Depth Structure Association for RGB-D Multi-Target Tracking

Shan Gao[1], Zhenjun Han[1], David Doermann[2], *Fellow, IEEE*, Jianbin Jiao[1*]
[1]University of Chinese Academy of Sciences Beijing, China.
[2]Institute for Advanced Computer Studies, University of Maryland, College Park.
[2]doermann@umiacs.umd.edu, [1*]jiaojb@ucas.ac.cn

***Abstract*—Multi-target tracking in outdoor scenes plays an important role in many computer vision applications. Most previous work on visual information based multi-target tracking does not incorporate depth information and the absence of depth information often leads to mismatching or tracking failures. In this paper, we propose a Depth Structure Association (DSA) approach for RGB-D data based multi-target tracking. DSA encodes depth information in a chain structure, the structure is used by DSA together with appearance and motion information to address object occlusion issues in outdoor scenes. Additionally, the use of DSA has the advantages of regulating a much smaller solution space, greatly reducing the computational complexity. Experimental results on three datasets demonstrate that our DSA approach can significantly reduce object mismatch and tracking failure for long term occlusions.**

I. Introduction

Object tracking is a fundamental problem in computer vision and human-computer interactions, and has many applications in robotics and intelligent vehicle systems. Recent visual multi-target tracking methods [1–10] have made great progress in addressing the problem, but these methods may fail when there exists complex backgrounds, uneven lighting, and in particular, serious multiple object occlusions.

One solution involves the use of RGB-D sensors, which can provide both range and image data (Fig. 1). Popular range sensors include stereo cameras [11–14], Microsoft-Kinect [15][16] and laser rangefinders [17]. The availability of range/depth data improves the tracking performance, but how to incorporate the image and depth data effectively is still an open problem.

For traditional RGB data based tracking, tracking-by-detection methods have gained popularity due to the improvement of object detection performance. These methods integrate cues such as appearance, motion, size, and location into an affinity model to measure similarity between detection responses or tracklets in an association optimization framework. To solve the optimization problem for RGB data, Zhang et al. [2] used flow network to find the minimum cost by defining a graph model; Berclaz et al. [3] used the k-shortest path algorithm for searching the solution space in a matching process and Pirsiavash et al. [4] proposed a globally-optimal greedy algorithm based on dynamic programming to search successive shortest paths in a residual graph. These methods can only solve the short-term and partial occlusion problems.

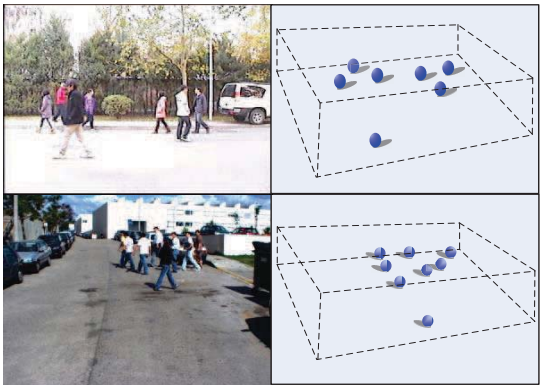


Fig. 1. Depth tells us much. The complex occlusion issues are not addressed among the multi-target in RGB image of left column. However, the D data in right column can provide us another dimension reference, and then we can obtain the depth structure of the targets in ground plan. In this paper, we explore how to utilize this depth structure variation coping with the multi-target association problem in tracking.

Datasets used in their experiments come from surveillance videos with a specific top-view angle, for which the partial occlusion problem is the most common issue. However, a heads-up view is the most common on a moving platform where full occlusion issue in dynamic background is unavoidable.

Another line of research for multi-target tracking focuses on the trajectory-level analysis. Typical approaches use trajectory-based CRF energy functions [5][6], continuous energy minimization [7], or trajectory-level discrete-continuous CRF models [8]. Although state-of-the-art results have been reported using these approaches, they perform poorly on dynamic backgrounds. Because moving platforms and dynamic backgrounds break the continuity of a trajectory, the objective function typically converges to the wrong local tracklet. Our method however does not impose additional trajectory-level constraints. In order to adapt to the dynamic background, we improve the trajectory-level continuity to the depth-level for the RGB-D data, which solves the occlusion problem.

To the best of our knowledge, this paper proposes the first general multi-target tracking association method for RGB-D data. We extend the traditional Multi-Dimensional Assignment (MDA) problem by combining the traditional 2D vision information and the depth relation. In particular, we reformulate the data association problem by adding the

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.711

depth variation relation to address complex full occlusions in dynamic backgrounds. We investigate the depth variation relation among the target observations in successive frames in addition to simple spatial-temporal relations based on the appearance and motion features. The complex tracking association problem can then be solved by forming a chain structure with depth variation in depth coordinates. Our method avoids the complex optimization problem by online updating of a depth structure matrix, which records the depth structure variation of the targets in multiple frames. In the matching procedure, we start the searching along a particular depth chain and incorporating the appearance and motion features to get a near optimal solution.

II. Depth Structure Association for RGB-D Multi-Target Tracking

A. RGB-D Data for Tracking

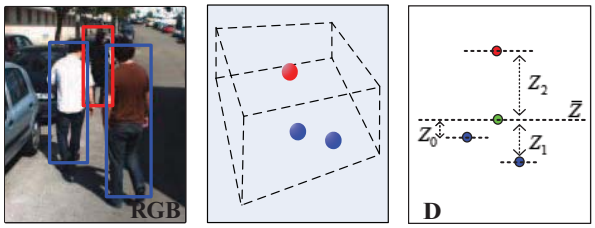
To acquire 3D information, the Microsoft Kinect RGB-D sensor provides a $640 * 480$ pixel resolution at 11 bits per pixel. Valid depth data of a target can be obtained after depth detection. The disparity vision plays a key role in stereo camera. Previous work [11][12] used a pair of forward-looking cameras to obtain the depth data. A laser rangefinder device can also be rotated about one of the main axes of sensor-based coordinate frame.

Generally, we consider a target observation o_i with 3D position feature $\pi_i = (u_i, v_i, z_i)$, appearance feature φ_i, and motion feature θ_i. π_i is the 3-dimensional spatial coordinates of the center of detection output, (u_i, v_i) denotes the observation's center in the image, and z_i denotes the depth value. The depth value can be calculated from the disparity in the stereo camera [11][12], the encoding depth technique in Kinect [15], and the coordinate transformation in rangefinder [17]. In addition, φ_i represents the HOGC feature [18] which is a unified vector combining the HOG and color features, and θ_i denotes the motion feature including velocity and orientation features. These features play key roles in our multi-target tracking method.

B. MDA Problem

Data association for multi-target tracking is the process of partitioning a set of observations into trajectories. Most of the previous approaches for this problem rely on the idea that different trajectories cannot claim the same observation. We define the piece of trajectory as a tracklet, corresponding to an observation. The MDA problem is known to be NP-hard for 3 or more frames association. For a sequence of k frames that consist of N observations each, the optimal partition of N trajectories can be mapped into a k-partite structure of MDA. Given N observations $\{\mathrm{o}_1, \mathrm{o}_2, \cdots, \mathrm{o}_N\}$, the problem is formulated as the following optimization problem:

$$\begin{aligned} &\max \sum_{i_i=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_k=1}^{N} \mathrm{A}_{i_1 i_2 \cdots i_k} X_{i_1 i_2 \cdots i_k} \\ &s.t. \quad \sum \sum_{I \backslash i_t} \cdots \sum X_{i_1 i_2 \cdots i_k} \leq 1; \;\; X_{i_1 i_2 \cdots i_k} \in \{0, 1\} \end{aligned} . \quad (1)$$



(a) (b) (c)

Fig. 2. The case of target observations in the same depth chain. (a) in the image domain; (b) in the 3D space; (c) in the ground domain.

Where $\mathrm{A}_{i_1 i_2 \cdots i_k}$ is an affinity measure matrix, in which the element $A_{ij} = 1$ denotes the affinity between observations o_i and o_j in adjacent frames, and $I \backslash i_t$ represents all observations in all frames other than frame t. $X_{ij} = 0$ denotes the observation o_j follows o_i in the assignment matrix $X_{i_1 i_2 \cdots i_k}$, which ensures that the linked tracklets have a high similarity over their observations. Note that the constraint on assignment matrix is less than or equal to 1, means the tracklets are not forced to add, end or cross with other trajectories. If X_{ij}=0, the observation o_j is not linked with any trajectory in the past and hence represents the start of a new trajectory. In the tracking process, this structure records the new trajectory of target adding and deleting, and changing the similarities over observations during the k frames.

C. Depth Structure Association in RGB-D Data

To solve the MDA problem for RGB-D data, our depth structure association method avoids the difficulties in the complex optimization compared with traditional MDA solutions. The classical vision-based approach attempts to assign the observation more distinguishing parameters from appearance, motion, and trajectory in multi-frame. We start the multi-target association from the depth structure among the observations in successive frames.

We define a matrix $D_i^{(t)}$ to describe the depth structure among the observations from frame $t-k$ to frame t, a total of $k+1$ frames. Note that the depth structure matrix reflects the depth variation of observations in successive $k+1$ frames. Assuming there are N observations labeled by an augmented index set, $O = \{o_i\}$; $0 \leq i \leq N$, we use a chain structure to describe this depth relation. $D_i^{(t)}$ records depth-chain label which the observation o_i belongs to. For example, when $N = 5$, $k = 3$:

$$D_i^{(t)} = \begin{bmatrix} 1 & 1 & 2 & 2 & 2 & 3 \\ 1 & 1 & 1/2 & 2 & 2 & 3 \\ 1 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 2 & 2/3 & 3 \end{bmatrix} . \quad (2)$$

The first element "1" means the observation o_1 belongs to chain 1. Note that one observation probably belongs to more than one chain as the elements "1/2" and "2/3" show. In order to define the depth chain structure, we introduce the root-mean-square deviation ε_i, to denote the offset from mean

3D position of a chain to 2D vision observation, (u_i, v_i):

$$\varepsilon_i = \sqrt{\frac{1}{|C_m|} \sum_{\pi_i \subset C_m} |\Phi(\pi_i) - \bar{\pi}_m|^2}. \quad (3)$$

Where $\Phi(\pi_i)$ is the translation function from image to world coordinates, $\bar{\pi}_m$ is average 3D position of the m-th chain in world coordinates, and C_m denotes the observation's 2D position (u_i, v_i) in m-th chain. This constraint includes not only the z_i value in depth coordinates, but also the (u_i, v_i) data in image coordinates. Using this definition, the target observations in a dense region have high probability of occluding each other, so these observations are divided into the same chain and we assign them each a discriminative depth factor dynamically to avoid ID-switching issue. The parameter offset ε_i is used to control the area of each chain and the overlapping area between the chains.

The depth factor matrix $S_i^{(t)}$ corresponds to the depth structure $D_i^{(t)}$, where each element records the depth score in each chain. We use the example in Eq. (2):

$$S_i^{(t)} = \begin{bmatrix} 0.5 & -0.5 & 0.8 & -0.4 & -0.5 & 0 \\ 0.7 & -0.2 & -0.4/0.7 & -0.3 & -0.5 & 0 \\ 0.5 & -0.5 & 0.5 & -0.5 & 0.5 & -0.5 \\ 0 & -0.2 & 0.4 & 0.2 & -0.4/0.5 & -0.5 \end{bmatrix}. \quad (4)$$

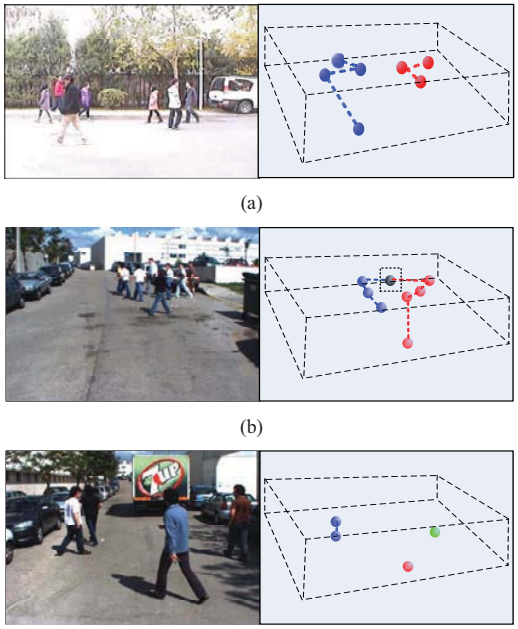
The matrix $S_i^{(t)}$ has the same scale as $D_i^{(t)}$ and the elements vary in the range $(-1, 1)$. We define the element s_i as a depth factor, which reflects the depth variation relation in each chain and records the score:

$$s_i = \log \frac{\beta_i}{1 - \beta_i}, \quad (5)$$

$$\beta_i = \left[1 + \exp\left(\frac{1}{\mathrm{m}} \sum_{j=1}^{m} z_j^{\mathrm{t}-1} - \hat{z}_i^t\right) \right]^{-1}. \quad (6)$$

Where $\hat{z}_i^t = z_i^t - \bar{z}^t$ denotes the relative depth value between an observation's depth z_i^t and chain's average depth $\bar{z}^t$. z_j^{t-1} denotes the depth value of the observation in the neighbor chain of the $(t-1)$-th frame. So β_i measures the depth relation of the observations at the same chain in adjacent frames. We consider the occlusion in Fig. 2 as an example. The depth parameters z_i^t, z_j^{t-1} of observations are known, and the green dashed line (in Fig. 2(c)) is the chain's average depth $\bar{z}^t$. Then according to the Eq. (6), the depth factor β_i, $i = 1, 2, 3$ can be obtained. In the sigmoid function, $\beta_i \in (0, 1)$. With some further analysis, we find that observations with the small depth value (marked with blue points in Fig. 2(c)) satisfy: $0.5 < \beta_i < 1$; and the observations with large depth value (marked with red points Fig. 2(c)) satisfy: $0 < \beta_i < 0.5$. Note that when we substitute β_i in the Eq. (5), a large depth value can get a negative depth factor s_i and the small depth can get a positive factor. In other words, the observation which is close to viewpoint can be given a large positive depth factor in association, vice versa.

There is another special case in depth chain: a chain with only one observation. This case is illustrated in Fig. 3(c). In this situation, the chain's average depth is, $\bar{z} = 0$, so the depth factor, $\beta_i \approx 0.5$. When we invoke Eq. (5), the depth factor is $s_i \approx 0$. This means that if a chain has only one observation, its depth factor has little influence on the structure variation.



(a) (b) (c)

Fig. 3. Illustration of three possible depth structures in RGB-D data. (a) shows two independent chain structures. (b) shows that two chains have the overlap observation. This case is common when the number of targets is large. (c) shows the case that a chain has only one observation as the red and green chains. It is common on the isolated target.

D. Multi-target Tracking via DSA

We propose an iterative approximation inference to solve the MDA problem given the depth structure matrix $D_i^{(t)}$, and depth factor matrix $S_i^{(t)}$. First, let us consider a 4-frame association problem. Suppose there are $n_a \sim n_d$ observations for frames $t_a \sim t_d$ respectively. Our objective function for the MDA of Eq. (1) can be rewritten as

$$\begin{aligned} &\max \sum_{a=1}^{n_a} \sum_{b=1}^{n_b} \sum_{c=1}^{n_c} \sum_{d=1}^{n_d} \mathrm{A}_{abcd} X_{abcd} \\ &s.t.\ X_{abcd} \in \{0, 1\}; \\ &\quad \sum_{b=1}^{n_b} \sum_{c=1}^{n_c} \sum_{d=1}^{n_d} X_{abcd} \le 1; \quad a = 1, 2, \cdots n_a. \\ &\quad \vdots \qquad\qquad\qquad\qquad \vdots \\ &\quad \sum_{a=1}^{n_a} \sum_{b=1}^{n_b} \sum_{c=1}^{n_c} X_{abcd} \le 1; \quad d = 1, 2, \cdots n_d. \end{aligned} \quad (7)$$

This k-partite structure (in this case, 4-partite) of the MDA enables enumeration over the trajectory and affinity function values by nested sums over the observations and frames. We leverage the k-partite structure of the problem by noting that

a trajectory (a,b,c,d) is formed by a series of traklet combinations $((a,b),(b,c),(c,d))$, so we can factor the decision matrix X_{abcd} as $y_{ab} \times g_{bc} \times h_{cd}$. This transformation reduces the number of decision variables to the order of $(k-1) \times n^2$ instead of n^k.

$$\begin{aligned} &\sum_a \sum_b \sum_c \sum_d A_{abcd} X_{abcd} \\ &= \sum_a \sum_b \sum_c \sum_d A_{abcd} y_{ab} g_{bc} h_{cd} \\ &= \sum_a \sum_b y_{ab} \sum_c g_{bc} \sum_d h_{cd} A_{abcd} \end{aligned} . \tag{8}$$

We further add the depth chain structure $D_i^{(t)}$ to factor the decision matrix. Again, we use the 4-partite example in Eq. (2) to illustrate the tracklet association process:

$$D_i^{(t)} = \begin{bmatrix} d_a & d_b & d_c & d_d \end{bmatrix}^T . \tag{9}$$

$$= \begin{bmatrix} 1 & 1 & 2 & 2 & 2 & 3 \\ 1 & 1 & 1/2 & 2 & 2 & 3 \\ 1 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 2 & 2/3 & 3 \end{bmatrix}$$

In the above matrix, the rows $d_a \sim d_d$ denote the distribution of the chain structure in frames $t_a \sim t_d$. Each element means the chain's label of one observation as in Eq. (2). We set the searching path y_{ab}, g_{bc}, h_{cd} along the elements with the same chain label. Here, we take the chain "3" as an example, the searching path follows the blue arrows.

Depth heuristic: It is important to know that we have decomposed the decision matrix X_{abcd} in Eq. (7) as y_{ab}, g_{bc}, h_{cd}, the affinity measure should be in accordance with this transformation. We define the affinity probability between two observations with the same label in adjacent frames as:

$$A_{ij} = \begin{cases} A_a\left(\varphi_i^{t-1}, \varphi_j^t\right) A_p\left(\theta_i^{t-1}, \theta_j^t\right) A_m\left(X_i^{t-1}, X_j^t\right), & if \;\; C_i^{t-1} = C_j^t \\ 0, & otherwise \end{cases} \tag{10}$$

where the affinity function $A(*)$ measures the affinity of appearance, position and motion feature, which are defined by Gaussian distributions [9]. The parameters C_i^{t-1} and C_j^t denote the observation's chain label in $t-1$ and t frames.

Previous greedy forward methods based only on information observed up to time t are "brittle" since once decision variables are obtained, the trajectory associations are fixed. These methods produce suboptimal solutions, because they ignore the sequential information, let alone the spatial RGB-D information in changing structure. In this paper, we consider the depth variation parameter $S_i^{(t)}$, and step through pairs of observations in the same chain of adjacent frames, updating the decision variables between them while keeping all other decision variables fixed. The depth factor $S_i^{(t)}$ can be added into the objective function in Eq. (8):

$$\begin{aligned} &\sum_a \sum_b \sum_c \sum_d (A_{abcd} + S_{abcd}) X_{abcd} \\ &= \sum_a \sum_b y_{ab}(A_{ab} + S_{ab}) \sum_c g_{bc}(A_{bc} + S_{bc}) \sum_d h_{cd}(A_{cd} + S_{cd}) \end{aligned} . \tag{11}$$

TABLE I
EVALUATION METRICS

Items	Definition
Recall(↑)	Correctly matched detections / total detections in ground truth
Prec.(↑)	Correctly matched detections / total detections in the tracking results
GT	Number of positions in ground truth
MT(↑)	The ratio of mostly tracked trajectories, that are tracked for more than 80%
ML(↓)	The ratio of mostly lost trajectories, that are tracked for less than 20%
PL(↓)	The ratio of partially lost trajectories, that are tracked in 20-80%
Frag.(↓)	fragments, the number of times that a ground truth trajectory is interrupted
IDS(↓)	ID switch, the number of times that a tracked trajectory changes its matched ID

Note: For the items with ↑, higher scores indicate better results, for those with ↓, lower scores indicate better results.

We find that the depth factor has been calculated as the affinity probability. We adopt an iterative strategy to search for a global optimal solution of Eq. (11), and find the best decision matrix X_{abcd}. We limit our search step along the descent direction given by the matching baseline in the same depth chain, this can exclude the invalid step connecting two chains with a large depth-span, and allow the search technique quickly converge to a near optimal solution. Moreover, the case that a chain has only one observation will decrease the complexity further. we then use switch labels to finish the observation matching at each iteration. For the selected observation, we try to switch labels with the observation in the same chain. If the new overall affinity probability is higher, we keep this change.

III. EXPERIMENTS

Datasets and metrics: To demonstrate the effectiveness of the proposed approach, we performed extensive experiments on three public datasets: the ISR-UC-SYNC dataset [18], the SDL dataset [19] and the SDL-Campus dataset [19]. Videos in these datasets are different from surveillance videos and are recorded using a non-static camera. The dynamic background and heads-up vision problems present challenges for tracking in these videos, and most occlusion cases in the real traffic scenes are full occlusions. These datasets consist of both videos from cameras and depth sequences from depth sensors. In our experiments, we use 4 frames for the association corresponding to a 4-partite structure. We normalize the appearance, position and motion similarity function in Eq. (10) to make them comparable.

To evaluate the tracking performance, we adopt the evaluation metrics [5][6] defined in Table I. These metrics have been widely used in previous literature for evaluating the performance of tracking. They measure not only the ID of each target, but also the long-term performance of the tracker. Although, frame rates, resolutions and densities are different in the three datasets, we use the same parameter

TABLE II
COMPARISONS OF EXPERIMENTS ON THREE DATASETS

Dataset	Method	Recall	Prec.	GT	MT	PL	ML	Frag.	IDS
SYNC	Berclaz et al. [3]	69.6%	74.8%	66	64.5%	22.7%	12.8%	45	23
	Andriyenko et al.[7]	73.4%	78.3%	66	69.7%	19.7%	10.6%	39	18
	Milan et al. [8]	75.6%	80.2%	66	71.2%	18.2%	10.6%	37	16
	NN	54.5%	64.3%	66	45.5%	30.3%	24.2%	52	31
	MDA	73.1%	78.6%	66	68.2%	15.2%	16.6%	39	17
	Our method	**85.0%**	**89.7%**	66	**80.3%**	**10.6%**	**9.1%**	**21**	**7**
SDL	Berclaz et al. [3]	68.9%	74.5%	92	60.9%	17.4%	21.7%	58	31
	Andriyenko et al. [7]	70.4%	76.4%	92	63.0%	20.7%	16.0%	51	29
	Yang et al. [5]	72.3%	77.8%	92	64.1%	21.7%	14.2%	47	26
	NN	59.4%	65.6%	92	43.5%	23.9%	32.6%	69	38
	MDA	67.0%	73.5%	92	59.8%	22.8%	17.4%	49	25
	Our method	**82.4%**	**87.3%**	92	**76.1%**	**15.2%**	**8.7%**	**28**	**14**
SDL-campus	Zhang et al. [2]	76.4%	79.8%	74	71.6%	18.9%	9.5%	30	16
	Milan et al. [8]	80.0%	84.5%	74	75.7%	16.2%	8.1%	26	14
	NN	67.7%	74.6%	74	60.8%	21.6%	17.6%	37	19
	MDA	78.3%	82.9%	74	73.0%	17.6%	9.4%	26	15
	Our method	**85.6%**	**89.3%**	74	**81.1%**	**12.2%**	**6.7%**	**14**	**8**

settings and ground truth data in our experiments and our method consistently outperforms previous methods on all three datasets. Therefore our method is not sensitive to parameters. Moreover, the DPM detector [20] we used in experiments is implemented in its generic, publicly available, pre-trained versions, which are not specifically trained for any dataset sequence.

Baselines: In order to verify the accuracy and efficiency of our method, we systematically compare the experimental results of our method against three groups of baseline methods (including the state-of-the-art methods). The first group of baselines are vision-based methods including Zhang et al.'s [2] and Berclaz et al.'s [3] network flow approach, Andriyenko et al.'s [7] continuous energy optimization, Yang et al.'s [5] online learned CRF model, and Milan et al.'s [8] detection and trajectory level exclusion method. The second group consists of depth-based methods. In particular, we use the depth value and the depth coordinates for association and use the Nearest Neighbor (NN) method to combine the motion and position features. The third group of baselines contains both vision and depth data and we adopt the proposed association method based on the appearance, motion, position features but without using the depth structure. In other words, all the target observations are associated in a traditional MDA method. TABLE II shows the comparison results of the above methods on three datasets.

Comparisons: All three datasets contain both vision and depth data. In order to scan targets at the waist level, the vision and depth sensors are mounted at a height of 0.9m. Long-term and frequent full occlusions occur in these sequences. Additionally, the dynamic background and illumination present challenges for solving the occlusion problem: false detection response. The "NN" and "MDA" methods based mainly on the motion and position features introduce false detection observations in trajectory. Vision cue based methods in the first group of baselines frequently make ID-switching errors. This is because the appearance model is easily confused by the recurring full occlusions, resulting in more fragmentation errors in tracking results.

In this case, the depth factor s_i which provides a reference in another dimension for the association problem is especially useful. A low affinity probability is given to the observation at the end of our depth chain, which makes it less likely to be associated with the occluding elements. We can see that our method has the lowest ML, Frag., and IDS errors in TABLE II. Recall and precision is greatly improved using our method. Meanwhile, the time complexity of our method is much lower than the MDA method, because the use of depth structure limits the search to a single chain, not all the observations. Fig. 4 shows the tracking samples on three datasets. The first and second rows of it show the complex occlusion issues. Although the cluttered and dynamic background brings in many challenges, our method can exclude the false detection responses, and keep the target ID correctly. So our method outperforms the methods in other groups of baselines. Therefore the depth information from the RGB-D data can provide another reference to the 2D image, and improve the accuracy and efficiency in multi-target tracking.

IV. CONCLUSION

Data association is a significant challenge for multi-target tracking. Existing approaches have used RGB-D data in detection and tracking, but most of them fail to form a complete association model with the depth data. In this paper, we implemented an RGB-D multi-target tracking method by integrating the depth and vision information into a depth-chain, which was formulated as a depth structure and solved by a depth heuristic searching. Extensive experiments on three public datasets demonstrated that our method is effective for complex tracking problems and advances the state-of-art. Although our data association is accurate and efficient, one limitation of our method is that the unstable

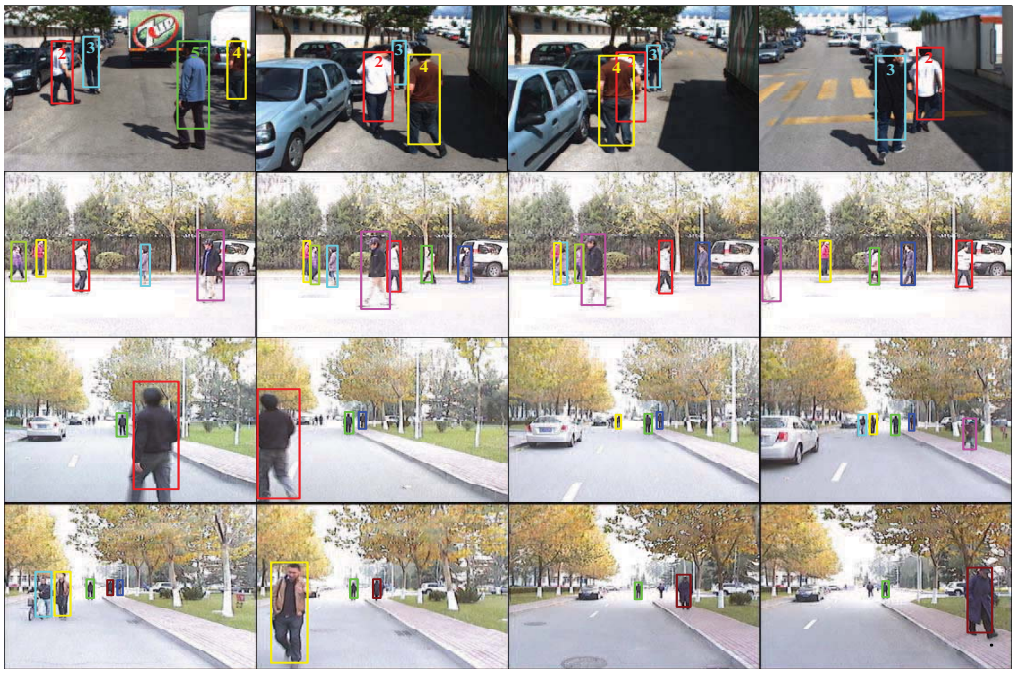


Fig. 4. Tracking examples on the three datasets.

detection responses may result in more false detection observations. We may use a detector based on RGB-D data to improve the detection accuracy and provide better detection response for tracking process.

ACKNOWLEDGMENT

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] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele, "Monocular visual scene understanding: Understanding multi-object traffic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 882–897, 2013.

[2] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[3] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1806–1819, 2011.

[4] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1201–1208, 2011.

[5] B. Yang and R. Nevatia, "An online learned crf model for multi-target tracking," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2034–2041, 2012.

[6] B. Yang, C. Huang, and R. Nevatia, "Learning affinities and dependencies for multi-target tracking using a crf model," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1233–1240, 2011.

[7] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1265–1272, 2011.

[8] A. Milan, K. Schindler, and S. Roth, "Detection-and trajectory-level exclusion in multiple object tracking," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 45, pp. 60–8, 2013.

[9] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," *Proc. Europeon Conference on Computer Vision*, pp. 788–801, 2008.

[10] R. T. Collins, "Multitarget data association with higher-order motion models," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1744–1751, 2012.

[11] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," *Proc. IEEE 11th International Conference on Computer Vision*, pp. 1–8, 2007.

[12] A. Ess, K. Schindler, B. Leibe, and L. Van Gool, "Object detection and tracking for autonomous navigation in dynamic environments," *The International Journal of Robotics Research*, vol. 29, no. 14, pp. 1707–1725, 2010.

[13] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *International journal of computer vision*, vol. 73, no. 1, pp. 41–59, 2007.

[14] A. Geiger, M. Lauer, and R. Urtasun, "A generative model for 3d urban scene understanding from movable platforms," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1945–1952, 2011.

[15] L. Spinello and K. O. Arras, "People detection in rgb-d data," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3838–3843, 2011.

[16] W. Choi, C. Pantofaru, and S. Savarese, "Detecting and tracking people using an rgb-d camera via multiple detector fusion," *Proc. IEEE International Conference on Computer Vision Workshops*, pp. 1076–1083, 2011.

[17] A. Broggi, P. Cerri, S. Ghidoni, P. Grisleri, and H. G. Jung, "A new approach to urban pedestrian detection for automatic braking," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 4, pp. 594–605, 2009.

[18] Z. Han, J. Jiao, B. Zhang, Q. Ye, and J. Liu, "Visual object tracking via sample-based adaptive sparse representation (adasr)," *Pattern Recognition*, vol. 44, no. 9, pp. 2170–2183, 2011.

[19] SDL, Dataset:http://www.ucassdl.cn/resource.asp.

[20] http://www.cs.berkeley.edu/rbg/latent/.