, using depth cameras to track people in the field of view from multiple cameras with real-time response becomes an important research issue.

The Kalman filter [22,23] and particle filter [24,25] are widely used for object tracking [26] to obtain the correspondences across frames from a monocular camera, based on the statistical properties of the observations. Furthermore, Gruenwedel et al. [27] utilized Kalman filter to track humans in a multi-camera environment for the indoor and meeting scenarios. In addition, Sharma and Moon [28] proposed a SIFT-based object tracking algorithm for video frames. Nevertheless, the long-term occlusion of the observation from a single camera causes a complete trajectory to be cut into broken slices of trajectories. Furthermore, the erroneously added trajectories from different cameras engender errors in cross-camera tracking.

Therefore, in this paper, multiple Kinect (depth) cameras mounted at side views with official sdk [11] are utilized for people detection [12] and the proposed pairwise trajectory matching for people tracking to deal with the occlusion problem occurred in the observed field of view.

3. Hand-gesture-triggered geometry calibration

In the initialization process of the proposed method, the geometries among multiple cameras must be calibrated. Hence, as shown in the top section of the flow chart in Fig. 2, the Kinect sdk [11] is applied to detect people and track the corresponding skeletal joints from the signals in the depth channel.

Joint $j^i_{t,c_k} = \left(x^i_{t,c_k}, y^i_{t,c_k}\right)$ represents the 2D position belonging to the ith joint from the kth camera c_k at time t in the captured 2D plane. In this hand-gesture-triggered multiple-camera calibration process, upward motion by the right hand triggers a signal in the server. For example, at time t at the kth camera c_k, upward motion by a right hand $g_{c_k}(t)$ is detected and the signal is triggered and sent to the server, which records the misalignment of the right hand joints. Upward motion of the right hand is expressed as follows:

$$g_{c_k}(t) = \begin{cases} 1, & \text{if } y^{i=RH}_{t,c_k} > y^{i=HD}_{t,c_k}; \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $y^{i=RH}_{t,c_k}$ represents the height of the right hand joint ($i = RH$) $j^{i=RH}_{t,c_k}$ from the kth camera c_k, and $y^{i=HD}_{t,c_k}$ represents the height of the head joint $j^{i=HD}_{t,c_k}$ belonging to the same camera. Hereafter, $j^{i=RH}_{t,c_k}$ is simplified as j^{RH}_{t,c_k}, and $y^{i=HD}_{t,c_k}$ is represented as y^{HD}_{t,c_k}.

The color image frames and the corresponding foreground detection results from the depth frames shown in Fig. 3 depicting

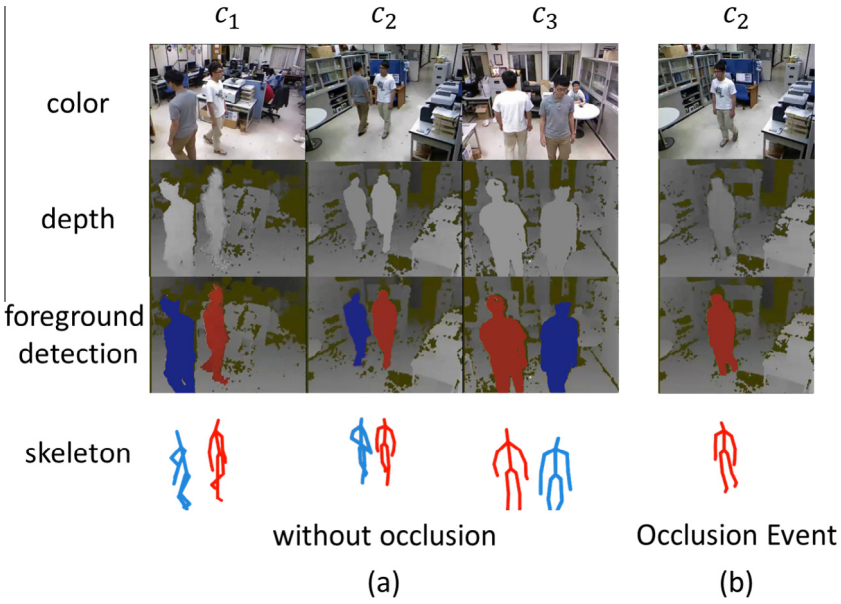


Fig. 1. The people detection and tracking results from the Kinect sdk [11] (a) without occlusion, and (b) an occlusion event occurs in c_2.

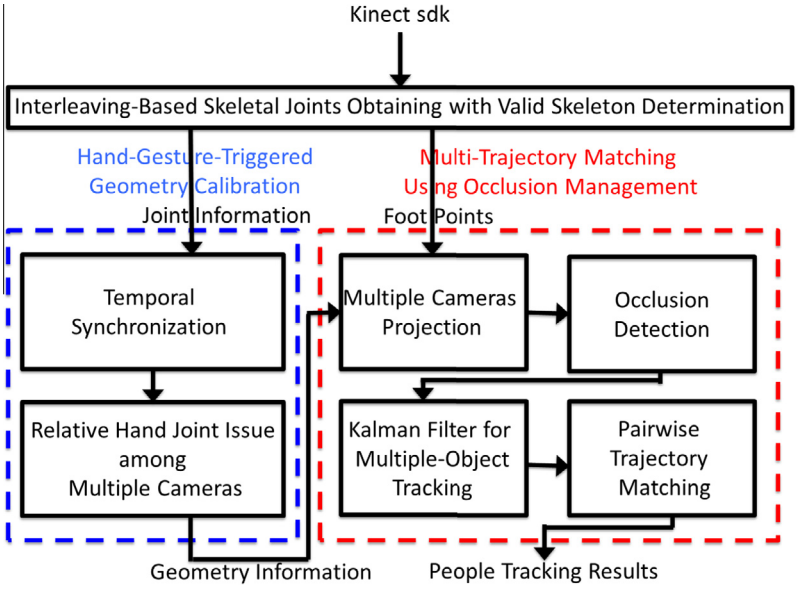


Fig. 2. The flowchart of the proposed people tracking system.

the motions captured from different cameras. The lifted hand of the user in camera c_2 is self-occluded by the user. Under this kind of situation, the trigger motion cannot be recognized from c_2, causing an invalid triggering issue. However, the upward motion can be properly recognized in all views of cameras, as shown by the rightmost column of Fig. 3. Therefore, in this paper, an upward motion is used for triggering the geometry calibration.

3.1. Temporal synchronization

The trigger signal defined in Eq. (1) is sent to the server through the network. When network traffic is heavy or the wireless environment is complex (doors and walls affect the wireless signal through reflection or diffusion), the arrival time of the triggered signal $g_{c_k}(t)$ from camera c_k to the server cannot be synchronized because of delays, jitters, or packet loss. To alleviate the temporal noise caused by environmental complexity, the successively received trigger signal $g_{c_k}(t)$ belonging to camera c_k is observed over a period of time in a sliding window with size w. The reception of a valid successive trigger signal from camera c_k is expressed as

$$s(t)=1,\quad g_{c_k}(t)=1\ \forall\left\{t-\frac{w}{2}<t<t+\frac{w}{2}\right\}\ \text{and}\ \forall k\in K, \tag{2}$$

where K is the entire camera set. Fig. 4 depicts an example of a valid trigger signal $g_{c_k}(t)$ from different cameras.

In Fig. 4, near time $t=270$, valid trigger signals are received from all cameras, because the upward motion of the right hand from each of the cameras could be detected. However, near times $t=400$, and $t=700$, c_2 and c_3, respectively, cannot receive valid trigger signals $s(t)$ because of the busy network environment. Near time $t=910$, all cameras can again receive trigger signals. Meanwhile, positions of the left foot joint j^{LF}_{t,c_k} and right foot joint j^{RF}_{t,c_k} from all cameras are sent to the server for further geometric calculation to calibrate the homography matrix and project the points among the cameras.

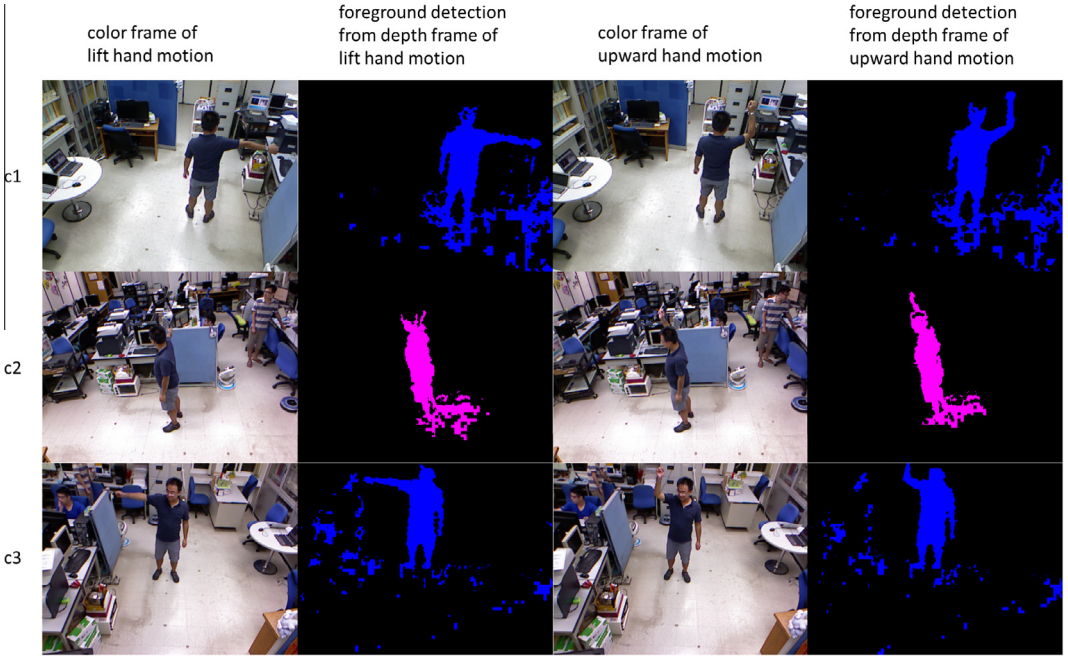


Fig. 3. Hand motion.

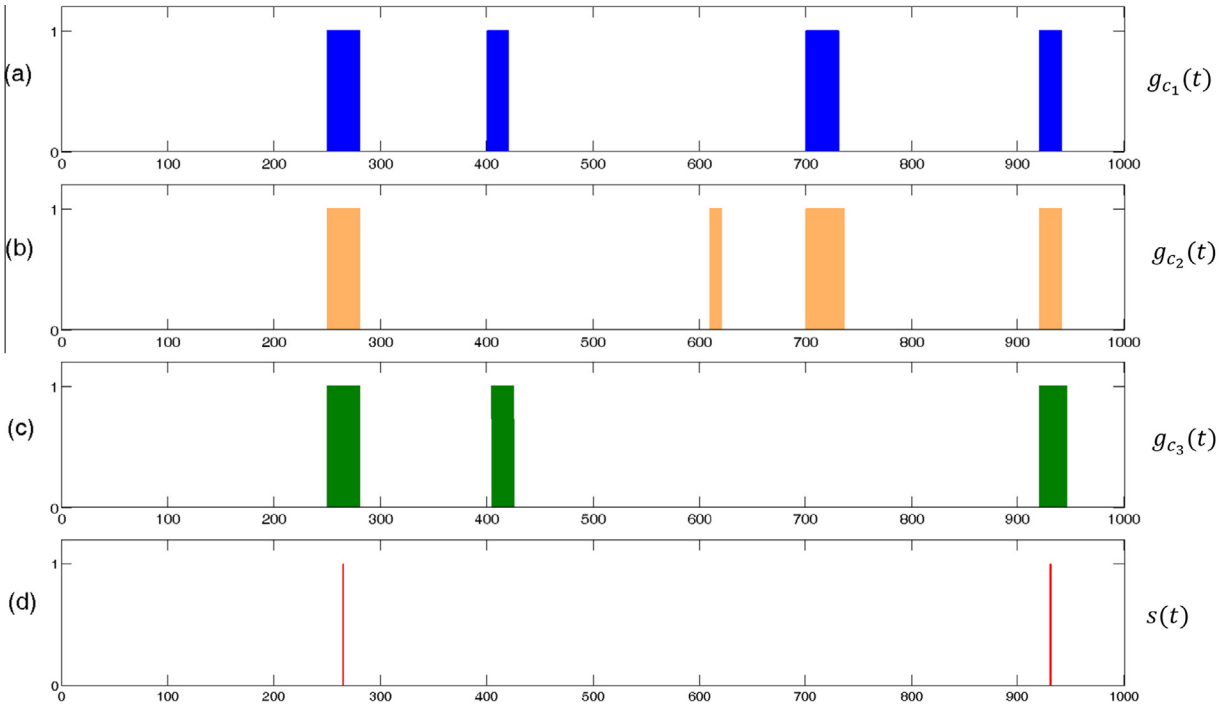


Fig. 4. The valid trigger signal function received from different cameras: (a) $g_{c_1}(t)$ from c_1, (b) $g_{c_2}(t)$ from c_2, (c) $g_{c_3}(t)$ from c_3, and (d) $s(t)$.

3.2. Relative hand joint issue among multiple cameras

In the proposed multiple-camera environment, each camera c_k can detect the left foot joint j^{LF}_{t,c_k} and right foot joint j^{RF}_{t,c_k}. However, when two cameras are mounted facing opposite directions, the coordinates are in opposing directions. Directly using the joints to calculate the geometries would lead to miscalculations caused by misalignment. Therefore, by calculating the centroid of the two joints, the foot joint position can be calculated as

$$j^{F}_{t,c_k} = \frac{j^{RF}_{t,c_k} + j^{LF}_{t,c_k}}{2} = \left(\frac{x^{RF}_{t,c_k} + x^{LF}_{t,c_k}}{2}, \frac{y^{RF}_{t,c_k} + y^{LF}_{t,c_k}}{2} \right) = \left(x^{F}_{t,c_k}, y^{F}_{t,c_k} \right), \tag{3}$$

when the final trigger signal is sent to the server (i.e., $s(t) = 1$ in Eq. (2)). Thus, misalignment among cameras can be avoided. Therefore, Eq. (1) should be modified to

$$g_{c_k}(t) = \begin{cases} 1, & \text{if } y^{RH}_{t,c_k} > y^{HD}_{t,c_k} \text{ or } y^{LH}_{t,c_k} > y^{HD}_{t,c_k}; \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

to send the trigger signal to the server. This hand-gesture-triggered geometry calibration process enables users to trigger the system according to a single hand raising gesture in the proposed system.

4. Proposed people tracking system

According to the foreground detection from the official Kinect sdk [11], not only the geometry among multiple cameras can be calibrated, but also the people moving in the field of view can be tracked. In this paper, a multi-trajectory matching using occlusion management is proposed, as depicted by the red dashed block of Fig. 2, including four major parts: multiple cameras projection, occlusion detection, Kalman filter for multiple-object tracking, and pairwise trajectory matching. These processes are described in detail in the following subsections.

4.1. Interleaving-based skeletal joints obtaining with valid skeleton determination

The official Kinect sdk [11] was adopted to obtain the frames in the color[1] channel (first row of Fig. 1), depth channel (second row), foreground detection (third row), and skeletons (fourth row). As shown in the second row of Fig. 1, a brighter luminance indicates that an object is closer to the depth camera, whereas a darker luminance signifies that an object is farther from the camera.

In an environment with multiple depth cameras, when a person is captured from different views (shown by the columns in Fig. 1), meanwhile, cross-camera interference [29] caused by multiple projections from the infrared emitters belonging to the multiple Kinect cameras in the overlapping area would not severely affect the person detection results obtained by applying the Kinect sdk in this study. Although the persons could be sequentially tracked by the Kinect sdk, partial or full occlusion over a long period of time resulted in mistracking. The occlusion effect can be addressed by using the proposed method.

Given the positions of the detected foot point defined in Eq. (3), $(x^F_{t,c_k}, y^F_{t,c_k})$, and the total number M of people detected by camera c_k, the foot point of the mth detected person is represented by

$$j^{F_m}_{t,c_k} = (x^{F_m}_{t,c_k}, y^{F_m}_{t,c_k}). \tag{5}$$

In addition, the position set is expressed as follows:

$$(X^F_{t,c_k}, Y^F_{t,c_k}) = \left\{ \left(x^{F_m}_{t,c_k}, y^{F_m}_{t,c_k} \right) : \ \forall 1 \leqslant m \leqslant M \right\}. \tag{6}$$

The set of detected foot points is used to generate the projective geometric matrix among various cameras.

4.1.1. Interleaving-based skeleton obtaining

Because the official Kinect sdk 1.8 [11] can provide up to six people tracking with centroids at each frame, but only with two of six people has the skeleton results, i.e., the number of tracked centroid of persons is larger than the number of tracked skeleton of persons. To extend the skeletons tracking capability to different persons, we propose an interleaving-based skeleton obtaining scheme using a time-sharing concept.

As shown in Fig. 5(a), the conventional skeleton obtained from Kinect sdk is locked to two of the detected persons according priority ordered by the entering time or the distance to the camera. Although the centroids of the other persons (circles with different colors) can be obtained, only the skeletons with black rectangle and red rectangle can be obtained at all time instances ($t1$–$t6$).

[1] For interpretation of color in Figs. 1–3, 5, 8–12, 16, and 17, the reader is referred to the web version of this article.

To share the occupied time instances from the skeleton tracked persons to other detected persons, we propose to randomly select two of the all detected persons for revealing the analytic skeleton results. As shown in Fig. 5(b), at each time instance, two of the whole detected persons are selected for revealing skeletons, e.g., black rectangle and green rectangle at $t1$ and black rectangle and blue rectangle at $t2$, etc. When observing from the time axis, the skeletons belonging to different persons can be obtained as a interleaving pattern, as shown in Fig. 5(c).

4.1.2. Moving average based valid skeleton determination

To reduce the effects caused by the noisy tracking results from unreliable skeletons, e.g., shortened or twisted skeletons belonging to a partial occluded person, we propose a moving average based valid skeleton determination scheme. For the obtained height of the skeleton h^m_t belonging to the person m at time t, the skeleton validation can be determined by observing the corresponding heights across the temporal axis, based on a moving average operation, with the observation sliding window size w as:

$$v^m_t = \begin{cases} 1, & \text{if } h^m_t > \alpha \cdot \left(\dfrac{\sum_{t-\frac{w}{2}}^{t-1} h^m_t + \sum_{t+1}^{t+\frac{w}{2}} h^m_t}{w} \right); \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where α is used as the height weighted coefficients for determining the validation of a skeleton. For different motions, e.g., fast moving, slow moving, turning around, of a tracked person, the influence of the weighted coefficients can be designed as symmetrical Gaussian distribution, Poisson distribution, and other distributions. However, to simplify the influence, the α is set as a constant coefficient, i.e. $\alpha = 0.5$, which is used for the experiments in this paper.

As shown in Fig. 6, for an obtained height of the skeleton, the past and next five heights are observed for determining the validation of the current skeleton h^m_t. When determining the validation of a skeleton, only the skeletons with valid flag $v^m_t = 1$ are used for observing the past side $(t - \frac{w}{2} \ldots t - 1)$. In other words, the invalid heights of skeletons are skipped for observation.

4.2. Multi-trajectory matching using occlusion management

Before mistracking can be corrected, cross-camera geometries is necessary to be built in the initialization process. The occlusion detected by each camera is used to identify missing parts clearly and relink to corresponding trajectories. Finally, trajectories of the same person from different cameras can be fused into a single trajectory, correcting the mistracking issue.

4.2.1. Multiple cameras projection

In the proposed system, a homography [16,17] technique is used to match corresponding objects among different views. Similar to multiple-camera people tracking systems [7–9], the positions of people detected by different cameras can be calculated according to homography matrices Hs.

Given the foot point captured by camera c_1 at time t of the mth person with the foot joint $j^{F_m}_{t,c_1} = \left(x^{F_m}_{t,c_1}, y^{F_m}_{t,c_1} \right)$, the projected foot point in camera c_2, $\left(x^{F_m}_{t,c_1 \to c_2}, y^{F_m}_{t,c_1 \to c_2} \right)$, can be calculated from $j^{F_m}_{t,c_1}$ and H as:

$$\begin{bmatrix} \beta \cdot x^{F_m}_{t,c_1 \to c_2} \\ \beta \cdot y^{F_m}_{t,c_1 \to c_2} \\ \beta \end{bmatrix} = H \begin{bmatrix} x^{F_m}_{t,c_1} \\ y^{F_m}_{t,c_1} \\ 1 \end{bmatrix}, \tag{8}$$

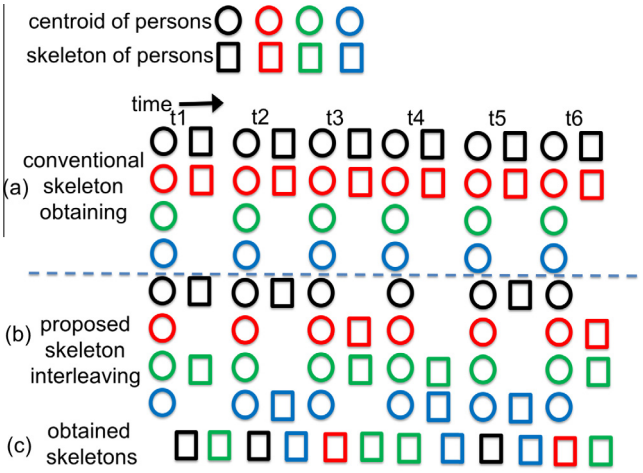


Fig. 5. Skeleton interleaving.

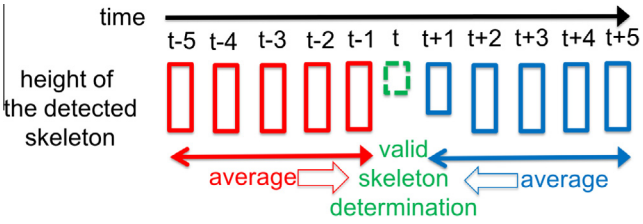


Fig. 6. Moving average based valid skeleton determination process, with the observation sliding window size: $w = 10$.

where β is a scalar. The homography matrices can be obtained by supplying the corresponding landmark points from different cameras [7–9] at the initial state to the virtual bird's eye view as the synergy point of view.

An example of the detected foot points of the proposed hand-gesture triggered geometry calibration for calculating the homography matrix H is shown in Fig. 7. When a upward motion is detected, the positions of the foot points from different views, e.g. c_1 and c_2 are captured for further generating the homography matrices. In addition, the occupied field of view in different cameras are mapped to a virtual bird's eye view. As a result, the homography matrices Hs can be correspondingly obtained.

4.2.2. Occlusion detection: multiple points in one region

In this study, the relationship among multiple points in one region proposed by Sun et al. [30] was considered to manage the occlusion in pairwise cameras. Region R is the detected foreground area in camera c_2 with left and right boundaries $x_{t,c_2}^{R_{min}}$ and $x_{t,c_2}^{R_{max}}$, respectively, in the x direction, and top and bottom boundaries $y_{t,c_2}^{R_{min}}$ and $y_{t,c_2}^{R_{max}}$, respectively, in the y direction (Fig. 8). The detected foot point projected from c_1 to c_2 of the mth person is expressed as

$$j_{t,c_1\to c_2}^{F_m} = \left(x_{t,c_1\to c_2}^{F_m}, y_{t,c_1\to c_2}^{F_m}\right) : \left\{x_{t,c_2}^{R_{max}} > x_{t,c_1\to c_2}^{F_m} > x_{t,c_2}^{R_{min}}, y_{t,c_1}^{R_{max}} > y_{t,c_2\to c_2}^{F_m} > y_{t,c_2}^{R_{min}}\right\}. \quad (9)$$

When multiple ($\geqslant 2$) projected foot points fall into the same region R, as the two blue circles $\left(x_{t,c_1\to c_2}^{F_m}, y_{t,c_1\to c_2}^{F_m}\right)$ and $\left(x_{t,c_1\to c_2}^{F_{m'}}, y_{t,c_1\to c_2}^{F_{m'}}\right)$ in the upper-left section of Fig. 8 do, the occlusion event is detected. Therefore, the camera by camera pairwise occlusion detection process can successfully reveal the occlusion events happened in each camera.

4.2.3. Kalman filter for multiple-object tracking

The foot point set $\left(X_{t,c_k}^{F}, Y_{t,c_k}^{F}\right)$ obtained in Eq. (6) and a conventional Kalman filter [31] is applied for people tracking in each camera. Detected foot points from different cameras are used to obtain trajectories for cameras as well as to project a synergistic virtual view (i.e., a bird's eye view). In the proposed method, detected foot points are projected from a real camera c_k to a virtual camera c_v. Based on the projected points $\left(X_{t,c_k\to c_v}^{F_m}, Y_{t,c_k\to c_v}^{F_m}\right)$ belonging to the mth person in the kth camera, according to Kalman filter, the trajectory is

$$\varphi_{c_k\to c_v}^{m} = \left\{j_{t,c_k\to c_v}^{F_m} = \left(x_{t,c_k\to c_v}^{F_m}, y_{t,c_k\to c_v}^{F_m}\right) : \forall t \in T\right\}, \quad (10)$$

where T is the period of observation time. For example, Fig. 9(a) depicts a trajectory in a spatio-temporal space. A set of detected trajectories is expressed as follows:

$$\Phi_{c_k\to c_v} = \left[\varphi_{c_k\to c_v}^{1}, \ldots \varphi_{c_k\to c_v}^{m}, \ldots \varphi_{c_k\to c_v}^{M'}\right], \quad (11)$$

where $\varphi_{c_k\to c_v}^{m}$ is the mth trajectory from the kth camera after tracking multiple objects for the total M' trajectories generated by Kalman filter technique. The number M' and M might not be the same in each frame because the number of detected people might be influenced by occlusion or environmental noises.

However, directly using the Kalman filter for tracking causes mistracking because of long-term occlusion. For example, the colored dots in Fig. 9(a) represent the tracking results from the Kalman filter. The blue and pink marks are trajectories in the spatio-temporal space of the same person from camera c_1; the marks are incorrectly interpreted as two independent trajectories. Camera c_2 exhibited a similar situation in which the orange and green marks were incorrectly assumed to be two trajectories. The multiple-trajectory matching scheme proposed in this paper reduces mistracking by reducing occlusion in an environment containing multiple depth cameras.

4.2.4. Pairwise trajectory matching

According to the occlusion detection scheme Eq. (9) and the trajectories obtained using the Kalman filter Eq. (11), the number of detected people can be determined according to a voting process based on observations from the temporal axis during a period which no occlusion occurred.

Given the obtained trajectories defined in Section 4.2.2, Gaffney's method [32] is applied in trajectory clustering using a linear regression mixtures model that enables continuous trajectory alignment in both time and space measurements. The trajectories obtained using the Kalman filter in Fig. 10 are depicted using colored markings. After applying Gaffney's clustering algorithm, the trajectories enclosed by the blue oval belong to one cluster, and the trajectories enclosed by the red oval belong to another cluster.

According the occlusion detection process defined in Eq. (9), the foot points with no occlusion are denoted as $j^{-F_n}_{t,c_k\to c_v}$. The number of cameras detecting occlusion is represented by O, and the fused foot point in the synergistic point of view is calculated as follows:

$$j'^{F_n}_{t,c_v} = \begin{cases} \dfrac{\sum_{c_k} j^{-F_n}_{t,c_k\to c_v}}{K-O}, & \text{if occlusion occurred;} \\ \dfrac{\sum_{c_k} j^{F_n}_{t,c_k\to c_v}}{K}, & \text{otherwise,} \end{cases} \quad (12)$$

where $n : 1 \leqslant n \leqslant N$ is the clustered trajectory index belonging to the nth person in a total of N clusters. The trajectory belonging to the nth person is represented by:

$$\varphi'^{m}_{c_v} = \left\{j'^{F_n}_{t,c_v} = \left(x'^{F_n}_{t,c_v}, y'^{F_n}_{t,c_v}\right) : \forall t \in T,\right\}, \quad (13)$$

where $j'^{F_n}_{t,c_v}$ is the 2D position calculated according to Eq. (12).

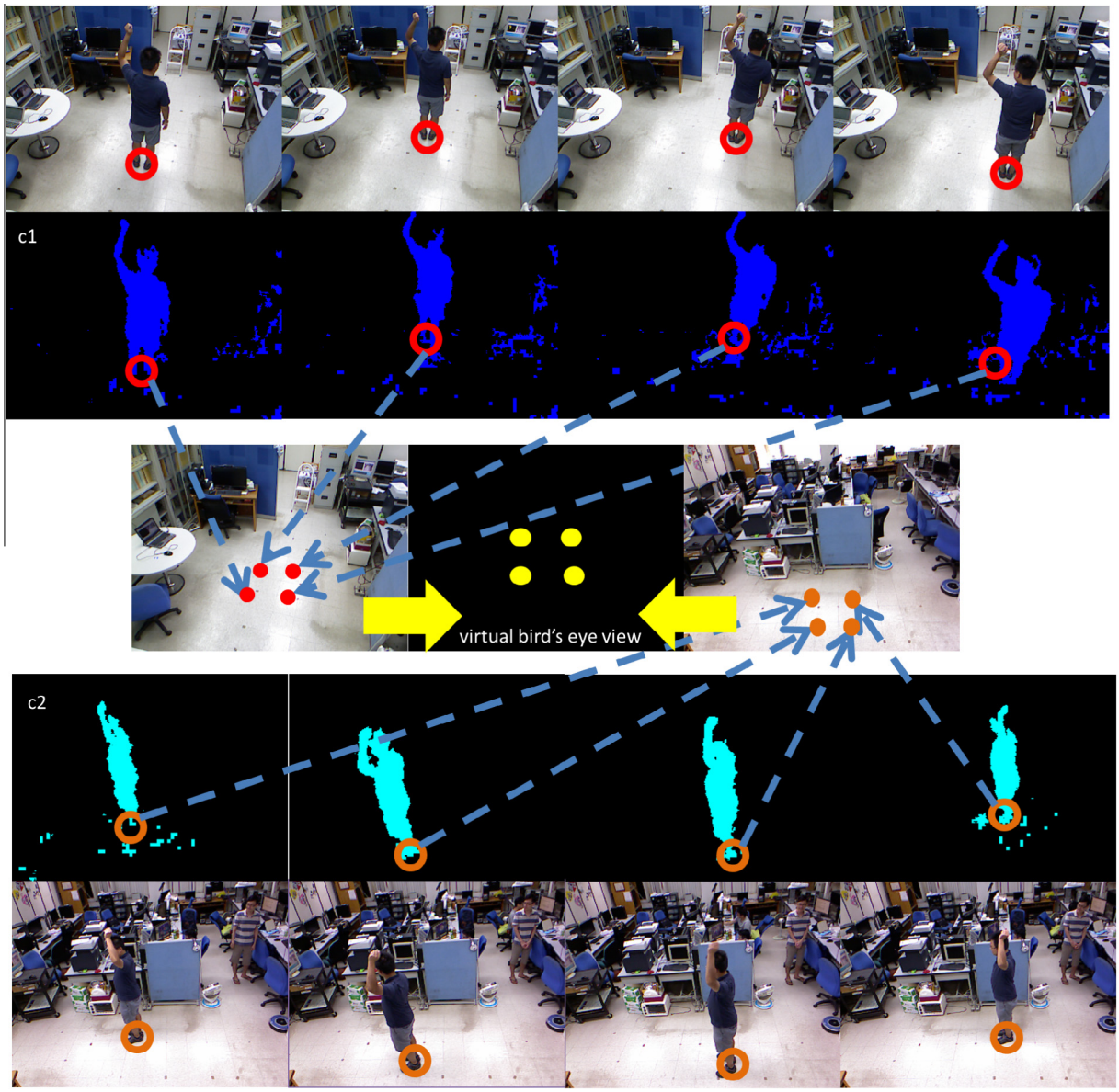


Fig. 7. Homography matrices generated from the proposed hand-gesture triggered geometry calibration from multi view cameras.

The red circles in Fig. 9(b) are the fused foot points in the synergistic view (i.e., $\varphi^{m}_{c_v}$ in Eq. (13)). Unlike the broken trajectories (four trajectories in Fig. 9(a) for the same person from two views) obtained using the conventional Kalman filter, the trajectory obtained using pairwise trajectory matching (depicted by the red circles in Fig. 9(b)) can be successfully fused into a single representative trajectory for one person.

5. Experimental results

In the experimental results, the official Kinect sdk 1.8 was adopted to obtain the skeleton analysis results from the built-in infrared depth cameras and to acquire tracking results for a maximum of two people with skeleton and joint information. The proposed method was evaluated under bright and dark lighting conditions in three settings: a studio, a laboratory, and a lobby (Fig. 11(a)–(c)). Table 1 shows the frame testing numbers of the scenes and time periods. Because the proposed method addresses the occlusion situations, only the successive frames before, during, and after occlusion were evaluated.

Three Kinect cameras with depth cameras were mounted 2.0 m high in each setting; the fields of view of the cameras overlapped in a 1.5 m × 1.5 m square on the ground plane. As shown in Fig. 11 (a)–(c), the positions of c_1, c_2, and c_3 are marked by blue, yellow, and green circles, respectively. Camera c_4 (the dashed black circle) represents the synergistic virtual view (bird's eye view) camera; no real camera was mounted at c_4, but the people detected by c_1, c_2, and c_3 were transformed into the synergistic virtual view (c_4). The green and red arrows in each view represent the movement trajectories of the two people.

5.1. Calibration

To track people by using cameras with different fields of view, the preprocessing step of the proposed system involves calibrating the geometries among the cameras according to the processes described in Section 3. The relative positions of the overlapping area and mounted cameras are depicted in the right parts of Fig. 11(a)–(c). For example, the foot joint j^{Fm}_{t,c_1} detected by camera c_1. According to the homography matrix calculated using the

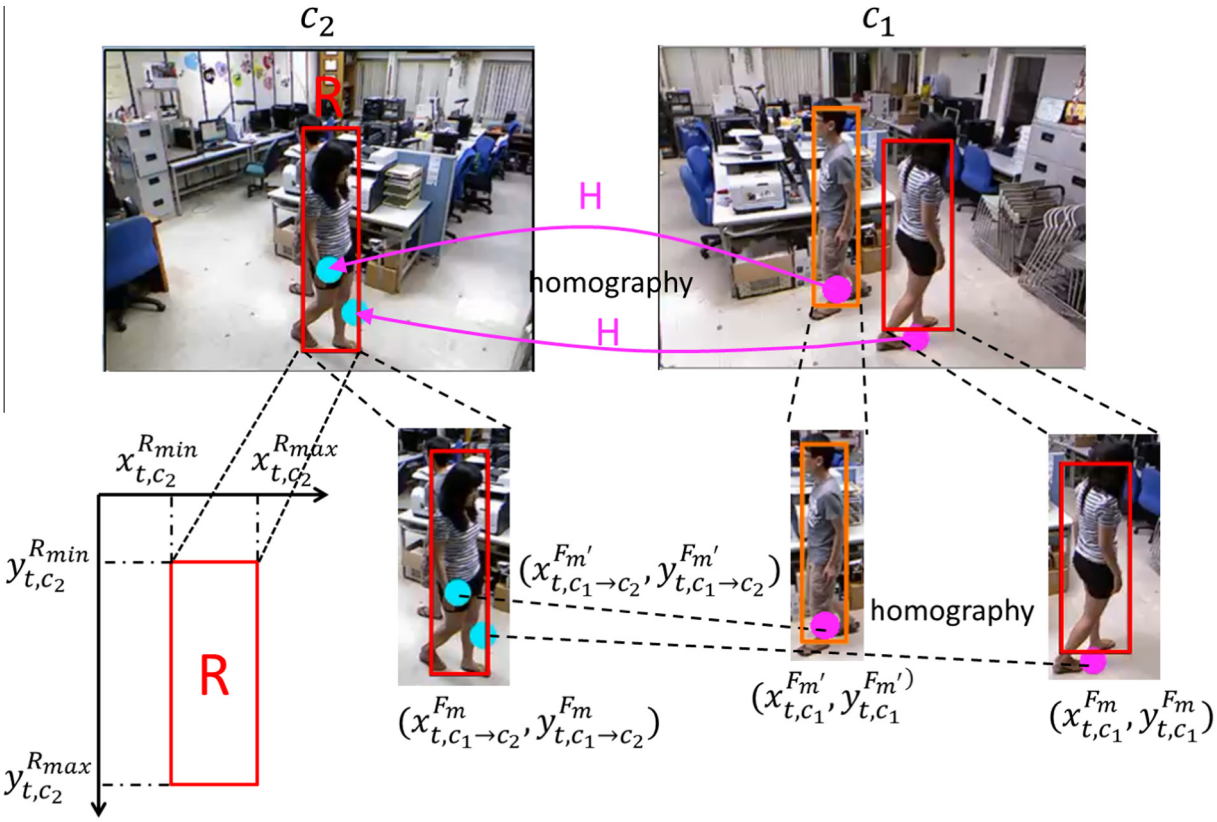


Fig. 8. Top: two example views of the person detection results by our system. The detected people are enclosed by rectangles. The calculated foot locations of different people are illustrated by different colored circles. Bottom: an occluded event is detected in the left example view where the two circles lie within the same rectangular region. In contrast to that, people associated to these two circles do not occlude each other in the example view on the right. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

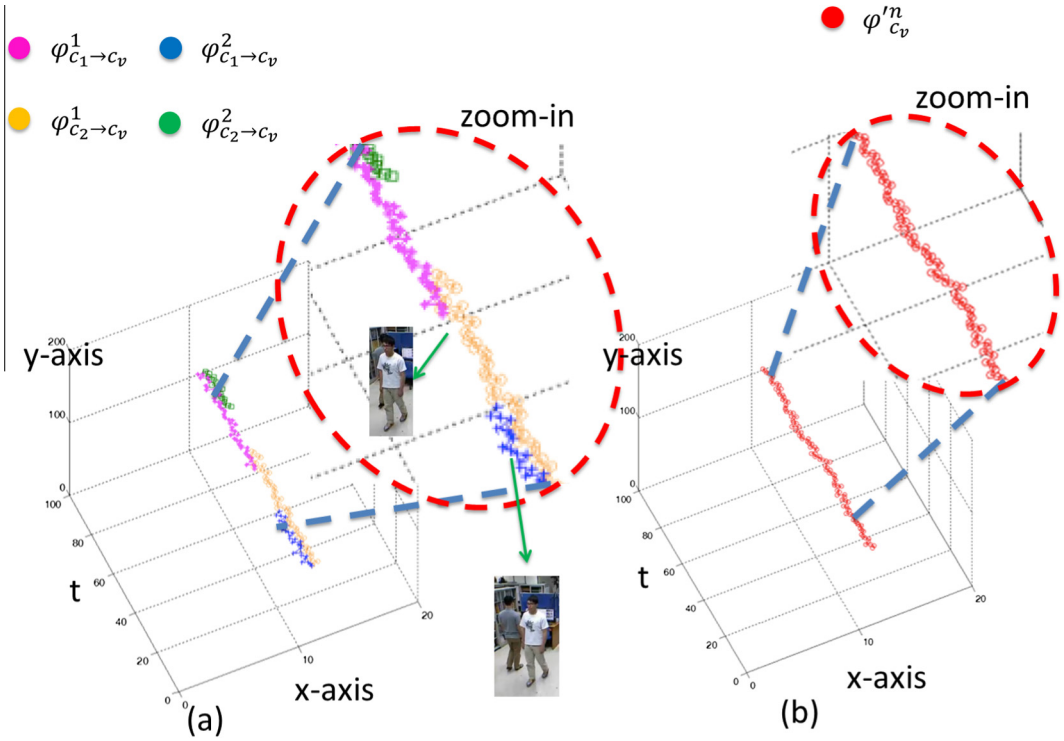


Fig. 9. The trajectory of an occlusion example shown in the synergistic virtual bird's eye view: (a) the pink marks and blue marks are the detected foot points from camera c_1 after applying Kalman filter tracking; the orange marks and green marks are the detected foot points from camera c_2. For example, the broken part between the pink trajectory and the blue trajectory is due to the occlusion event happened, and (b) the obtained fused trajectory (the red) marks which applied the proposed pairwise trajectory matching scheme. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

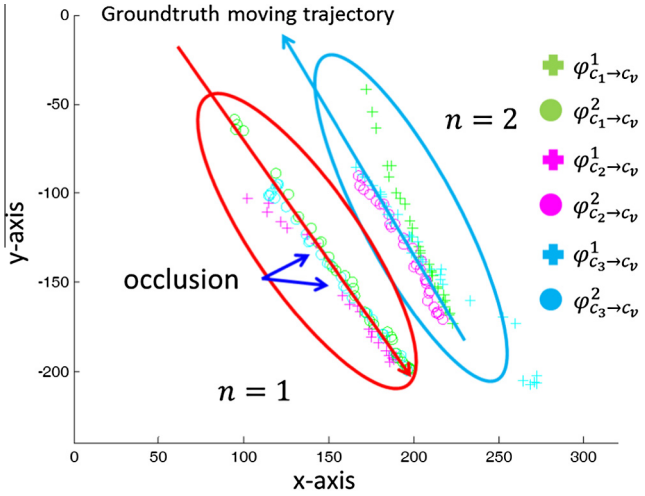


Fig. 10. Trajectory clustering results according to Gaffney's method [32].

procedure described in Section 4.2.1, the foot joint $j^{F_m}_{t,C_2\to c_1}$ belonging to the same person was transformed from camera c_2.

To evaluate the proposed hand-gesture-triggered calibration process, the foot points recorded from all cameras sending to the server were temporally synchronized according to the procedure described in Section 3.1. As shown in Fig. 11(d), the right foot joint j^{RF}_{t,c_k} detected by each camera when an upward hand motion was captured was used in a *one foot* scenario to generate the homography matrix used for matching the corresponding points among cameras. When the RANdom SAmple Consensus[33] (RANSAC) algorithm was applied to remove the outliers, distortion of the *one foot* results was reduced to a *one foot + ransac* result. Determined according to the process proposed in Section 3.2, the *spatial* result prevented relatively less distortion. Furthermore, when the RANSAC algorithm was applied, the distortion decreased further.

A common method for automatic identifying the corresponding points among different views is to use the SIFT [14] algorithm; however, when this method was applied, the distortion was high, because the SIFT feature points could not determine the corresponding points from the cameras, which were mounted too far from each other or the feature points were not is the same plane. As shown by the blue bar, the manually selected corresponding points among different views can achieve the lowest distortion with highest accuracy for calibration, however, for a flat and clean plane or the plane with repeated textures, it is still time consuming to manually identify the feature points in the area, shown in Fig. 12. Besides manually selected feature points scheme, to effectively obtain the corresponding points in a few seconds, in the evaluation, the proposed *spatial* calibration using RANSAC achieved the most favorable performance and the less severe distortion (average: *10.13* pixels, shown by the purple bar); this calibration process was adopted in all of the subsequent tests.

5.2. Tracking

Fig. 13 shows a representative test conducted to evaluate the proposed tracking method. In Fig. 13, two people are moving toward each other; the male is occluded by the female in cameras c_2 and c_1 (Fig. 13(b) and (c), respectively). The second row of Fig. 13 shows the results obtained by using the Kalman filter directly. The occluded person was wrongly assigned to a different trajectory in a false positive situation. However, when the proposed method was applied, the people were correctly assigned two trajectories (the third row), even when occlusion occurred. Moreover, when the setting was dark, because the infrared depth cameras on the Kinect cameras could detect the people's motion, the people were still correctly tracked (Fig. 14).

5.2.1. Performance comparisons

The performance of the proposed method was compared using the methods of Berclaz et al. [9], Ozturk et al. [13], and Baum et al. [15]. Three cameras were used to record frames from three points

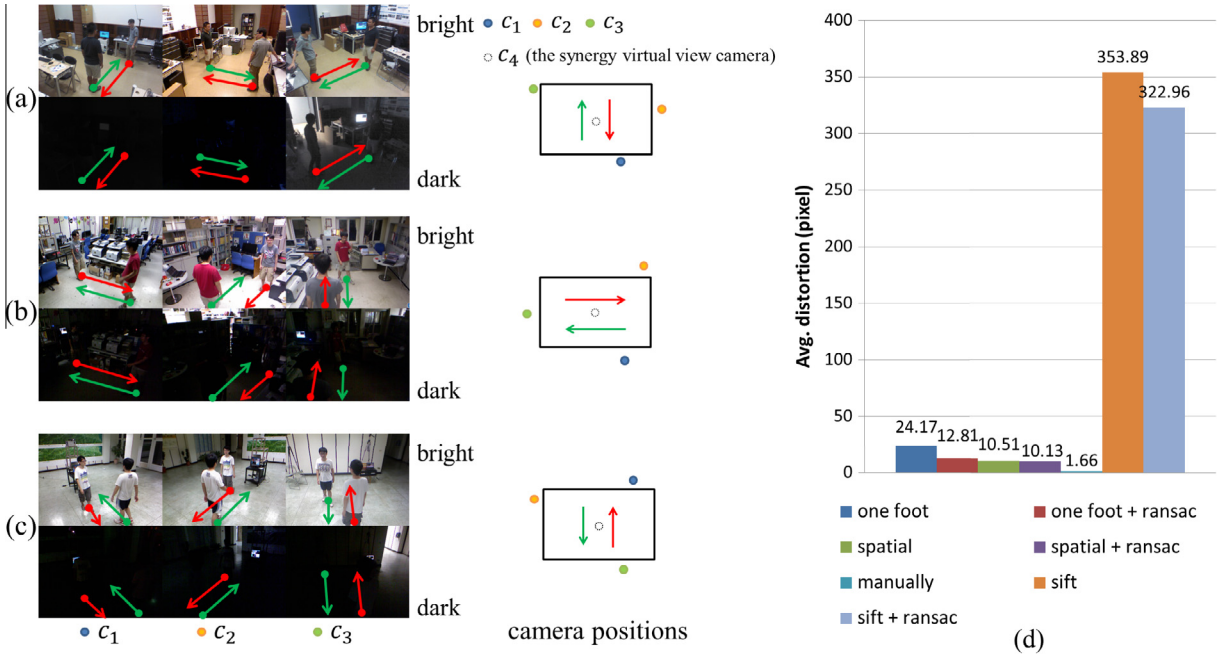


Fig. 11. The camera setting and in three testing scenes and the corresponding people movement trajectories: (a) *Studio*, (b) *Lab*, (c) *Lobby*. c_1, c_2, c_3 are the locations of the mounted Kinect cameras at the side view, and c_4 (the black dashed circle) is location of the synergistic virtual view (bird's eye view) camera. (d) Calibration results of the average distortion from 20 trials of different feature point obtaining approaches: *one foot*, *spatial*, *manually*, and *SIFT*, without/with the RANSAC outlier removal process.

Table 1
The testing number of frames for different scenes with bright and dark lighting conditions of the successive frames with before, during, and after occlusion situations.

	Number of frames	Period of time (s)
Bright Lab	32	7.11
Dark Lab	34	7.56
Bright Studio	34	7.56
Dark Studio	41	9.11
Bright Lobby	37	8.22
Dark Lobby	40	8.89
Bright Lab-E	1617	240
Dark Lab-E	1358	240
Bright Sidewalk-E	1503	240
Dark Sidewalk-E	1389	240
Bright Lab-Crowded	175	52

of view. Fig. 15(a) shows frames obtained from cameras in the color channel that were used for people detection. Fig. 15(b) illustrates that the depth information corresponded to information recorded by Kinect cameras equipped with depth cameras, enabling people detection and tracking. A multiple-camera people tracking method [9] was applied in comparison and implemented in the color channel according to the source code of the authors [34] (Fig. 15(c)). To compare the performance of the approach using a single bird's-eye-view camera, conventional frame differencing in the color channel was implemented according to the foreground detection process described by Ozturk et al. [13]; Fig. 15(d) depicts a representative captured image. By adopting the depth channel to perform foreground object detection based on depth frame differencing at the same position, the method of Baum et al. [15] was implemented for comparison (Fig. 15(e)).

5.2.2. Subjective evaluation

The dataset listed in Table 1 was subjected to a subjective evaluation (studio, lab, and lobby settings with bright and dark lighting conditions). At first, Fig. 16(a) shows the movement of the people, and the ground truth is labeled in Fig. 16(b). Next, Fig. 16(c) shows the respective detection and tracking results of the proposed method and comparative methods. The proposed method reliably detected and tracked people, yielding results similar to those of the bird's eye view approach [15] in the depth channel. In some applications, mounting a camera on the ceiling to obtain a bird's eye view (e.g., [13,15]) is unfeasible. The mounted side-view depth cameras in the proposed method, however, produced satisfactory tracking results, comparing to the other methods (Fig. 16(d)–(f)).

The limited detection observed when using the method of Berclaz et al. [9] (POM) was caused by the tradeoff for the grid setting (20 cm × 20 cm) in the POM source code. In addition, in the bright lab setting (Fig. 16(d)), the dots in cyan and blue indicate that the detected results were incorrectly divided into two trajectories, whereas the results obtained using the proposed method were able to correctly generate only one trajectory per person, even when occlusion occurred. When the color-channel approach of Ozturk et al. [13] was applied, cameras could not detect people in dark lighting settings.

5.2.3. Objective evaluation

Fig. 17 shows a comparison of the average distortion observed in the obtained detection results with the ground truth. After the Euclidean distance was calculated, the overall results indicated that the proposed method exhibited the lowest distortion (Fig. 17(a)); Fig. 17(b) shows the corresponding detailed results. The lack of bars (method of Ozturk et al. [13]) in Fig. 17(b) signifies that the cameras could not detect people in the color channel in situations that were too dark. The scene could be captured from the bird's eye view compared to the ground truth by using the methods described in [13,15]. The detected foot points calculated according to the centroid of the blobs (people) could not provide adequately precise results, because of perspective distortion by the camera. By contrast, the results obtained using the multiple-camera approach exhibited greater distortion caused by the grid-setting property.

Because the multiple-camera approach proposed by Berclaz et al. [9] is the most related to the proposed method, the accuracy of results obtained by applying these two method in people detection were evaluated. The false positive rate (*FPR*) and false negative rate (*FNR*) were calculated as follows:

$$FPR = \frac{FP}{FP + TN}, \qquad FNR = \frac{FN}{TP + FN}, \tag{14}$$

where *FP* (false positive) is the number of falsely detected foot points while ground truth does not contain any points, *FN* (false negative) is the number of points where tracking results does not contain any point while ground truth contains at least one point, *TP* (true positive) is the number of correctly detected foot points where both ground truth and tracking results agree on the presence of people, and *TN* (true negative) is the number of points where both ground truth and tracking results agree on the absence of any people. The accuracy can be calculated as follows:

Fig. 12. A flat and clean floor plane (with repeated patterns) on a stage, captured from different views (a) far view, (b) close view, and (c) the plane with a landmark.

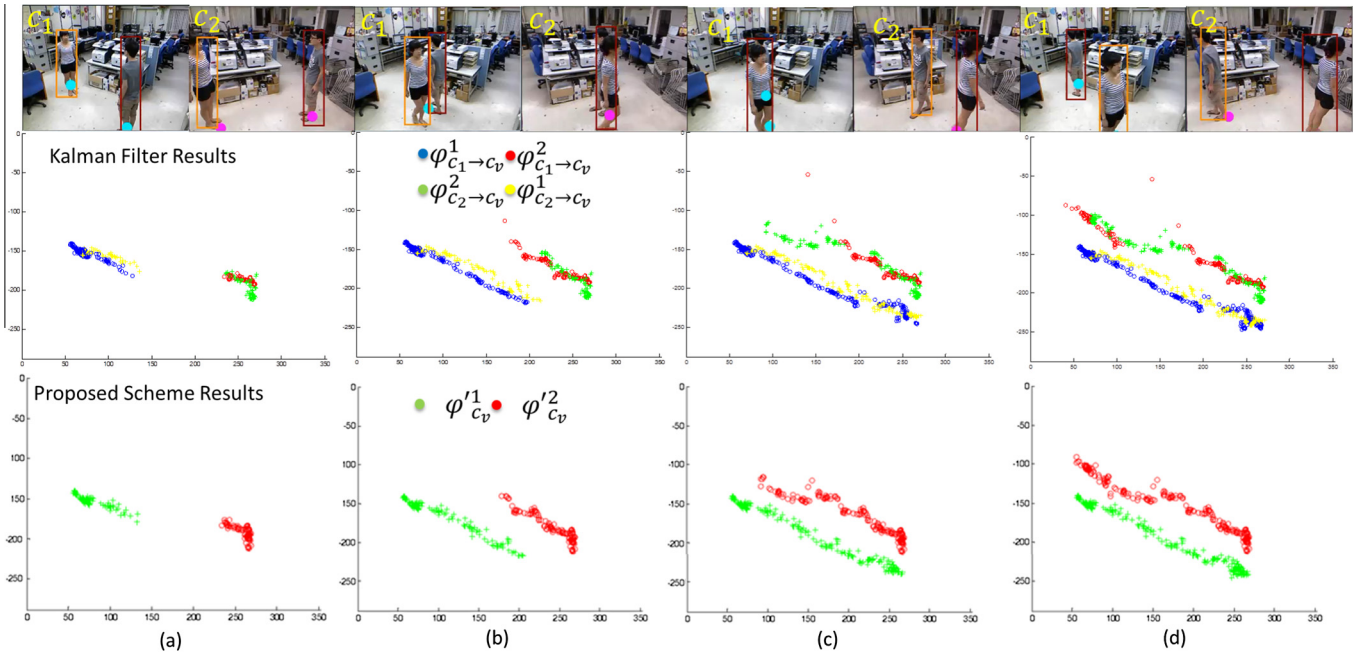


Fig. 13. The results for the bright scenes: (a) before occlusion, (b) occlusion in c_2, (c) occlusion in c_1, and (d) the obtained trajectories after occlusion.

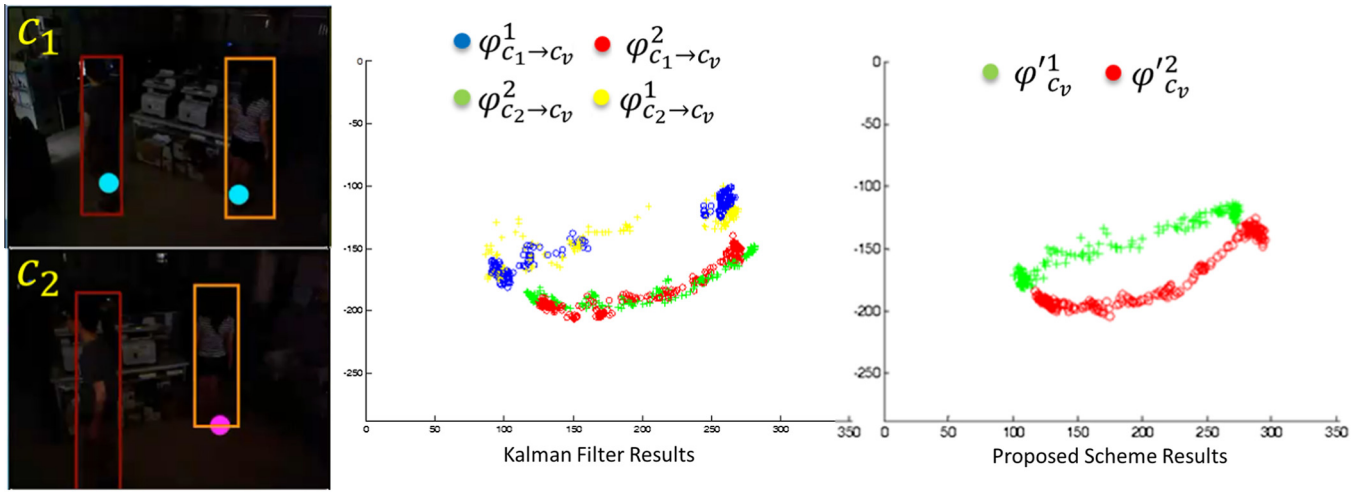


Fig. 14. The results for the dark scene after occlusion. The proposed scheme can completely obtain two trajectories belonging to two persons.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}. \quad (15)$$

Fig. 18(a) shows that, on average, the proposed method provided both a low *FPR* and *FNR* as well as high accuracy. Fig. 18(b) depicts the detailed results obtained at each setting. It is obvious that, by comparing to Berclaz et al.'s [9] method, the proposed method achieves much lower *FPR* and *FNR* simultaneously, with much higher accuracy.

Fig. 19(a) shows the overall foreground people and trajectory detection results. The ground truth in the tests was two people walking in the field of view in two trajectories. The people were accurately detected using the side-view infrared depth cameras in the Kinect cameras by fusing the trajectories among different views. A false positive occurred when the method of Ozturk et al. [13] was applied because the people were separately detected before and after occlusion to generate additional trajectories. The false negative results obtained when the method of Ozturk et al. [13] was applied were due to inability when using a conventional frame differencing technique to detect people from the color channel in an extremely dark situation; Fig. 19(b) depicts the detailed results. Both the proposed method and that developed by Baum et al. [15] yielded accurate results. The proposed method can be used to capture people in the field of view from the side and manage occlusion, whereas that by Baum et al. [15] can capture images from a bird's eye view with no occlusion. In applications for which mounting cameras to provide a bird's eye view is infeasible, the proposed method can be used to detect and track people moving within the field of view. In addition, similar to the performance evaluation of multi-camera human detection/tracking in [27], the representative people detection results in three different scenes with bright and dark lighting conditions are shown in Fig. 20. It is obvious that the proposed method achieves the best performance, with almost zero people detection error in most of the cases. The detection errors

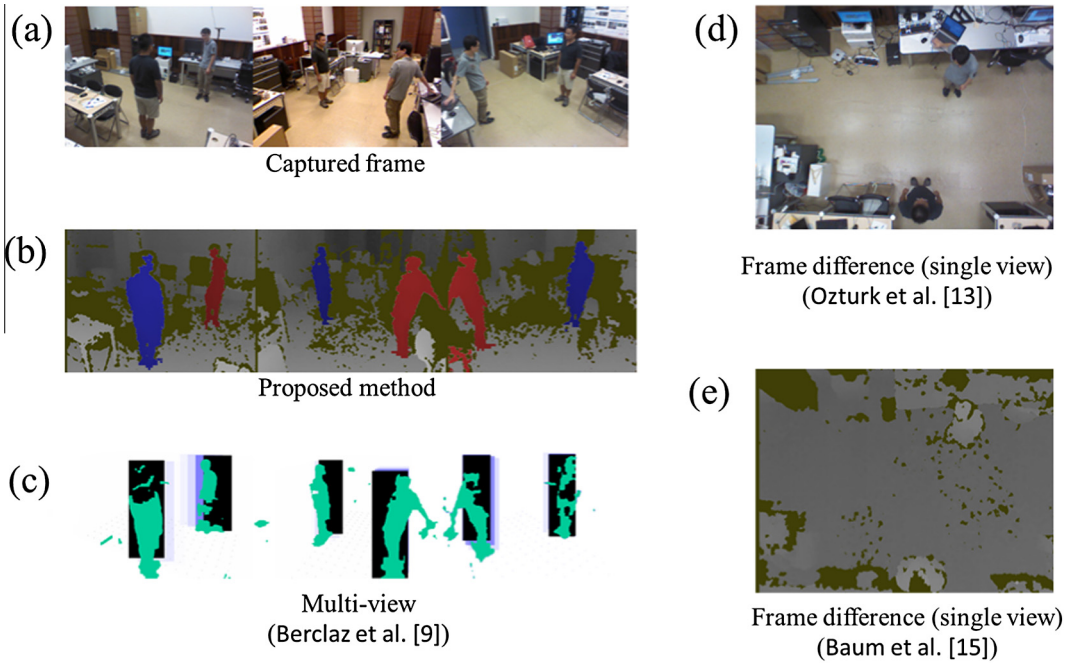


Fig. 15. Comparison to the state-of-the-art methods (a) captured frame in the color channel, (b) proposed method based on Kinect foreground detection results, (c) Berclaz et al. [9], (d) Ozturk et al. [13], and (e) Baum et al. [15]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

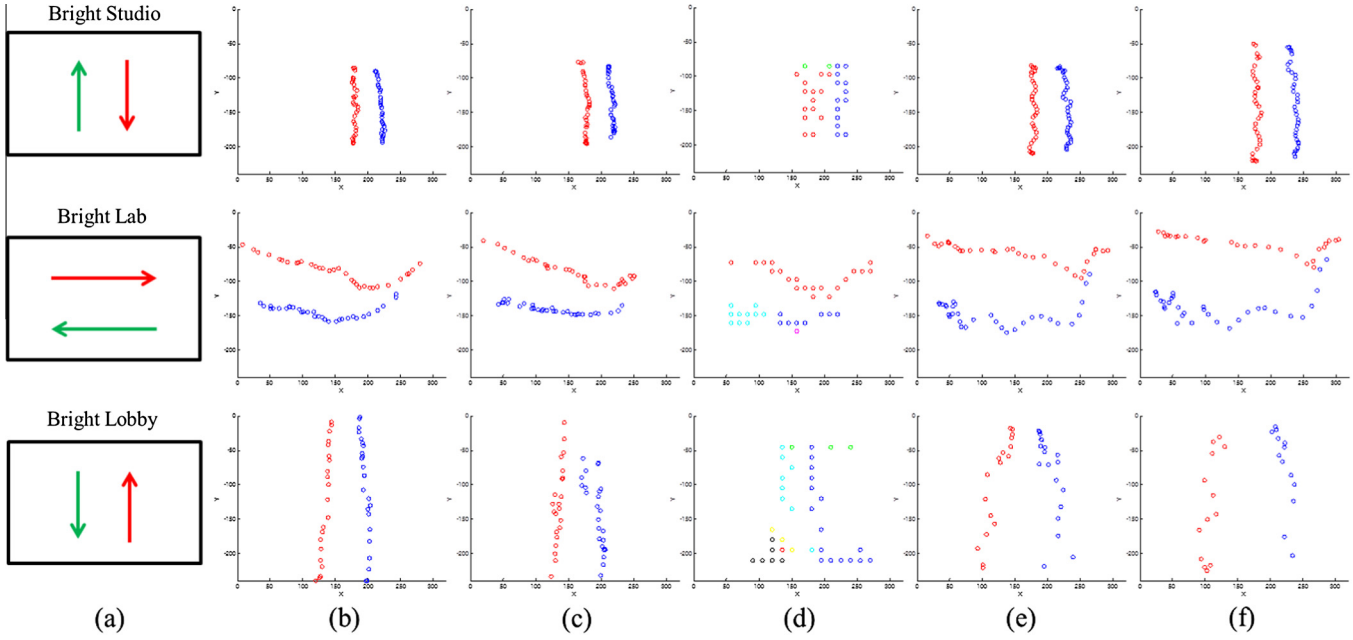


Fig. 16. The detected foot point $j_{t,c_v}^{'F_m}$ depicted in the synergy point of view: (a) movement trajectory, (b) manually labeled ground truth, (c) proposed method, (d) Berclaz et al. [9], (e) Ozturk et al. [13], and (f) Baum et al. [15].

happened in [9,13,15] are caused by too dark lighting condition, occlusion, and too close situations.

5.2.4. Extensive tests

Because the method of Berclaz et al. [9] also mounted the multi-cameras at the side view, we arrange people to walk into the field of view from one person to four persons moving in the area. The testing environment, people moving field of view, and lighting conditions are shown in Fig. 21.

For the people staying very close situations, our method is compared with the method proposed by Berclaz et al. [9], as shown in Fig. 22(a) and (b). One of persons is severely occluded by another from frame 65 to frame 74 in view c_2, but the tracking issue can still be compensated by the other views without occlusion situations according to the proposed skeleton-based pairwise trajectory matching scheme. However, with satisfactory foreground detection results in non-occlusion views c_1 and c_3 of [9], the final tracking results in Fig. 22(c) of the proposed method and [9] can both provide separately correct people tracking trajectories in the field of view.

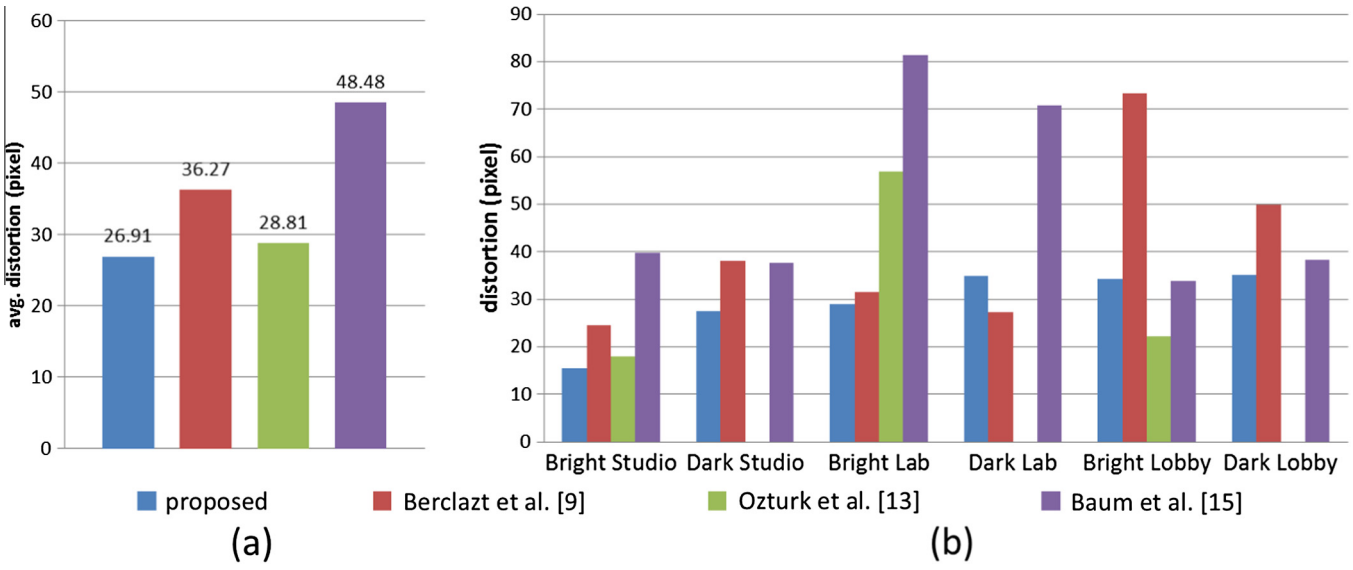


Fig. 17. The distortion results of the detected person's foot point from the ground truth in the synergy point of view to the detected position $j_{t,c_p}^{i^{F_m}}$ in the pixel level accuracy: (a) the overall results, and (b) the detail results in different scenes.

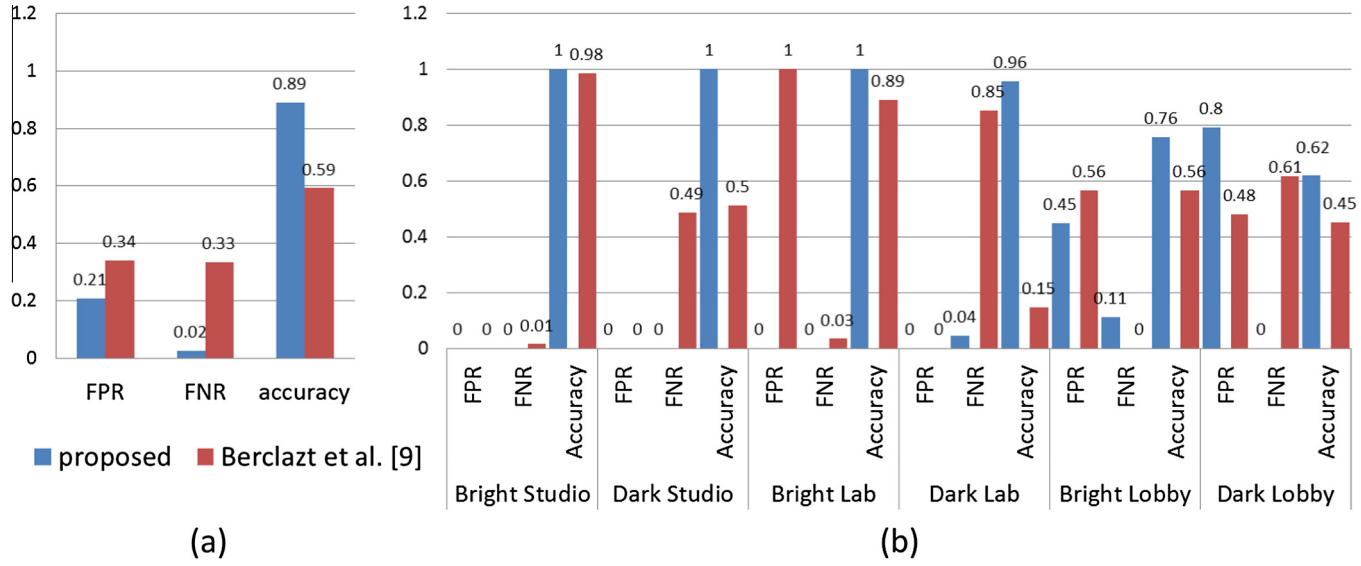


Fig. 18. The results false negative rate (FNR), false positive rate (FPR), and accuracy in the tests: (a) the overall results, and (b) the detail results in different scenes.

When six persons staying in the field of view from the beginning, as shown in Fig. 23(a) and (b), the six persons can be separately detected by the depth camera c_1, many people partially occluded in c_3, and severely occluded in c_2 from frame 54 to frame 64 (without any skeletons), for the proposed method. By compensating from the non-occlusion views to the occlusion views, the proposed method can still provide satisfactory results, as shown in Fig. 23(c). However, the severe occlusion issue cause [9] cannot properly detect people from the color frames in different views, and the people detection, tracking, and trajectories are affected by imperfect foreground detection.

On the other hand, the extensive overall results of the tests in bright conditions and dark conditions are shown in Fig. 24. Because of using the infrared-base Kinect camera for obtaining the depth data in the proposed method, as shown in Fig. 24(b), even in the dark lighting condition, the proposed can still successfully detect the moving people. We should notice that, the conventional Kinect sdk 1.8 can only track two persons with skeletons. In the six persons test, although the freely moving persons are very easily occluded by each other (Fig. 23(a)) when moving to different places, even with severe occlusion situations in the filed view, the proposed pairwise trajectory matching scheme can compensate the tracking results from the other views without occlusion issues for generating reliable trajectories, as shown in the rightmost figure of Fig. 24(c). Furthermore, with the proposed pairwise trajectory matching scheme, in overall, the proposed method has higher accuracy with lower FNR and FPR, as shown in Fig. 24 (d).

5.3. Time complexity

The comparison of computational complexity is shown by frame per second (fps): the proposed method (28.51 fps, greatest number of fps), Berclaz et al. [9] (1.82 fps), Ozturk et al. [13] (16.67 fps), and Baum et al. [15] (10.29 fps) in the overall results. Only the proposed method and [9] are multiple cameras scenarios.

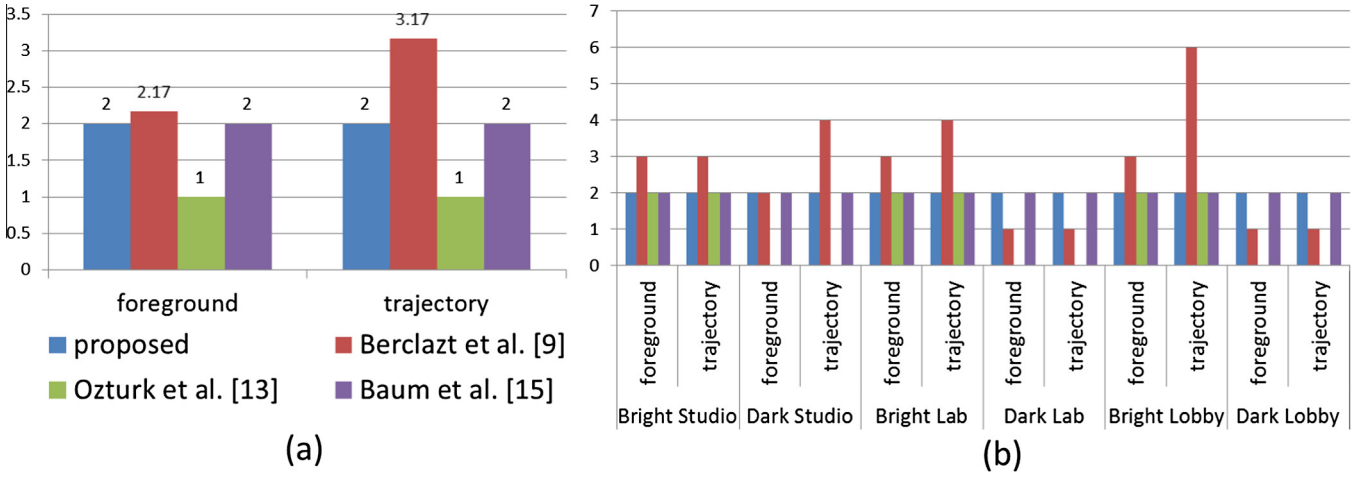


Fig. 19. The results of the detected trajectories with the ground truth: 2 persons in the field of view as the foreground objects and with 2 trajectories. (a) The overall results, and (b) the detail results in different scenes.

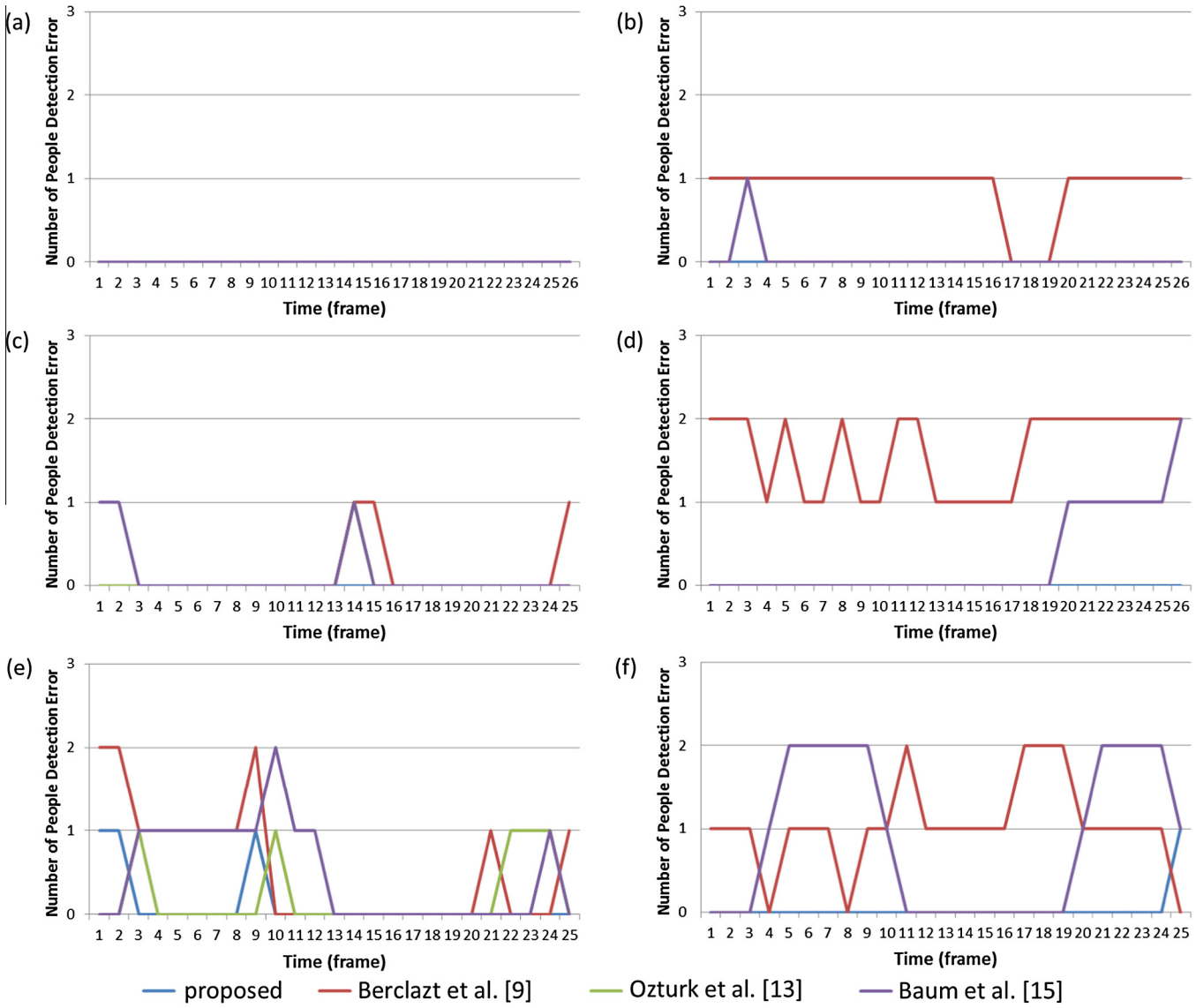
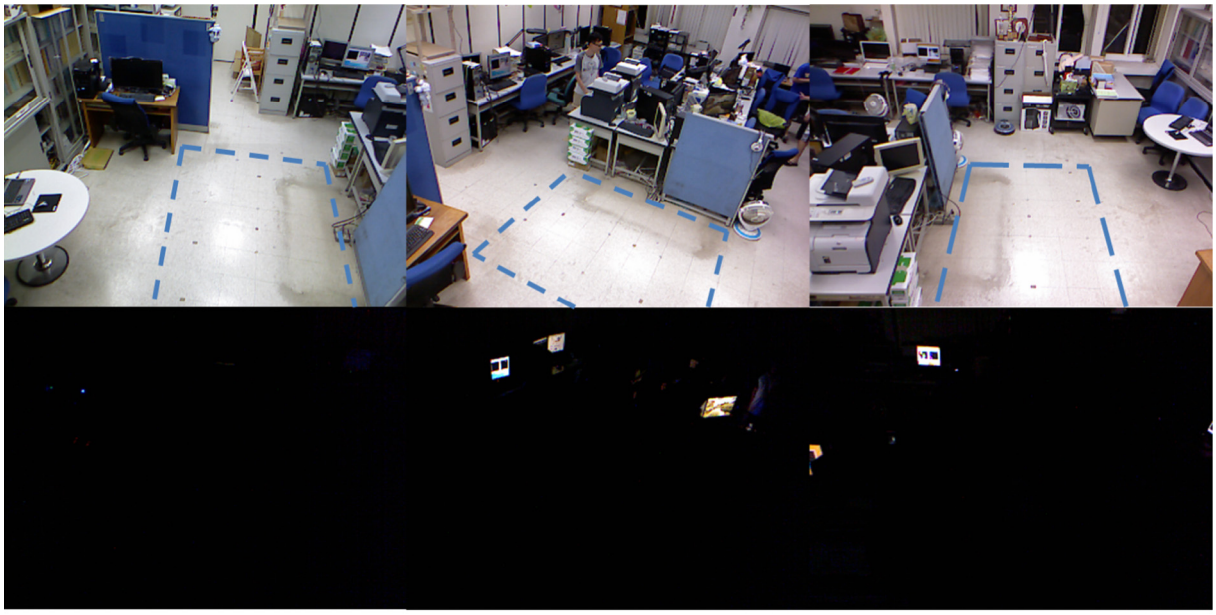


Fig. 20. Results for number of people detection error in different scenes from the most severe consecutive frames in the test video sequences: (a) Bright Studio, (b) Dark Studio, (c) Bright Lab, (d) Dark Lab, (e) Bright Lobby, and (f) Dark Lobby.



(a) lab scene, **Bright Lab – E**, **Bright Lab – Crowded** (upper), and **Dark Lab – E** (bottom)

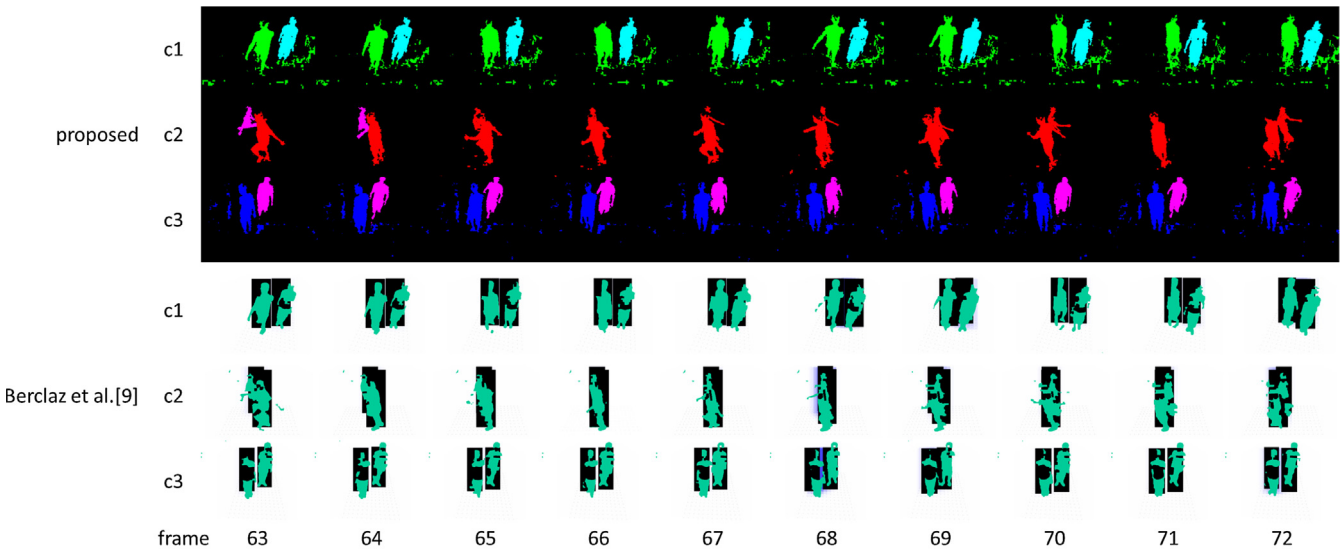
(b) sidewalk scene, **Bright Sidewalk – E** (upper) and **Dark Sidewalk – E** (bottom)

Fig. 21. Extensive tests for **Lab** and **Sidewalk** scenes from the color channels. Dashed blue rectangle are with the size 2.0 m × 2.0 m and 2.3 m × 2.4 m, correspondingly. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

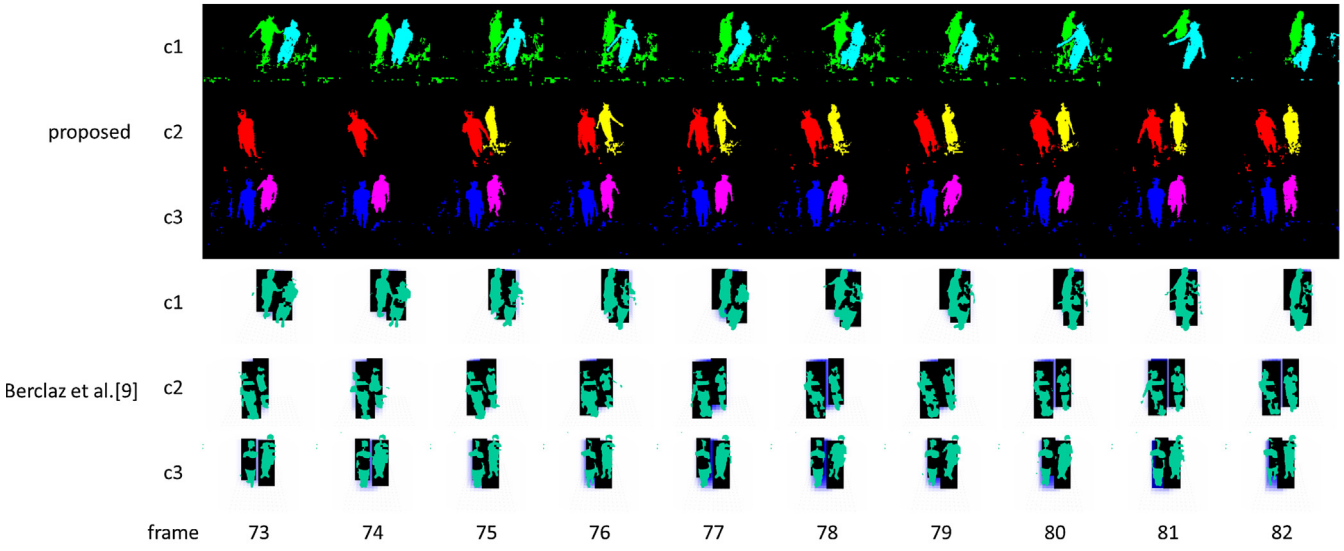
The computational complexity experiments were performed using a computer with an Intel Core i7, a 2.67-GHz CPU, and an 8-GB RAM. Therefore, according to the results, the proposed method is suitable for realtime applications.

6. Conclusions

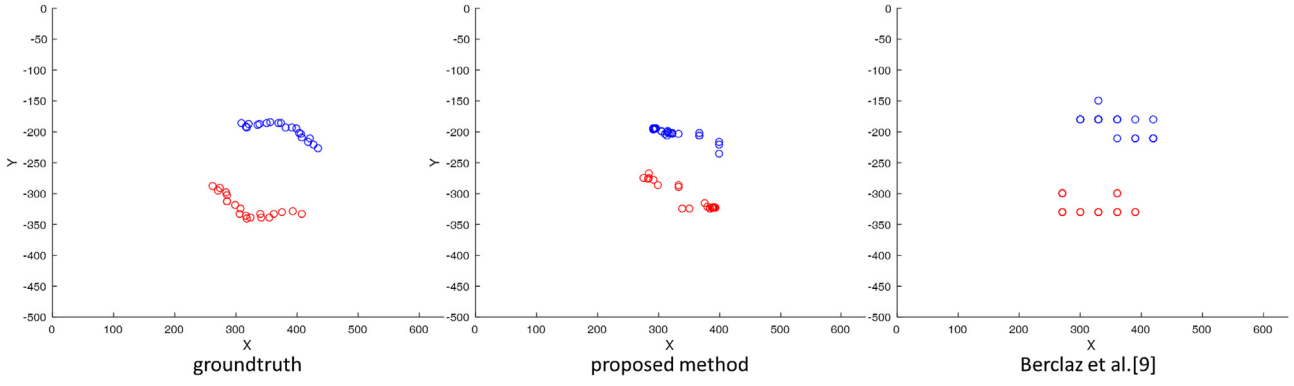
This paper proposes a pairwise trajectory matching scheme that involves fusing the detected trajectories from multiple depth cameras to reduce mistracking during occlusion. Based on the skeleton and joints of a person analyzed using a Kinect camera, the foot points (joints) can be used to track people when cameras have overlapping fields of view. Using homography transformation among views with Kalman filter for people tracking, the proposed pairwise trajectory matching method can compensate for occlusion in the synergistic virtual bird's eye view. The contribution of this paper is trifold: (1) A hand-gesture-triggered calibration process, with a natural user interface, allows general users (not computer vision experts) to effectively create geometries among multiple infrared depth cameras, through a temporal synchronization to achieve cross-camera calibration. (2) To extend the number of tracked persons with skeletons, we proposed an interleaving-based skeleton obtaining and moving average based valid skeleton determination. (3) Occlusion is satisfactorily managed by using the proposed pairwise trajectory matching scheme. In addition, in the crowded scene with extensive tests, the



(a) frame 63-72

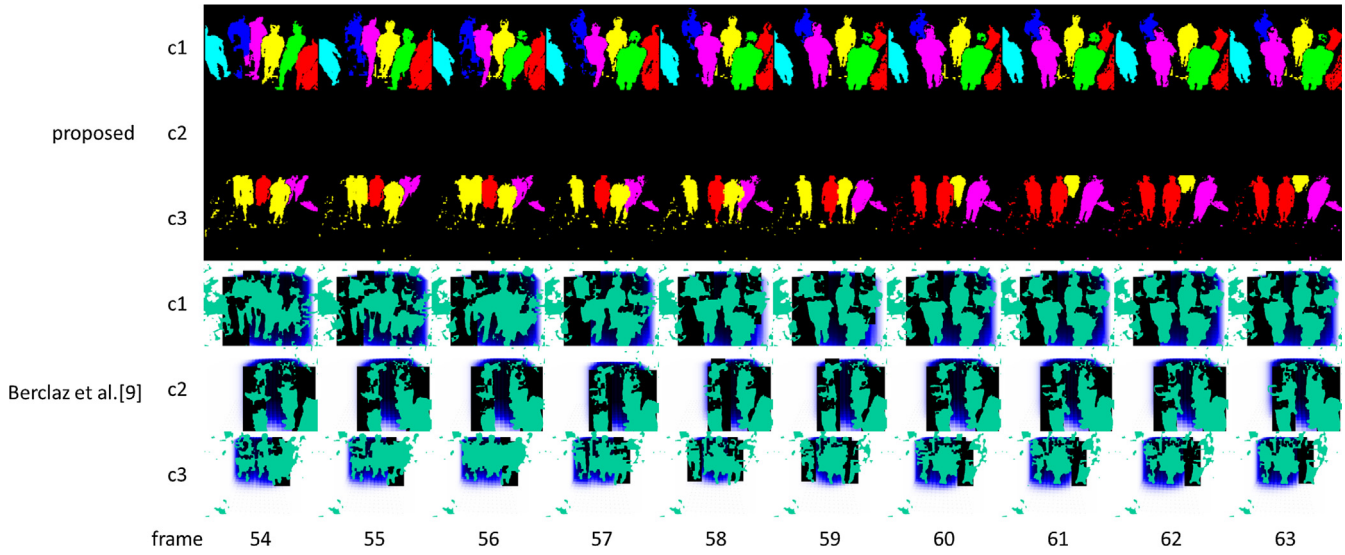


(b) frame 73-82

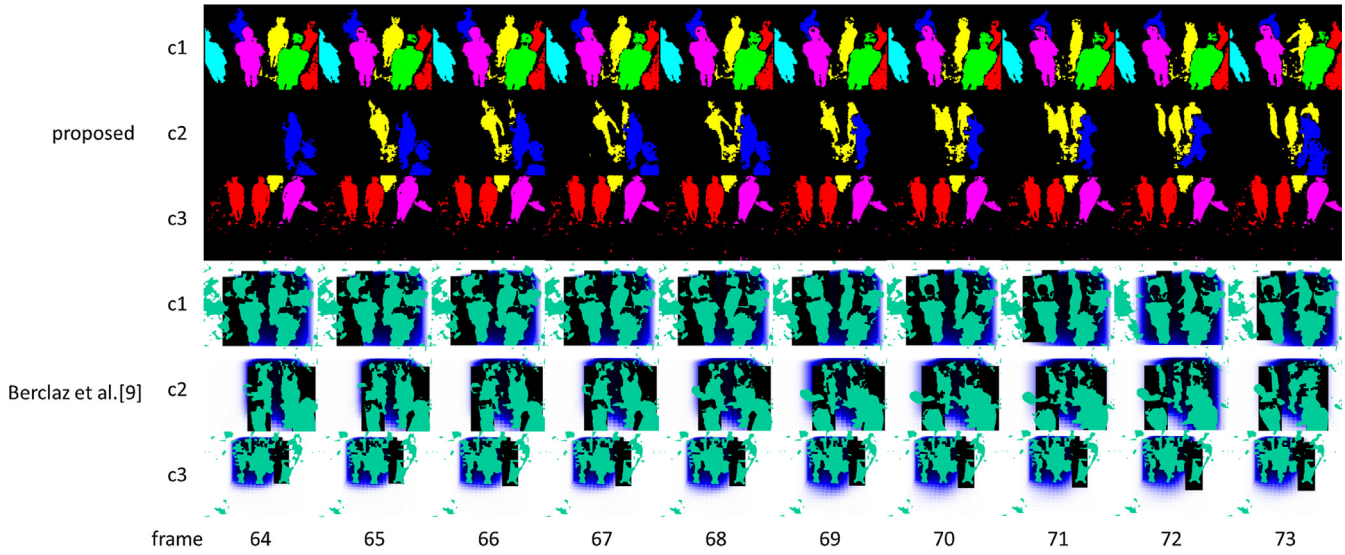


(c) the detection results in the field of view

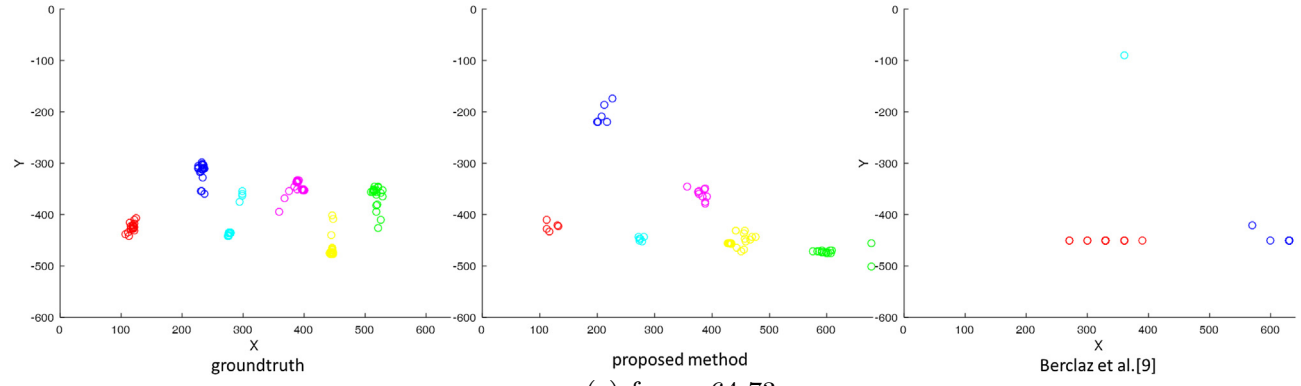
Fig. 22. The tracking results of one person occluding another, from the test video sequence **Bright Lab-E**.



(a) frame 54-63



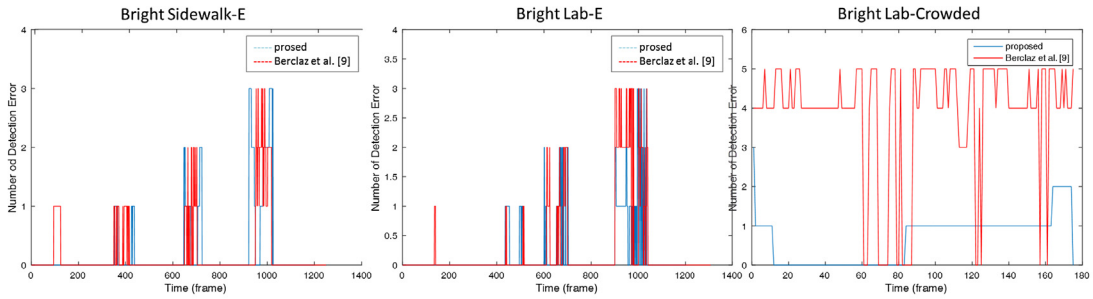
(b) frame 64-73



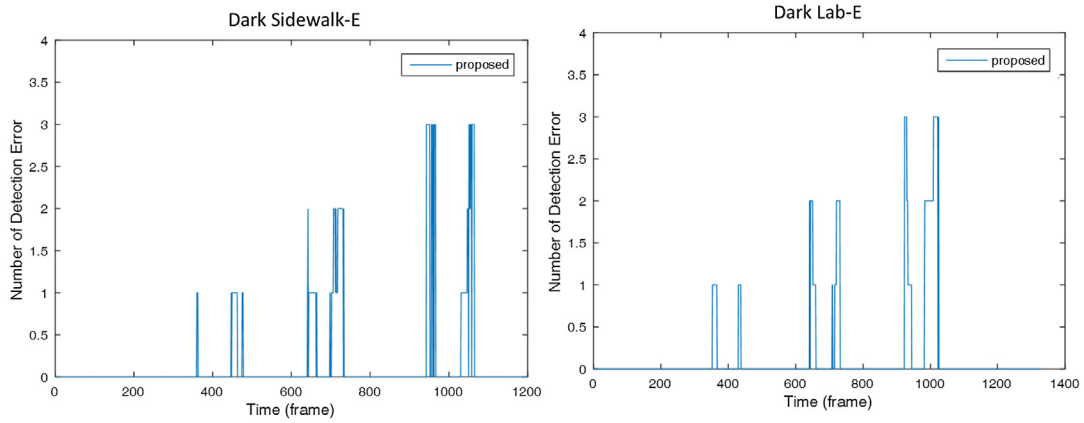
(c) frame 64-73

Fig. 23. Results of six persons simultaneously freely moving, from the test video sequence **Bright Lab-Crowded**.

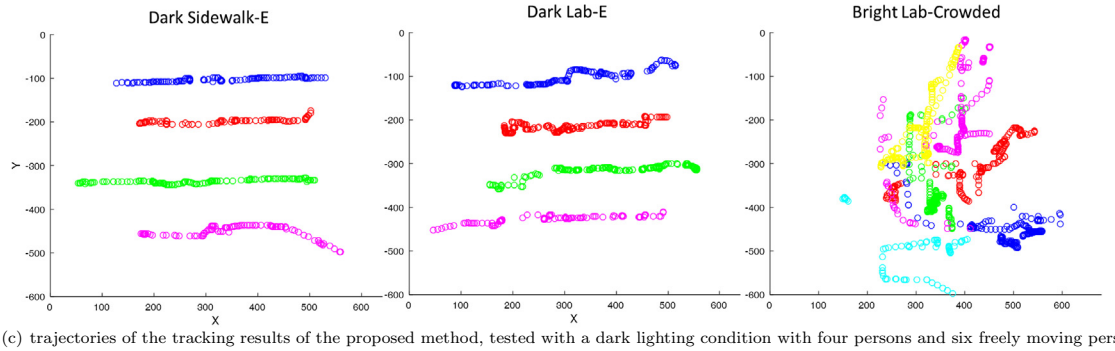
proposed method can compensate the tracking capability from the non-occluded views to the occluded views, with low computational complexity, which is suitable for realtime applications. Moreover, by using infrared depth cameras, people can be tracked closely from bright to extremely dark environments, even when occlusion occurs.



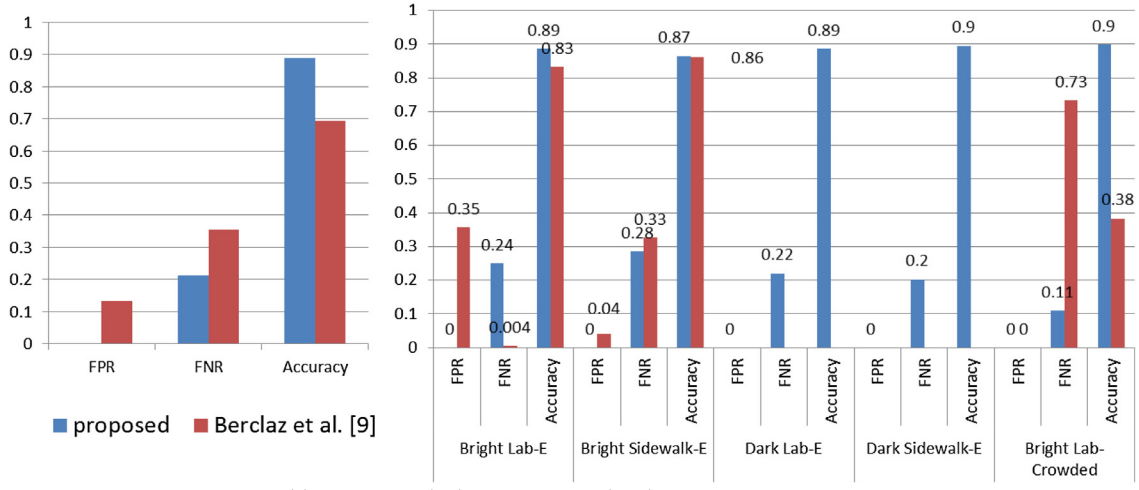
(a) results of the bright lighting condition



(b) results of the dark lighting condition (Method of Berclaz et al. [9] has poor results because of using the color information)



(c) trajectories of the tracking results of the proposed method, tested with a dark lighting condition with four persons and six freely moving persons



(d) overall results (left) and detail results (right) of FPR, FNR, and Accuracy

Fig. 24. People detection results of bright/dark lighting conditions.

Acknowledgment

This research is partially supported by Ministry of Science and Technology, Taiwan, under Grant No. MOST 104-2221-E-119-001.

References

[1] I. Haritaoglu, D. Harwood, L. Davis, W4: Who? when? where? what? a real time system for detecting and tracking people, in: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 222–227.
[2] R.T. Collins, Mean-shift blob tracking through scale space, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2003, pp. II-234-40.
[3] M. Han, W. Xu, H. Tao, Y. Gong, An algorithm for multiple object trajectory tracking, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2004, pp. I-864–I-871.
[4] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2000, pp. 142–149.
[5] S. Khan, M. Shah, Tracking people in presence of occlusion, in: Asian Conference on Computer Vision, 2000, pp. 1132–1137.
[6] Q. Cai, J.K. Aggarwal, Automatic tracking of human motion in indoor scenes across multiple synchronized video streams, in: Sixth International Conference on Computer Vision, 1998, pp. 356–362.
[7] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, S. Mayban, Principal axis-based correspondence between multiple cameras for people tracking, IEEE Trans. Pattern Anal. Mach. Intell. 28 (4) (2006) 663–671.
[8] S.M. Khan, M. Shah, Tracking multiple occluding people by localizing on multiple scene planes, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 505–519.
[9] J. Berclaz, F. Fleuret, E. Turetken, P. Fua, Multiple object tracking using k-shortest paths optimization, IEEE Trans. Pattern Anal. Mach. Intell. 33 (9) (2011) 1806–1819.
[10] Kinect. <http://www.microsoft.com/en-us/kinectforwindows>.
[11] Kinect sdk. <http://www.microsoft.com/en-us/kinectforwindowsdev/Downloads.aspx>.
[12] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1297–1304.
[13] O. Ozturk, T. Yamasaki, K. Aizawa, Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis, in: IEEE 12th International Conference on Computer Vision Workshops, 2009, pp. 1020–1027.
[14] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.
[15] M. Baum, F. Faion, U. Hanebeck, Tracking ground moving extended objects using rgbd data, in: IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems, 2012, pp. 186–191.
[16] K. Bradshaw, I. Reid, D. Murray, The active recovery of 3d motion trajectories and their use in prediction, IEEE Trans. Pattern Anal. Mach. Intell. 19 (3) (1997) 219–234.
[17] L. Lee, R. Romano, G. Stein, Monitoring activities from multiple video streams: establishing a common coordinate frame, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 758–767.
[18] F. Fleuret, J. Berclaz, R.L.P. Fua, Multi-camera people tracking with a probabilistic occupancy map, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 267–282.
[19] T. Bagautdinov, F. Fleuret, P. Fua, Probability occupancy maps for occluded depth images, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015.
[20] O. Jafari, D. Mitzel, B. Leibe, Real-time rgb-d based people detection and tracking for mobile robots and head-worn cameras, in: Proceedings of IEEE International Conference on Robotics and Automation, 2014, pp. 5636–5643.
[21] M. Munaro, E. Menegatti, Fast rgb-d people tracking for service robots, J. Auton. Robots 37 (3) (2014) 227–242.
[22] I. Mikic, S. Santini, R. Jain, Video processing and integration from multiple cameras, in: Proceedings of the Image Understanding Workshop, 1998, pp. 183–187.
[23] J. Black, T. Ellis, P. Rosin, Multi view image surveillance and tracking, in: Proceedings of Workshop on Motion and Video Computing, 2002, pp. 169–174.
[24] J. Giebel, D.M. Gavrila, C. Schnorr, A bayesian framework for multi-cue 3d object tracking, in: Proceedings of European Conference on Computer Vision, 2004, pp. 241–252.
[25] K. Smith, D. Gatica-Perez, J. Odobez, Using particles to track varying numbers of interacting people, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 962–969.
[26] S. Lee, Real-time camera tracking using a particle filter combined with unscented Kalman filters, J. Electron. Imaging 23 (1) (2014). 013029-1–013029-18.
[27] S. Gruenwedel, V. Jelaca, J.O. Nino-Castaneda, P.v. Hese, D.v. Cauwelaert, D.v. Haerenborgh, P. Veelaert, W. Philips, Low-complexity scalable distributed multicamera tracking of humans, ACM Trans. Sen. Networks 10 (2) (2014) 24:1–24:32.
[28] K. Sharma, I. Moon, Improved scale-invariant feature transform feature-matching technique-based object tracking in video sequences via a neural network and Kinect sensor, J. Electron. Imaging 22 (3) (2013). 033017-1–033017-14.
[29] A. Maimone, H. Fuchs, Reducing interference between multiple structured light depth sensors using motion, in: IEEE Virtual Reality, 2012, pp. 51–54.
[30] S.W. Sun, H.Y. Lo, H.J. Lin, Y.S. Chen, F. Huang, H.Y.M. Liao, A multiple structured light sensors tracking system that can always select a better view to perform tracking, in: Asia–Pacific Signal and Information Processing Association Annual Summit and Conference, 2009.
[31] G. Welch, G. Bishop, An Introduction to the Kalman Filter, Tech. Rep. 95-041, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 1995.
[32] S.J. Gaffney, Probabilistic Curve-aligned Clustering and Prediction with Regression Mixture Models, Ph.D. Thesis, Department of Computer Science, Department of Computer Science, University of California, Irvine, 2004.
[33] M.A. Fischler, R. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[34] http://cvlab.epfl.ch/software/pom.