STRUCTURE-AWARE MULTI-OBJECT DISCOVERY FOR WEAKLY SUPERVISED TRACKING

Yuankai Qi, Hongxun Yao, Xiaoshuai Sun, Xin Sun, Yanhao Zhang, Qingming Huang

School of Computer Science & Technology, Harbin Institute of Technology
qykshr@gmail.com, {h.yao, xiaoshuaisun, sunxin, yhzhang}@hit.edu.cn, qmhuang@jdl.ac.cn

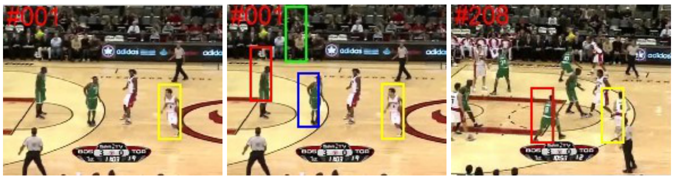
ABSTRACT

Recent progress on tracking has focused on designing robust statistical model or proposing effective appearance features to improve precision. This paper addresses another problem, namely the discovery and tracking of generic multi-object which have the similar appearance and motion pattern based on limited human annotations. We present a model-free tracking method that can automatically discover and track multi-object sharing the same spatial and motion structure, and update the structure during the tracking without prior acknowledge. The candidate objects are first selected by a SVM classifier trained on histogram-of-gradient (HOG) features. Then a segment algorithm is exploited to decide the suitable sizes of tracking boxes. The structure constrains are updated in a real-time manner according to the motion measure among the specified object and corresponding candidates. Experimental results reveal significant convenience and remarkable performance of our approach for the task of structure preserving multi-object discovery and tracking.

***Index Terms*—** Automatically discover, adaptive structure, multi-object tracking

1. INTRODUCTION

Multi-object tracking plays a key role in plenty of applications ranging from vision-based surveillance to automatic aiming. Considerable progress has been made in the development of tracking specific objects such as hands [1], humans [2] and rigid objects [3] over the last few years. Specifically, model-free tracking methods for generic objects have received increasing focus [4][5][6]. However, in model-free tracking, the interesting objects are manually marked in the first frame and the tracker will trace them throughout the remains of the video. In our work, we proposed a novel joint framework for object structure discovery and tracking where only a limited number of source object annotations are required. The unlabeled objects that share the similar visual appearance and motion patterns with the annotated objects will be automatically discovered, integrated and traced. Figure 1 gives a glimpse of our work. Figure 1(a) shows the manual annotation with yellow box. Figure 1(b) shows the discovered objects with red, blue and green boxes that have similar appearance. And Figure 1(c) shows one tracking result at the 208 frame, in which objects with different motion pattern from the annotation had been filtered out.



(a) (b) (c)

Fig. 1. Objects discovery and tracking with limited annotations.

The proposed framework consists of four modules: object discovery, structure construction, motion discovery and structure-preserving tracking. Although object detection is challenged by various factors such as illumination, appearance deformation, and viewpoint changes, we design a joint model to solve all the above mentioned problems. The first step for object discovery is appearance based classification. Since we want to track generic objects, Dala-Triggs detector [7] is exploited here to capture the appearance information. In case of false classification in some complex environments caused by this simple classifier, we integrate position-gradient similarity to further filter out the false candidates.

The spatial structure constraints can improve tracking accuracy when the objects have similar appearance or are occluded [8][9]. But the structure configuration nodes are fixed in existing approaches. In our work, the configuration of discovered structure is updated during the tracking process. Objects will be removed from the configuration if their motion patterns do not match that of the marked object. Since objects may have various scales and larger boxes may contain more background information which may lead false classification, we further design a segment process to determine the size of bounding boxes.

Due to the motion patterns are collected as the video goes, the movements often show instable characteristics. For example, a static man we saw in a video may move slightly between frames. This makes motion pattern definition infeasible frame

This work was supported in part by National Natural Science Foundation of China: 61133003, 61025011 and 61332016.

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

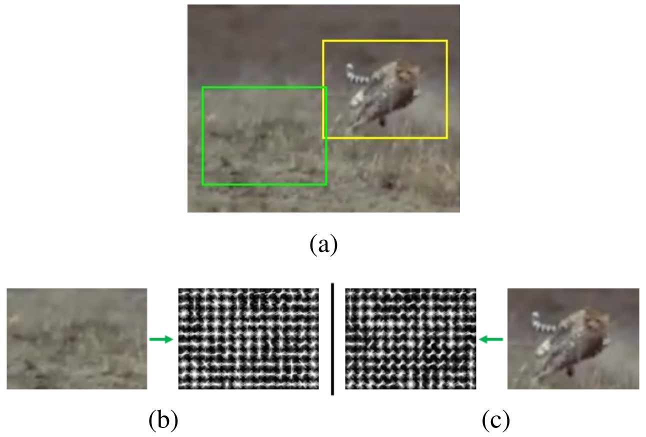


Fig. 2. False cheetah classification based on HOG features.(a)high scored grass (b)grass and its HOG features (c)cheetah and its HOG features.

by frame. Empirically, we gather motion information in every seven continuous frames. We designed a function to measure the movement and the similarity between two objects. Its output is used in structure updating.

For structure tracking, we use a model-free method [8], which incorporates spatial constraints, to trace the objects of the discovered structure.

The rest contents are organized as bellows. In section 2, we describe each component of our framework in detail. Section 3 presents the experiments and evaluation results. Finally, we conclude the paper in section 4.

2. THE PROPOSED METHOD

The proposed method can be divided into three stages which include four modules. First, we adopt appearance based object discovery to produce possible tracking candidates. Second, we construct the object structure based on the discovered candidates. Finally, multi-object tracking is performed and the motion patterns are simultaneously collected, which is used to update the spatial structure. All the processes are conducted on gray-scale values.

2.1. Object Discovery

According to Dalal Triggs detector [7], we use the SVM scheme and HOG features to discover the candidates. HOG features can reduce the illuminative disturbance and catch multiple direction edge information. They are extracted upon 8-by-8 pixel cell size, 2-by-2 cell block size and 9 directions (unsigned). The SVM classifier is formulated as:

$$S_{svm} = \mathrm{W}^T \varPhi_i + b \tag{1}$$

where $\varPhi_i$ is the HOG feature of the ith candidate, W is the weight for the feature, and b is the bias of the hyperplane.

The SVM is trained on 50 positive examples and 200 negative examples. In the first frame, we sample positive examples within an annular region a few pixels around the annotation and the negative ones far from the annotation (at most 50 percent overlap with the annotation). The advantage of using negative examples which contain both background and part of the annotation is that the classifier gives a small score to the candidate containing only partial appearance of the annotation.

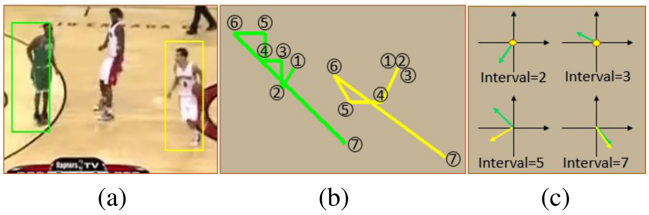


Fig. 3. A movement case of basketball players in seven continuous frames (a)two tracking objects (b)movement traces (c)motion vectors with different intervals.

We use overlapped sliding windows on the first frame to obtain candidates. To get the similar appearance with various size, we draw candidates with different scales (0.5x, 1x, and 2x are considered in our experiments). All the train samples and test samples are normalized to the same size (40-by-40 in our experiments).

As HOG features are position-independent, different appearance may present very similar statistic result [10]. Figure 2(a) shows a result of cheetah classification in which a bunch of grass gets a high score. Figure 2(b) and (c) reveal the reason that the cheetah and the grass in green box have similar statistic HOG features. To overcome this drawback, we exploit position-gradient similarity measurement.

Gradient magnitudes of each candidate image region are calculated and then converted to a vector Ψ_i. We use the intersection function to reflect the gradient similarity in terms of position by

$$S_{pg} = \sum_{i=1}^{M} \min(\Psi_z, \Psi_i), \tag{2}$$

where Ψ_z is the position-gradient vector of the annotation region and M denotes the number of candidates. The score of each candidate is given by

$$S = S_{svm} + S_{pg}. \tag{3}$$

Candidates with scores higher than a specified threshold are picked out. Together with the manually marked object, they constitute the initial tracking objects.

2.2. Structure Construction

Recent studies [8][9] show that either object detection or object tracking can obtain considerable improvement when considering structure information. Based on the success of [11], we construct the structure for initial tracking objects determined in previous step (section 2.1).

Before constructing the structure, a segment process on the candidate image region is conducted to get a more precise bounding box. In this process, we use Canny operator to get edges. And then in the top, bottom, left and right directions, we find the first continuous edges respectively. The first points of these edges can be seen as the outmost edge points. Bounding boxes are determined based on the outmost points with 10 pixels margin.

We denote V to represent the initial tracking objects set and $\mathrm{x}_i = (x_i, y_i)$ to represent the center location of each object $i \in V$. Inspired by [11], a minimum spanning tree structure can be generated on these objects. If x_i and x_j are connected in the tree, we denote the edge as $\mathrm{e}_{ij} = \mathrm{x}_i - \mathrm{x}_j$. All the edges determined by the tree structure form the edge set E.

2.3. Object Tracking and Motion Discovery

When initial tracking objects are selected and their spatial structures are built, we conduct a model-free multi-object tracking [8], which uses an SVM to predict object presence incorporating spatial constraints.

2.3.1. Object Tracking

The tracking model consists of two terms: (1)an appearance similarity measurement and (2)an edge deformation measurement for the tracking objects as below:

$$s = \sum_{i \in V} \mathrm{W}_i^T \varPhi_i - \sum_{(i,j) \in E} \lambda_{ij} \|(\mathrm{x}_i - \mathrm{x}_j) - \mathrm{e}_{ij}\|^2 \tag{4}$$

where $\varPhi_i$ denotes the HOG features of object i, W_i is the linear weight to feature $\varPhi_i$, λ_{ij} controls the trade-off between the two terms (in our experiment λ_{ij}=0.001) and e_{ij} is the edge between object i and object j in the last frame.

2.3.2. Motion Discovery and Structure Updating

We try to find the objects which can maximize equation (4) and define

$$\mathrm{m}_i^t = \left(\mathrm{x}_i^{t+1} - \mathrm{x}_i^t\right) / \left\|\mathrm{x}_i^{t+1} - \mathrm{x}_i^t\right\|_2 \tag{5}$$

as the motion vector for object $i \in V$ at the tth frame. When vectors are normalized, the dot product can be used to measure their similarity as the following

$$\left(\widetilde{S}_m\right)_{zi}^t = \left\langle \mathrm{m}_z^t, \mathrm{m}_i^t \right\rangle, \tag{6}$$

where $\widetilde{S}_m$ is the motion similarity between object i and the manual marked object z. The value of $\left(\widetilde{S}_m\right)_{zi}^t$ ranges from -1 to 1 with the most similar closely to 1.

However, note that the frame by frame motion vector dot product is too sensitive to describe the similarity. For example, as shown in Figure 3(a) the green and the yellow bounded players are confrontational in seven consecutive frames. Figure 3(b) shows the movement traces frame by frame. Figure 3(c) presents motion vectors with $interval$=2,3,5,7 respectively. When $interval$=2,3, the yellow bounded player stood originally, so he had a totally different motion pattern from the green bounded player. When $interval$=5, there is a big difference between their motion directions. But the truth is that, the motion patterns of the two players are very similar throughout the seven frames. This fact can be relatively precisely described by $interval$=7. We also observed motion vectors with $interval$=8,9,10, and their comparation states $interval$=7 is optimal globally.

So we observe motion similarity on seven consecutive frames and define the follow measure

$$(S_m)_{zi}^t = \begin{cases} 1, & \left(\widetilde{S}_m\right)_{zi}^t > thred, t = t, \cdots, t+6 \\ 0, & otherwise \end{cases} . \tag{7}$$

In our experiment, we set $thred = 0.5$.

When $(S_m)_{zi}^t = 0$, we delete object i from objects set V, and reconstruct the tree structure as section 2.2 discribed.

3. EXPERIMENTS

We conduct experiments on five challenging image sequences. Two of them (*Shaking* and *Basketball*) were already used in [12]; the other three were used in [8]. The shortest length of the videos is 279 frames and the longest is 2249 frames. Their average length is 1121 frames. All the ground-truth used in our experiments is obtained from [8].

We evaluate the performance of the tracking system with two indicators, which are commonly used in tracking field: (1)average distance error (Err.): the average distance of center points between the tracked box and the ground-truth box and (2)precision (Prec.): the average percentage of frames for which the overlap between the identified boxes and the ground-truth boxes higher than 50 percent. For each image sequence, the all tracking methods run five times separately. The two measurements are calculated on all target objects.

Comparisons are made with OAB[13], TLD[5] and SPOT[8] trackers. OAB, TLD and SPOT trackings are initiated with tracking targets used in [8], which proposed SPOT. Our tracking method initiated with only one of the initial targets of SPOT, and the other targets in our tracking are automatically discovered.

The visualized tracking results are presented in Figure 4. The first column shows the manual annotation in each tracking. The second column shows the remaining tracking objects discovered by our method. The rest three columns give some results during the whole tracking. In the head four tracking experiments(first four rows in Figure 4), we use dotted line to denote objects that do not have similar motion pattern to the manual annotation and we turn the show off during the fifth tracking experiment.

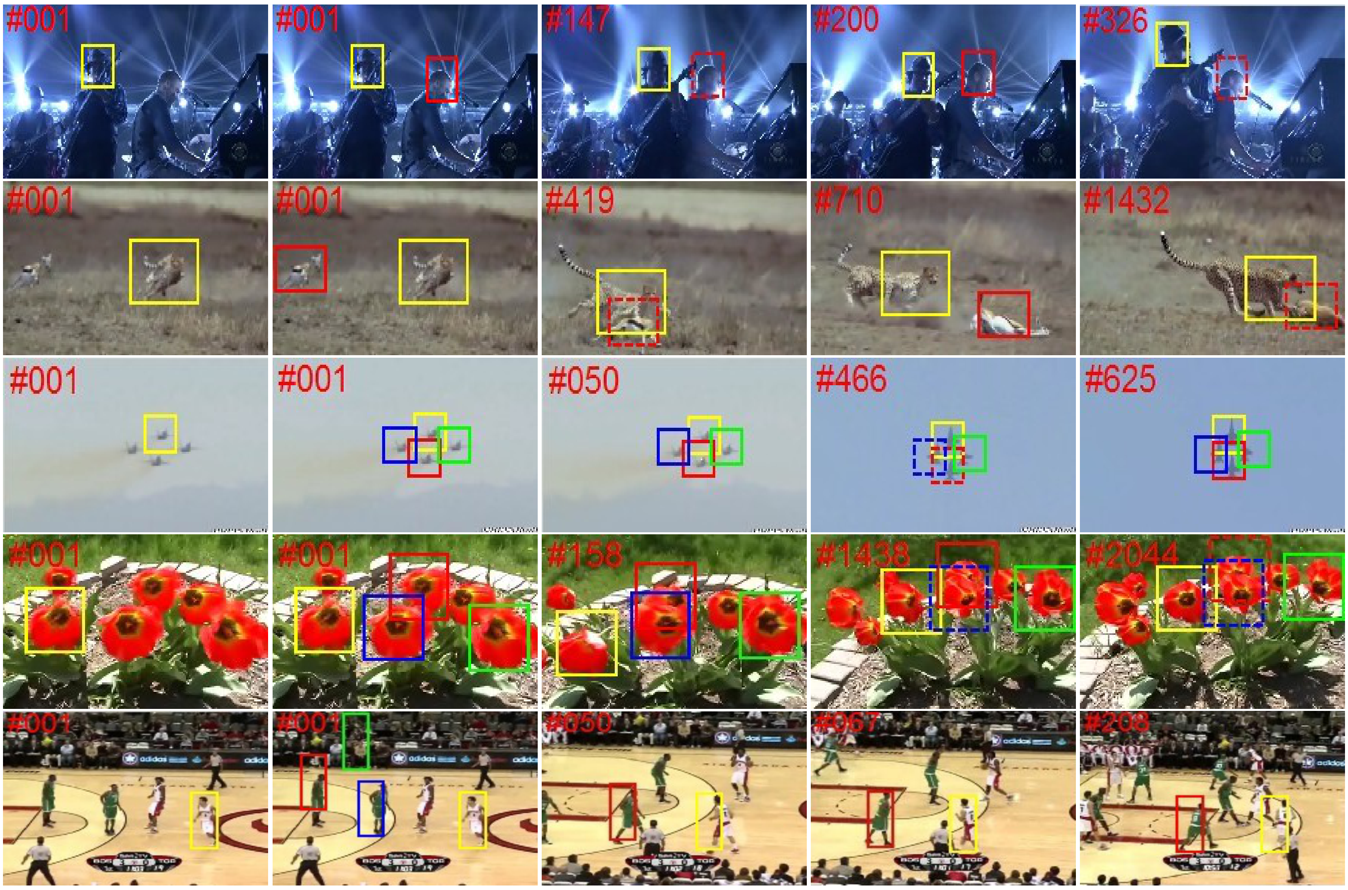


Fig. 4. Tracking results obtained by our method in environments including heavy occlusion (red flowers), camera shake (air show), motion blur (hunting and basketball) and illumination change (shaking). The manual marked tracking object is bounded in yellow box. The remaining bounded objects are discovered by our method. In first four experiments (rows), dotted boxes denote objects having a different motion pattern from manual annotation. And in the last experiment (row), we delete them instead of denoting them in dotted boxes.

The quantitive performance of four model-free trackers is presented in Table 1. Because our method is based on the SPOT, in a few cases the results are very close. The results show that (1)the spatial structure are effective indeed and (2)our method usually gains a highest precision but a lager center point distance error compared with the runner up, namely SPOT. The difference between the two methods are: (1)very close but different initial tracking position and (2)whether updating spatial structure (our method updates the configuration nodes of the structure, but SPOT not), so the bigger distance errors on center point are caused by different initial tracking position and the higher precision of our method which means higher overlap with the ground-truth shows that structure discovery and updating are useful.

Table 1. Performance of four model-free trackers. The best results of each video are boldfaced.

	OAB [13]		TLD [5]		SPOT [8]		Ours	
	Prec.	Err.	Prec.	Err.	Prec.	Err.	Prec.	Err.
Flowers	0.09	79.7	0.30	33.3	**0.99**	**8.2**	0.98	10.1
Air Show	0.86	9.3	0.53	31.3	0.92	**6.9**	**0.97**	9.5
Hunting	0.25	104.9	0.08	166.4	0.87	**17.9**	**0.97**	19.4
Shaking	0.47	61.9	0.47	14.3	**0.97**	**9.8**	0.80	24.9
Basketball	0.63	24.4	0.67	15.6	0.85	**12.7**	**0.88**	15.9

4. CONCLUSION

We have proposed a method to automatically discover tracking objects and spatial structure according to appearance and motion pattern. The experiments demonstrate that the motion structure of multiple objects can be effectively discovered with limited human interactions. The spatial constrains provided by such structure can further help us achieve more accurate object tracking results. In our future work, we will consider a more robust target discovery method which is unsupervised and hence totally free of human interactions.

5. REFERENCES

[1] Robert Y Wang and Jovan Popović, "Real-time hand-tracking with a color glove," in *ACM Transactions on*

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

Graphics (TOG). ACM, 2009, vol. 28, p. 63.

[2] Tao Zhao and Ramakant Nevatia, "Tracking multiple humans in complex situations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1208–1221, 2004.

[3] Frank Dellaert and Chuck E Thorpe, "Robust car tracking using kalman filtering and bayesian templates," in *Intelligent Systems & Advanced Manufacturing*. International Society for Optics and Photonics, 1998, pp. 72–83.

[4] Bernt Schiele, "Model-free tracking of cars and people based on color regions," *Image and Vision Computing*, vol. 24, no. 11, pp. 1172–1178, 2006.

[5] Zdenek Kalal, Jiri Matas, and Krystian Mikolajczyk, "Pn learning: Bootstrapping binary classifiers by structural constraints," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 49–56.

[6] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie, "Robust object tracking with online multiple instance learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, pp. 1619–1632, 2011.

[7] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[8] Lu Zhang and Laurens van der Maaten, "Structure preserving object tracking," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 1838–1845.

[9] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg, "Who are you with and where are you going?," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1345–1352.

[10] Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, and Antonio Torralba, "Hog-gles: Visualizing object detection features," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013.

[11] Xiangxin Zhu and Deva Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[12] Junseok Kwon and Kyoung Mu Lee, "Visual tracking decomposition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1269–1276.

[13] Helmut Grabner, Michael Grabner, and Horst Bischof, "Real-time tracking via on-line boosting.," in *BMVC*, 2006, vol. 1, p. 6.