Robust depth-based object tracking from a moving binocular camera

Article *in* Signal Processing · September 2014

DOI: 10.1016/j.sigpro.2014.08.041

4 authors, including:

Liujuan Cao
Xiamen University
34 PUBLICATIONS **148** CITATIONS
SEE PROFILE

Cheng Wang
Xiamen University
157 PUBLICATIONS **894** CITATIONS
SEE PROFILE

Jonathan Li
University of Waterloo
252 PUBLICATIONS **3,240** CITATIONS
SEE PROFILE

Signal Processing 112 (2015) 154–161

Contents lists available at ScienceDirect

Signal Processing

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

Robust depth-based object tracking from a moving binocular camera

Liujuan Cao*, Cheng Wang, Jonathan Li

School of Information Science and Engineering, Xiamen University, China

ARTICLE INFO

Article history:
Received 18 March 2014
Received in revised form
20 August 2014
Accepted 25 August 2014
Available online 10 September 2014

Keywords:
3D Object tracking
Depth information
Local patch-based appearance model

ABSTRACT

Depth is a rich source of information and has been successfully utilized in numerous computer vision applications. However, it is often ignored in object tracking. In this paper, in contrast to traditional 2D image-based tracking method, we propose a novel 3D object tracking method from a moving binocular camera. To effectively handle the deformable targets, a target is first represented by a local patch-based appearance model. Then, to handle the partial occlusions, we design a simple yet effective scheme to detect and recovery occlusions using depth information obtained from a moving binocular camera. Therefore, the proposed method can simultaneous capture target appearance changes and alleviate the drifting problem. The experimental results demonstrate the effectiveness of the proposed method.

1. Introduction

Object tracking is one of the basic tasks in numerous computer vision applications, such as intelligence video surveillance, human machine interfaces, special effects in motion pictures, indexing for multimedia, and so on. Robust object tracking will greatly improve the performance of object intelligence video surveillance, human machine interfaces and activity analysis. However, designing robust object tracking methods is still an open issue, especially considering various complicated variations that may occur in dynamic scenes, e.g., illumination variations, pose changes, background clutters, occlusions, etc.

To achieve the goal of robust object tracking, a large number of 2D image-based methods using different features and learning schemes have been proposed over the years. However, the 2D image-based tracking methods are easily corrupted by the noises and cannot effectively handle the occlusions. Differently, depth information obtained from a binocular moving camera can provide potentially useful information to deal with the occlusions.

In this paper, we present a novel 3D object tracking method from a moving binocular camera that learns a robust patch-based target appearance model and explicitly handles the occlusions by using depth information. The key idea is firstly to utilize a local patch-based appearance model to represent the target. Since the target is represented by a local patch-based appearance model, the proposed tracking method can effectively handle the deformable targets. Then, to handle the partial occlusions, we use a simple yet effective scheme to detect and recovery occlusions using depth information obtained from a moving binocular camera, which is more robust than existing 2D image-based tracking methods. Experimental results on challenging video sequences demonstrate the robustness of the proposed 3D tracking method by comparing it with several state-of-the-art tracking methods.

The rest of the paper is organized as follows: Section 2 reviews the related work. The overview of the proposed

* Correspondence to: Haiyun Park, Xiamen University, 361005, Xiamen, Fujian, China. Tel.: +86 592 6162556.
E-mail address: caoliujuan@xmu.edu.cn (L. Cao).

http://dx.doi.org/10.1016/j.sigpro.2014.08.041
0165-1684/© 2014 Elsevier B.V. All rights reserved.

tracking algorithm is described in Section 3. The local patch-based target appearance model is described in Section 4. The tracking process and occlusion detection scheme are presented in Section 5. A summary of the proposed tracking method is given in Section 6. In Section 7, experimental results are given. Finally, the conclusions are drawn in Section 8.

2. Related work

To achieve robust object tracking even in complexity dynamic scenes, a number of tracking methods have been proposed.

In the tracking literature, one popular technique is to track object using fixed appearance models [1–3]. These methods assume that object will look nearly identical in each new frame. Thus, an appearance model of the object from the first frame can be always used to describe object appearance. However, these fixed appearance model-based tracking methods cannot achieve long-term robust tracking in dynamic scenes, which often requires addressing difficult target appearance update problem. To handle this problem, a number of authors have formulated the problem of visual tracking as an online learning problem, in which the target appearance is updated adaptively using the images tracked from the previous frames. Collins et al. [4] present a method to adaptively select one color feature from several different color spaces to construct adaptive appearance models, which can best discriminate the object from the current background. In Ref. [5], Avidan proposes a method using an adaptive ensemble of classifiers for object appearance model maintenance and tracking. Unfortunately, one inherent problem of online learning-based trackers is drift, a gradual adaptation of the tracker to non-targets.

To alleviating the drifting problem, a number of top-performing tracking methods have recently been proposed. Matthews et al. [6] propose a method by making sure the current tracker does not stray too far from the initial appearance model. Within the semi-supervised learning framework, Grabner et al. [7] treat all incoming samples as unlabeled data. One of the key limitations of the above methods is that very large changes would cause the failures. In [8] and [9], a multiple instance boosting based technique and a co-training based technique are respectively proposed to deal with the drifting problem. In [10] and [11], a structured output support vector machine is applied to object tracking. In [12], Gall et al. propose a Hough forests-based visual tracking method. Lu and Hager [26] propose a model adaptation method driven by feature matching and feature distinctiveness that is robust to drift. Oron et al. [27] develop another method to deal with the drift problem by automatically estimating the amount of local (dis)order in an object. In [28], a discriminative metric is learned for robust visual tracking. In addition, with the popularity of low-rank subspaces and sparse representations in image processing and machine learning, a variety of low-rank and sparse representations based tracking methods have been recently proposed [29–31].

Rather than using only 2D images, a number of authors have used 3D or stereo-based information to improve the performances of the computer vision systems [14–17,32–34,36–38]. For object tracking literature, Hu et al. [13] propose a principal axis-based stereo tracking method. Ess et al. [14] propose a robust multi-person tracking method from a mobile platform, in which depth information is used to verify object candidates obtained by object detection. Some authors focus on tracking the human body by using RGB-D cameras [15–17]. In [32], Choi and Christensen propose a RGB-D object tracking method using a particle filter on GPU. Kooa et al. [33] propose a novel model-free approach for tracking multiple objects from RGB-D point set data. In [34], Ren et al. propose a probabilistic framework for simultaneous tracking and reconstruction of 3D rigid objects using an RGB-D camera. However, their performances are severely impaired in an outdoor environment due to complicated illumination changes. Moreover, Kinect requires a minimum and a maximum distance from objects to the cameras in order to obtain accurate depth values.

Please refer to [18–21] for more complete reviews on the tracking methods.

3. Overview of the proposed tracking method

In this section, we develop our depth-based tracking algorithm from a moving binocular camera. The flowchart of the proposed tracking method is shown in Fig. 1.

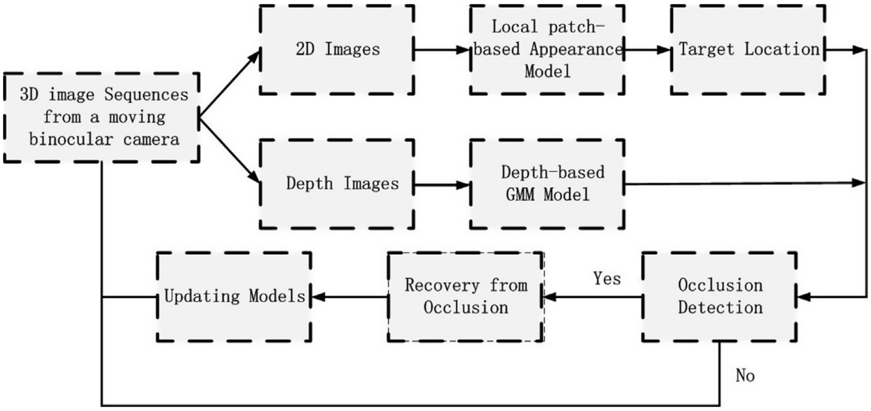

Fig. 1. The flowchart of the proposed tracking method.

Specifically, the proposed 3D tracking algorithm works as follows: The object of interest is manually selected in the first frame by a bounding box. Then, we construct a local patch-based appearance model and a depth-based model for the interested object. For one incoming video frame t, we first obtain a set of likelihood maps that are produced by matching each foreground patch model. Then, we design a robust estimator to fuse the likelihood maps and make the final decision on the new target position. According to the depth-based GMM model at frame t-1, the occlusion is detected and recovered. Finally, the local patch-based target appearance model and the depth-based model are adaptively updated respectively to capture the target appearance and depth variations. The tracking procedure continues in this iterative fashion until the end of video.

Below we give a detailed description about each component of our method.

4. The local patch based target appearance model

The appearance model is a basic issue to be considered in the object tracking problem. In this section, we propose an effective appearance model, which represent the target by a set of patches. Specifically, in our tracking algorithm, a target is represented by a local patch-based adaptive appearance model to explicitly handle partial occlusions. As illustrated in Fig. 2(a) and (b), multiple local patches are used to capture the local information of the target. Specifically, as illustrated in Fig. 2(a), the bounding box of a target is first divided into a set of non-overlapping horizontal patches with $(\beta \times W) \times (\gamma \times H)$ pixels, where W and $H$are the width and height of the bounding box respectively. β and γ are scale factors and typically set as 0.4 and 0.15 respectively. The two factors can be perturbed with little effect on performance. Similarly, as illustrated in Fig. 2(b), the bounding box of the target is then divided into a set of non-overlapping vertical patches with $(\gamma \times W) \times (\beta \times H)$ pixels.

A good feature is one of the key factors for a well-designed computer vision and pattern recognition system. Feature issues include: what feature is desirable for the recognition of a pattern and how to effectively extract the feature from the original input image. In this paper, to more clearly show the advantages of the proposed depth-based tracking method, we adopt intensity histograms to represent the local patches. The intensity histograms can be extracted in an efficient manner by using the integral image technique. It is important to note that other (potentially more efficient and robust) features may also be considered.

5. Tracking and occlusion detection

5.1. Tracking

The key part of the proposed 3D tracking algorithm proceeds by first computing a set of likelihood maps $M = M_i(\cdot, \cdot | i = 1, 2, \ldots, N_p)$ that serve as proposal solutions, and then optimally fusing them using a robust estimator. N_p denotes the number of effective patches and is adaptive updated in the tracking process. Given a patch, if its appearance variation between consecutive frames is above a predefined threshold, we consider the patch is unstable. Otherwise, the patch is considered as stable. We use the N_p stable patches to represent the object of interested. In this paper, we choose the intensity histogram intersection to calculate the appearance variations and the predefined threshold is set as 0.7. One element $M(x, y)$ of a likelihood map $M_i(\cdot, \cdot)$ is obtained by calculating the histogram intersection distance between *ith* patch's model to that of a candidate image patch centered at location x and y.

To fuse the tracking maps M and make the final decision, a robust estimator is designed

$$(x^*, y^*) = \text{argmin}_{(x,y)} S(x, y) \tag{1}$$

where $S(x, y) = \rho\%$ value in the sorted set $M_i(x, y) | i = 1, 2, \ldots, N_p$. Typically, ρ is set as 15. The intuitive idea is to filter the most similar and dissimilar patches of the candidate to the target. The most similar patches of the candidate to the target cannot discriminate foreground from background whereas the most dissimilar patches of the candidate to the target have high possibility to suffer a sudden appearance change or occlusion, etc.

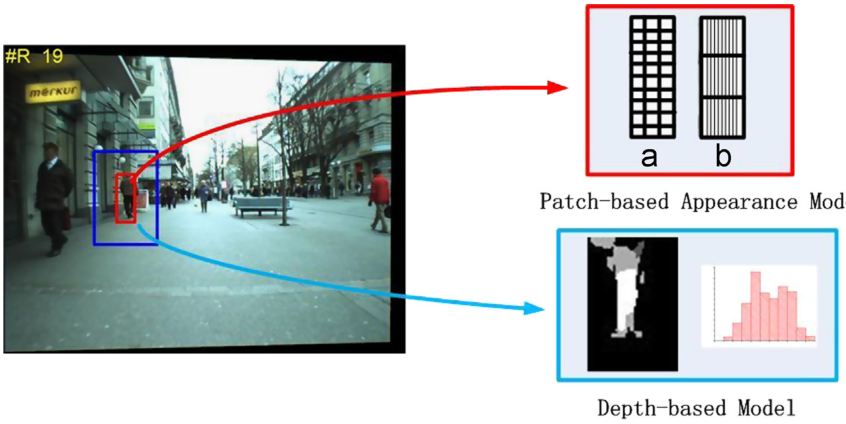

Fig. 2. Illustration of a local patch-based appearance model of the target and its depth-based model.

Table 1
A summary of the proposed tracking algorithm.

Algorithm 1 Depth-based object tracking from a moving binocular camera

```
Initialization:
1. Acquire one manually labeled frame.
2. Initialize a patch-based appearance model for the interested object.
3. Initialize a depth-based model for the interested object.
for t=2 to the end of the video
    1. Generate a set of likelihood maps via a rich set of each foreground patch's model.
    2. Fuse the likelihood maps and make the final decision on the new object position.
    3. Occlusions detection and recovery.
    4. Update the patch-based target appearance model.
    5. Update the depth-based model.
end for
```

5.2. *Occlusion detection and recovery*

Occlusion Detection: To detect the occlusion, we assume that the object depth is dominant within the bounding box that encompasses the object. Thus, the dominant depth value changes when an occlusion appearances. In this paper, given the dense depth information from a moving binocular camera, we employ the mixture of Gaussian model (GMM) [35] to construct the depth-based model, which greatly compensates the depth noises

$$p(d_t) = \sum_{i=1}^{K} w_{i,t} * \eta(d_t, \mu_{i,t}, \Sigma_{i,t}) \tag{2}$$

where K is the number of Gaussian components and η is a Gaussian probability density function

$$\eta(d_t, \mu_t, \Sigma_t) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_t|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (d_t - \mu_t)^T \Sigma_t^{-1} (x_t - \mu_t) \right\} \tag{3}$$

$w_{i,t}$, $\mu_{i,t}$ and $\Sigma_{i,t}$ are the time adaptive mixture coefficients, mean and variance, respectively, of the *ith* Gaussian of the mixture associated with d_t. At each time instant, the Gaussian components are evaluated in descending order with respect to w/Σ to find the first matching with d_t (a match occurs if the value falls within $2.5 \times \Sigma$ of the mean of the component). The first B components are chosen as the target depth-based model, where

$$B = \operatorname{argmin}_b \left(\sum_{k=1}^{b} w_k > T \right) \tag{4}$$

where T is the minimum portion of the target depth-based model. The weight $w_{i,t}$ is adjusted as follows:

$$w_{i,t} = (1-\alpha) w_{i,t-1} + a(M_{i,t}) \tag{5}$$

where a is the learning rate and $M_{i,t}$ is 1 for the model which is matched and 0 for others.

According to the depth-based GMM model, occlusion detection is done after getting the new target position. First, the pixels within the bounding box located at the new target position are classified as occlusion if they do not adhere to the model of the depth-based GMM model. Then, occlusion is detected if the ratio of occluded pixels and total pixels within the target bounding box is more than 0.8. Otherwise, the target is deemed to non-occluded.

Recovery from occlusion: The local patch-based target appearance model and depth-based GMM model are fixed when entering the occlusion state. Then, for one incoming video frame t+1, we first draw testing samples from a search window centered in previous target position. Furthermore, we compute the distances between the samples and the local patch-based target appearance model. The occlusion is recovered if the minimum distance among these distances is smaller than a given threshold.

6. Summary of the proposed tracking algorithm

A summary of the proposed tracking algorithm is described in Table1.

7. Experiments

The performance of the proposed depth-based tracking method is evaluated in this section. The algorithm is implemented using C++, on a computer with Intel-Core 2 2.86 GHz processor. It achieves the processing speed of 2 fps at the resolution of 320 × 240 pixels. The dense depth images from the moving binocular camera are generated by the well-known belief propagation algorithm [25].

We carry out experiments on five challenging sequences in the well-known binocular video datasets [14]. The challenges of these five sequences include large illumination variation, small target, drastic change in scale, occlusion and background clutters. We compare the performance of the proposed method to several state-of-the-art tracking methods, e.g., the Fragments-based Tracker (**FT**) [3], the context-based tracker [22] (**CT**), the Tracking-Learning-Detection-based Tracker (**TLD**) [23], and the depth driven tracker [24] (**DDT**). We use the same parameters as the authors have given on their papers and websites for all of our experiments.

To clearly show the tracking results obtained by the proposed method, we give the tracking results of the proposed tracker in the five testing image sequences in Fig. 3–7. It can be seen from these five figures that the proposed tracking method can achieve robust tracking even in the challenging conditions, such as occlusion, appearance changes, and size changes etc.

The quantitative comparison results of the trackers (i.e., our method, the **FT**, the **CT**, the **TLD**, and **DDT**) are listed in Fig. 8 and Table 2. The quantitative performance is measured by the center location errors (pixels) in each frame and average center location errors in the whole sequences. The ground truth is achieved by manually labeling all frames from the video sequences. It is easily to see that, due to using fixed appearance models, FT cannot effectively handle the target appearance variations. The **CT** can handle the drift

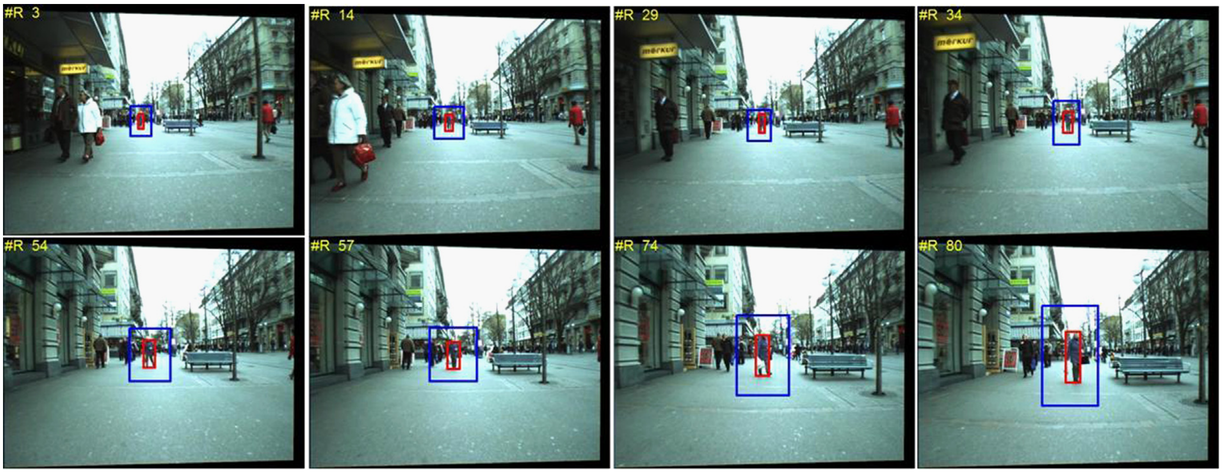

Fig. 3. Tracking results of the proposed tracking method on the first challenging video sequence.

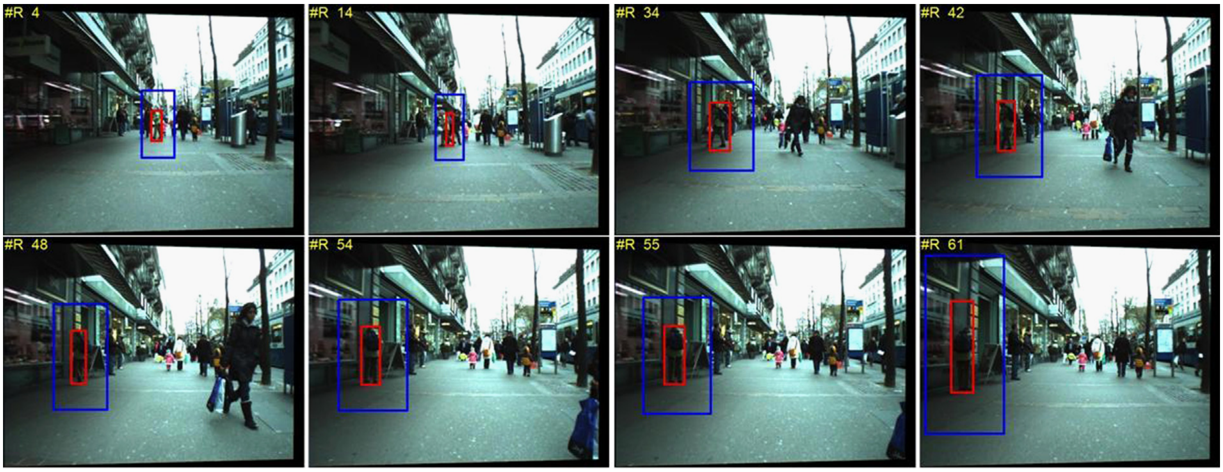

Fig. 4. Tracking results of the proposed tracking method on the second challenging video sequence.

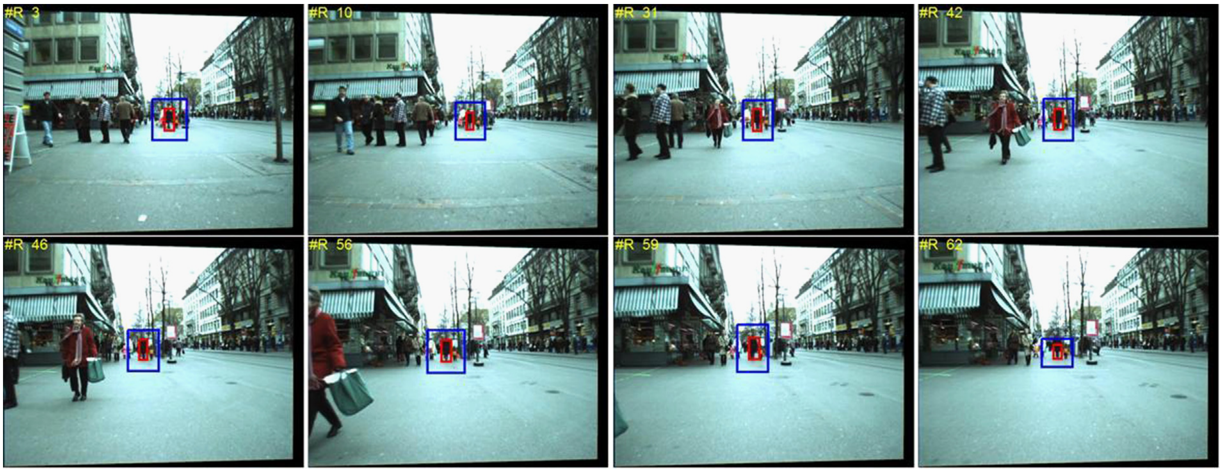

Fig. 5. Tracking results of the proposed tracking method on the third challenging video sequence.

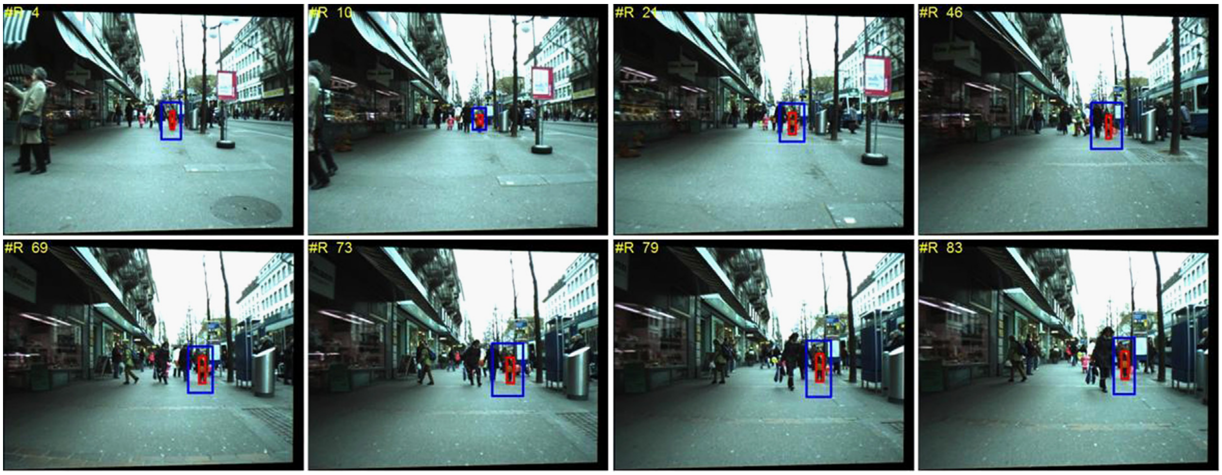

Fig. 6. Tracking results of the proposed tracking method on the fourth challenging video sequence.

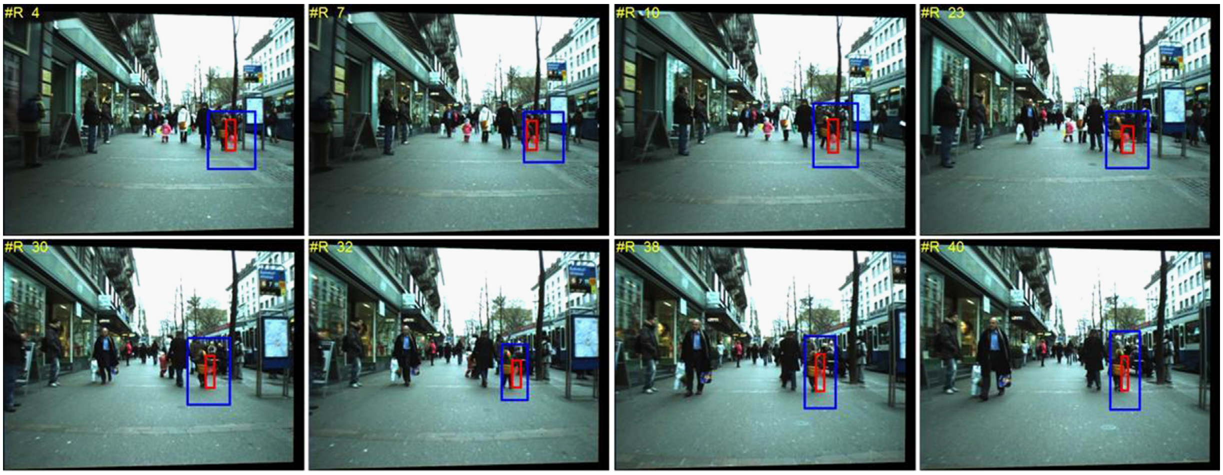

Fig. 7. Tracking results of the proposed tracking method on the fifth challenging video sequence.

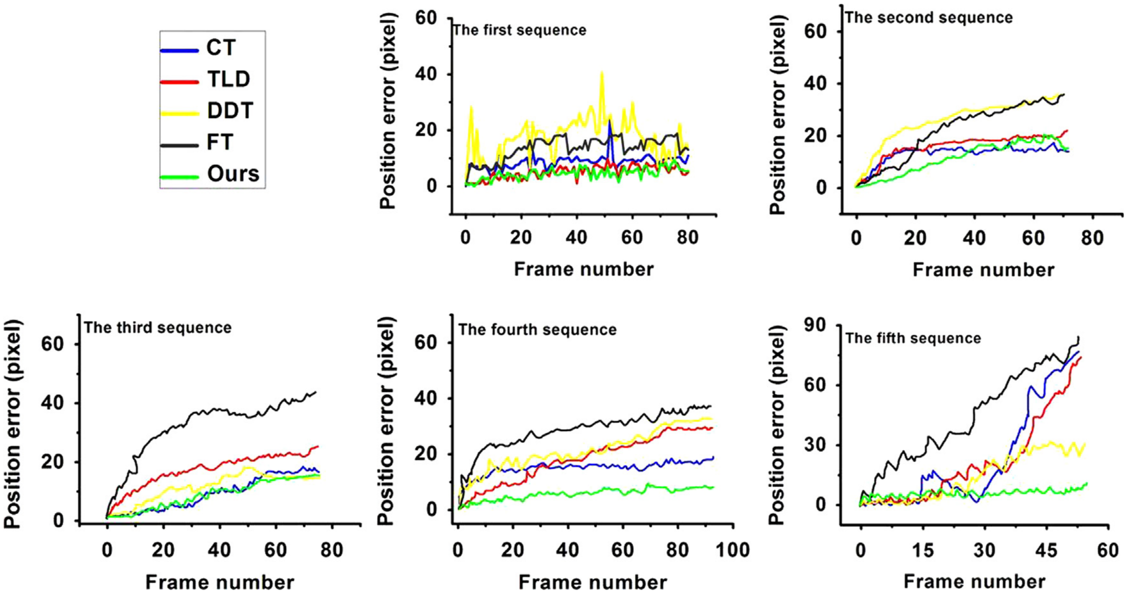

Fig. 8. Position error curves for five image sequences we tested on.

Table 2
Average center location errors (pixels). Quantitative comparison results on five challenging sequences by our method, the **FT**, **CT**, **TLD**, and **DDT** respectively.

Image sequence	**FT**	**CT**	**TLD**	**DDT**	**Ours**
1	12	7	6	21	5
2	20	12	15	22	10
3	25	12	20	10	11
4	22	13	17	18	6
5	52	46	45	23	9

and occlusion problem relatively well in some sequences due to using the context information. The **TLD** does not perform well in case of occlusions and dramatic figure/ground appearance pattern changes. The **DDT** cannot effectively track a target when the depth information is inaccurate. Based on powerful local patch-based target appearance model and depth-based occlusion detection, the proposed tracking method can track the targets for almost the full length of all these sequences.

8. Conclusion

In this paper, we have proposed a robust depth-based tracking method from a moving binocular camera. The local patch-based target appearance model is first used to handle the deformable targets. Then, given the dense depth information obtained from a moving binocular camera, a depth-based GMM model is used to detect occlusion. Therefore, our method can simultaneous capture target appearance changes and alleviate the drifting problem. Extensive comparison experiments on several challenging video sequences demonstrate the advantage of our method.

Acknowledgment

This work is supported by the Natural Science Foundation of China (No. 61402388 and No. 61373076).

References

[1] M. Isard, A. Blake., Condensation-conditional density propagation for visual tracking, International Journal of Computer Vision 29 (1) (1998) 5–28.

[2] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (5) (2003) 564–577.

[3] A. Adam, E. Rivlin, I. Shimshoni, Robust fragments-based tracking using the integral histogram, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 798–805.

[4] R. Collins, Y. Liu, M. Leordeanu, Online selection of discriminative tracking features, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (10) (2005) 1631–1643.

[5] S. Avidan, Ensemble tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2) (2007) 261–271.

[6] I. Matthews, T. Ishikawa, S. Baker, The template update problem, IEEE transactions on pattern analysis and machine intelligence 26 (6) (2004) 810–815.

[7] H. Grabner, C. Leistner, H. Bischof, Semi-supervised on-line boosting for robust tracking, European Conference on Computer Vision, 5302, 2008, pp. 234–247.

[8] B. Babenko, M. Yang, S. Belongie, Robust object tracking with online multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1619–1632.

[9] Q. Yu, T. Dinh, G. Medioni, Online tracking and reacquisition using co-trained generative and discriminative trackers, European Conference on Computer Vision 5303 (2008) 678–691.

[10] S. Hare, A. Saffari, P. Torr., Struck: structured output tracking with Kernels, International Conference on Computer Vision, 2011, pp. 263–270.

[11] R. Yao, Q.F. Shi, C.H. Shen, Y.N. Zhang, A.V.D. Hengel, Part-based visual tracking with online latent structural learning, 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2363–2370.

[12] J. Gall, A. Yao, L. Van, V. Lempitsky., Hough forests for object detection, tracking, and action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (11) (2011) 2188–2202.

[13] W.M. Hu, M. Hu, T.N. Tan, J.G. Lou, S. Maybank, Principal axis-based correspondence between multiple cameras for people tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4) (2006) 663–671.

[14] A. Ess, B. Leibe, K. Schindler, L.V. Gool, Robust multi-person tracking from a mobile platform, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (10) (2009) 1831–1846.

[15] S.R. Song, J.X. Xiao, Tracking revisited using RGBD Camera: unified benchmark and baselines, 2013 IEEE International Conference on Computer Vision (ICCV), pp. 233–240.

[16] M. Luber, L. Spinello, K.O. Arras, People tracking in RGB-D data with on-line boosted target models, IROS (2011). 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3844–3849.

[17] L. Spinello, M. Luber, K.O. Arras, Tracking people in 3Dusinga bottom-up top-downdetector, IEEE International Conference on Robotics and Automation (ICRA), pp. 1304–1310.

[18] A. Yilmaz, O. Javed, M. Shah, Object tracking: a survey, ACM Computing Surveys (CSUR) 38 (4) (2006) 13.

[19] H.X. Yang, L. Shao, F. Zhen, L. Wang, Z. Song, Recent advances and trends in visual tracking: a review, Neurocomputing 74 (18) (2011) 3823–3831.

[20] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, A. van den Hengel., A survey of appearance models in visual object tracking, ACM Transactions on Intelligent Systems and Technology (TIST) 4 (4) (2013) 58.

[21] Y. Wu, J.W. Lim, M.H. Yang, Online object tracking: a benchmark, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2411–2418.

[22] T.B. Dinh, N. Vo, G. Medioni, Context tracker: exploring supporters and distracters in unconstrained environments, 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1177–1184.

[23] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (7) (2012) 1409–1422.

[24] C. Li, L. Lu, G.D. Hager, J.Y. Tang, H.Z. Wang, Robust object tracking in crowd dynamic scenes using explicit stereo depth, Proceeding ACCV'12, Proceedings of the 11th Asian conference on Computer Vision Volume Part III, 2012, pp. 71–85.

[25] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient belief propagation in early vision, Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 1, 2004, pp. I-261–I-268.

[26] L. Lu, G. Hager, A Nonparametric treatment for location/segmentation based visual tracking, Computer Vision and Pattern Recognition (2007) 1–8.

[27] S. Oron, A.B. Hillel, D. Levi, S.Avidan, Locally orderless tracking, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1940–1947.

[28] N. Jiang, W. Liu, Y. Wu, Learning adaptive metric for robust visual tracking, Image Processing, IEEE Transactions on 20 (8) 2288-2300.

[29] T.X. Bai, Y.F. Li, Robust visual tracking with structured sparse representation appearance model, Pattern Recognition 45 (6) (2012) 2390–2404.

[30] X. Mei, H. Ling, Y. Wu, E. Blasch, L. Bai, Minimum error bounded efficient L1 tracker with occlusion detection, 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1257–1264.

[31] K.H. Zhang, L. Zhang, M. H.Yang, Real-time, Compressive tracking, Proceeding ECCV'12 Proceedings of the 12th European conference on Computer Vision Part III 2012, pp. 864-877.

[32] C. Choi and H. I. Christensen RGB-D Object Tracking: A Particle Filter Approach on GPU. The Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2013, pp. 1084-1091.

[33] S. Kooa, D. Leeb, D. Kwona, Incremental object learning and robust tracking of multiple objects from RGB-D point set data, J. Vis. Communication and Image Represent. Journal of Visual Communication and Image Representation archive 25 (1) (2014) 108–121.

[34] C.Y. Ren, V.A. Prisacariu, D.W. Murray, I.D. Reid., STAR3D: simultaneous tracking and reconstruction of 3D objects using RGB-D data, International Conference on Computer Vision (ICCV) (2013) 1561–1568.

[35] C. Stauffer, W.E.L. Grimson, Learning patterns of activity using real-time tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 747–757.

[36] Y. Gao, J.H. Tang, R.C. Hong, S.C. Yan, Q.H. Dai, N.Y. Zhang, T.S Chua., Camera constraint-free view-based 3D object retrieval, IEEE Transactions on Image Processing 21 (4) (2012) 2269–2281.

[37] Y. Gao, M. Wang, Z.J. Zha, Q. Tian, Q.H. Dai, N.Y. Zhang, Less is more: efficient 3d object retrieval with query view selection, IEEE Transactions on Multimedia 13 (5) (2011) 1007–1018.

[38] Y. Gao, M. Wang, Z.J. Zha, J.L. Shen, X.L. Li, X.D. Wu., Visual-textual joint relevance learning for tag-based social image search, IEEE Transactions on Image Processing 221 (14) (2013) 363–376.