Real-Time Multi-pedestrian Tracking Based on Vision and Depth Information Fusion

Shan Gao, Zhenjun Han, Ce Li, and Jianbin Jiao*

University of Chinese Academy of Sciences, Beijing, China
{gaoshan10,lice09}@mails.ucas.ac.cn,
{hanzhj,jiaojb}@ucas.ac.cn

Abstract. Visual object tracking plays an essential role in vision based applications. Most of the previous Multi-pedestrian Tracking has limitations due to considering each pedestrian with the same motion and appearance model in a uniform observation space, leading to tracking failures in complex occlusions. To address this problem without losing real-time performance, we propose a graph based approach for multi-pedestrian tracking using fused vision and depth data in this paper, where one main contribution is devoted in terms of the consideration of pedestrians with different priori probability in distinguishing observation space divided based on vision and depth information. Then we formulate the tracking model using an Improved Bipartite Graph (IBG), which is then optimized with a heuristic algorithm. Experiments on three datasets of fused vision and depth data demonstrate robust tracking results of the proposed approach.

Keywords: Multi-pedestrian Tracking, Bipartite Graph, Data Fusion.

1 Introduction

Multi-pedestrian tracking can provide fundamental information to lots of applications including video content analysis and driving assistant systems, which makes it one of the most active topics in the areas of computer vision and image processing in recent years [1, 2].

Given a video sequence, the task of multi-pedestrian tracking is to locate all pedestrians in each frame and correctly associate pedestrians over frames. However, the task is still an open problem for the following challenges. First, the outputs from a pedestrian detector are not always reliable, especially when vision data is captured with a moving camera [10, 13, 14]. Second, various occlusions among pedestrians are common [3, 6, 16].

From the perspective of image based tracking methods [11, 17, 18], people tend to find proper appearance models that can distinguish one object from others or the background [4, 5]. Some association based methods focus on tracking multiple pedestrians of a pre-known class simultaneously [6, 7]. They typically associate

* Corresponding author.

B. Huet et al. (Eds.): PCM 2013, LNCS 8294, pp. 708–719, 2013.

detection responses produced by a pre-trained detector into long tracks, and then pursuit a global optimal solution for all targets. Appearance models are often pre-defined [8] or online learned to distinguish multiple targets [9]. In addition, motion models, such as linear models on tracklet pairs [10, 16], are also adopted to constrain smoothness of motion.

It is noticed that most previous methods assume that all the track associations are independent. However, in a context of serious occlusion, the observations for pedestrians are not precise and trajectories can be mixed with each other. This make the independent assumption does not hold in a tracking process of serious occlusion. So in this paper, we divide the trajectories into distinguishing observation space, and give them different priori probability.

In recent years, benefiting from the reliable depth information of a laser rangefinder, tracking of pedestrians with occlusions achieves more promising results. However, there are also limitations when using depth information from rangefinder. In common, a laser rangefinder cannot discriminate different pedestrians of same depth. Furthermore, when there is inter-pedestrian occlusion, the mixture of depth information makes it difficult to discriminate pedestrians in the depth domain, and consequently tracking failure happens.

In this paper, considering the mutual complementary of vision and depth data, fusion of different sensing devices helps to overcome the limitation of each sensing technology, leading to enhance the performance of multi-pedestrian tracking. We formulate the data association progress from the perspective of Maximum A Posteriori (MAP) by proposing an IBG matching approach. We also propose an occlusion model cooperated with the depth information from laser rangefinder in consequent frames, which dynamically outputs the discriminated weight cost in IBG. The proposed approach can process independent observations and trajectories in a long-term tracking with complicated occlusion problem. Consequently, the robustness and accuracy of multi-pedestrian tracking using vision and depth can be improved.

The remainder of this paper is organized as follows: the proposed IBG matching approach is presented in section 2. Experiments are presented in section 3. We conclude the paper in section 4.

2 IBG Matching

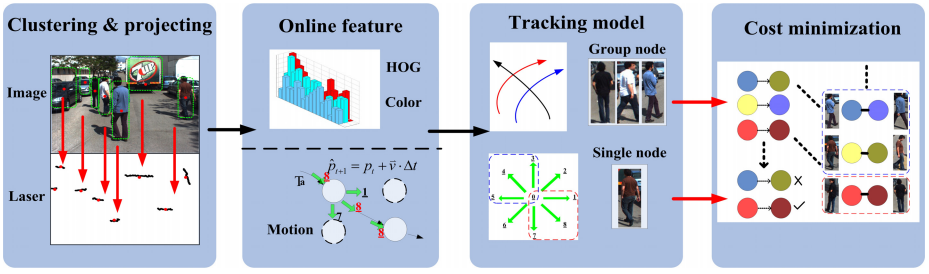


Fig. 1. Framework of our method

Framework: Given an input video sequence with the pedestrians' initialization after laser pattern classifier, clustering and projecting job, we can obtain the candidate regions both laser and image planes. And then the features are extracted to describe the pedestrians in terms of their appearances and motion characteristics, which include their position, velocity and moving orientation information. An online adaptive feature pool, named adaptive HOGC [19], is used to represent pedestrians' temporal association, considering the temporal consistence property of tracking. After that we reformulate the MAP estimation problem as a new tracking model with fused online updating information, and divide the observation space into two parts: group node and single node. Finally, we minimize the cost function using a heuristic searching algorithm to near optimality enlightened by IBG matching, in which the cost edges can be dynamically updated by the spatial and temporal relation from group and single nodes, and then we track each pedestrian and keep its identity.

2.1 Cost Function

Like some previous MAP work formulation [20, 21], observations of the *i*-th pedestrian at frame *t* are formulated as $T_{i_t} = \{pos_i, ori_i, vel_i, f_i\}$, indicating its position, orientation, velocity and appearance feature. Therefore the whole trajectory hypothesis of the *i*-th pedestrian can be written as $s_i = (T_{i_1}, T_{i_2}, \cdots, T_{i_n})$. Association hypothesis is a set $S = \{s_i\}$ composed by all the single trajectory hypotheses. The objective of data association is to maximize the posteriori probability of *S*, given the observation set $T = \{T_{i_t}\}$:

$$S^* = \arg\max_S P(S \mid T) \propto \arg\max_S P(T \mid S) P(S). \tag{1}$$

Assuming that motion of each pedestrian is independent, we can decompose Eq. (1) as

$$S^* \propto \arg\max_S \prod_i P(T_i \mid S) \prod_{s_k \in S} P(s_k). \tag{2}$$

Supposing that $S = -\log S^*$, the Eq. (2) can be equivalent to rewrite as

$$S = \arg\min_S \sum_{s_k \in s} -\log P(s_k) + \sum_i -\log P(T_i \mid S), \tag{3}$$

where $P(T_i \mid S)$ is formulated as the Bernoulli distribution.

$$P(T_i \mid S) = \begin{cases} p_i & if\ s_k \in S, T_i \in s_k \\ 1 - p_i & otherwise \end{cases}, \tag{4}$$

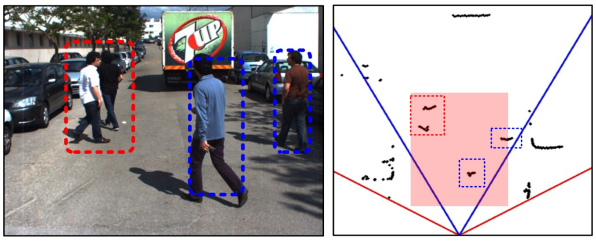
p_i will be given in detail in section 2.2. And in Eq. (3), $P(s_k)$ is modeled as the similarity function in each frame:

$$P_k(T_{i_{k-1}} \mid T_{i_k}) = P_{app}(T_{i_{k-1}} \mid T_{i_k}) P_{motion}(T_{i_{k-1}} \mid T_{i_k}) P_{ori}(T_{i_{k-1}} \mid T_{i_k}). \tag{5}$$

where P_{app}, P_{motion} and P_{ori} represent the similarity of the appearance, motion model and moving orientation respectively. And the similarity function is modeled as a Gaussian function.

The observation space S can be divided into two parts $S=\{S^{\alpha}+S^{\beta}\}$ by the affinity relation of trajectory hypothesis among pedestrians, and then a new priori probability $P(S)$ can be calculated in a new observation space, which is used to measure the trajectory relation among all the pedestrians. S^{α}, called group nodes, denotes the trajectories of pedestrians with the priori probability $P(S^{\alpha})$, who walk close each other (shown in the red rectangle in Fig.2a), and S^{β} denotes the trajectories of isolate pedestrians with the priori probability $P(S^{\beta})$ (shown in the blue rectangle in Fig.2a), known as single nodes. Based on this reasonable division, we rewrite Eq. (1) as:

$$S^{*} \approx \arg\max_{S} \prod P(S^{\alpha}) \prod P(T \mid S^{\alpha}) + \prod P(S^{\beta}) \prod P(T \mid S^{\beta}). \tag{6}$$



(a) (b)

Fig. 2. (a) Pedestrians in the image domain; and (b) pedestrians in depth domain

In Eq. (6), there are four terms in total. Since the values of $P(S^{\alpha})P(T \mid S^{\beta})$ and $P(S^{\beta})P(T \mid S^{\alpha})$ are small, these two terms can be omitted in our MAP formulation. To couple the non-overlap constraints with the objective function (7), then we define two 0-1 indicator variables $t_{j,i}$ and t_i as

$$t_{j,i} = \begin{cases} 1 & if\ T_j\ is\ right\ after\ T_i \\ 0 & otherwise \end{cases}, \tag{7}$$

$$t_i = \begin{cases} 1 & if\ T_i \in S_k \\ 0 & otherwise \end{cases}. \tag{8}$$

It's easy to see that these variables are determined for a given association hypothesis S, and vice versa. S is non-overlap if and only if

$$t_i = \sum_j t_{j,i} \le 1. \tag{9}$$

Based on two 0-1 variables, a new objective cost function is obtained when substituting Eq. (3) into Eq. (6), as follows:

$$S = \arg\min_{S^{\alpha}} \sum_{i,j} C_{j,i} t_{j,i} + \sum_{i} C_i t_i + \arg\min_{S^{\beta}} \sum_{i,j} C_{j,i} t_{j,i}. \tag{10}$$

Subject to:

$$C_{j,i} = -\log P_k \left(T_{j_{k-1}} \mid T_{i_k} \right), \tag{11}$$

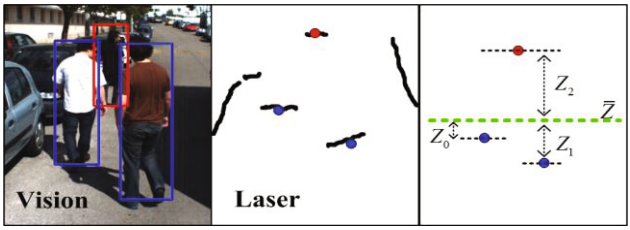
$$C_i = -\log P(T_i \mid S) = -\log \frac{p_i}{1-p_i} = \log \frac{1-p_i}{p_i}. \tag{12}$$

To solve the objective function Eq.(10), subjecting to the constrains Eq.(7-9), is equivalent to find a min-cost path in a fully connected graph model: The object candidate is represented by the node of the graph in successive frames, C_i and $C_{j,i}$ donate the cost of edges in graph, and the 0-1 indicator t_i and $t_{j,i}$ donate the connection relation of the nodes. In this paper, our goal is to find the min-cost path combination among the different weighed edges of graph model.

2.2 Occlusion Model

The cost of edge C_i and $C_{j,i}$ in the graph play a key role in graph model, which leads the discriminating capacity among objects candidates (nodes). In our method, an occlusion model along the depth dimension is defined to discriminate pedestrians by their depth ordering. The cost C_i can be dynamically updated, which are defined in Eq. (12). Then an updated weight factor as a cost C_i can be obtained in each frame. We definite the factor in Eq. (12) as

$$p_i = \left[1 + \exp(\bar{Z} - Z_i) \right]^{-1}. \tag{13}$$



(a) (b) (c)

Fig. 3. (a) Pedestrians in image domain; (b) pedestrians in depth domain; (c) occlusion measure with depth information

Eq. (13) can be solved by depth value from the laser rangefinder. As shown in Fig.3a, three pedestrians can be divided into S^{α}, also known as the group nodes. Firstly, we calculate the average depth $\bar{Z}$ (the green line in Fig.3c) of the objects (red and blue points in Fig.3c). Then we can obtain the relative depth Z_i between each object and $\bar{Z}$. Finally, we can get the probability p_i, which can be invoked by Eq. (12) and then we obtain C_i. Based on the definition, the pedestrians with bigger C_i is more likely to be occluded, such as the red point with positive C_i is apt to be occluded by

the blue points with negative C_i. Certainly, a single pedestrian in S^{β} always holds the zero relative depth and then has a zero C_i.

2.3 Optimization of IBG

The formulation in Eq. (10) can be mapped into an improved bipartite graph G, as shown in Fig.4. Each space-time location j, or equivalently pedestrian observation at time t-1, corresponds to a pair of nodes $(n_{t-1,j}, n'_{t-1,j})$ connected by an edge of cost C_i. Each putative pedestrian observation at time t, corresponds to a node $n_{t,j}$ .The transition from $n'_{t-1,j}$ to $n_{t,j}$ in two frames is an edge $(n'_{t-1,j}, n_{t,j})$ with cost $C_{j,i}$.

To optimize the IBG, we adopt an iterative strategy. It is known that a good initial solution can yield convergence to a better optimum. Therefore, we use a greedy algorithm to get initial solution. The greedy algorithm will find out the shortest path on the IBG and remove the selected cost edge from the graph until there is no C_i edge in the residue graph. Given an initial solution, the cost minimization algorithm is described in Algorithm 1.

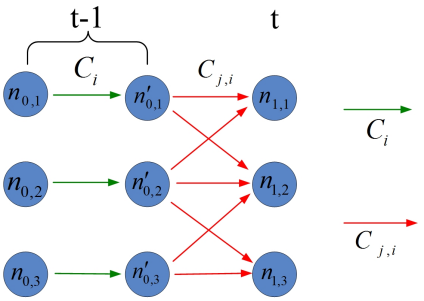


Fig. 4. Improved bipartite graph (IBG)

Algorithm 1. Switching labels with low cost

Input: IBG G; cost$\{C_i\}$, $\{C_{j,i}\}$; observation S;
Initial: Finding the label set L' of lowest cost by a greedy algorithm and evaluate its overall cost S^* by (10);
1 **For** $i < n$ **do**
2 Set maximum cost $S' = +\infty$;
3 **For** $j = i, \cdots, n$ that $C_{i,j} \in G$ **do**
4 –switch label of $n'_{t-1,j}$ and $N_{t,j}$ under constrains in Eq.10 and evaluate new cost S'';
5 –if $S'' < S'$, $S' = S''$;
6 **end**
7 If $S' < S^*$, $S^* = S'$, update L with this switch;
8 **end**
Output: Label set L of G.

3 Experiments

To demonstrate the effectiveness of the proposed approach, we performed comprehensive experiments on three public datasets. One is ISR-UC-imglidar-sync dataset [11]. The other two are SDL-1 and SDL-2 datasets [12], which are collected on a robot platform with a camera and a laser rangefinder. The platform moves in a campus, captures two video sequences of pedestrians occluded with each other. To evaluate the tracking performance, we adopt evaluation metrics defined in Table 1.

Table 1. Evaluation metrics

Items	Definition
GTP	Number of positions in ground truth
LF (↓)	Label failure, different with the ground truth
PF (↓)	Position failure, different with the ground truth
GTI	Number of ID in ground truth
MT(↑)	Mostly tracked trajectories, the number of trajectories that are tracked for more than 80%
ML(↓)	Mostly lost trajectories, the number of trajectories that are tracked for less than 20%
PL (↓)	Partially lost trajectories, the number of trajectories that are tracked in 20-80%
Sp (↑)	Speed, frame per second (fps)

For items with ↑, higher scores indicate better results, for those with ↓, lower scores indicate better results. The first three items (**GTP**, **LF**, **PF**) measure the accuracy between tracking result and ground truth in each frame. The following four items (**GTI**, **MT**, **ML**, **PL**) reflect the overall discrimination capacity in the occurrence period of targets. **Sp** shows the calculating speed of different algorithms, which includes observation detecting time in laser coordinates.

Table 2. Comparisons of four methods on three datasets

Dataset	Method	GTP	PF	LF	GTI	MT	ML	PL	Sp(fps)
SYNC	Dep (NN)	2360	612	566	10	4	2	4	**60-80**
	Dep+Fea	2360	229	312	10	7	1	2	25-30
	Dep+Mot	2360	268	347	10	6	2	2	25-35
	Dep+IBG	2360	**0**	**0**	10	**10**	**0**	**0**	20-25
SDL-1	Dep (NN)	384	189	189	4	1	1	2	**60-80**
	Dep+Fea	384	68	68	4	2	1	1	25-30
	Dep+Mot	384	72	72	4	2	1	1	25-35
	Dep+IBG	384	**6**	**6**	4	**4**	**0**	**0**	20-25
SDL-2	Dep (NN)	1954	512	495	31	13	8	10	**55-70**
	Dep+Fea	1954	274	294	31	18	6	7	25-30
	Dep+Mot	1954	240	286	31	17	7	7	25-30
	Dep+IBG	1954	**43**	**46**	31	**25**	**3**	**3**	18-25

Four relevant methods are evaluated on these datasets for comparison and results are listed in Table 2. “Dep(NN)” denotes the “nearest neighbor” method only applying depth information from laser rangefinder. “Dep+Fea” means the combination of the first method and the feature matching. “Dep+Mot” is the combination of the first method and motion estimate and “Dep+IBG” is the ultimate results of our method. It can be seen in Table 2 that our method has the best performance in all of the three datasets.

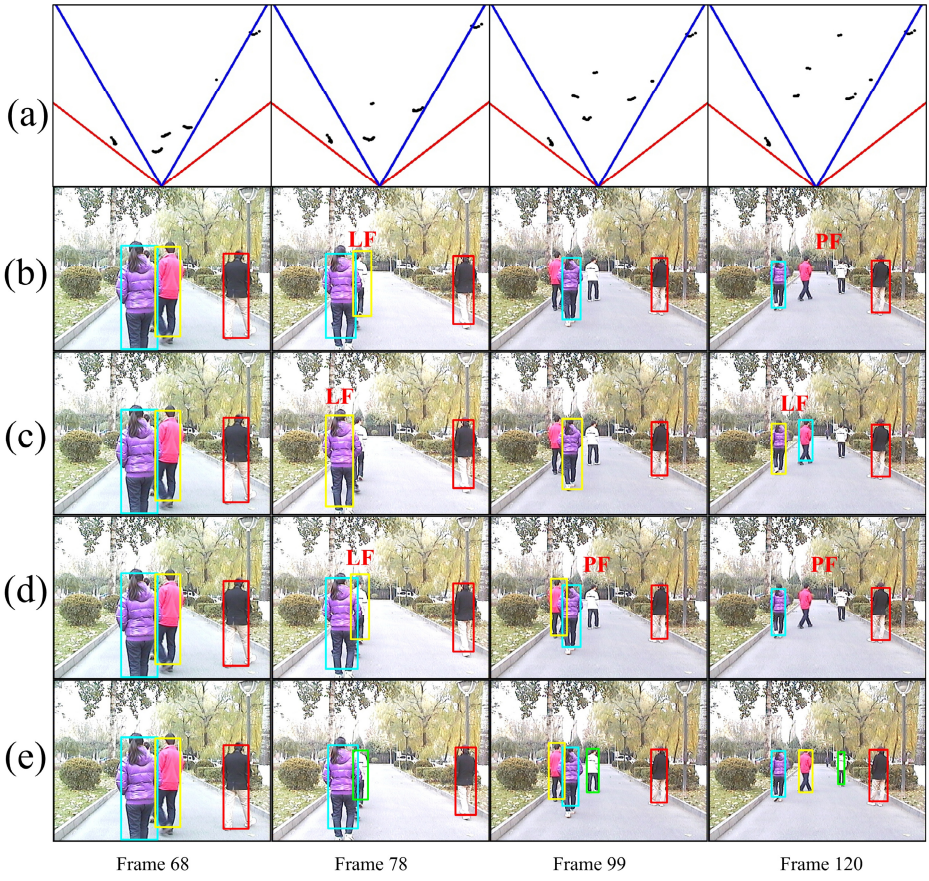


Fig. 5. Tracking examples on the dataset SDL-1. The row (a) denotes “depth picture” obtained by laser rangefinder; the row (b)-(e) show the results of “Dep (NN)”, “Dep+Fea”, “Dep+Mot” and “Dep+IBG” based tracking methods respectively. The **LF** and **PF** mean “label failure” and “position failure”.

SDL Datasets: Fig.5 and Fig.6 show the tracking results by our online tracking model approach. The first rows (a) in both figures are “depth picture” obtained by our laser rangefinder. Traditional methods with motion and appearance model or just the simple combination of these two models can only get the cost edge of “single-node”

in the improved bipartite graph. Regardless of the affinity relation among the group pedestrians, the other three methods ("Dep (NN)" in row (b), "Dep+Fea" in row (c), "Dep+Mot" in row (d)) result in mislabeling and losing pedestrians (**LF**, **PF**) during long-term and complex occlusion. However, by considering the group-node in tracking model, the pedestrians can get greatly different cost edges in our objective function, which can explain the conspicuous improvements on **MT**. Particularly, when the number of pedestrian increasing, our tracking model can keep great accuracy and robustness in the last row (e)of Fig.5 and Fig.6.

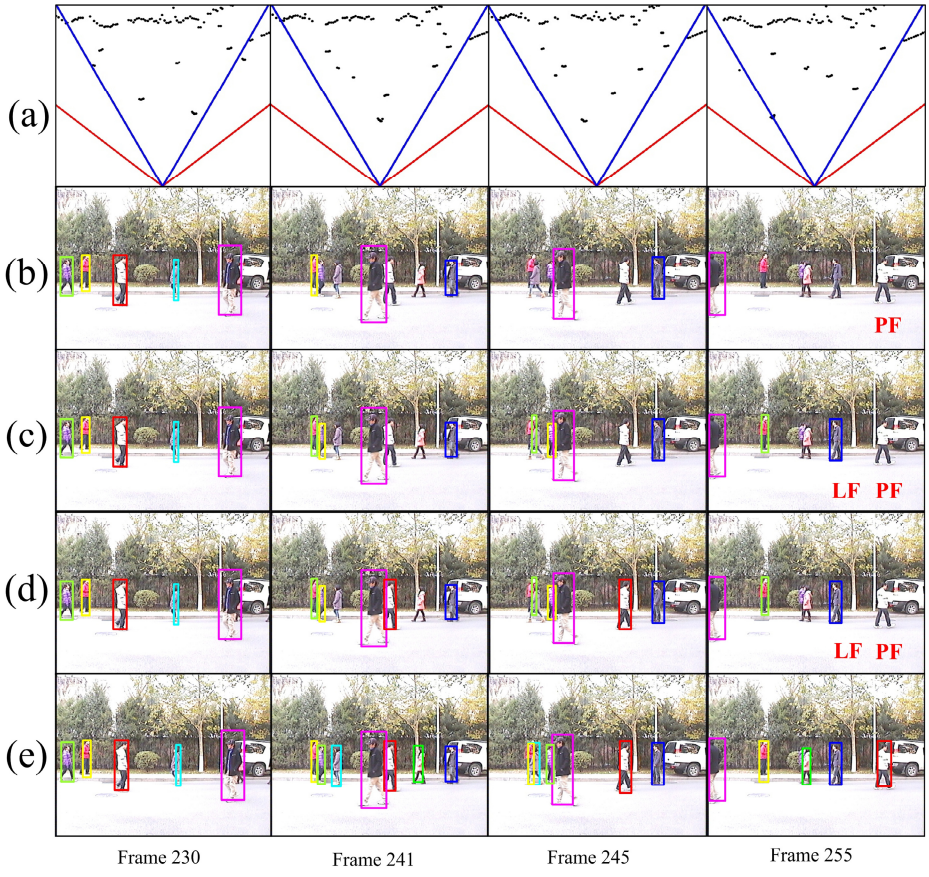


Fig. 6. Tracking examples on the dataset SDL-2. The row (a) denotes "depth picture" obtained by laser rangefinder; the row (b)-(e) show the results of "Dep (NN)", "Dep+Fea", "Dep+Mot" and "Dep+IBG" based tracking methods respectively. The **LF** and **PF** mean "label failure" and "position failure".

SYNC Dataset: as shown in Fig.7, only using the position information of laser rangefinder can not discriminate the pedestrians, especially when the merge and split happened in pedestrians' moving process, as shown in row (b), frame 351, Fig.7.

In the laser level, pedestrians and cars are treated as the same cluster when the pedestrian is so close to the car. As the result, the algorithm of "Dep (NN)" makes the **PF** and **LF** errors increase at the same time. As the appearance model added, we can finish the online feature extracting work using the "Dep+Fea" method in row (c) of Fig.7. Mostly like the "Dep+Fea" method, the motion model "Dep+Mot" can predict the motion state when pedestrians get out of occlusion. So the **PF** and **LF** errors in these two methods given in row (c) and (d) decrease greatly compared with the method of "Dep (NN)". Then we further apply tracking model "Dep+IBG" to associate the candidates in observations, the errors have dropped greatly compared with the first three methods. Because our tracking model makes full use of the fused depth and vision information among tracked pedestrians as discriminated costs in objective function, which make it easier to find the min-cost path in IBG.

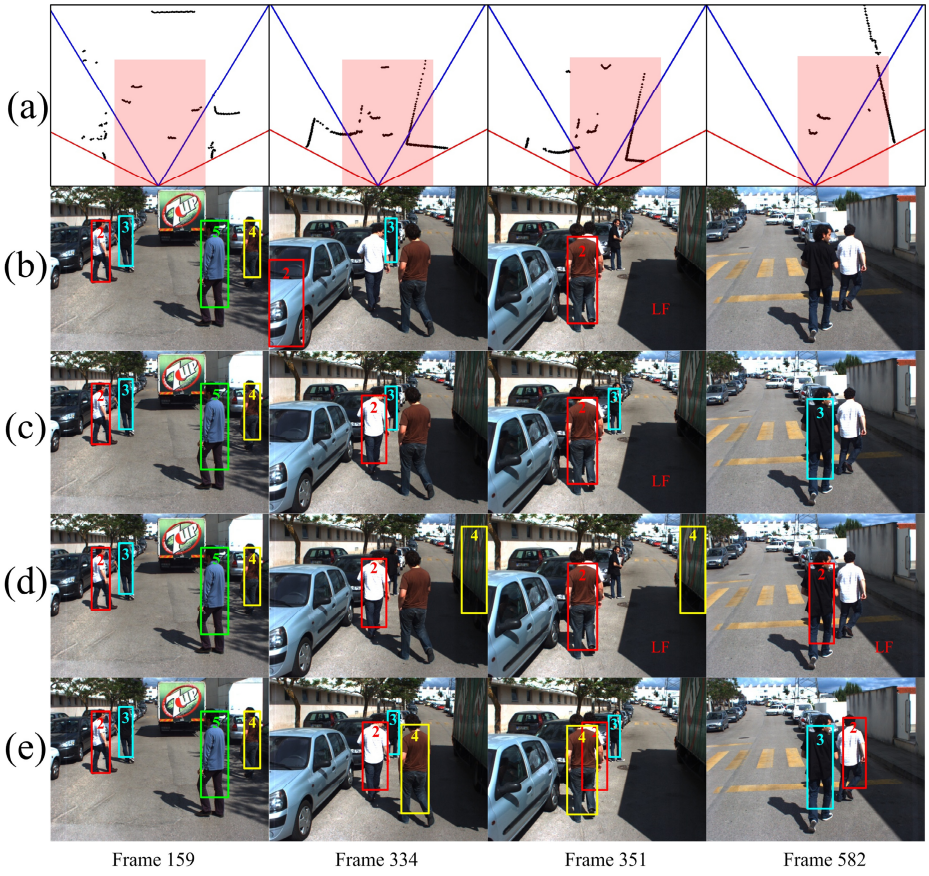


Fig. 7. Tracking examples on dataset SYNC. The row (a) denotes "depth picture" obtained by laser rangefinder; the row (b)-(e) show the results of "Dep (NN)", "Dep+Fea", "Dep+Mot" and "Dep+IBG" based tracking methods respectively. The **LF** and **PF** mean "label failure" and "position failure".

4 Conclusions

Data association is a primary issue of pedestrians tracking. In this paper, we implemented a real-time pedestrians tracking approach by integrating the depth and vision information into an IBG, which was formulated as the association cost function, and then solved by a heuristic algorithm. The extensive evaluation based on the public data sets demonstrates that our approach is powerful for complex tracking problems compared with the existing work. However, there is a limitation in our work. We use three kinds of patterns in the laser coordinates to classify the objects. Even though our data association is accurate and efficient, the unstable detection responses probably bring in more false detection observation. We will cope with this limitation in the future work.

Acknowledgments. This work is supported in part by National Basic Research Program of China (973 Program) with No. 2011CB706900, 2010CB731800, and National Science Foundation of China with No. 61039003, 61271433 and 61202323.

References

1. Okuma, K., Taleghani, A., de Freitas, N., Little, J.J., Lowe, D.G.: A boosted particle filter: Multitarget detection and tracking. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 28–39. Springer, Heidelberg (2004)
2. Zhao, T., Nevatia, R.: Tracking multiple humans in crowded environment. In: Proc. IEEE Int. Conf. CVPR (2004)
3. Andriyenko, A., Roth, S., et al.: An Analytical Formulation of Global Occlusion Reasoning for Multi-Target Tracking. In: Proc. IEEE Int. Conf. ICCV (2011)
4. Li, Y., Ai, H., Yamashita, T., et al.: Tracking in low frame rate video: A cascade particle filter with discriminative observers of different lifespan. In: Proc. IEEE Int. Conf. CVPR (2007)
5. Wang, S., Lu, H., Yang, F., Yang, M.-H.: Superpixel tracking. In: Proc. IEEE Int. Conf. ICCV (2011)
6. Song, B., Jeng, T.-Y., Staudt, E., Roy-Chowdhury, A.K.: A stochastic graph evolution framework for robust multi-target tracking. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 605–619. Springer, Heidelberg (2010)
7. Andriyenko, A., Schindler, K.: Multi-target tracking by continuous energy minimization. In: Proc. IEEE Int. Conf. CVPR (2011)
8. Pirsiavash, H., Ramanan, D., et al.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: Proc. IEEE Int. Conf. CVPR (2011)
9. Shitrit, H.B., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: Proc. IEEE Int. Conf. ICCV (2011)
10. Breitenstein, M.D., Reichlin, F., Leibe, B., et al.: Online multi-person tracking-by-detection from a single, uncalibrated camera. IEEE Transactions on PAMI 33(9), 1820–1833 (2011)
11. Oliveira, L., Nunes, U.: Semantic fusion of laser and vision in pedestrian detection. In: Pattern Recognition, pp. 3648–3659 (2010)
12. SDL Dataset: `http://www.ucassdl.cn/resource.asp`

13. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile Vision System for Robust Multi-Person Tracking. In: Proc. IEEE Int. Conf. CVPR (2008)
14. Wojec, C., Walk, S., Roth, S., et al.: Monocular Visual Scene Understanding: Understanding Multi-Object Traffic Scenes. IEEE Transactions on PAMI 35(4), 882–896 (2013)
15. Andriyenko, A., Schindler, K., Roth, S.: Discrete-Continuous Optimization for Multi-Target Tracking. In: Proc. IEEE Int. Conf. CVPR (2012)
16. Yang, B., Nevatia, R.: An online learned CRF model for multi-target tracking. Proc. IEEE Int. Conf. CVPR (2012)
17. Ess, A., Leibe, B., Van Gool, L.: Depth and Appearance for Mobile Scene Analysis. In: Proc. IEEE Int. Conf. CVPR (2008)
18. Bajracharya, M., Moghaddam, B., et al.: A Fast Stereo-based System for Detecting and Tracking Pedestrians from a Moving Vehicle. Journal of Robotics Research 2009 (2009)
19. Han, Z.J., Jiao, J.B., Zhang, B.C., Ye, Q.X., Liu, J.Z.: Visual Object Tracking via Sample-Based Adaptive Sparse Representation (AdaSR). Pattern Recognition (44), 2170–2183 (2011)
20. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: Proc. IEEE Int. Conf. CVPR (2008)
21. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple Object Tracking Using K-Shortest Paths Optimization. IEEE Transactions on PAMI 33(9), 1806–1819 (2011)