International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.9 (2015), pp.101-120
http://dx.doi.org/10.14257/ijsip.2015.8.9.11

Tracking Multiple-person using Sparse Stereo Information

Keli Hu[1*], Yuzhang Gu[2], Shigen Shen[1], Cheng Zhang[2], Yunlong Zhan[2]

[1]*Department of Computer Science and Engineering, Shaoxing University, Shaoxing 312000, Zhejiang, China*
[2]*Key Laboratory of Wireless Sensor Network & Communication, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China*
**E-mail: ancimoon@gmail.com*

Abstract

In this study, we address the problem of multi-person detection and tracking in challenging scenes using sparse stereo information. In each frame, only a sparse set of object feature points are extracted. All these feature points are then projected onto a plan-view map, and grouped into several clusters by employing the biometric information, the optical flow information of object feature points, as well as the width of a person. By producing clusters, the location of a possible person can be determined. In addition, a Modified Joint Probabilistic Data Association Filter (MJPDAF) is proposed for improving the performance of measurements association during the people tracking process. Compared to the traditional JPDAF, the methods for the construction of the validation matrix and the calculation of association probabilities are improved. Experiments on challenging datasets demonstrate that the proposed algorithm is robust for people detection and tracking through fixed stereo vision.

Keywords: *stereo, binocular, detection, tracking, JPDAF, optical flow*

1. Introduction

Due to many practical applications [1,2], multi-person detection and tracking has attracted much research interest in the area of computer vision. In fact, a great many algorithms [1,2] have been proposed over the past decades. However, it remains to be a challenging problem on account of complicated surveillance scenes.

Specifically, in pedestrian detection, the classification-based algorithms are often employed [3,4]. However, such methods frequently fail when occlusion exists. Besides, they are also very sensitive to the camera's oblique angle. If the angle (Figure 1) changes a lot, the detector has to be retrained before being put into use. For object tracking, a class of successful approach defines multiple people tracking to be a global optimization problem over the complete sequence [5-7] and shows good results. However, it is highly depends on the performance of the pedestrian detector, and it is definitely very time-consuming because all the detection results of the pedestrian detector have to be input through the whole sequence. Thus, it is not quite suitable for a real-time surveillance system.

All the algorithms mentioned above are mainly based on the monocular system, in which both occlusion and illumination invariance are difficult to deal with. However, when considering a stereo vision system with short baseline, the problems are no longer so hard because the 3D information of the surveillance scene can be recovered. To our knowledge, although 3D reconstruction is still a challenging problem [8,9] especially for surveillance, many stereo algorithms have still been proposed [10-16]. Generally, 3D reconstruction can be divided into two categories: one is dense reconstruction; the other corresponds to sparse reconstruction. Most existing stereo algorithms are based on dense

ISSN: 2005-4254 IJSIP

reconstruction. For instance, in [10], disparity information were employed for tracking a rigid object. By using two stereo cameras, people in [11] were detected by employing the background subtraction scheme in both depth and color space with the help of blob clustering, and tracked by applying color histogram. The biometric information of a person was applied for people detection in [13,14]. Meanwhile, the track matching process in [13] was achieved by formulating the process as a weighted bipartite graph and using a weighted maximum cardinality matching scheme. In addition, ref [16] proposed a particle filter algorithm for tracking people in the fused plan-view maps which were generated from reconstruction information of multiple stereo cameras. The performance of all these dense reconstruction based methods highly depends on the accuracy of the depth map. Unfortunately, the existing methods for recovering dense 3D information are not reliable enough due to textureless or occlusion even with a high computational cost [8]. In general, compared to dense reconstruction, sparse feature points based reconstruction are not only more cost effective, but also more accurate [9]. An object detection and tracking algorithm based on the sparse depth map was given in [17]. However, it only applied the stereo optical flow for object segmentation and tracking, which is not suitable for tracking people in complex scenes. In addition, to our knowledge, few work have been done on sparse reconstruction based people detection and tracking. Hence, it is essential to find an effective and efficient way to address such a challenging problem.

1.1 Proposed Contribution

We present a novel algorithm based on detecting and tracking people through fixed stereo vision. In this work, only a sparse set of feature points is extracted at each time. The proposed algorithm exhibits two novel features. First, a new method for clustering sparse feature points is proposed for people detection. The biometric information, optical flow information corresponding to the sparse feature points, as well as person width is employed to assist clustering. Secondly, for people tracking, a Modified Joint Probabilistic Data Association Filter (MJPDAF) [18-21] is proposed. In MJPDAF, two different validation gates are employed for improving the construction of the validation matrix, and the optical flow information is incorporated into the calculation of association probabilities. From our experimental results, it can be verified that the proposed method for people detection and tracking produces excellent results in real-world video data.

The remainder of the paper is organized as follows: In Section 2, system overview is given. In Section 3, preliminaries such as the method of sparse 3D feature calculation and the method for making a plan-view map are presented. In Section 4, the algorithm for people detection is proposed. In Section 5, method for people tracking is given. Experimental evaluations and discussions are presented in Section 6, and Section 7 is the conclusion.

2. System Overview

In this work, we have designed a short baseline binocular stereo vision system, in which each camera is placed at an over-head position with an oblique angle. Figure 1 gives an intuitive graph illustration. Such a installation method can greatly reduce occlusion on the image plane. In this work, we assume that there exists a flat ground plane in the surveillance scene. Both people detection and tracking are processed and achieved on that ground.

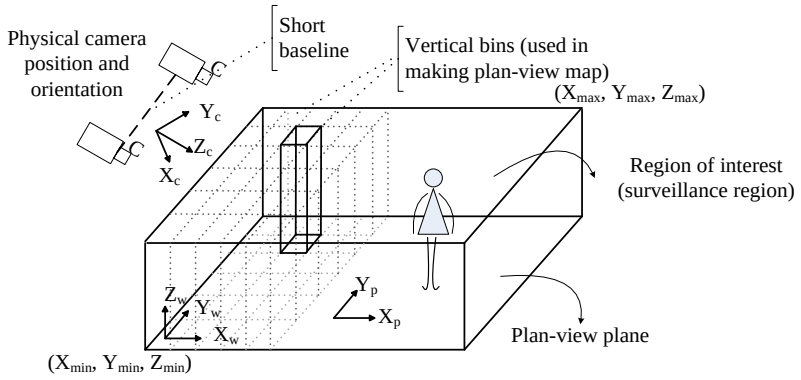

Figure 1. System Setup

As shown in Figure 2, the method for people detection and tracking proposed here is mainly based on sparse 3D features. People detection problem is solved by clustering sparse feature points. People tracking results are achieved by using the information which the cluster set provides. Details of each block in Figure 2 can be found in the following sections.

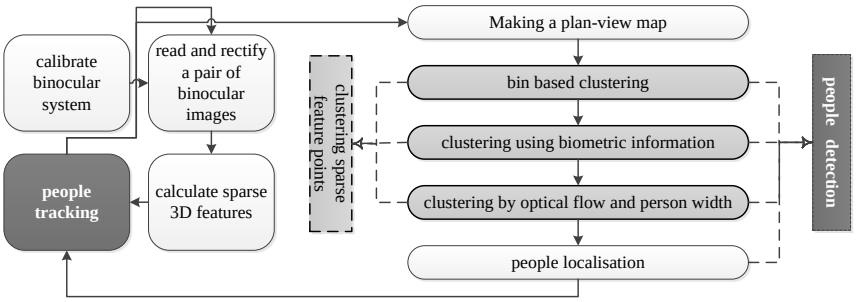

Figure 2. Flowchart of people detection and tracking algorithm

3. Preliminaries

3.1. Sparse 3D Feature Calculation

In our system, only a sparse set of feature points is required at each time. With the short baseline between two cameras, the point correspondence in the two views can be easily established. While a variety of methods for obtaining stable correspondences are available [8,9], the method proposed in [9] is applied here owing to its good performance.

By using this method, a set of feature points $\mathbf{P}=\left\{\mathbf{P}^i \middle| \mathbf{P}^i=\left(u_i,v_i,d_i\right)\right\}, i=1,...,N$ is then achieved. Each point holds a disparity d_i and its location (u_i,v_i) in the left rectified image. Let $\mathbf{P}_c=\left\{\mathbf{P}_c^i, i=1,...,N \middle| \mathbf{P}_c^i=\left(X_c^i,Y_c^i,Z_c^i\right)\right\}$ denote the corresponding 3D coordinates set in the "camera" reference system, which is given by

$$Z_c=\frac{bf_u}{d}, X_c=\frac{Z_c\left(u-u_o\right)}{f_u}, Y_c=\frac{Z_c\left(v-v_o\right)}{f_v} \tag{1}$$

where f_u and f_v are the horizontal and vertical focal length of the left camera, b is the length of the baseline, u_0 and v_0 are the displacement (away from the optic axis) of the center of coordinates on the projection screen.

3.2. Making a Plan-view Map

The decision of orthogonally projecting sparse feature points into a plan-view map is supported by the fact that people do not tend to be overlapped on the floor plane as much as they are in the originally captured images, which is important for the robustness of people detection and tracking algorithm.

To make a plan-view map, we first translate the coordinates of the extracted feature points to a "world" reference system which is placed at ground level and parallel to it, denoted by $\mathbf{P}_w$.

In our system, a ROI (region of interest) is defined. As shown in Figure 1, a cuboid ROI is applied here.

A plan-view map divides the region of the floor plane of the ROI into a set of cells of fixed size (δ_x, δ_y). In this work, the origin of the plan-view coordinate coincides with the "world" reference system's original position. Hence, the cell (x^i, y^i) into which the 3D point $\mathbf{P}^i$ is projected can be calculated as

$$x^i = X_w^j / \delta_x ; y^i = Y_w^j / \delta_y \tag{2}$$

The set of feature points projected into each cell can be expressed as

$$\mathrm{M}^i = \left\{ \mathbf{P}^j \middle| \begin{array}{l} X_w^j / \delta_x = x^i, Y_w^j / \delta_y = y^i, \\ X_w^j \in [X_{\min}, X_{\max}], Y_w^j \in [Y_{\min}, Y_{\max}], Z_w^j \in [Z_{\min}, Z_{\max}] \end{array} \right\} \tag{3}$$

where $(X_{\min}, Y_{\min}, Z_{\min})$ and $(X_{\max}, Y_{\max}, Z_{\max})$ are two boundary points of the cuboid ROI.

In addition to the plan-view map, we create a height map H, which indicates the maximum height of the points projected into each cell (x^i, y^i). Mathematically, it is defined to be

$$\mathrm{H}^i = Z_w^\alpha, \alpha = \underset{\mathbf{P}^j \in \mathbf{M}^i, \alpha = j}{\arg\max} \left(Z_w^j \right) \tag{4}$$

4. People Detection

On the plan-view map, a person can be simply represented by an ellipse with a unique angle of rotation (see Figure 3). The projected points belonging to a person are close to each other and should be grouped into clusters to indicate the possible locations of each people. However, the number of people located in the ROI is unknown at each time. In other words, the number of clusters is not fixed over time. This uncertainty causes a critical problem in classical clustering algorithms which require the number of cluster in prior. Thus, a novel and effective clustering algorithm is proposed in this section. Such an algorithm consists of three steps (see Figure 2), details of each step are described below.

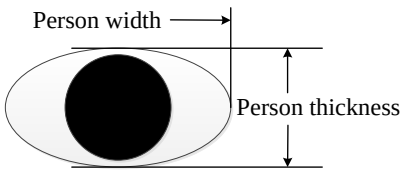

Figure 3. The Projection of a Person from the Top View

4.1 Bin based Clustering

In Figure 1, vertical bins are used to make the plan-view map. Meanwhile, those bins are also applied for constructing the basic unit of cluster, which called bin-cluster here. Feature points projected into the same bin are grouped into a specific bin-cluster

$$\mathrm{B}^i = \left\{ \mathbf{P}^j \middle| \mathbf{P}^j \in \mathrm{M}^{\,i}, \mathrm{M}^{\,i} \neq \varnothing \right\} \tag{5}$$

4.2 Clustering using Biometric Information

In this subsection, we try to group bin-clusters based on the fact that the projected feature points belonging to a person are close to each other. In other words, if the distance between two bin-clusters is small enough, it can be thought that those two bin-clusters are from the same person. Then such a clustering problem turns to finding a distance threshold to judge whether any two of these bin-clusters are from one person or not.

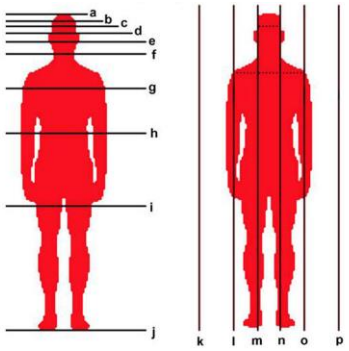

Figure 4. Golden Ratio: the Left One is Vertical, the Other is Horizontal

The biometric information which has been applied for assisting region clustering [13,14] is employed here to tackle this problem. The number ϕ ($\phi \sim 1.618$), which is known as the Golden Ratio, was employed to build the biometric information of a person. Figure 4 shows how a body is segmented using ϕ. As defined in [14], $|aj|$ is the height of a human body, $|ai| = |aj|/\phi$, $|ah| = |ai|/\phi$, $|lo| \equiv |ag| = |ah|/\phi$. Then we can get $|lo| = |aj|/\phi^3$, where $|lo|$ is the width of the shoulder. In this work, only the ratio between the height of a person and the width of its shoulder is used to group those bin-clusters into a bio-cluster set O.

Each bin-cluster B^i is considered as a virtual person with H^i to be its height, and $\xi = \mathrm{H}^i/\phi^3$ to be its shoulder width. A relationship called adjacent bin-cluster is defined as $\mathrm{B}^m \xleftrightarrow{Adjac} \mathrm{B}^n$, ($m \neq n$), which means B^m is an adjacent bin-cluster of B^n, and vice versa. Accordingly, we obtain

$$\left\| \mathbf{L}^{\mathrm{B}^m}, \mathbf{L}^{\mathrm{B}^n} \right\| < \frac{\max\left(\mathrm{H}^m, \mathrm{H}^n\right)}{2\phi^3} \Leftrightarrow \mathrm{B}^m \xleftrightarrow{Adjac} \mathrm{B}^n \tag{6}$$

where $\mathbf{L}^{\mathrm{B}^m} = \left(\delta_x x_{PV}^{\mathrm{B}^m}, \delta_y y_{PV}^{\mathrm{B}^m}\right)$, $\left(x_{PV}^{\mathrm{B}^m}, y_{PV}^{\mathrm{B}^m}\right)$ is the location of B^i in the plan-view map.

Suppose all bin-clusters B have been grouped into K bio-clusters $\mathrm{O} = \left\{\mathrm{O}^i\right\}, i = 1,...,K$. For each bio-cluster,

$$\exists\, \mathrm{B}^m \xleftrightarrow{Adjac} \mathrm{B}^n, \mathrm{B}^m \subseteq \mathrm{O}^i \wedge \mathrm{B}^n \subseteq \mathrm{O}^i \wedge m \neq n \tag{7}$$

Similarly, for each pair of different bio-clusters, none of the bio-clusters should meet the criteria

$$\exists\, \mathrm{B}^m \xleftrightarrow{Adjac} \mathrm{B}^n, \mathrm{B}^m \subseteq \mathrm{O}^i \wedge \mathrm{B}^n \subseteq \mathrm{O}^j \wedge m \neq n \wedge i \neq j \tag{8}$$

4.3 Clustering by Optical Flow and Person Width

Looking into the method discussed in section 4.2, it can be find easily that feature points belonging to different people with small distance will be grouped into one cluster,

which results in wrongly identifying two people to be one. Two kinds of typical position relations between people from the top view are shown in Figure 5. Suppose the feature points belonging to each individual person are already grouped together. When the distance $d < \max(h_A, h_B)/2\phi^3$ (Eq. (6)), person A and B in Figure 5 will be wrongly merged, where h_A and h_B are the height of the corresponding bin-clusters respectively. It is reasonable that we set max(h_A, h_B) as 2.0 meters, because almost every person is lower than that. Then we can get $\max(h_A, h_B)/2\phi^3 = 0.236$ meter. Under normal circumstance, a person will not follow another one too close (less than 0.236 meter). As a result, for the pedestrian monitoring system, there is no problem if we assume that people with the position relation shown in Figure 5(b) will not merged as one person by using the method mentioned in Section 4.2. However, the case shown in Figure 5(a) must be seriously considered, because people frequently walk together shoulder by shoulder.

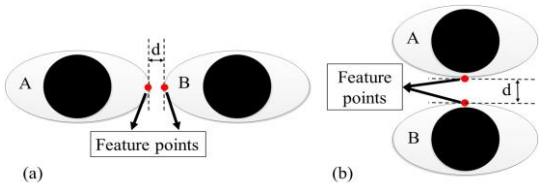

Figure 5. Typical Position Relations between People from the Top View

As shown in Figure 6, both optical flow and person width are employed to address such a problem. Optical flow is used to incorporate people detection and tracking results at time *t*-1 into the judgment of large bio-cluster at time *t*. Person width is employed to ensure there are no clusters much larger than the size of a real person. Details for how to use those two kinds of information are given below.

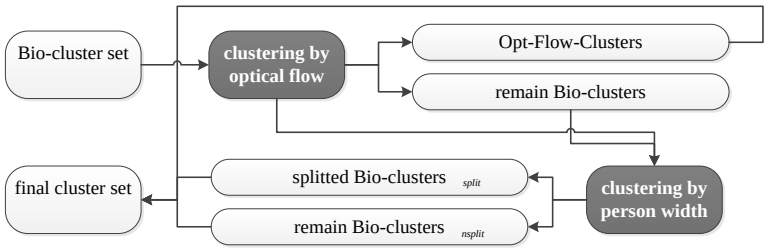

Figure 6. Flowchart of the Method of Clustering by Optical Flow and Person Width

4.4 Clustering by Optical Flow

Suppose R persons are tracked at time $t-1$, denoted by a set P_{t-1}. Each person P_{t-1}^{i} holds the feature points belonging to it on the left image plane. In this paper, only the feature points at time $t-1$ are considered, denoted by $\mathbf{P}^{\mathbf{P}_{t-1}^{i}}$. The optical flow for $\mathbf{P}^{\mathbf{P}_{t-1}^{i}}$ is calculated by using the iterative Lucas-Kanade method with pyramids. $\mathrm{F}\left(\mathbf{P}^{\mathbf{P}_{t-1}^{i}}\right)$ is the corresponding points set of $\mathbf{P}^{\mathbf{P}_{t-1}^{i}}$ after the process of optical flow. Then the convex hull $\mathrm{V}\left(\mathrm{F}\left(\mathbf{P}^{\mathbf{P}_{t-1}^{i}}\right)\right)$ of each optical flow points set is calculated [22]. Feature points located in $\mathrm{V}\left(\mathrm{F}\left(\mathbf{P}^{\mathbf{P}_{t-1}^{i}}\right)\right)$ will be assigned a ID with the value i. Once the feature points are projected, the bin-cluster is assigned with the same ID as the corresponding feature points. Bin-clusters with more than one ID are out of the consideration. Finally, the convex hulls

of bin-clusters with the same ID are also calculated on the plan-view map. Bin-clusters located in each hulls will be given a new ID, which is equal to the ID corresponding to the convex hull. Up to now, each bin-cluster has been assigned a unique ID, which signifies which person it may belong to.

Figure 7 gives the illustration of this method. The solid circles stand for the feature points, and the hollow circles denote the points generated by the optical flow. As shown in Figure 7(b), points in the optical flow convex hulls are assigned one or more ID. As discussed above, bin-clusters with more than one ID are out of consideration when constructing convex hulls on the plan-view map (see Figure 7(c-d)). Finally, all the bin-clusters located in the same convex hulls are labeled with a unique ID in Figure 7(d).

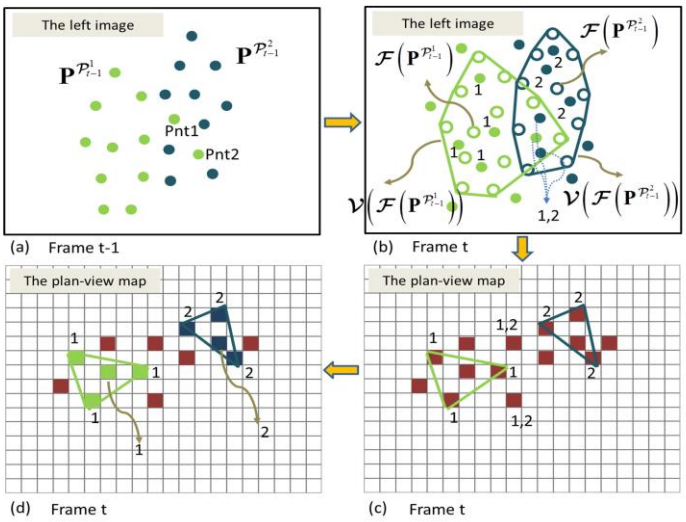

Figure 7. Processing Steps for Labeling Bin-clusters at Time t using the Optical Flow Information

For convenience, a new concept named optical-flow-projection-cluster (OFPC) is defined here. For each bio-cluster O^i, all the bin-clusters belonging to it with the same ID constitute an OFPC. The OFPC set for O^i is denoted by $\mathrm{OFP}_{\mathrm{O}^i}$. Suppose $\mathrm{OFP}_{\mathrm{O}^i}^c \subseteq \mathrm{OFP}_{\mathrm{O}^i}$, then the superscript c equals to the value of the ID which the corresponding OFPC holds. Now we get

$$\mathrm{OFP}_{\mathrm{O}^i}^c = \left\{ \mathrm{B}^j \middle| ID^{\mathrm{B}^j} = c, \mathrm{B}^j \subseteq \mathrm{O}^i \right\} \tag{9}$$

where ID^{B^j} is the ID of the bin-cluster B^j. Assuming there are B bin-clusters in $\mathrm{OFP}_{\mathrm{O}^i}^c$. The distance between a bin-cluster B^j and $\mathrm{OFP}_{\mathrm{O}^i}^c$ is given by

$$\mathrm{D}\left(\mathrm{B}^j, \mathrm{OFP}_{\mathrm{O}^i}^c\right) = \min\left(\left\| \mathbf{L}^{\mathrm{B}^j}, \mathbf{L}^{\mathrm{B}^k_{\mathrm{OFP}^c_{\mathrm{O}^i}}} \right\| \right), k = 1, \ldots, B \tag{10}$$

where $\mathrm{B}^k \subseteq \mathrm{OFP}_{\mathrm{O}^i}^c$, $\mathbf{L}^{\mathrm{B}^k_{\mathrm{OFP}^c_{\mathrm{O}^i}}} = \left(\delta_x x_{PV}^{\mathrm{B}^k}, \delta_y y_{PV}^{\mathrm{B}^k} \right)$, which is the location of the k^{th} bin-cluster belonging to $\mathrm{OFP}_{\mathrm{O}^i}^c$. Similarly, the distance between B^j and $\mathrm{OFP}_{\mathrm{O}^i}$ is defined as

$$\mathrm{D}\left(\mathrm{B}^j, \mathrm{OFP}_{\mathrm{O}^i}\right) = \min\left(\mathrm{D}\left(\mathrm{B}^j, \mathrm{OFP}_{\mathrm{O}^i}{}^k\right) \right), k = 1, \ldots, P \tag{11}$$

where P is the number of OFPC that $\mathrm{OFP}_{\mathrm{O}^i}$ contains.

Considering a bio-cluster O^i, if $P>1$, and there are less than θ_1 bin-clusters B^j which satisfy the condition $D\left(B^j, OFP_{O^i}\right)>\theta_2$, where $B^j \subseteq O^i$. Then we can draw the conclusion that O^i needs to be divided into P opt-flow-clusters, denoted by OF. The corresponding splitting method is presented below:
Given the bio-cluster O^i, $OFP_{O^i} \subseteq O^i$.

(1) Create P opt-flow-clusters $OF_{O^i}=\left\{OF_{O^i}^k\right\}, k=1,...P$, $OF_{O^i}^k=OFP_{O^i}{}^k$.

(2) Considering each bin-cluster B^j, $B^j \subseteq O^i \wedge B^j \not\subset OFP_{O^i}$. Add B^j to $OF_{O^i}^m$, where $m=\arg\min_{m=k}\left\{D\left(B^j, OF_{O^i}^k\right), k=1,...,P\right\}$.

The calculation of the distance between a bin-cluster and an opt-flow-cluster is the same as the method defined in Eq. (10). Finally, we get a set of opt-flow-clusters $OF=\left\{OF_{O^i}\right\}$, where O^i is the bio-cluster which meet splitting conditions.

For other bio-clusters which do not meet such conditions, another judgment is proposed in the following subsection.

4.5 Clustering by Person Width

As seen in Figure 3, a person can be simply represented by an ellipse with a unique angle of rotation from the top view. The algorithm used in [23] is applied here for calculating an ellipse. As shown in Figure 8(b), the fitting algorithm performs well even there are some noise points (pointed by the arrow) in the cluster. Take the bio-cluster O^i for example, the ellipse for O^i is shown in Figure 8(a). The long axis of the ellipse is employed as the width of the bio-cluster, denoted by a, and the short axis is applied as the thickness b. Both the width and thickness are measured in cell with a fixed size $\left(\delta_x, \delta_y\right)$. The real length of a and b, denoted by cw and ct, which are measured in meter, can be calculated as

$$\begin{cases} cw=\sqrt{\left(a\delta_x\cos\alpha\right)^2+\left(a\delta_y\sin\alpha\right)^2} \\ ct=\sqrt{\left(b\delta_x\sin\alpha\right)^2+\left(b\delta_y\cos\alpha\right)^2} \end{cases} \tag{12}$$

where α is the rotation angle.

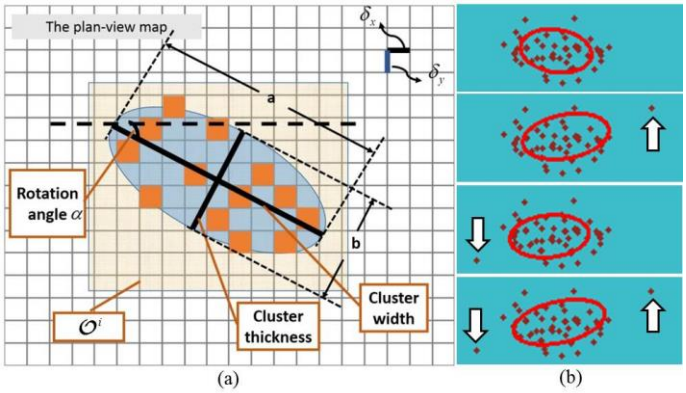

Figure 8. (a) Ellipse Fitting, (b) Ellipse Fitting Examples

As discussed in subsection 4.3, the feature points belonging to people with the similar position relation shown in Figure 5(b) will not construct a large cluster. Thus, with the width and the thickness information, only the width of a bio-cluster is applied for judging whether a remain cluster is large or not. To improve system's robustness, the optical flow

information is also incorporated into the judgment. The number S, which signifies how many small bio-clusters should be constructed from the corresponding large bio-cluster, can be calculated as

$$S = \max\left(P, \lfloor cw/W \rfloor + f\left(\frac{cw - W\lfloor cw/W \rfloor}{W}\right)\right), f(x) = \begin{cases} 1 & x > \theta_3 \\ 0 & else \end{cases} \tag{13}$$

where W is the width of a person, P is the number of OPFC belonging to this bio-cluster. If $S > 1$, then O^i needs to be divided into S small bio-clusters using the splitting method presented below:

Given the bio-cluster O^i, $\mathrm{B}^j \subseteq \mathrm{O}^i$.

(1) Define a point set $\mathbf{P} = \left\{\mathbf{L}^{\mathrm{B}^j} \middle| \mathrm{B}^j \subseteq \mathrm{O}^i\right\}$.

(2) Employ K-means algorithm (use kmeans++ center [24]) to divide **P** into S plan-view clusters.

(3) Bin-clusters corresponding to each plan-view cluster are combined to be a new bio-cluster.

Both the splitting judgment and the splitting method mentioned above are iteratively repeated until no more bio-clusters can meet the conditions for splitting.

Till now, clusters we've got consist of the opt-flow-clusters OF, the splitted bio-clusters O_{split}, and the remain bio-clusters which do not satisfy any splitting condition, denoted by O_{nsplit}. All these clusters constitute the final cluster set $\mathrm{C} = \left\{\mathrm{OF}, \mathrm{O}_{split}, \mathrm{O}_{nsplit}\right\}$, while clusters with less than 5 bin-clusters are ignored.

For each cluster C^i, the width cw_{C^i} and thickness ct_{C^i} are calculated. The center of the corresponding ellipse measured in cell is employed as the cluster location on the plan-view map, denoted by $\mathbf{L}^{\mathrm{C}^i}$. The OFPC set of C^i is denoted by $\mathrm{OFP}_{\mathrm{C}^i}$. In addition, feature points on the left image plane which have been projected into C^i are represented as $\mathbf{P}^{\mathrm{C}^i}$, and the convex hull $\mathrm{V}\left(\mathbf{P}^{\mathrm{C}^i}\right)$ of $\mathbf{P}^{\mathrm{C}^i}$ is also calculated. Finally, by employing the method used in [25], the corresponding color histogram $\mathbf{\Gamma}_{\mathrm{C}^i}$ is given by

$$\mathbf{\Gamma}^u_{\mathrm{C}^i} = D_0 \sum_{i=1}^{n} k\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \delta[b(\mathbf{x}_i - \mathbf{x}) - u] \tag{14}$$

where D_0 is a normalization constant, u is the index of the histogram bin, h is the scale of the bounding box of $\mathbf{P}^{\mathrm{C}^i}$, $\mathbf{x}$ is the center of the bounding box, n is the number of the pixels located in the area of $\mathrm{V}\left(\mathbf{P}^{\mathrm{C}^i}\right)$, and $\mathbf{x}_i$ represents the corresponding pixel location. Details for the function $b(\mathbf{x}_i - \mathbf{x})$ can be found in [25]. $k(\mathbf{x})$ is a kernel function, which is defined as

$$k(\mathbf{x}) = \begin{cases} \frac{1}{2} c_d^{-1}(d+2)\left(1 - \|\mathbf{x}\|^2\right) & \|\mathbf{x}\| < 1 \\ 0 & otherwise \end{cases} \tag{15}$$

where c_d is the volume of the unit d-dimensional sphere.

4.6 People Localization

The people tracking block described in Section 0 assigns several clusters to people in tracking process. Suppose C_t^i is one of the remain clusters at time t. Considering R

persons are tracked, $C^{P_t^j}$ is the cluster assigned to the j^{th} person. If C_t^i satisfies the condition

$$\lambda_0 W > cw_{C_t^i} > \lambda_1 W \wedge \lambda_0 T > ct_{C_t^i} > \lambda_1 T \wedge D\left(C_t^i, C^{P_t^j}\right) > T, j = 1,...R,$$

the cluster C_t^i will be considered as a new target, where T is the thickness of a person, λ_0 and λ_1 can be set to 1.3 and 0.7 respectively, and $D\left(C_t^i, C^{P_t^j}\right)$ is defined as the distance between cluster centers.

5. People Tracking

In our system, when a person is totally occluded by another one or no feature points belonging to this person are detected, cluster corresponding to this person will disappear. In addition, feature points belonging to one person may be projected into more than one cluster because of error projection (the precision of the 3D location will get worse and worse with the increase of the distance between people and camera). Due to this, there may be no measurements or multiple measurements for a single person. To solve this problem during the tracking process, MJPDAF is proposed here. The JPDAF [18] is an existing extension to the PDAF [21], which has shown to be very effective in handling clutter and missed detections [18-20].

Considering R persons are tracked. Let $\mathbf{Z}(t) = \{\mathbf{z}_1(t),...,\mathbf{z}_{m_t}(t)\}$ denote the measurement set which consists of all clusters detected at time t, where $\mathbf{z}_j(t) = \mathbf{L}_t^{C^j}$, m_t equals to the number of clusters we've got at time t.

The main difference between the MJPDAF and JPDAF is the method for constructing the validation matrix and calculating the association probabilities. We mainly discuss the difference below. Details for other information of JPDAF can be found in [18].

Traditionally, the following validation matrix is defined [18]

$$\mathbf{\Omega}_t^{'} = \left[\omega_{ji}^{'}\right], j = 1,...,m_t, i = 0,1,...,R \tag{16}$$

with binary elements to indicate if measurement j lies in the valid area for target i. Normally, the valid area is a "g-sigma" ellipsoid $\left\{\boldsymbol{\gamma}_j^{i\prime}\mathbf{S}^{-1}\boldsymbol{\gamma}_j^i < G\right\}$, where $\boldsymbol{\gamma}_j^i$ is the innovation vector [18], $\mathbf{S}$ is the covariance matrix [18], G is the validation gate. Index i =0 stands for "no target" and the corresponding column of $\mathbf{\Omega}_t^{'}$ has all units (each measurement could have originated from clutter or false alarm).

A proper validation gate helps to filter out most of the measurements which are not originated from a specific person. However, in this study, due to the noise of 3D reconstruction, locations of clusters are not stable enough. Some valid measurements may be wrongly discarded if only one gate is employed. Therefore, in MJPDAF, two different validation gates are defined, denoted by G_0 and G_1, $G_0 < G_1$ ($G_0 \in [1.0, 6.0], G_1 \in [6.1, 70.0]$ based on experience). Those measurements localised in the small validation gate are considered to be credible, but those localised between G_0 and G_1 cannot be easily seen as valid measurements for a person. To solve such a problem, both the color and optical flow information of clusters are taken into consideration. In the following, details for the MJPDAF are given.

In MJPDAF, the validation matrix $\mathbf{\Omega}_t = \left[\omega_{ji}\right]$ holds the same structure as $\mathbf{\Omega}_t^{'}$. Steps for constructing $\mathbf{\Omega}_t$ are presented below:

Given information of each person at time $t-1$, the clusters $\mathbf{C}_t = \{\mathbf{C}_t^j\}$, $j = 1,...,m_t$, $\mathbf{C}_t^j$ is the cluster corresponding to $\mathbf{z}_j(t)$.

(1) Initialize $\mathbf{\Omega}_t$, set $\omega_{j0} = 1, \omega_{ji} = 0, j = 1,...,m_t, i = 1,...R$.

(2) If the measurement $\mathbf{z}_j(t)$ lies in the small validation gate G_0, two cases are considered:
   a) If $nm_t^i < \theta_4$, we set $\omega_{ji} = 1$, where nm_t^i means that measurements originated from $\mathbf{P}^i$ are not detected in the past nm_t^i consecutive frames.
   b) If $nm_t^i < \theta_4 \wedge cw_{\mathbf{C}_t^j} > 0.7W \wedge ct_{\mathbf{C}_t^j} > 0.7T$, ω_{ji} is set as 1, and ω_{j0} is set as 0, which means $\mathbf{z}_j(t)$ cannot be a noise measure for $\mathbf{P}^i$.

(3) When the measurement $\mathbf{z}_j(t)$ lies in G_1, but out of G_0, and $nm_t^i < \theta_4$, if $\mathbf{z}_j(t)$ satisfies the condition

$$\mathrm{PN}\left(\mathbf{OFP}_{\mathbf{C}_t^j}^i\right) > 0.5\mathrm{PN}\left(\mathbf{C}_t^j\right) \vee \mathrm{L}\left(\mathbf{\Gamma}_{\mathbf{C}_t^j}, \mathbf{\Gamma}_{\mathbf{P}^i}\right) > \theta_5,$$

then ω_{ji} is set as 1. $\mathrm{PN}\left(\mathbf{OFP}_{\mathbf{C}_t^j}^i\right)$ is the number of the feature points projected into $\mathbf{OFP}_{\mathbf{C}_t^j}^i$, and $\mathbf{OFP}_{\mathbf{C}_t^j}^i$ is the OFPC with the ID i corresponding to $\mathbf{P}^i$. $\mathrm{PN}\left(\mathbf{C}_t^j\right)$ is the number of the feature points projected into $\mathbf{C}_t^j$. $\mathbf{\Gamma}_{\mathbf{P}^i}$ is the color histogram of $\mathbf{P}^i$ which was calculated when it was first detected, using the same method as $\mathbf{\Gamma}_{\mathbf{C}_t^i}$ (Eq. (14)). The likelihood between $\mathbf{\Gamma}_{\mathbf{C}_t^i}$ and $\mathbf{\Gamma}_{\mathbf{P}^i}$ is given as

$$\mathrm{L}\left(\mathbf{\Gamma}_{\mathbf{C}_t^i}, \mathbf{\Gamma}_{\mathbf{P}^i}\right) = \sum\nolimits_u \sqrt{\mathbf{\Gamma}_{\mathbf{C}_t^i}^u \mathbf{\Gamma}_{\mathbf{P}^i}^u} \tag{17}$$

Suppose β_j^i is the traditional association probability, which denotes the probability that the measurement $\mathbf{z}_j(t)$ belongs to $\mathbf{P}^i$. To introduce cluster association probability between consecutive frames in the image space into measurement association, the optical flow information is employed here. The newly defined association probability is given by

$$\beta_j^{i\,new} = D_1\left(\alpha P_{opt}^{ij} + (1-\alpha)\beta_j^i\right) \tag{18}$$

where D_1 is a normalization constant, and P_{opt}^{ij} is the cluster association probability calculated by using the optical information in the image space, which is given by

$$P_{opt}^{ij} = \mathrm{PN}\left(\mathbf{OFP}_{\mathbf{C}_t^j}^i\right) \Big/ \mathrm{PN}\left(\mathbf{C}_t^j\right) \tag{19}$$

After the process of MJPDAF, the information for each person $\mathbf{P}^i$ needs to be updated. If no measurements originated from $\mathbf{P}^i$ are detected, we set $\mathbf{P}^{\mathbf{P}_t^i} = \mathbf{F}\left(\mathbf{P}^{\mathbf{P}_{t-1}^i}\right)$ when $nm_t^i \le \theta_4$. On the contrary, if $\mathbf{P}^i$ owns valid measurements at time t, let $\beta_m^{i\,new}$ denote the maximum association probability and $\beta_n^{i\,new}$ represent the second maximum one, $m \ne n$, then two different cases are considered:

a) If $\beta_m^{i\,new} > 0$, the feature points projected into $\mathbf{C}_t^m$ are assigned to $\mathbf{P}^i$ as new feature points at time t.
b) If $\beta_n^{i\,new} / \beta_m^{i\,new} > \theta_6$, both the feature points belong to $\mathbf{C}_t^m$ and $\mathbf{C}_t^n$ are assigned to $\mathbf{P}^i$.

6. Experiments

To the best of our knowledge, there are few public databases prepared for testing algorithms based on the short baseline binocular system. So we collected several challenging video sequence sets from our own stereo vision system. Our stereo vision system consists of two cameras with approximate parallel optical axes (Figure 1). All the challenging video sequence sets are captured at 25 frames per second with resolution 704x576.

Scene 1. A pathway aside a wall Scene 2. A corner of a pathway

Figure 9. Scenes for Experiments

Two different outdoor scenes are taken into our consideration. First, Scene 1, as shown in Figure 9, the binocular system is mounted 2.5 meters high from the ground with a small oblique angle (about 15 degree), and the baseline is about 1m. The pathway is set to be the surveillance ROI. Due to small oblique angle, it is more challenging when multi-people appears in the ROI, because persons far away from the camera will be frequently and seriously occluded by humans close to the stereo vision system. Second, in the Scene 2, our stereo vision system is placed at a rather high position with a large oblique angle (about 3 meters high, 50 degree), and the baseline is shorter (about 50cm). Thus, people can hardly be wholly occluded, then corresponding feature points belonging to partly occluded people can be extracted. From this point, it is more suitable for multi-people tracking than Scene 1. However, it is still quite challenging for existed detection and tracking algorithms. In addition, challenges such as illumination variation, shadow interference often happen in this Scene.

In the following, we will mainly discuss the performance of our approach dealing with single person detection and tracking, multi-person detection and tracking with and without large pose changes. Several captures of the detection and tracking results and some quantitative analysis of our method are given.

For comparison, we implemented HJRMT (Hypergraphs for Joint Multi-View Reconstruction and Multi-Object Tracking) [7]. It has been proved that HJRMT achieved results which significantly exceed the current state of the art by testing on the PETS 2009 dataset, even if detections from only one view were used [7]. In this paper, comparisons between our method and HJRMT of Scene 1 are given (object detector in Scene 2 performs poorly due to the large oblique angle). Specifically, only detections from the left view are employed for HJRMT (small difference between the two views of our system). For object detection, we also use discriminatively trained deformable part models [26] as HJRMT did in [7]. For the corresponding detection code, we use the version provided by OpenCV Library [Available at http://opencv.org/].

6.1 Setting Parameters

The width of a person W is set as 0.4 meter, and the thickness T is set as 0.3 meter. θ_1, θ_2 and θ_3 mentioned in Section 4.3 are set as 8, $0.2W$ and 0.75 respectively. In addition, we set $\theta_4 = 10$, $\theta_5 = 0.7$ and $\theta_6 = 0.7$, which are mentioned in Section 5. The cell size of the plan-view map is set as $(0.02, 0.01)$ measured in meter. The ratio α

defined in Eq. (18) is set as 0.5. It is worth emphasising that all these parameters are kept for all experiments. For ROI, with the same "world" reference system which is shown in Figure 1, we set $X \in [0,14]$, $Y \in [0,2.8]$, $Z \in [0.2,2]$ for Scene 1 (measured in meter), and $X \in [0,6.5]$, $Y \in [0,3.5]$, $Z \in [0.2,2]$ for Scene 2.

6.2 Experimental Results

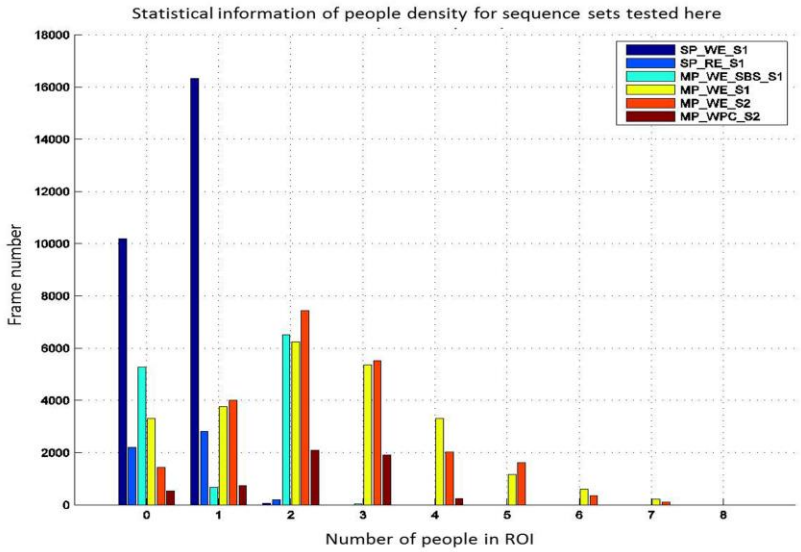

Figure 10. People Density for Each Sequence Set

Table 1 lists the information of all 6 testing sequence sets. The first four sequence sets are captured from Scene 1 and the last two from Scene 2. In addition, Figure 10 gives a quantitative analysis of people density for each sequence set.

Table 1. Information of Testing Sequence Sets

Name	Brief	Total Frames
SP_WE_S1	Single person, walking erect	26591
SP_RE_S1	Single person, running erect	5208
MP_WE_SBS_S1	Multi-person, walking erect, shoulder by shoulder	12489
MP_WE_S1	Multi-person, walking erect	23970
MP_WE_S2	Multi-person, walking erect	22492
MP_WPC_S2	Multi-person, some people change their poses	5492

6.2.1. Single Person Experiments: As shown in Figure 11, a single person can be robustly detected and tracked using our method. The hollow circles on the image plane denote the locations of target's optical flow points, and the solid circles correspond to the target's feature points (see Figure 12). As shown in Figure 11-Figure 15, all points on the background are filtered out by using the preset ROI and the people detection method proposed here. Each white image, located on the left side of the left rectified image, is the plan-view image corresponding to the plan-view map. Each cell of the plan-view map corresponds to a pixel on the plan-view image. The rectangle with a greater size represents the boundary of the ROI projected onto the plan-view map. For convenience, each plan-view image is resized into an appropriate size for showing. The small rectangle on the upper left stands for the area on the plan-view map projected by a cube with the

volume $1m^3$ corresponds to the "world" reference system. This rectangle is applied here for demonstrating the real size of the corresponding plan-view image before scaling.

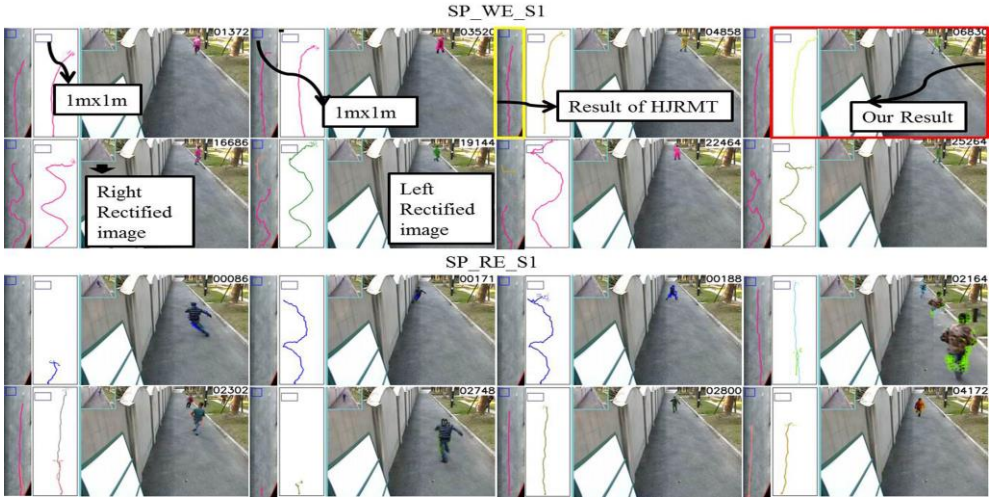

Figure 11. Single Person Detection and Tracking Result (captures for SP_WE_S1, SP_RE_S1)

Owing to our improved MJPDAF, in which both optical flow and color information are used, our method performs well even when people running through the ROI (See row 3-4 in Figure 11). Specifically, as shown in the first three images in the 3rd row, even the person keeps changing his moving direction while running, our method can still perform very well.

As shown in the first row in Figure 11, HJRMT performs well when people walk straightly. However, when people change their directions while walking, HJRMT often fails. What is worse, HJRMT even does not start to track when people run very fast. That is mainly because the object detector [26] frequently fail to detect people when the pose of people changes too much while running.

Figure 12. Multi-person Detection and Tracking Result (captures for MP_WE_SBS_S1)

6.2.2 Multi-person Detection and Tracking: Most object detection and tracking algorithms are sensitive to the distance between objects, i.e., they may regard different objects as a single one when they are close to each other. In this subsection, we show the performance of HJRMT and our algorithm under this condition (see Figure 12). For HJRMT, Identity Switches (IDS) often happens. However, as we can see, our algorithm can successfully finish detection and tracking. This means, by clustering the feature points using optical flow and person width, we have successfully resolved the problem occurred in Figure 5. Besides the results shown in Figure 12, targets 1-3 shown in frame #12030 (2nd row in Figure 14) are also well detected and tracked though they are walking closely.

Figure 13. Multi-person Detection and Tracking Result (Captures for MP_WE_S1)

For a multi-person detection and tracking system, the occlusion among persons is very common yet challenging. We try to solve this problem by projecting the feature points onto a plan-view map. However, sometimes few feature points belonging to targets, even no feature points can be detected because of severe occlusion. Such a problem may affect the performance of our algorithm. Therefore, the sequence sets MP_WE_S1 and MP_WE_S2 are collected for testing.

As seen in Figure 13, HJRMT does not perform very well (IDS, and Track Fragments). The prime reason for the poor performance is that HJRMT is quite sensitive to the detection accuracy (the image coordination of the detection is employed to localise the target location in the "world" reference system), and it becomes more sensitive when target moves far away from our binocular system. In addition, the detector [26] often fail in such challenging scenes (severe occlusion), or cannot produce detections accurate enough. In contrast to HJRMT, our method achieves better results. As shown in the 1st row in Figure 13, let us focus on the target which the arrow points to. Notice that although part of this target is visible in the image captured from the left camera, it has already been wholly occluded in the view of the right camera. This leading to no detection of feature points from frame #1409 to #1417. However, our algorithm successfully associates the target in frame #1418 owing to optical flow information. Since frame #1420, due to the failure in the target's detection of feature points, only the prediction scheme of JPDA can be used through frame #1420 to #1433. Unfortunately, the trajectory has already begun to deviate from the real path. Nevertheless, our algorithm performs well by using color information when feature points are detected, though the new detection is out of the range of the small validation gate. Captures in the 2nd row of Figure 13 show the robustness of our algorithm for dealing with the problem of cross walk, even one of the targets (target 2 in frame #1948, target 2 in frame #2838) is wholly occluded during the process.

Figure 14. Multi-person Detection and Tracking Result (Captures for MP_WE_S2)

As shown in Figure 13 and Figure 14, our algorithm perform well when multi-person appear in the surveillance ROI. However, from several frames such as #6254 and #8222 in Figure 13, it can be discovered that the trajectories are sometimes not smooth enough. This is mainly because feature points are not stable when the corresponding person is frequently occluded. Such a problem seldom happens in Scene 2 due to the installation attributions (relatively large height, oblique angle and short baseline), as shown in Figure 14.

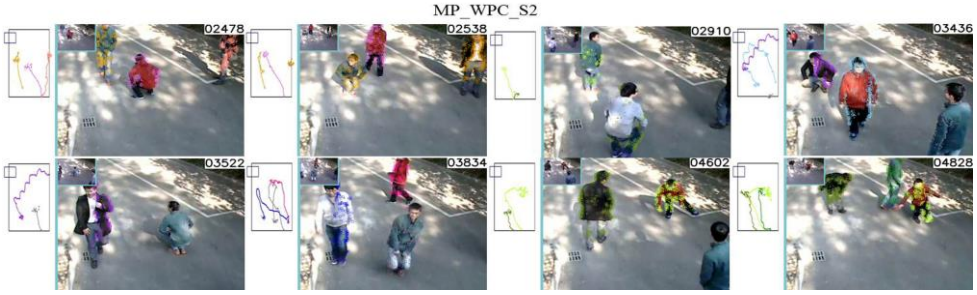

Figure 15. Multi-person Detection and Tracking Result (Captures for MP_WPC_S2)

Throughout our detection and tracking method, it can be easily found that the biometric information plays an essential role. In addition, people are supposed to walk erect (ellipse assumption from the top view). Thus, it is of great importance to check if our method can perform well when people change poses while walking. As shown in Figure 15 and frame #15356 in Figure 14, targets can be robustly tracked even they are getting down or bending over .etc. Specifically, it is the most challenging in frame #4602 in Figure 15, because a person changes its pose greatly. Generally speaking, our algorithm can successfully deal with such challenging scenarios mainly owing to the proposed clustering method with the help of optical flow information.

Scenes with illumination variation, shadow interference are very challenging for tracking algorithms[1,2]. Thanks to the robustness of sparse feature points, our algorithm performs well even such challenging things always happen in Scene 2 (Figure 14, Figure 15).

6.2.3 Quantitative Result: Some quantitative results of our method for each testing sequence set are shown in Table 2. Metrics consist of Ground Truth (GT), Mostly Tracked (MT), Partly Tracked (PT), Mostly Lost (ML), Track Fragments (FM) and Identity Switches (IDS) [6] are employed here.

Table 2. Quantitative Result for Each Sequence Set

Sequence	GT[%]	MT[%]	PT[%]	ML[%]	Frag	IDS
SP_WE_S1	55	100.0	0.0	0.0	0	0
SP_RE_S1	23	100.0	0.0	0.0	0	0
MP_WE_SBS_S1	44	100.0	0.0	0.0	0	0
MP_WE_S1	153	98.69	1.31	0.0	7	4
MP_WE_S2	232	96.12	3.88	0.0	19	6
MP_WPC_S2	40	95.0	5.0	0.0	4	3

7. Conclusions

In this paper, we present a new method for people detection and tracking through fixed stereo vision. Concerning robustness and efficiency, only a sparse set of object feature points are extracted. The biometric and optical flow information, as well as the width of a person is used to group 3D feature points into several clusters, with each cluster corresponding to a potential person. To enhance the robustness of the method for people tracking, the MJPDAF is proposed here. By testing our algorithm in complex scenes, experimental results show that our algorithm is effective to solve the challenging detection and tracking problems, even when a target disappears from camera views for a short while. In future, we will make efforts to further enhance the system robustness, and try to tracking multi-person in scenes full of obstacles (e.g. office).

Acknowledgements

This work is supported by the Startup Project of Doctor scientific research of Shaoxing University under Grant No. 20145026, the ‘Strategic Priority Research Program’ of the Chinese Academy of Sciences (CAS) under Grant No. XDA06020300, and National Natural Science Foundation of China under Grant No. 61272034.

References

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey", ACM Computing Surveys, vol. 38, no. 4, **(2006)**.
[2] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review", Neurocomputing, vol. 74, no. 18, **(2011)**.
[3] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, **(2012)**.
[4] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of Pedestrian Detection for Advanced Driver Assistance Systems", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 7, **(2010)**.
[5] H. Chang, L. Yuan, and R. Nevatia, "Multiple Target Tracking by Learning-Based Hierarchical Association of Detection Responses", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 4, **(2013)**.
[6] B. Yang, and R. Nevatia, "Online Learned Discriminative Part-Based Appearance Models for Multi-human Tracking", European Conference on Computer Vision (ECCV), **(2012)**.
[7] M. Hofmann, D. Wolf, and G. Rigoll, "Hypergraphs for Joint Multi-View Reconstruction and Multi-Object Tracking", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), **(2013)**.
[8] H. Xiaoyan, and P. Mordohai, "Evaluation of stereo confidence indoors and outdoors", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), **(2010)**.
[9] A. Geiger, M. Roser, and R. Urtasun, "Efficient Large-Scale Stereo Matching", Asian Conference of Computer Vision (ACCV), **(2011)**.
[10] O. Zoidi, N. Nikolaidis, and I. Pitas, "Appearance based object tracking in stereo sequences", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), **(2013)**.
[11] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for EasyLiving", Proceedings of the 3rd IEEE International Workshop on Visual Surveillance, **(2000)**.

[12] R. Muñoz-Salinas, E. Aguirre, and M. García-Silvente, "People detection and tracking using stereo vision and color", Image and Vision Computing, vol. 25, no. 6, **(2007)**.
[13] P. Kelly, N.E. O'Connor, and A. F. Smeaton, "Robust pedestrian detection and tracking in crowded scenes", Image and Vision Computing, vol. 27, no. 10, **(2009)**.
[14] P. Kelly, E. Cooke, N. O'Connor, and A. Smeaton, "Pedestrian Detection Using Stereo and Biometric Information", in: A. Campilho, and M. Kamel, (Eds.), Image Analysis and Recognition, Springer Berlin / Heidelberg, **(2006)**, pp. 802-813.
[15] K. Umeda, T. Nakanishi, Y. Hashimoto, K. Irie, and K. Terabayashi, "Subtraction stereo-a stereo camera system that focuses on moving regions", IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, **(2009)**.
[16] R. Muñoz-Salinas, R. Medina-Carnicer, F. J. Madrid-Cuevas, and A. Carmona-Poyato, "People detection and tracking with multiple stereo cameras using particle filters", Journal Of Visual Communication And Image Representation, vol. 20, no. 5, **(2009)**.
[17] P. Lenz, J. Ziegler, A. Geiger, and M. Roser, "Sparse Scene Flow Segmentation for Moving Object Detection", IEEE Intelligent Vehicles Symposium, **(2011)**; Baden-Baden, Germany.
[18] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association", IEEE Journal of Oceanic Engineering, vol. 8, no. 3, **(1983)**.
[19] I. Cox, "A review of statistical data association techniques for motion correspondence", International journal of computer vision, vol. 10, no. 1, **(1993)**.
[20] D. Svensson, M. Ulmke, and L. Danielsson, "Joint probabilistic data association filter for partially unresolved target groups", Proceedings of the 13th Conference on Information Fusion (FUSION), **(2010)**.
[21] Y. Bar-Shalom, and E. Tse, "Tracking in a cluttered environment with probabilistic data association", Automatica, vol. 11, no. 5, **(1975)**.
[22] J. Sklansky, "Finding the convex hull of a simple polygon", Pattern Recognition Letters, vol. 1, no. 2, **(1982)**.
[23] A. W. Fitzgibbon, and R. B. Fisher, "A buyer's guide to conic fitting. Proceedings of the 6th British conference on Machine vision, **(1995)**; Birmingham, United Kingdom.
[24] D. Arthur, and S. Vassilvitskii, "k-means++: the advantages of careful seeding", Proceedings of the 8th annual ACM-SIAM symposium on Discrete algorithms, **(2007)**; New Orleans, Louisiana.
[25] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 5, **(2003)**.
[26] P. F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, **(2010)**.

Authors

Keli Hu, he received his bachelor degree in communication engineering, in Hangzhou Dianzi University, Hangzhou, China, in 2009, and received his Ph.D. degree in information and communication engineering in Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China, in 2014. From May 2011 to the present, he is a teacher in the Department of Computer Science and Engineering, Shaoxing University, Zhejiang, China. His research interests include artificial intelligence, pattern recognition, computer vision and image processing.

Yuzhang Gu, he received the bachelor's and master's degree in Department of Optical Engineering, Zhejiang University in 1997 and 2000, and received his Ph.D. degree in Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology in 2009. From May 2011 to the present, he is an Associate Professor at Internet of Things System Technology Laboratory, Shanghai Institute of Microsystem and Information Technology of Chinese Academy of Sciences. He has long been engaged in the research of Bionic Robot Vision and Binocular Vision Sensing Technology.

Shigen Shen, he received the B.S. degree in fundamental mathematics from Zhejiang Normal University, Jinhua, China, in 1995, the M.S. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2005, and the Ph.D. degree in pattern recognition and intelligent systems from Donghua University, Shanghai, China, in 2013. He is currently a Professor with the Department of Computer Science and Engineering, Shaoxing University, Shaoxing, China. He is also with the College of Mathematics, Physics and Information Engineering, Jiaxing University, Jiaxing, China. His current research interests include wireless sensor networks, cloud computing and game theory.

Cheng Zhang, he received his bachelor degree in Electronic and Information Engineering, in Hangzhou Dianzi University, Hangzhou, China, in 2009. He is currently working towards the Ph.D. degree in information and communication engineering in Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China. His research interests include artificial intelligence, pattern recognition, computer vision and image processing, 2D and 3D face recognition.

Yunlong Zhan, he received his bachelor degree in Electronic Engineering, in Shanghai Jiao Tong University, Shanghai, China, in 2011. He is currently working towards the Ph.D. degree in information and communication engineering in Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China. His current research is stereo vision.

