BUAA-PRO: A Tracking Dataset with Pixel-Level Annotation

Annan Li[1]
liannan@buaa.edu.cn
Zhiyuan Chen[2]
dechen@buaa.edu.cn
Yunhong Wang[1]
yhwang@buaa.edu.cn

[1] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, China.

[2] School of Computer Science and Engineering, Beihang University, Beijing, China.

Abstract

In computer vision tasks, the position and size of a target object can be represented by a rectangle bounding box. Because its efficiency and simplicity, bounding box is widely used. However, using this simple method cannot deal with the diversity of shape. An obvious drawback is non-target elements are inevitably included in the box, which is a crucial issue in object tracking. Tracking-by-segmentation is an effective way of overcoming the limitation of bounding box. However, related benchmark dataset is still rare. To address this problem, a dataset of 150 video sequences with fine foreground mask is presented in this paper. It provides a preciser benchmark for object tracking and can be also used for video segmentation.

1 Introduction

Object tracking is an important computer vision task. To tackle this problem, many approaches have been proposed in the past few decades [4]. In most cases, a target object is represented by a rectangle bounding box. The effectiveness of a tracker is also evaluated by the overlap between estimated and ground truth bounding boxes. As a simplified representation method, bounding box provides a rough approximation to the size and position of the target. Since the image data structure is a two dimensional array, rectangle bounding box, which is actually a set of array indexes, is computational efficient and very easy to use.

Although in many scenarios, bounding box is sufficient for representing the target, as shown in Figure 1, a simple rectangle cannot cover the variations of shape. In object tracking, drift is a fundamental challenge. To successfully track a moving object, an ideal tracker should be able to do two things: 1) it can effectively discriminate target from background or other moving objects; 2) its model can be adapted to the appearance change of the target. Tracker drift is caused by the gradual erosion of the target model during update. Since the bounding rectangle is usually bigger than the actual contour, non-target elements are inevitably included in the target model. In other words, the target model is polluted at the very beginning. Therefore, it is not surprising that it drifts from the real object as the model updates over time.

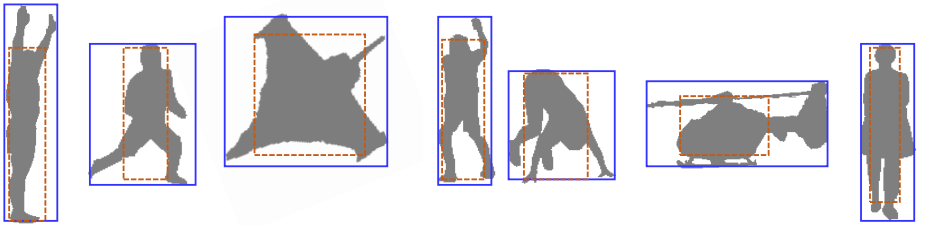

Figure 1: Limitation of bounding box. Non-target elements are inevitably included in both boundary based (solid rectangles) and torso based (dashed rectangles) bounding box.

The limitation of bounding box has been already recognized in the literature of object tracking. Ren and Malik [29] first treat tracking as a repeated foreground/background segmentation problem. The introduction of segmentation provides a better separation between target and background. Therefore, tracking-by-segmentation becomes one of the main stream approaches in object tracking [3, 6, 10, 33, 36, 40]. However, despite a decade's efforts on segmentation based tracking algorithms, main evaluation benchmarks are still bounding box based [18, 32, 37, 39]. To make comparison, target segments have to be converted into bounding boxes.

According to [18], on a dataset of 305 images, the average overlap between manual bounding box and the true shape is 0.5679 for torso based bound box (see Figure 1), and for boundary based bound box the mean overlap is only 0.4875. It leads to a situation that, if not converted back to bounding box, an ideal segmentation of the target might be judged as a failure based on the commonly used threshold 0.5. It implies that the quantitative evaluation in existing tracking benchmarks is an accurate statistic of inaccurate results.

Recently, learning from large-scale dataset by convolutional neural networks (ConvNets) achieves great success in computer vision and such approach has been successfully adopted to both tracking [23] and segmentation [20]. Tracking-by-segmentation strategy can achieve comparable results to the state-of-the-art before ConvNets emerges. However, to our knowledge, no competitive ConvNets based tracking-by-segmentation method has been yet proposed. Lacks of dataset with pixel-level annotation is a possible reason.

To address the above-mentioned issues, a large-scale tracking dataset with pixel-level annotation is presented in this paper. Extensive experiments and comparisons on this datasets show that it provides a preciser benchmark for visual tracking. This dataset can also be used for develop and evaluate video segmentation algorithms.

The remainder of this paper is organized as follows. Section 2 gives a brief review of related works. In Section 3, we elaborate the details of the proposed dataset. A thorough experimental evaluation of state-of-the-art object tracking algorithms is presented in Section 4. We conclude this paper and discuss future work in Section 5.

2 Related Work

2.1 Object Tracking

For quite a long time, evaluations of object tracking algorithms are independently performed using limited number of self-collected video sequences. Using different datasets makes it difficult to compare different tracking approaches. This situation has been changed recently

by introducing large-scale public benchmarks [18, 28, 32, 37].

ALOV++ [32] is a dataset consists of 305 video sequence, including 250 new sequences, 65 sequences from the PETS workshop [1] and several self-collected sequences used in previous studies. ALOV++ has 64 different types of targets, including human face, person, ball, octopus, microscopic cells, plastic bag etc. The total number of frames is 89364.

OTB [37, 39] dataset contains 100 video sequence collected from existing studies. It is mainly composed by sequences of human (36 body and 26 face sequences). Besides the bounding box, OTB also provides 11 attributes of the sequences: illumination variation, scale Variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out-of-plane rotation, out-of-view, background clutters and low resolution.

NUS-PRO [18] dataset provides 365 new video sequences collected from YouTube. There are four main categories: 1) rigid objects (airplane, boat, car, helicopter and motorcycle); 2) face; 3) pedestrians; 4) six kinds of sportsman (basketball, gymnastics, handball, racing, soccer and tennis). Besides them five long sequences with more than 2,000 frames are also included. Like OTB, NUS-PRO also provides 12 attributes. A notable feature of NUS-PRO is additional occlusion status are provided for every frame.

YouTube-BoundingBoxes (YTBB) [28] is the largest publicly available dataset for object tracking, which has 240,000 videos and 5.6 million frames are annotated with bounding boxes. However, this dataset is not originally designed for object tracking. The bounding box is not densely marked. The interval is one second, which includes around 30 frames. Moreover, since YTBB is marked by the *Amazon Mechanical Turk*, the quality is not well controlled.

All the above-mentioned large datasets are annotated by bounding boxes. And there is only one target in a video sequence.

2.2 Video Segmentation

Although pixel-level annotation for video is rare in the literature of object tracking, similar datasets have been already constructed for the purpose of video segmentation research.

DAVIS [2, 26] is the most relevant dataset. The latest release consists 150 sequences in which 10474 frames are annotated. Similar to OTB, NUS-PRO and YTBB, attributes like scale variation, background clutter are also provided. Although it is originally designed for the so called video object segmentation task, it can be also used for evaluating object tracking algorithms.

Freiburg-Berkeley Motion Segmentation (FBMS) [24] is designed for motion segmentation tasks such as optic flow. It contains 59 sequences with 720 annotated frames. Though most scenarios are rather simple, FBMS provides segmentation mark for all appearing objects at an interval of 20 frames. The marked objects are mainly animals.

SegTrack [35] is a small dataset consists of 6 videos of humans and animals, in which the challenges are background-foreground color similarity, fast motion and complex shape deformation. Due to the small number, it is insufficient for evaluating object tracking algorithms.

Other datasets for video segmentation include FlyingThings3D [22], KITTI [12] and Scene Flow [21]. Although point trajectories are available, since no object is separated, they cannot be used for evaluating object tracking methods.

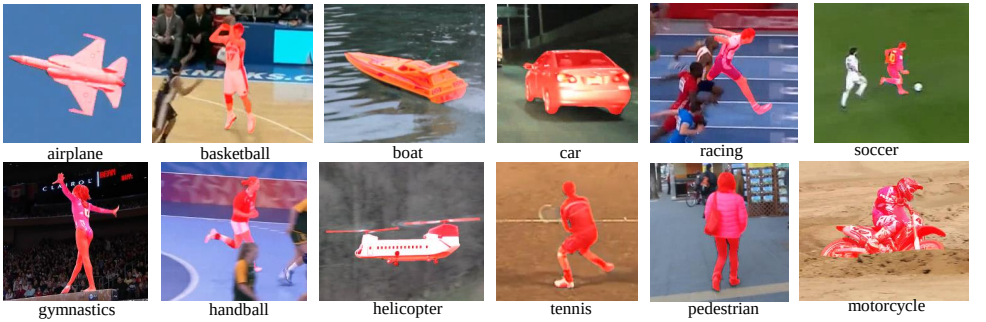

Figure 2: Exemplar annotations (red mask) of the proposed dataset. Best viewed in color.

Figure 3: Detailed issues in annotation.

Table 1: Statistics of the proposed dataset.

Category		Seq. No.	Marked frames		
			Min	Max	Mean
rigid object	airplane	10	39	41	40.8
	boat	10	47	61	56.7
	car	10	47	101	72.5
	helicopter	10	61	73	36.9
	motorcycle	10	38	73	52
sportsman	basketball	20	35	70	46.6
	gymnastics	10	62	391	125.9
	handball	5	36	50	43.8
	racing	5	51	84	69
	soccer	5	43	61	50.2
	tennis	5	45	11.6	75.6
pedestrian		50	35	71	49.42

3 The Dataset

As discussed in Section 1, although bounding box has obvious limitations, if take computational cost into consideration, it is still a practical representation method for object tracking and detection. Considering that quite a number computer vision studies focus on predicting mask from bounding boxes, a dataset with both bounding box and segmentation mask annotation could be useful in many aspects. Bounding box can be automatically generated from mask boundaries. However, in such box shape deformation of the real target can cause dramatic change of the bounding box area. The real outputs of real detector or tracker are actually not that sensitive. Observing that the torso based bounding box (see Figure 1) provides a better simulation of real detector or tracker, we build our dataset from NUS-PRO [13]. The new dataset is referred as BUAA-PRO.

3.1 Dataset Characteristics

The dataset is constructed by marking 150 video sequences in NUS-PRO, which correspond to the first half of category *rigid object*, *sportsman* and *pedestrian*. The segmentation mask is manually marked at an interval of five frames. A frame will be skipped if severe or full occlusion presents. Consequently, totally 8,714 frames are annotated with segmentation

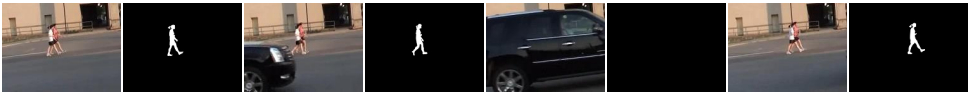

Figure 4: Temporal full occlusion.

mask. Examples of the mask can be found in Figure 2. Detailed statistics of the dataset are shown in Table 1.

The key difference between the proposed dataset and video object segmentation datasets like DAVIS is the *occlusion*. Specifically, it lies in two asptects. First, inherited from NUS-PRO, the proposed dataset contains occlusion status annotation, i.e. no occlusion, partial occlusion and full occlusion, for every frame, which is an important factor that is relevant to the area change of target mask. Secondly, as can be seen from Figure 4, temporal full occlusion is common in the proposed dataset. The memory mechanism that maintains the appearance model of the target and re-detects it in the incoming frame is a basic and necessary element of a tracker. However, this mechanism is not explicitly considered in the video object segmentation task. Therefore, though sharing some similarities, the problems are different.

3.2 Data Annotation

The masks of target object are manually achieved according to following criteria:

Details Marking the fine foreground mask of a target object is very time-consuming. As shown in Figure 3, trivial details like fingers and hairs of a target person bring in a large amount of annotation work but contribute little to the total evaluation accuracy. Balancing between efficiency and accuracy, these details are overlook during the annotation process.

Compression Artifacts We found that the block effects caused by video compression are common in the annotated video sequences. It leads to a dilemma in marking small objects: some parts of the target deviate from the main body in the image, if following the image fact the real-world fact will be violated. In this case, we choose to overlook the drifting parts and complete the whole object by estimation.

Motion Blur Due to fast movement of the target, motion blur also presents frequently. It leads to a phenomenon that some parts of the object become transparent in a short period. In this case, we choose to only mark the visible part.

Accessories The accessories on target person can be divided into two categories: 1) those addicted to target person throughout the whole sequence, such as bags on a pedestrian; 2) the balls played by sportsman. The former is considered as a part of the target, and the latter is ruled out from the target.

Propeller In the *helicopter* sequence, rotating propeller is a special phenomenon. When the revolving speed is high it becomes transparent, but it is visible when the speed is slow. Only the visible propellers are marked in this dataset.

Occlusion In NUS-PRO the missing part is completed by estimation in partial occluded frames. However, as shown in Figure 3, such completion is unfeasible for a fine mark. Thus, we only annotate the visible part of the target. When severe occlusion, such as the spindrift in Figure 3, presents, bounding box can be estimated from adjacent frames, but drawing a fine mask is impossible. Such frames are regarded as full occluded in this dataset (see the spindrift).

Table 2: Three criteria for computing the overlap ratio. B denotes the estimated bounding box and M is the manual mask.

Criterion	Occlusion		
	NA	Partial	Full
I	$\frac{area(B \cap M)}{area(B \cup M)}$	$\frac{area(B \cap M)}{area(B \cup M)}$	$\frac{area(B \cap M)}{area(B \cup M)}$
II	$\frac{area(B \cap M)}{area(B \cup M)}$	$\frac{area(B \cap M)}{area(B \cup M)}$	-
III	$\frac{area(B \cap M)}{area(B \cup M)}$	$\frac{area(B \cap M)}{area(B)}$	-

Table 3: Evaluated tracking algorithms.

Color-Based Probabilistic Tracking (CPF) [27]	Kernel-based Mean-Shift (KMS) [8]
Locally Orderless Tracking (LOT) [25]	Fragments-based tracking (Frag) [1]
Incremental Visual Tracking (IVT) [30]	On-line AdaBoost (OAB) [13]
Adaptive Structural Local Appearance model (ASLA) [16]	Semi-supervised Tracking (SemiT) [14]
Sparsity-based Collaborative Model (SCM) [44]	Semi-supervised Tracking with Adaptive Prior (BSBT) [22]
L1 Accelerated Proximal Gradient (L1APG) [3]	Multiple Instance Learning (MIL) [2]
Multi-Task Tracking (MTT) [43]	Compressive Tracking (CT) [42]
Local Sparse appearance model with K-Selection (LSK) [19]	Track-Learning-Detection method (TLD) [20]
Online Robust Image Alignment (ORIA) [33]	Circulant Structure tracking with Kernels (CSK) [15]
Distribution Fields Tracking (DFT) [31]	Context Tracking (CXT) [4]

3.3 Evaluation Methodology

Since the original images in our dataset are derived from the NUS-PRO, we adopt the three-criteria evaluation method introduced in [18], which is based on commonly used overlapping ratio defined in the PASCAL VOC challenge [11] with considerations of three kinds of occlusion status (see Table 2). It should be pointed out that, in partial occluded frames, the overlap will be much smaller using mask annotation, since missing parts are ruled out from the segmentation mask. In bounding box based annotation, occluded parts are completed by estimation.

4 Experiments

An extensive experimental survey of 20 popular trackers, which are summarized in Table 3, is presented. The implementations of these trackers are based on the publicly available code library [39] with default parameter settings. Evaluation of them is performed by comparing the manual segmentation mask with the rectangle mask given by the bounding box using the described in Section 3.3.

Results in term of the so-called threshold-response relationship (TRR) curves are shown in Figure 5, corresponding area under curves (AUC) are shown in Figuere 6. As can be seen, compared with results by bounding box vs. bounding box comparison, the overall overlap ratio drops but surprisingly not that significant. Considering that the area of corresponding bounding box is around two times of that of segmentation mask, the curves are expected much lower than those calculated from box-vs.-box comparison. A possible explanation is that, in most frames, the evaluated trackers actually failed. There is no difference using box or mask if a tracker completely drifts from the target. In other words, the difference is diluted.

We compare the results obtained by box-vs.-box comparison in [18]. With the same inputs, using fine mask gives different results of ranking. This phenomenon shows that bounding box based evaluations are not necessarily consistent with mask based evaluations.

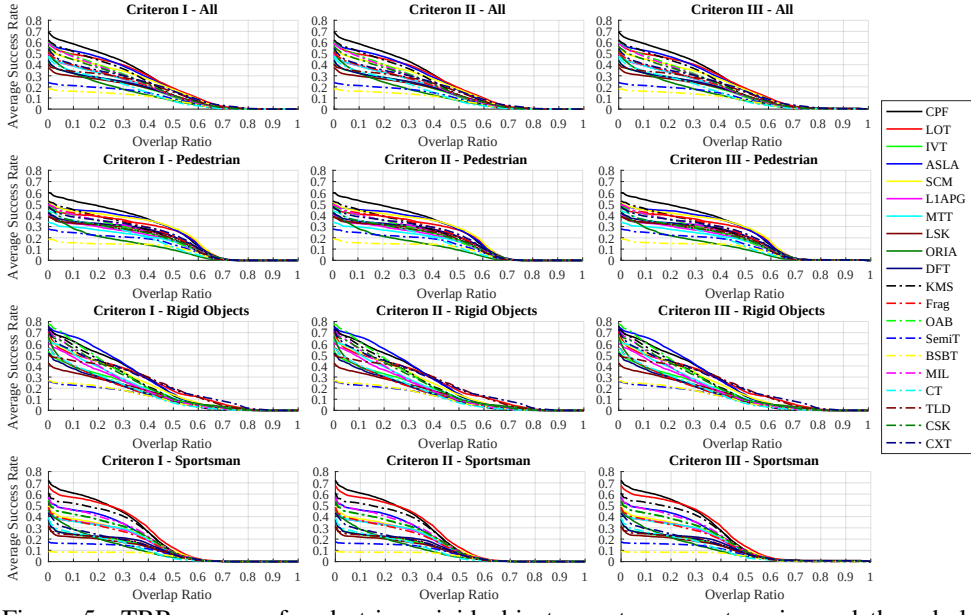

Figure 5: TRR curves of pedestrian, rigid object, sportsman, categories and the whole database (best viewed on high resolution displays).

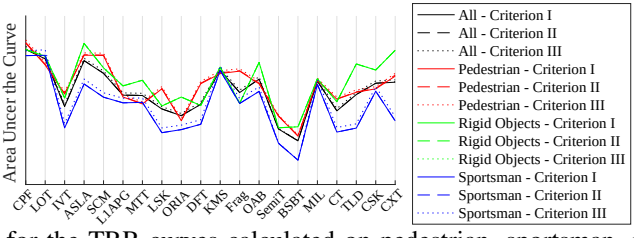

Figure 6: AUC for the TRR curves calculated on pedestrian, sportsman, rigid objects sequences and the whole database.

It is interesting that, though the detailed ranks are different, box based AUCs and mask based AUCs share the same union set of outstanding trackers. Which implies that the general conclusions are consistent.

The task of object tracking is defined as a general problem for common objects. However, it is unrealistic to deal with all kinds of challenges by a single tracker. Therefore, *Ad Hoc* evaluations for specific problem is more valuable in practice. To this end, attribute based evaluations for the 20 trackers are also conducted.

Experimental results in TRR curves and corresponding AUCs are shown in Figure 7, 8, 9 and 10 respectively. The general results are similar, a notable phenomenon is for fast background change the overall overlap by mask is higher than that by box. Since the targets are mainly rigid cars in this scenario, bounding boxes and mask are comparable in size and area. Masks bring in smaller intersection but may also reduce the union area.

The CPF, LOT and ALSA methods perform well in handling image sequences with scale change, shape deformation, partial occlusion and clutter background. For challenging videos containing flash and similar objects, the top ranked methods are LOT, KMS and CPF. The correlations can be explained from two aspects, namely, the image data and similarities in

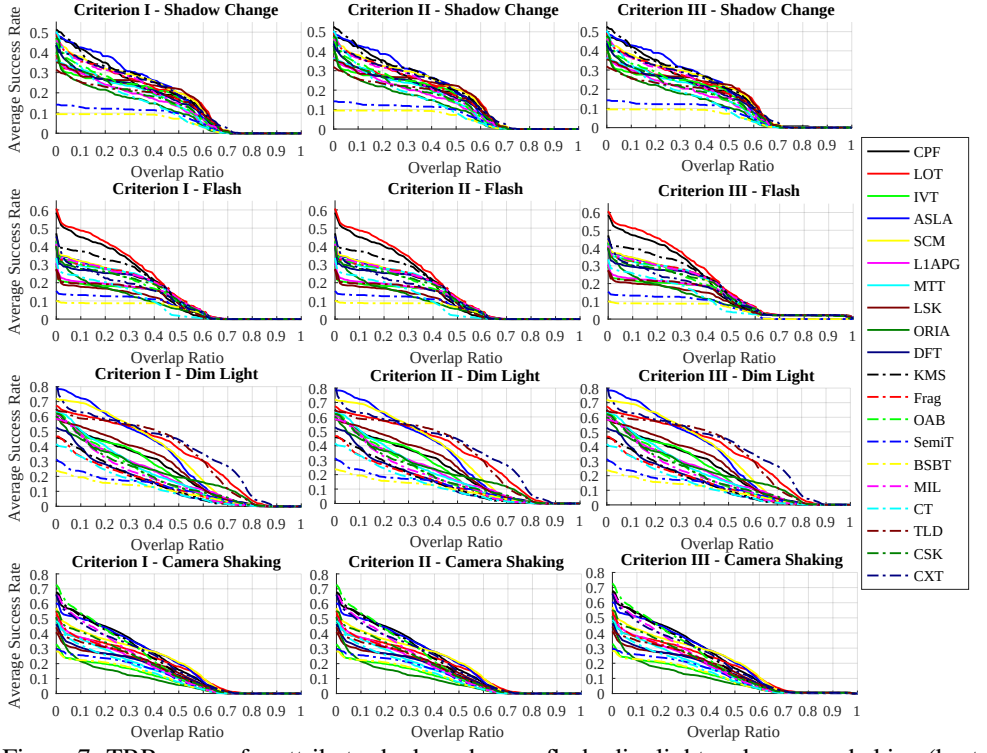

Figure 7: TRR curves for attribute shadow change, flash, dim light and camera shaking (best viewed on high resolution displays).

algorithmic properties. All the sequences with the challenging flash factor and two thirds of the videos containing similar objects are in the basketball category. Thus, the reasons why some methods perform well in dealing with flash and similar objects can be better accounted for by data correlation. On the other hand, scale change, occlusions and clutter background are common challenging factors in various categories. Therefore, the tracking results may be accounted by the similar algorithmic properties of the CPF, ASLA and SCM methods.

5 Conclusion

The essence of object tracking is how to separate foreground target from cluttered background. Compared with the diversity of shape, simple rectangle bounding box is insufficient for representing the target object. Its limitation has been recognized and studied in the tracking community, however, benchmark dataset for object tracking with pixel-level annotation is still rare. In this paper, we propose a large-scale dataset with carefully marked foreground masks , which provides a preciser benchmark for object tracking. It can be also used for studying the problem of video segmentation. The dataset will be release online in the near future for research purpose.

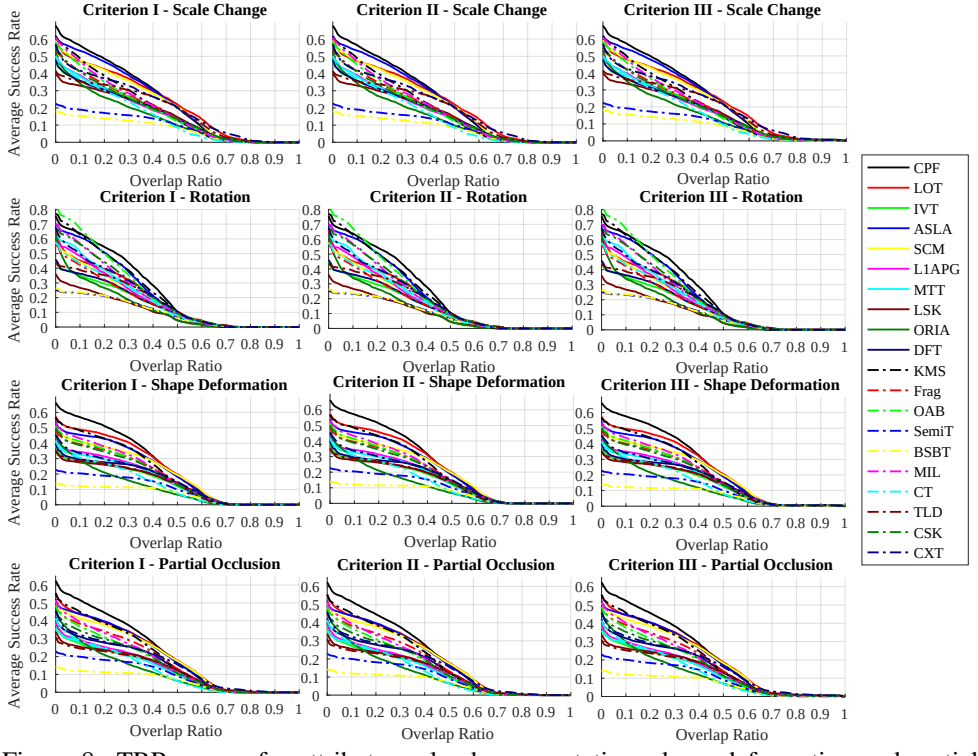

Figure 8: TRR curves for attribute scale change, rotation, shape deformation and partial occlusion (best viewed on high resolution displays).

Acknowledgment

This work was supported by the The National Natural Science Foundation of China under Grant 61573045.

References

[1] http://www.cvg.rdg.ac.uk/PETS2009/a.html.

[2] A. Adam, E. Rivlin, and I. Shimshoni. Robust Fragments-based Tracking using the Integral Histogram. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 798–805, 2006.

[3] C. Aeschliman, J. Park, and A. C. Kak. A Probabilistic Framework for Joint Segmentation and Tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1371–1378, 2010.

[4] B. Babenko, M.-H. Yang, and S. Belongie. Robust Object Tracking with Online Multiple Instance Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1619–1632, 2011.

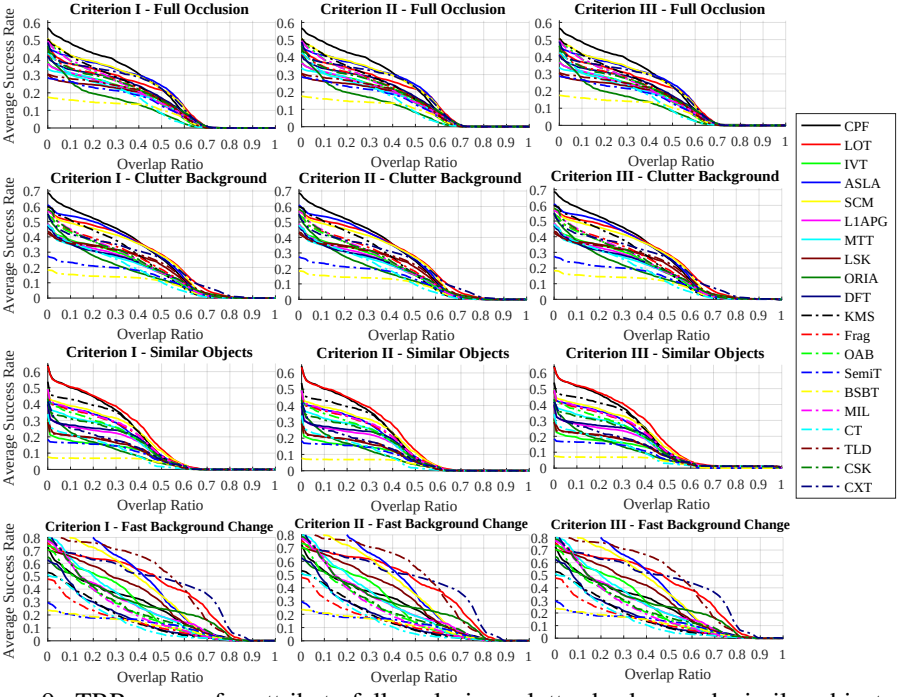

Figure 9: TRR curves for attribute full occlusion, clutter background, similar objects and fast background change (best viewed on high resolution displays).

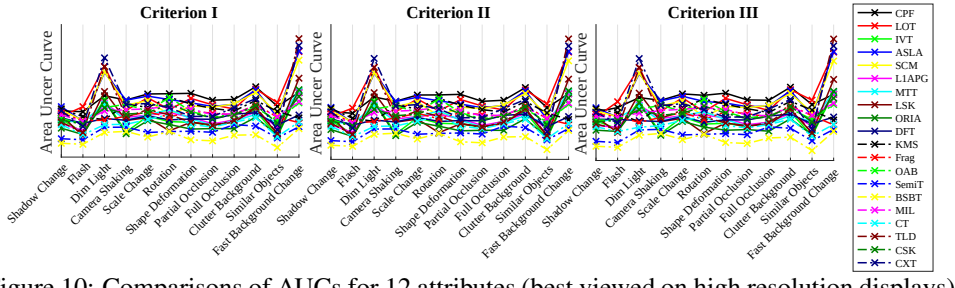

Figure 10: Comparisons of AUCs for 12 attributes (best viewed on high resolution displays).

[5] Chenglong Bao, Yi Wu, Haibin Ling, and Hui Ji. Real Time Robust L1 Tracker Using Accelerated Proximal Gradient Approach. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1830–1837, 2012.

[6] V. Belagiannis, F. Schubert, N. Navab, and S. Ilic. Segmentation Based Particle Filtering for Real-time 2D Object Tracking. In *European Conference on Computer Vision*, pages 842–855, 2012.

[7] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 DAVIS Challenge on Video Object Segmentation. *arXiv preprint arXiv:1803.00557*, 2018.

[8] D Comaniciu, V Ramesh, and P Meer. Kernel-Based Object Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[9] Thang Ba Dinh, Nam Vo, and G. Medioni. Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1177–1184, 2011.

[10] S. Duffner and C. Garcia. Pixeltrack: A Fast Adaptive Algorithm for Tracking Non-rigid Objects. In *IEEE International Conference on Computer Vision*, pages 2480–2487, 2013.

[11] M. Everingham, Luc Van Gool, C.K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010.

[12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[13] H Grabner, M Grabner, and H Bischof. Real-Time Tracking via On-line Boosting. In *British Machine Vision Conference*, pages 6.1–6.10, 2006.

[14] Helmut Grabner, Christian Leistner, and Horst Bischof. Semi-supervised On-Line Boosting for Robust Tracking. In *European Conference on Computer Vision*, pages 234–247, 2008.

[15] J. F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In *European Conference on Computer Vision*, pages 702–715, 2012.

[16] Xu Jia, Huchuan Lu, and M.-H. Yang. Visual Tracking via Adaptive Structural Local Sparse Appearance Model. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1822–1829, 2012.

[17] Zdenek Kalal, Jiri Matas, and Krystian Mikolajczyk. P-N Learning: Bootstrapping Binary Classifiers by Structural Constraints. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 49–56, 2010.

[18] A. Li, M. Lin, Y. Wu, M. H. Yang, and S. Yan. NUS-PRO: A New Visual Tracking Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 335–349, Feb 2016.

[19] Baiyang Liu, Junzhou Huang, Lin Yang, and Casimir Kulikowsk. Robust Tracking using Local Sparse Appearance Model and K-Selection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1313–1320, 2011.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[21] N. Mayer, E. Ilg, P. HÃd'usser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.

[22] N. Mayer, E. Ilg, P. HÃd'usser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.

[23] H. Nam and B. Han. Learning Multi-Domain Convolutional Neural Networks for Visual Tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4293–4302, 2016.

[24] P. Ochs, J. Malik, and T. Brox. Segmentation of Moving Objects by Long Term Video Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6): 1187–1200, 2014.

[25] Shaul Oron, Aharon Bar-Hillel, Dan Levi, and Shai Avidan. Locally Orderless Tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1940–1947, 2012.

[26] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.

[27] Patrick Pérez, Carine Hue, Jaco Vermaak, and Michel Gangnet. Color-Based Probabilistic Tracking. In *European Conference on Computer Vision*, pages 661–675, 2002.

[28] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke. YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7464–7473, 2017.

[29] Xiaofeng Ren and Jitendra Malik. Tracking as Repeated Figure/Ground Segmentation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

[30] D.A. Ross, J. Lim, R.S. Lin, and M.-H. Yang. Incremental Learning for Robust Visual Tracking. *International Journal of Computer Vision*, 77(1):125–141, 2008.

[31] Laura Sevilla-Lara and Erik Learned-Miller. Distribution Fields for Tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1910–1917, 2012.

[32] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual Tracking: An Experimental Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1442–1468, 2014.

[33] J. Son, I. Jung, K. Park, , and B. Han. Tracking-by-Segmentation Using Online Gradient Boosting Decision Tree. In *IEEE International Conference on Computer Vision*, pages 3056–3064, 2015.

[34] Severin Stalder, Helmut Grabner, and Luc Van Gool. Beyond Semi-Supervised Tracking: Tracking Should Be as Simple as Detection, but not Simpler than Recognition. In *IEEE International Conference on Computer Vision (Workshops)*, pages 1409–1416, 2009.

[35] David Tsai, Matthew Flagg, Atsushi Nakazawa, and James M. Rehg. Motion Coherent Tracking Using Multi-label MRF Optimization. *International Journal of Computer Vision*, 100(2):190–202, 2012.

[36] S. Wang, H. Lu, F. Yang, and M.-H. Yang. Superpixel Tracking. In *IEEE International Conference on Computer Vision*, pages 1323–1330, 2011.

[37] Y. Wu, J. Lim, and M. H. Yang. Object Tracking Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015.

[38] Yi Wu, Bin Shen, and Haibin Ling. Online Robust Image Alignment via Iterative Convex Optimization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1808–1814, 2012.

[39] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online Object Tracking: A Benchmark. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[40] Donghun Yeo, Jeany Son, Bohyung Han, and Joon Hee Han. Superpixel-based Tracking-by-Segmentation using Markov Chains. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1812–1821, 2017.

[41] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object Tracking: A Survey. *ACM Computing Surveys*, 38(4), December 2006.

[42] Kaihua Zhang, Lei Zhang, and M.-H. Yang. Real-time Compressive Tracking. In *European Conference on Computer Vision*, pages 866–879, 2012.

[43] Tianzhu Zhang, Bernard Ghanem, Si Liu, and Narendra Ahuja. Robust Visual Tracking via Multi-Task Sparse Learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2042–2049, 2012.

[44] Wei Zhong, Huchuan Lu, and M.-H. Yang. Robust Object Tracking via Sparsity-based Collaborative Model. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1838–1845, 2012.