This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

NUS-PRO: A New Visual Tracking Challenge

Annan Li, *Member, IEEE,* Min Lin, *Member, IEEE,* Yi Wu, *Member, IEEE,* Ming-Hsuan Yang, *Senior Member, IEEE,* and Shuicheng Yan *Senior Member, IEEE*

Abstract—Numerous approaches on object tracking have been proposed during the past decade with demonstrated success. However, most tracking algorithms are evaluated on limited video sequences and annotations. For thorough performance evaluation, we propose a large-scale database which contains 365 challenging image sequences of pedestrians and rigid objects. The database covers 12 kinds of objects, and most of the sequences are captured from moving cameras. Each sequence is annotated with target location and occlusion level for evaluation. A thorough experimental evaluation of 20 state-of-the-art tracking algorithms is presented with detailed analysis using different metrics. The database is publicly available and evaluation can be carried out online for fair assessments of visual tracking algorithms.

Index Terms—Object tracking, performance evaluation, benchmark database.

1 Introduction

Object tracking is one of the most important tasks in computer vision [1], and a considerable number of approaches have been proposed in the past few decades. To demonstrate the merits and effectiveness of these methods, experimental evaluations have often been carried out on a few datasets. In most cases, these datasets are constructed for specific goals, and thereby limited in terms of variability, scale and annotations. All these factors affect soundness and completeness of performance evaluations on different object tracking algorithms.

Existing image sequences commonly used in object tracking are generally collected in two ways. Image sequences are collected from surveillance cameras, such as the CAVIAR [2], TRECVID [3] and PETS [4] datasets where objects typically appear at a distance in static background. However, only humans or pedestrians are annotated in these datasets. In addition to surveillance videos, image sequences are collected from consumer cameras where typically one or a few objects appear in the scenes [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. In [16], experimental comparisons of 11 tracking algorithms on 15 sequences are presented. Nevertheless, the number of widely used sequences is limited (less than 30) and most existing algorithms use a subset of them for experimental validation. As different subset selection makes it difficult to have fair comparisons of tracking algorithms [17], [18], it is of critical importance to construct a large and challenging database for thorough performance evaluation. To address the above-mentioned issues, we collect a large-scale database containing 365 image sequences, most of which are captured with moving cameras. The proposed NUS People and Rigid Objects (NUS-PRO) database is constructed for thorough performance evaluation on single object tracking. In addition, we notice an issue regarding parameter overfitting for specific sequences commonly occurred in the tracking literature. This problem is addressed in this work by using an online evaluation system which provides experimental results by withholding the ground-truth annotations of test sequences.

The remainder of this paper is organized as follows. Section 2 gives technical details of the NUS-PRO database. In Section 3, we elaborate the evaluation metrics for the NUS-PRO database. Section 4 describes the online evaluation system. A thorough experimental evaluation of state-of-the-art object tracking algorithms on the NUS-PRO database is presented in Section 5. We conclude this paper and discuss future work in Section 6.

- *A. Li is with is with Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. This work was done when he first author was a research fellow in National University of Singapore. E-mail: lia@i2r.a-star.edu.sg*
- *M. Lin and S. Yan are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. E-mail: {linmin,eleyans}@nus.edu.sg*
- *Yi Wu is with the School of Information and Control Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China . E-mail: ywu.china@gmail.com*
- *M.-H. Yang is with the School of Engineering, University of California, Merced, CA 95344 USA. E-mail: mhyang@ucmerced.edu*

2 The NUS-PRO Database

We describe the details of the NUS-PRO database in this section including characteristics and annotation issues of the collected image sequences. As a key feature of the NUS-PRO database, the occlusion level annotation is discussed for thorough evaluation.

2.1 Database Characteristics

The NUS-PRO database consists of 365 image sequences collected from YouTube. The frame number of a sequence ranges from 146 to 5,040, and the median number

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

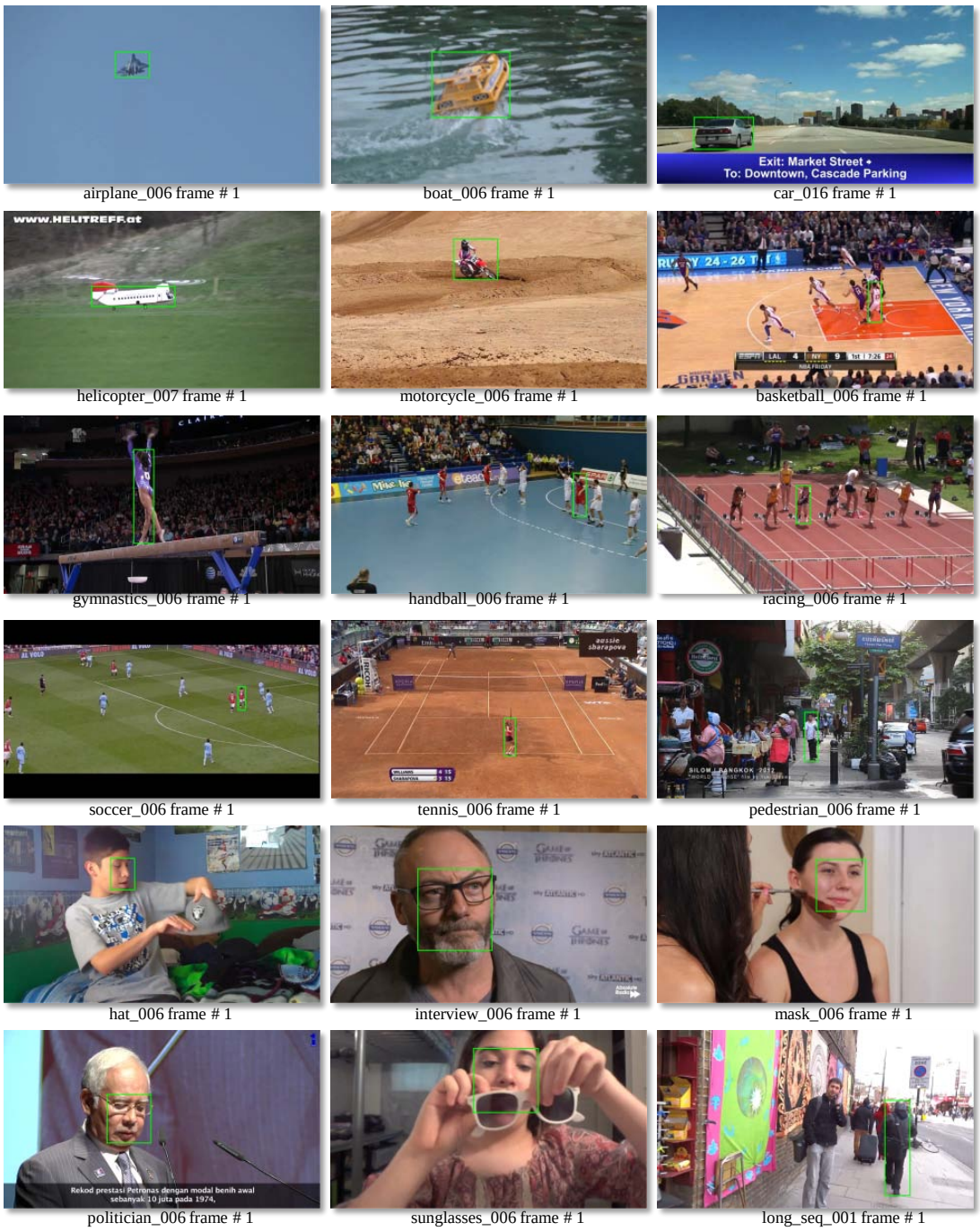


Fig. 1. Exemplar frames of the *airplane*, *boat*, *car*, *helicopter*, *motorcycle*, *basketball*, *gymnastics*, *handball*, *racing*, *soccer*, *tennis* and *pedestrian*, *hat*, *interview*, *mask*, *politician*, *sunglasses* and *long_seq* sequences of the NUS-PRO database.

of frames is 300. The image sequences of the NUS-PRO database are divided into 5 categories including *face*, *pedestrian*, *sportsman*, *rigid object* and *long sequences*. The *rigid object* category is further divided into 5 subcategories including *airplane*, *boat*, *car*, *helicopter* and *motorcycle* [1]. The *sportsman* category is divided into 6 subcategories, including *basketball*, *gymnastics*, *handball*, *racing*, *soccer* and *tennis*. The *face* category is further divided into 5 subcategories including *hat*, *mask*, *interview*, *politician* and *sunglasses*. Consequently, there are 17 kinds of objects in the NUS-PRO database, of which exemplar images and characteristics are shown in Figure 1 and Table 1.

Various factors affect the performance of a tracking algorithm. For the NUS-PRO database, we summarize and divide these factors into 12 categories, as shown in Figure 2, in which 2 categories, namely *camera shake* and *full occlusion*, are less addressed in the literature. There are

1. These objects are considered based on the holistic appearance despite some local deformation change such as rotating propellers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

Fig. 2. The NUS-PRO database is challenging due to 12 factors including shadow change, flash, dim light, clutter background, fast background change, rotation, shape deformation, scale change, partial occlusion, full occlusion, similar objects and camera shake.

mainly two differences between the NUS-PRO database and existing datasets. First, many image sequences in the NUS-PRO database are recorded by hand-held cameras, and thus contain sudden object movements caused by hand shake. Second, we annotate sequences with partial or full occlusions for further analysis in this work.

Figure 3 shows sequence statistics of challenging factors where there are more videos with *scale change*, *shape deformation*, *partial occlusion* and *clutter background*. For a specific challenging factor, the proportions of each type of object are illustrated in different colors. Typically, there are several challenging factors in one image sequence. In the NUS-PRO database, the number of challenging factors in a single image sequence ranges from 0 to 6 as shown in Figure 4.

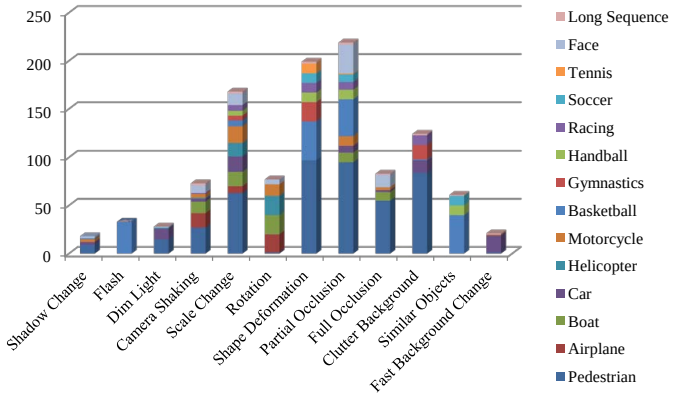


Fig. 3. Number of sequence for each factor.

In the NUS-PRO database, all images are of the same size, i.e., 1280×720 pixels. Due to diversity of shape, we compare the size of target objects by the bounding box areas. The bounding box sizes are computed from the whole 365 sequences, while the target sizes based on contour masks are measured from 305 sequences excluding the *face* class. Figure 5 shows the statistics of the target size by the square root of the target area, which roughly corresponds to the side length of a target

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1
Statistics of the NUS-PRO database.

Category		Sequence Number	Frame Number		
			Min	Max	Mean
rigid object	airplane	20	200	300	250
	boat	20	280	300	299
	car	20	233	600	383
	helicopter	20	280	360	307
	motorcycle	20	190	360	268
sportsman	basketball	40	172	360	237
	gymnastics	20	220	1960	551
	handball	10	180	503	292
	racing	10	220	460	331
	soccer	10	210	400	295
	tennis	10	223	683	436
pedestrian		100	200	460	269
face	hat	10	146	300	266
	interview	20	500	500	500
	mask	10	200	500	382.7
	politician	10	460	500	494
	sunglasses	10	167	500	364.7
long_seq		5	2133	5040	3834.6

Fig. 4. Challenging factors in an image sequence.

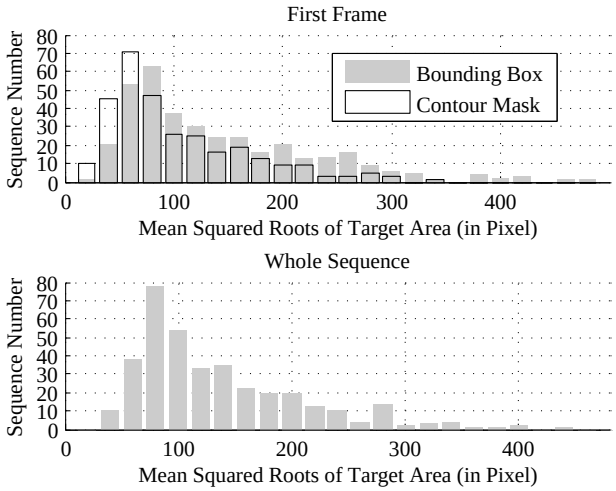


Fig. 5. Statistics of target size. The size of a target is measured by the square root of the area of its bounding box, which roughly corresponds to the side length of a target object.

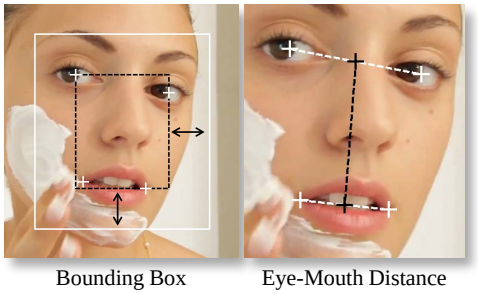


Fig. 6. A bounding box of an image, illustrated by white solid lines, is obtained by expanding the bounding box of four facial fiducial points (eye centers and mouth corners) by 40% of the eye-mouth distance (the black dashed line in the right image).

object. A few bounding boxes of objects with large shape variation are presented in Figure 7. The median target size of the first frame and the whole sequence are both around 98 pixels.

2.2 Bounding Box Annotation

While the fine foreground mask of a target object should ideally be annotated for accurate evaluation, it entails time-consuming pixel-level segmentation tasks. On the other hand, the position of a target for object tracking is usually represented by a rectangle bounding box for convenience. We note that most publicly available datasets do not clearly define the bounding box of a target object [6], [5], [7], [8], [9], [10], [13], [14], [15], which may significantly affect evaluation results of tracking algorithms.

A bounding box is defined as the boundary of a target object in the CAVIAR dataset [2], which is refereed as *boundary based bounding box* in this work. Such annotation is effective for those objects with compact shapes. However, the bounding box may contain a significant amount of background pixels if the shape of the object cannot be described compactly by a rectangle as illustrated in Figure 7.

In the NUS-PRO database, some objects (e.g., airplane, boat, car and motorcycle) have compact shapes, but many others (e.g., helicopter, sportsman and pedestrian) have more complex variations. Thus, the annotated boundary based bounding boxes are less effective. To obtain more accurate annotations, one trade-off is to enclosing most foreground parts of a target object (e.g., torso) without considering its entirety via pixel-based segmentation process. In this work, we annotate the bounding box of an object with *torso based bounding box* using the following rules:

- For a face image sequence, the bounding box is obtained by expanding an inner bounding (See Figure 6). The inner bounding box is defined by the boundary of 4 fiducial points, i.e., eye centers and mouth corners. The horizontal and vertical expansions are both 40% of the eye-mouth distance which is between the middle points of eye centers and mouth corners.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

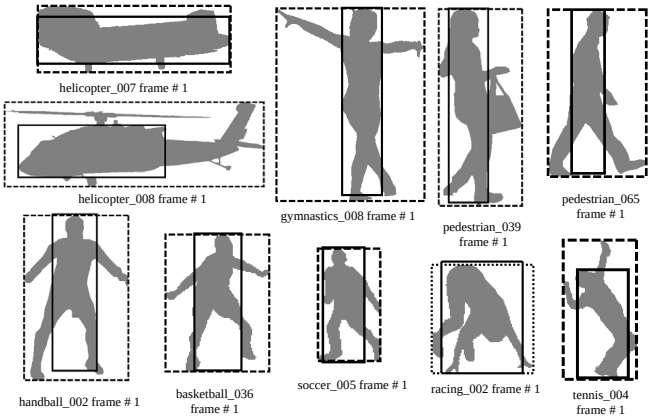


Fig. 7. Sample annotated boundary (dotted line) and torso (solid line) based bounding boxes. For a non-rigid object, the torso based annotation contains fewer background pixels, which helps increase the overlap ratio [19].

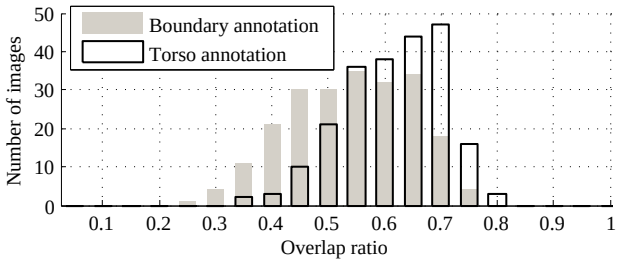


Fig. 8. Overlap ratio for the boundary based and torso based bounding box in the first frames of helicopter, sportsman and pedestrian sequences.

- For a helicopter sequence, the top and bottom extents of the bounding box are defined as the top and bottom horizontal boundaries of the main body excluding the propeller. If the helicopter has one propeller, the left and right extents of the bounding box are defined by the head and the engine vents (See Figure 7). Otherwise, the left and right extents are defined by the helicopter head and tail.
- For a sportsman or pedestrian sequence, the top extent of the bounding box is defined by the top of head, and torso refers to the body excluding the arms and calves (See Figure 7). The bottom extent of a bounding box is defined as the sole or the lowest position of the torso if it is lower than the feet. For the left and right extents, they are defined by the left-most (or right-most) one of the three positions: the shoulder boundary, the knee position, and the torso boundary.

Figure 7 shows that torso based bounding boxes significantly reduce the amount of background pixels included in these annotations. The torso based bounding boxes overlap with the ground truth contour masks better and provide higher quality annotations. For quantitative comparisons based on two types of bounding boxes, we manually label the fine foreground mask of the target in the first frames of each sequence (See Figure 7). Evaluations are carried out based on the overlap ratio, R^o, between the bounding box B and the ground truth mask G, which is used in the PASCAL VOC challenge [19]:

$$R^o = \frac{area(B \cap G)}{area(B \cup G)}. \quad (1)$$

As a trade-off between accuracy and convenience, 220 sequences (helicopter, basketball, gymnastics, racing, soccer, tennis and pedestrian) in the NUS-PRO database are annotated by torso based bounding boxes. The airplane, boat, car and motorcycle sequences are labeled by boundary based bounding boxes. Figure 8 shows the distribution of the overlap ratios computed based on the first frames of 220 sequences. With fixed overlap ratio higher than 0.5 (which is often used in the tracking literature), the number of images by the the torso annotation (white bars) is higher than that of images with boundary annotation (shaded bars). That is, the torso based bounding boxes provide more accurate annotations than the boundary based bounding boxes. The mean and standard variation of the overlap ratios obtained by torso based bounding box are 0.5679 and 0.0889 respectively, while those of the boundary based bounding box are 0.4875 and 0.1092. It should be noted that we do not annotate the rotation angle of a bounding box as the target position and extent are the most important factor for object tracking.

2.3 Occlusion Annotation

Occlusion is one of the main challenges for object tracking, and consequently how occluded objects is annotated are important for performance evaluation. For objects that are fully occluded for some frames, it is not clear whether the tracking results in that period should be included in performance evaluation. Therefore, the occlusion level in such frames should be annotated for different evaluation criteria. To the best of our knowledge, little attention has been paid to consider this factor in the tracking literature. For example, the *girl* sequence [20] contains frames with full occlusion, but this factor is not considered in the reported results. Recently, benchmarks for pedestrian detection [21] and stereo matching [22] use occlusion labels for assessment. In the NUS-PRO database, we annotate the occlusion level of each image sequence for more detailed performance evaluation.

In the NUS-PRO database, the occlusion level of each frame is annotated and classified into 3 categories: no occlusion, partial occlusion and full occlusion. In addition, it is necessary to consider the visible and invisible parts of an occluded object for evaluation. One approach is to label only the visible parts, while the other is to annotate the entire object by inferring the occluded parts. For each target in the NUS-PRO database, we annotate the full extent which includes both the visible and invisible parts.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

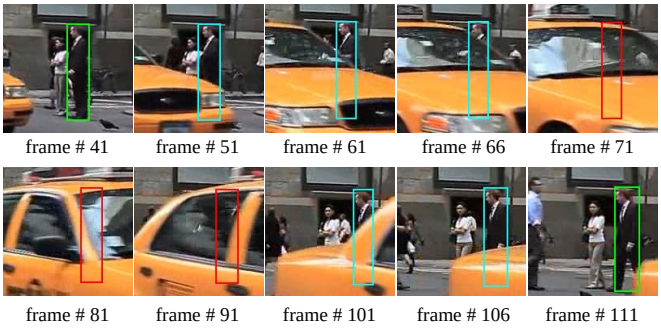


Fig. 9. Sample frames of partial occlusion (cyan rectangles), full occlusion (red rectangles) and no occlusion (green rectangles).

The invisible parts are inferred (by visual inspection) and annotated according to the movement of the objects (i.e., forward and backward prediction). Figure 9 shows some annotated examples of the sequence *pedestrian_076* in which the pedestrian is occluded by a passing taxi. Since most target objects have stable motion patterns during the transient occluded period, the annotations of their full extents are usually good approximates of the ground truth data. The annotated occlusion level provides additional information for performance evaluation.

3 Evaluation Methodology

Two evaluation metrics have been widely used for performance evaluation of object tracking algorithms. One is the center location error, in which tracking methods are assessed by the Euclidean distances between the predicted and annotated target centers. The other metric is the overlap ratio [19], which is computed based on Equation 1. An object is usually considered as being successfully tracked if the ratio is above 0.5, and tracking algorithms are evaluated by the number of successes.

The center location error metric does not consider the scale difference among image sequences, thereby making it less effective for performance evaluation. Furthermore, the center location is only effective to describe object position for certain compact and convex objects (e.g., squares and circles). It has been shown that the center location error is not a good metric for evaluating object tracking algorithms especially for contour based methods [23].

For the aforementioned two reasons, we evaluate tracking algorithms using the overlap ratio with the NUS-PRO database in this work. Table 2 shows three criteria based on the overlap ratio and level of occlusion in which only Criterion I includes fully occluded frames in evaluation. For objects that are partially occluded, there are two ways to predict the target location (bounding box), i.e., estimating the full extent and only the visible parts. Since each annotation in the NUS-PRO database is based on the whole object, the overlap ratio defined in Equation 1 is modified for performance evaluation based on the visible parts. That is, the union area is replaced by the predicted area in Criterion III for compensation, which is similar to the evaluation methodology of excluding ambiguous regions in recent benchmark on pedestrian detection algorithms [21].

TABLE 2
Three criteria for computing the overlap ratio.

Criterion	No occlusion	Partial occlusion	Full occlusion
I	$\frac{area(B \cap G)}{area(B \cup G)}$	$\frac{area(B \cap G)}{area(B \cup G)}$	$\frac{area(B \cap G)}{area(B \cup G)}$
II	$\frac{area(B \cap G)}{area(B \cup G)}$	$\frac{area(B \cap G)}{area(B \cup G)}$	-
III	$\frac{area(B \cap G)}{area(B \cup G)}$	$\frac{area(B \cap G)}{area(B)}$	-

Based on the overlap ratio of each frame, the percentage of successfully tracked frames is computed for evaluation. In addition to different criteria for computing the overlap ratio, we use different thresholds to determine whether a frame is successfully tracked or not. Consequently, the performance of an object tracking method is better analyzed with curves based on the dose-response relationship [44]. In this work, we refer such plots as the threshold-response relationship (TRR) curves.

4 Online Performance Evaluation

Performance evaluation of a tracking algorithm using the NUS-PRO database is carried out via an online system. Once the tracking results are submitted, the overlap ratios defined in Table 2 for each frame will be computed and reported. As described in Section 1, most commonly used datasets for object tracking are composed of image sequences with publicly available annotations, which may lead to parameter overfitting. Similar to the PASCAL VOC challenges [19], this problem is alleviated by withholding most of the annotations of the test data with a small portion for public use. In the NUS-PRO database, the sequence number of each category is the multiples of 10, and one tenth of the annotations that cover all kinds of objects are provided. The publicly available subsets of the NUS-PRO database are summarized as below:

- The whole database contains 365 image sequences.
- The bounding boxes and occlusion level annotations of the first frame of each sequence are provided.
- The challenging factor label of each sequence is presented.
- The foreground masks of non-face objects and the fiducial points of face images in the first frames are annotated.
- The complete bounding boxes and occlusion level annotations of 73 sequences are available for algorithm development purpose.

The NUS-PRO database and the online evaluation system can be accessed at http://www.lv-nus.org/pro/nus_pro.html. The evaluation is carried out by the complete database including the one tenth with publicly available annotations. When a zip archive with the bounding box locations (the coordinates of 4 corner

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 3
Evaluated tracking algorithms.

Method	Representation	Search Model
Color-Based Probabilistic Tracking (CPF) [24]	L, IH	PF
Locally Orderless Tracking (LOT) [25]	L,color	PF
Incremental Visual Tracking (IVT) [26]	H, PCA, GM	PF
Adaptive Structural Local Appearance model (ASLA) [27]	L, SR, GM	PF
Sparsity-based Collaborative Model (SCM) [28]	L, SR, GM+DM	PF
L1 Accelerated Proximal Gradient (L1APG) [29]	H, SR, GM	PF
Multi-Task Tracking (MTT) [30]	H, SR, GM	PF
Local Sparse appearance model with K-Selection (LSK) [31]	L, SR, GM	LOS
Online Robust Image Alignment (ORIA) [32]	H, T, GM	LOS
Distribution Fields Tracking (DFT) [33]	L, T	LOS
Kernel-based Mean-Shift (KMS) [34]	H, IH	LOS
Fragments-based tracking (Frag) [35]	L, IH	DS
On-line AdaBoost (OAB) [36]	H, Haar, DM	DS
Semi-supervised Tracking (SemiT) [37]	H, Haar, DM	DS
Semi-supervised Tracking with Adaptive Prior (BSBT) [38]	H, Haar, DM	DS
Multiple Instance Learning (MIL) [39]	H, Haar, DM	DS
Compressive Tracking (CT) [40]	H, Haar, DM	DS
Track-Learning-Detection method (TLD) [41]	L, BP, DM	DS
Circulant Structure tracking with Kernels (CSK) [42]	H, T, DM	DS
Context Tracking (CXT) [43]	H, BP, DM	DS

points) of each tracking result is submitted, the corresponding overlap ratios for all frames defined in Table 2 are returned.

5 Experiments

We present the evaluation results of state-of-the-art tracking algorithms on the NUS-PRO database with detailed analysis in this section.

5.1 Evaluated Algorithms

We evaluate 20 state-of-the-art tracking methods on the NUS-PRO database using the publicly available source codes, a recent code library [23], or executable files. Table 3 summarizes the differences of the evaluated object tracking algorithms in terms of the representation and search models. For the representation model, L and H stand for local and holistic; T, IH and BP are template, intensity histogram and binary pattern; Haar, PCA, SPCA, SR, DM and GM represent Haar-like features, principal component analysis (PCA), sparse PCA, sparse representation, discriminative model and generative model respectively. For the search model, PF, LOS and DS denote particle filter, local optimum search and dense sampling.

For fair assessments of tracking algorithms, we use the default parameters from the original implementations and fix them for all the 300 image sequences in our experiments. For the tracker with a re-detection module (e.g., TLD), no tracking results are returned if it loses track of the target object. In such scenarios, the tracking result of last tracked position is used for the frames that the tracker loses track of the target object.

5.2 Evaluation by Challenging Factors

We first present the TRR curves of the evaluated tracking algorithms based on the criteria defined in Table 2 and the area under the curve (AUC) which is also used in the PASCAL VOC challenge [19].

As a tracking algorithm can be categorized by its representation and search model, it is of great interest to analyze the effects of these modules especially for the top performing methods. To analyze whether a method is robust to each challenging factor discussed in Section 2.1, we plot the TRR curves using the corresponding sequences in Figure 10-12 and the corresponding AUCs in Figure 13. The top three tracking methods with the largest AUCs are summarized in Table 4.

The best performing 7 tracking algorithms for handling each challenging factor are as shown in Table 4 including the CPF [24], LOT [25], ASLA [27], SCM [28], KMS [34], TLD [41] and CXT [43] methods. Since there are 12 challenging factors and 3 criteria, a tracking approach can appear at most 36 times in Table 4, and the top performing 7 methods appear 24, 9, 22, 24, 9, 5 and 15 times respectively. The results also show that no single tracking algorithm outperforms all the other methods in dealing with all types of challenging scenarios. Overall, the CPF, ASLA and SCM methods, which appear 24, 22 and 24 times respectively, perform well in dealing with various challenging factors.

The CPF, ASLA and SCM methods perform well in handling image sequences with scale change, partial occlusion, full occlusion and clutter background. For challenging videos containing flash and similar objects, the KMS, LOT and CPF methods rank among the top. The correlations can be explained from two aspects, namely, the image data and similarities in 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 fac-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

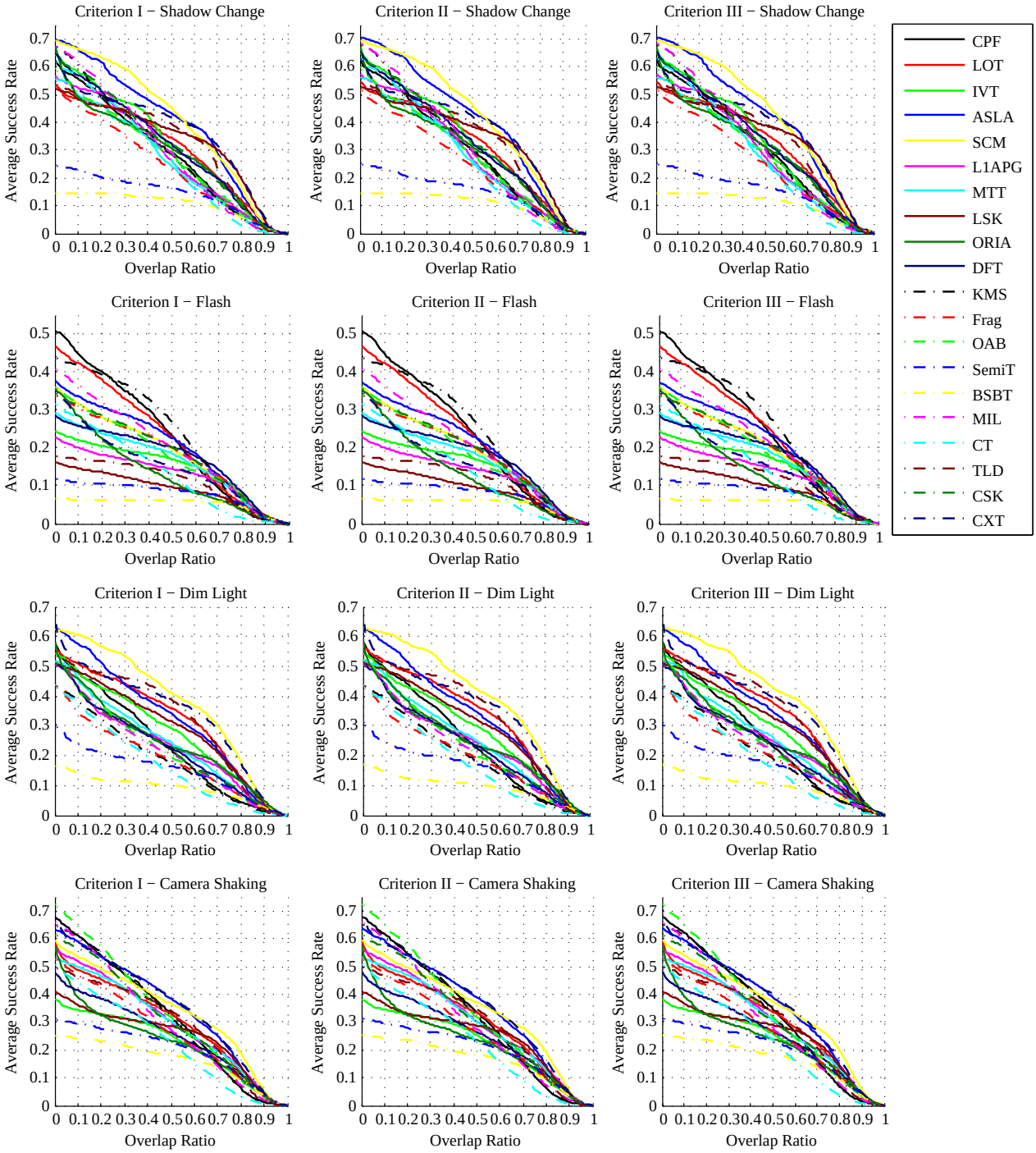


Fig. 10. TRR curves of sequences containing shadow change, flash, dim light and camera shake challenges (best viewed on high resolution displays).

tors in various categories of the NUS-PRO database. Therefore, the tracking results may be accounted by the similar algorithmic properties of the CPF, ASLA and SCM methods.

As illustrated in Table 3, the CPF, ASLA and SCM methods have two similar components, i.e., a local (L) appearance representation and a particle filter (PF) search model. Table 5 shows the top 3 performing methods in terms of the adopted representation and search models (where $\times$ indicates that there is no consistent component can be found among the tracking methods). Overall, there is no representation or search model can be consistently found in all the evaluation results except that the combination of L and PF appear frequently, which suggests tracking algorithms with this combination tend to better handle sequences with scale change, shape deformation, occlusions and cluttered background.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

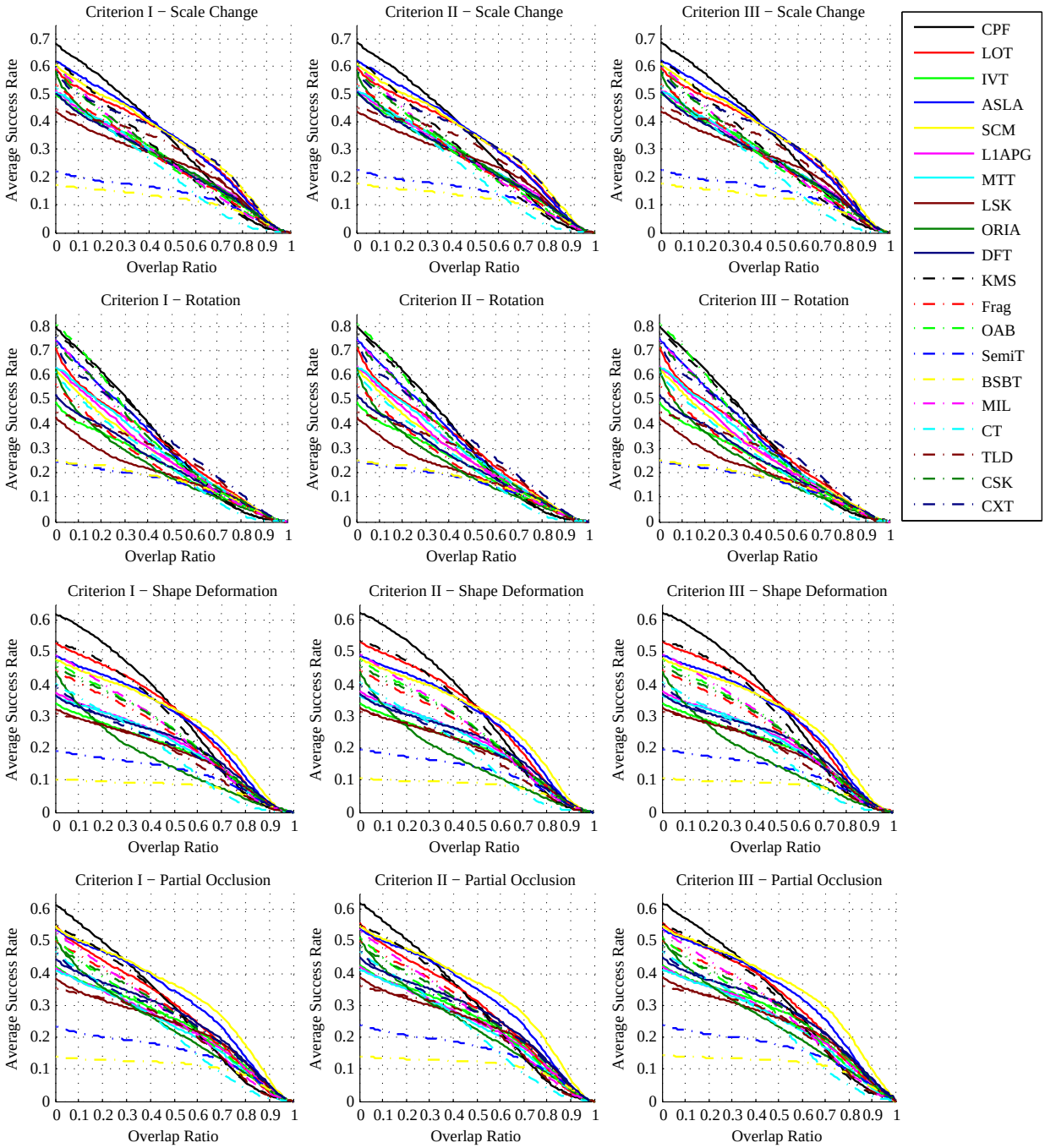


Fig. 11. TRR curves of sequences with scale change, rotation, shape deformation and partial occlusion challenges (best viewed on high resolution displays).

5.3 Evaluation by Object Categories

In this section, we present how the tracking algorithms perform for different object categories in the NUS-PRO database. Figure 14 shows the TRR curves of the 4 main object categories from the whole database. The corresponding AUC plots are presented in Figure 15 and Table 6. Overall, the SCM [28], CXT [43], CPF [24] and CXT methods perform well in the pedestrian, rigid object, sportsman and face sequences, while the ASLA algorithm achieves the best results on the entire database. The results are consistent with the findings of Section 5.2 (in which the ASLA method performs well in several categories) as the NUS-PRO database consists of a large number of sequences with challenging factors of scale change, shape deformation, partial occlusion and clutter background.

The AUCs of the TRR curves in Figure 15 show that the *sportsman* category is the most challenging among 4 main object types in the NUS-PRO database, followed by the classes of *pedestrians*, *rigid objects* and *faces*. The results can be accounted by large appearance variation due to shape deformations in the *sportsman* sequences.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

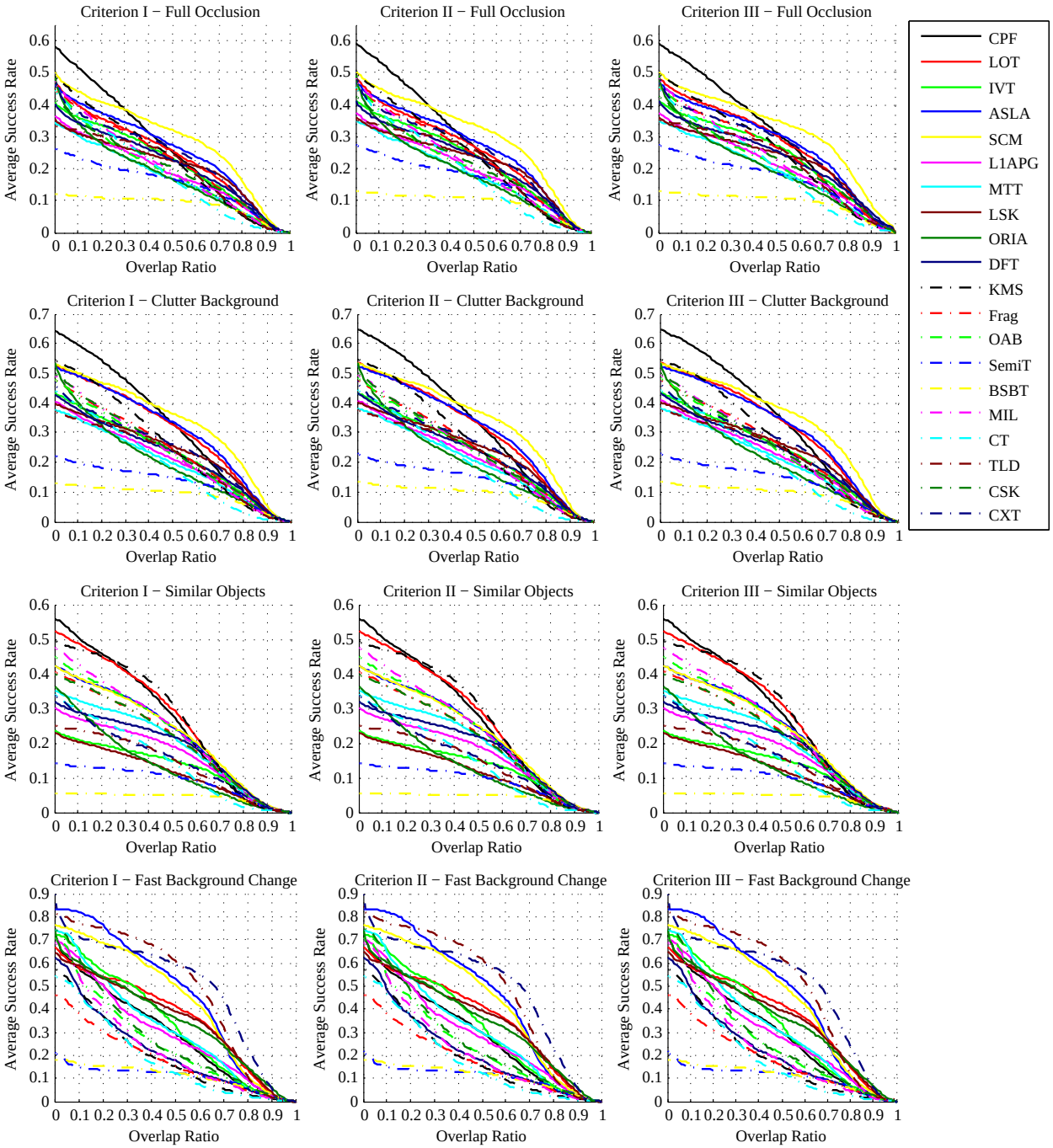


Fig. 12. TRR curves of sequences containing temporal full occlusion, clutter background, similar objects and fast background change challenges (best viewed on high resolution displays).

5.4 Analysis of Temporal Full Occlusion

We analyze the evaluation results based on the annotated occlusion level available in the NUS-PRO database. For each of the 83 sequences with temporal full occlusion, we split it into two subsequences based on the frame in which the first full occlusion occurs. Figure 16 shows the TRR curves of the subsequence before and after the first full occlusion. The performance of all the algorithms drops significantly, which indicates that existing methods are not effective in handling temporal full occlusion. Although this problem may be alleviated with off-line forward-backward motion prediction, the focus of this work is on methods without resorting to offline inference (e.g., graph matching of tracklets [45]).

We note that most of the evaluated methods are not equipped with schemes of detecting as well as handling full occlusion. However, the motion or search models of the evaluated trackers provide some temporal and spatial information, which help alleviate the short-term drift problems caused by full occlusion. The results suggest that the evaluated trackers may have drifted away before full occlusion occurs. Figure 16 (right) shows that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

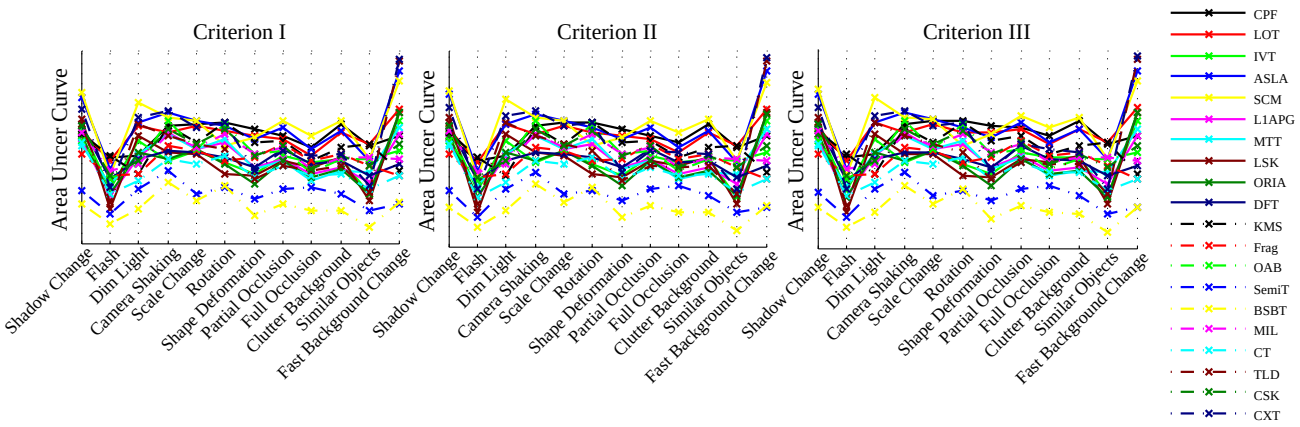


Fig. 13. Area under the curve (AUC) for different challenging factors (best viewed on high resolution displays).

TABLE 4
Top 3 methods with the largest AUCs for 12 factors.

Challenge	Criterion	Approaches
shadow change	I	SCM, ASLA, CXT
	II	SCM, ASLA, CXT
	III	SCM, ASLA, CXT
flash	I	KMS, CPF, LOT
	II	KMS, CPF, LOT
	III	KMS, CPF, LOT
dim light	I	SCM, CXT, ASLA
	II	SCM, CXT, TLD
	III	SCM, CXT, TLD
camera shake	I	CXT, ASLA, SCM
	II	CXT, ASLA, SCM
	III	CXT, ASLA, SCM
scale change	I	ASLA, SCM, CPF
	II	ASLA, SCM, CPF
	III	SCM, ASLA, CPF
rotation	I	CXT, CPF, KMS
	II	CPF, CXT, KMS
	III	CPF, CXT, KMS
shape deformation	I	CPF, LOT, SCM
	II	CPF, LOT, SCM
	III	CPF, LOT, SCM
partial occlusion	I	SCM, ASLA, CPF
	II	SCM, ASLA, CPF
	III	SCM, ASLA, CPF
full occlusion	I	SCM, CPF, ASLA
	II	SCM, CPF, ASLA
	III	SCM, CPF, ASLA
clutter background	I	SCM, CPF, ASLA
	II	SCM, CPF, ASLA
	III	SCM, CPF, ASLA
similar objects	I	KMS, LOT, CPF
	II	KMS, LOT, CPF
	III	LOT, KMS, CPF
fast background change	I	CXT, TLD, ASLA
	II	CXT, TLD, ASLA
	III	CXT, TLD, ASLA

the mean overlap ratios of the first frame where full occlusion occurs. In most scenarios, the target objects are usually partially occluded before full occlusions happen. The overlap ratios drop fast before full occlusions, while after the full occlusions they drop slowly and approach to some low values. The experimental results confirm that heavy occlusions have greater effect on tracking performance.

TABLE 5
Representation and search models of the top 3 algorithms. (where $\times$ indicates that there is no consistent component can be found among the tracking methods).

Challenge/Category	Representation	Search Model
shadow change	$\times$	$\times$
flash	$\times$	$\times$
dim light	$\times$	$\times$
camera shake	$\times$	$\times$
scale change	L	PF
rotation	$\times$	$\times$
shape deformation	L	PF
partial occlusion	L	PF
full occlusion	L	PF
clutter background	L	PF
similar objects	$\times$	$\times$
fast background change	$\times$	$\times$
pedestrian	L	PF
sportsman	$\times$	$\times$
rigid objects	$\times$	$\times$
all	L	PF

TABLE 6
Top 3 approaches with the largest AUCs obtained on the 4 main object categories and the whole database.

Category	Criterion	Approaches
all	I	ASLA, SCM, LOT
	II	ASLA, SCM, LOT
	III	ASLA, SCM, LOT
pedestrian	I	SCM, CPF, ASLA
	II	SCM, CPF, ASLA
	III	SCM, CPF, ASLA
rigid object	I	CXT, ASLA, CPF
	II	CXT, ASLA, CPF
	III	CXT, ASLA, CPF
sportsman	I	CPF, LOT, KMS
	II	CPF, LOT, KMS
	III	CPF, LOT, KMS
face	I	CXT, ORIA, SCM
	II	CXT, ORIA, SCM
	III	CXT, ORIA, ORIA

These results also explain why the TRR curves based on three criteria defined in Table 2 are similar (See Figure 10, 11, 12 and 14) as most trackers drift away quickly when partial occlusions occur. The performance differences in the following frames are not significant

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

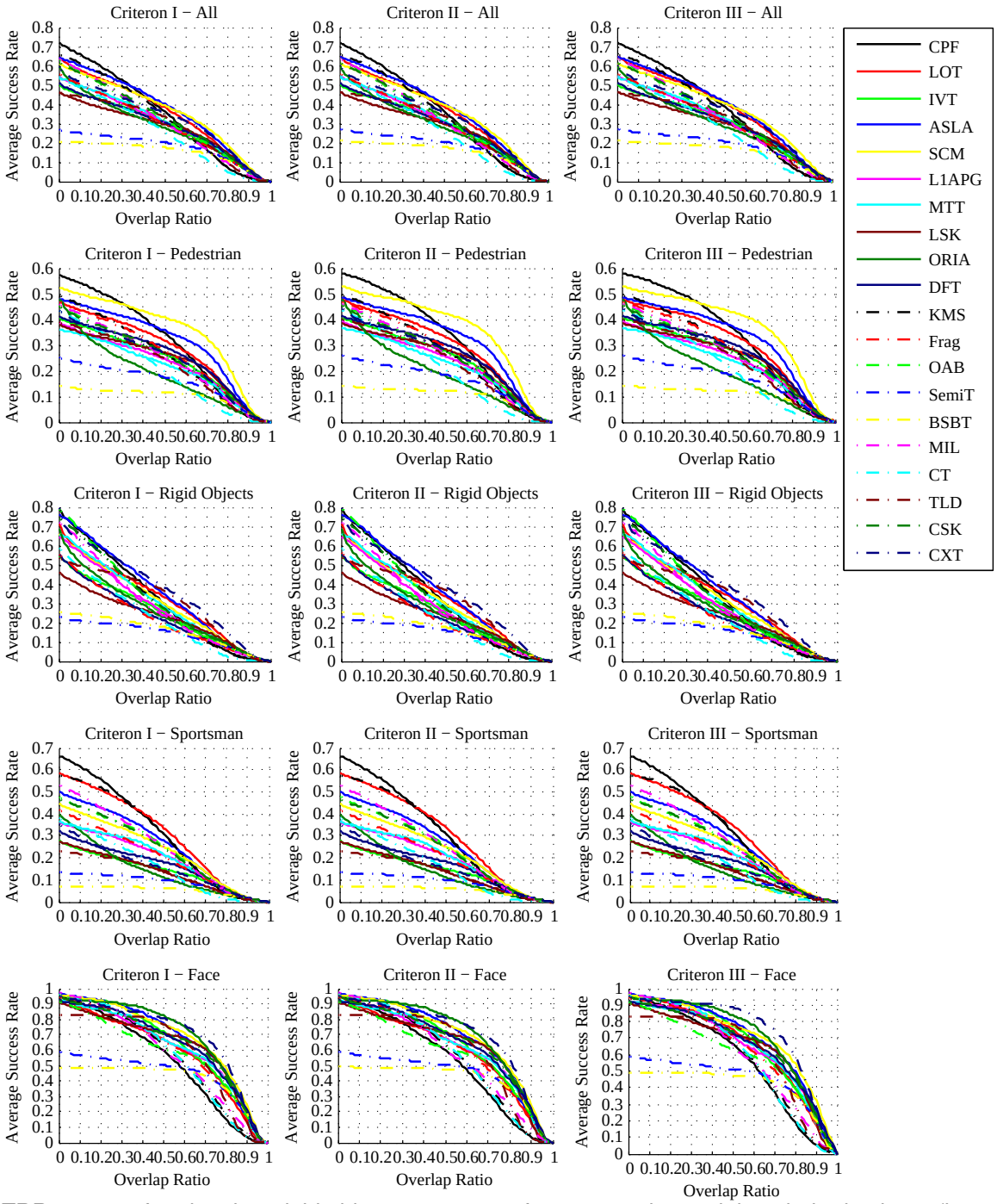


Fig. 14. TRR curves of pedestrian, rigid object, sportsman, face categories and the whole database (best viewed on high resolution displays).

as the overlap ratios are low. Consequently, the performance of each evaluated tracking method is mainly determined by the period before occlusions occur. As the three criteria are the same for the case without occlusion (first column of Table 2), the corresponding TRR curves are similar.

5.5 Analysis of Long-Term Tracking

In addition to performance assessment discussed above, long-term tracking is another important evaluation criterion. For certain vision tasks such as surveillance and video analysis, performance evaluation of long sequences is important as objects are likely to appear in the scene for a long period of time. Toward this end, 5 long sequences whose mean length is 3,835 frames are included in the proposed dataset. The TRR curves of the 20 trackers on these sequences are shown in Figure 17 (a)-(c), and the corresponding AUCs are illustrated in Figure 17 (d). Overall, the AUCs of the ASLA, SCM and OAB methods are larger than those of the other trackers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

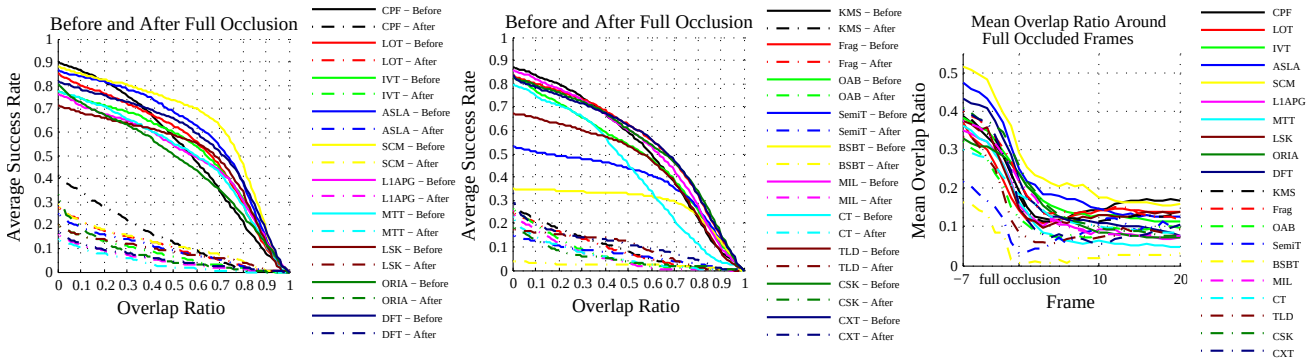


Fig. 16. Effects of full occlusions. The left and middle plots show the TRR curves computed from the subsequences before and after full occlusions occur, while the right plot illustrates the mean overlap ratios around the first time when full occlusions occur.

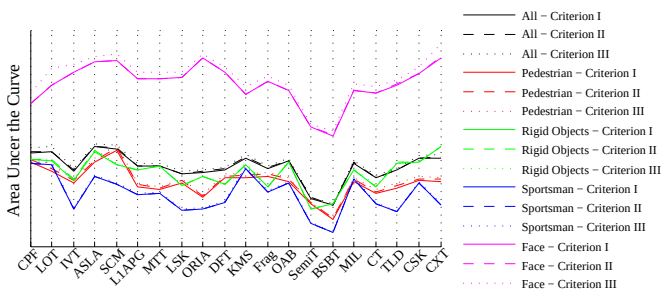


Fig. 15. AUC for the TRR curves calculated on pedestrian, sportsman, rigid objects sequences and the whole database.

In addition to TRR and AUC, another mesure for long-term performance is the duration for an object being tracked successfully, i.e., the frame number from the beginning to where a tracker fails. An object is considered to be successfully tracked in one frame if the overlap ratio is above 0.5. To avoid bias caused by sudden change, we use the average overlap ratio of a short duration instead of the overlap ratio of a single frame (i.e., 21 frames including this frame and the 10 frames before and after it in this work). Using the criteria defined Table 2, three overlap ratios for successful tracking are computed and the longest one is used for evaluation.

Figure 17 (e) shows the mean and median length of successful tracking on the 5 long sequences. The ASLA, SCM and IVT methods perform well with mean length of 1,147.2, 1,109.0, and 829.8 frames,and median length of 414.0, 413.0, and 308.0, respectively. We also present the length of successful tracking in the 360 short sequences in Figure 17 (f). The ASLA, SCM and CXT algorithms perform well with mean length of 122.5, 113.9 and 112.1 frames respectively. In addition, the ASLA, SCM and MIL methods outperform the other approaches in terms of median length (61.5, 61.0, and 53.0 frames, respectively).

6 Conclusions

In this paper, we present a large-scale video database and evaluation metrics for object tracking. In the NUS-PRO database, we annotate each sequence with object location, occlusion level and challenging factors. We propose three criteria for detailed performance evaluation, and carry out experiments using 20 state-of-the-art tracking algorithms on the NUS-PRO database. Extensive experimental evaluation for each challenging factor, full occlusions and object size are presented with detailed analyses. While the NUS-PRO database is developed for performance evaluation of object tracking, it can be used for other tasks such as optic flow, object detection, and object classification, as it contains image sequences of various object categories with ground-truth annotations.

Acknowledgments

This work was done when he first author was a research fellow in National University of Singapore. The authors would like to thank Quanhong Fu for her help in English writing. This work is partially supported by National Natural Science Foundation of China (No. 61328205). M.-H. Yang is supported in part by the National Science Foundation CAREER Grant 1149783 and IIS Grant 1152576.

References

[1] A. Yilmaz, O. Javed, and M. Shah, "Object Tracking: A Survey," *ACM Computing Surveys*, vol. 38, Dec. 2006.
[2] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.
[3] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation Campaigns and TRECVid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pp. 321–330, 2006.
[4] http://www.cvg.rdg.ac.uk/PETS2009/a.html.
[5] http://www.ces.clemson.edu/~stb/research/headtracker/seq/.
[6] http://www.cs.technion.ac.il/~amita/fragtrack/fragtrack.htm.
[7] http://www.cs.toronto.edu/vis/projects/adaptiveAppearance.html.
[8] http://www.cs.toronto.edu/~dross/ivt/.
[9] http://www.dabi.temple.edu/~hbling/code_data.htm.
[10] http://vision.ucsd.edu/~bbabenko/projectmiltrack.shtml.
[11] http://www.ymer.org/amir/software/milforests/.
[12] http://www.cise.ufl.edu/~smshahed/tracking.htm.
[13] http://cv.snu.ac.kr/research/~vtd/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2417577, IEEE Transactions on Pattern Analysis and Machine Intelligence

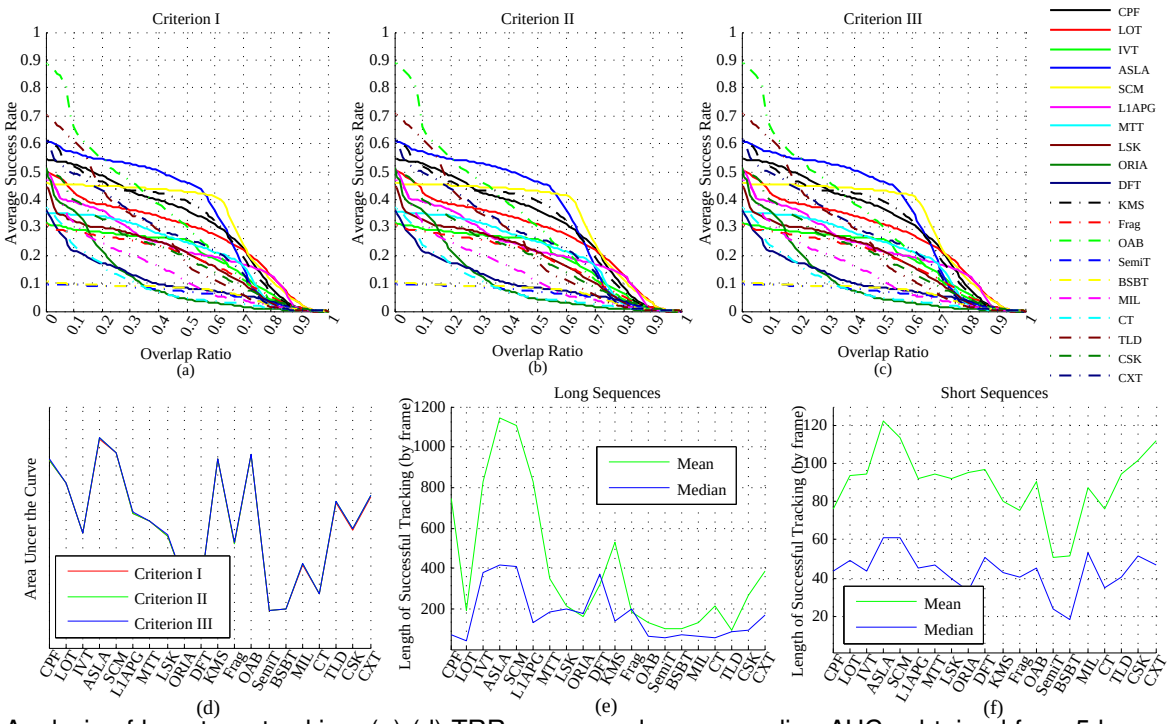


Fig. 17. Analysis of long-term tracking. (a)-(d) TRR curves and corresponding AUCs obtained from 5 long sequence. (e) The length of successful tracking on long sequences. (f) The length of successful tracking on short sequences.

[14] http://gpu4vision.icg.tugraz.at/index.php?content=subsites/prost/prost.php.

[15] http://www4.comp.polyu.edu.hk/~cslzhang/CT/CT.htm.

[16] Q. Wang, F. Chen, W. Xu, and M.-H. Yang, "An Experimental Comparison of Online Object-Tracking Algorithms," in *Society of Photo-Optical Instrumentation Engineers (SPIE)*, vol. 8138, p. 81381A, 2011.

[17] Y. Pang and H. Ling, "Finding the Best from the Second Bests-Inhibiting Subjective Bias in Evaluation of Visual Tracking Algorithms," in *IEEE International Conference on Computer Vision*, pp. 2784–2791, 2013.

[18] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Cehovin, G. Nebehay, G. Fernandez, T. Vojir, A. Gatt, *et al.*, "The Visual Object Tracking VOT2013 Challenge Results," in *IEEE International Conference on Computer Vision Workshops*, pp. 98–111, 2013.

[19] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[20] S. Wang, H. Lu, F. Yang, and M.-H. Yang, "Superpixel Tracking," in *IEEE International Conference on Computer Vision*, pp. 1323–1330, 2011.

[21] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[22] 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*, pp. 3354–3361, 2012.

[23] Y. Wu, J. Lim, and M.-H. Yang, "Online Object Tracking: A Benchmark," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[24] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-Based Probabilistic Tracking," in *European Conference on Computer Vision*, pp. 661–675, 2002.

[25] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan, "Locally Orderless Tracking," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1940–1947, 2012.

[26] D. Ross, J. Lim, R. Lin, and M.-H. Yang, "Incremental Learning for Robust Visual Tracking," *International Journal of Computer Vision*, vol. 77, no. 1, pp. 125–141, 2008.

[27] X. Jia, H. 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*, pp. 1822–1829, 2012.

[28] W. Zhong, H. Lu, and M.-H. Yang, "Robust Object Tracking via Sparsity-based Collaborative Model," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1838–1845, 2012.

[29] C. Bao, Y. Wu, H. Ling, and H. Ji, "Real Time Robust L1 Tracker Using Accelerated Proximal Gradient Approach," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1830–1837, 2012.

[30] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust Visual Tracking via Multi-Task Sparse Learning," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2042–2049, 2012.

[31] B. Liu, J. Huang, L. Yang, and C. Kulikowsk, "Robust Tracking using Local Sparse Appearance Model and K-Selection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1313–1320, 2011.

[32] Y. Wu, B. Shen, and H. Ling, "Online Robust Image Alignment via Iterative Convex Optimization," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1808–1814, 2012.

[33] L. Sevilla-Lara and E. Learned-Miller, "Distribution Fields for Tracking," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1910–1917, 2012.

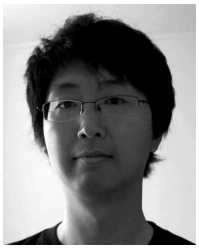
[34] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-Based Object Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003.

[35] 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*, pp. 798–805, 2006.

[36] H. Grabner, M. Grabner, and H. Bischof, "Real-Time Tracking via On-line Boosting," in *British Machine Vision Conference*, pp. 6.1–6.10, 2006.

[37] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised On-

Line Boosting for Robust Tracking," in *European Conference on Computer Vision*, pp. 234–247, 2008.
[38] S. Stalder, H. Grabner, and L. 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)*, pp. 1409–1416, 2009.
[39] B. Babenko, M.-H. Yang, and S. Belongie, "Robust Object Tracking with Online Multiple Instance Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619–1632, 2011.
[40] K. Zhang, L. Zhang, and M.-H. Yang, "Real-time Compressive Tracking," in *European Conference on Computer Vision*, pp. 866–879, 2012.
[41] Z. Kalal, J. Matas, and K. Mikolajczyk, "P-N Learning: Bootstrapping Binary Classifiers by Structural Constraints," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 49–56, 2010.
[42] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the Circulant Structure of Tracking-by-Detection with Kernels," in *European Conference on Computer Vision*, pp. 702–715, 2012.
[43] T. B. Dinh, N. Vo, and G. Medioni, "Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1177–1184, 2011.
[44] http://en.wikipedia.org/wiki/Dose-response_relationship.
[45] A. Andriyenko and K. Schindler, "Multi-target Tracking by Continuous Energy Minimization," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1265–1272, 2011.

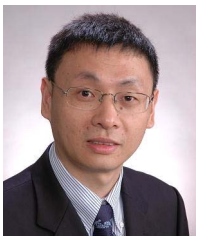


Annan Li received the B.S. and M.S. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2003 and 2006, and the Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011. From 2011 to 2013, he was a postdoctoral research fellow at National University of Singapore. He is currently a research scientist in Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. His research interests include computer vision, pattern recognition, and statistical learning. He is a member of IEEE.

Min Lin is currently a Ph.D. Candidate at the NUS Graduate School for Integrative Sciences and Engineering at National University of Singapore. He's currently a member of Learning and Vision lab (http://www.lv-nus.org). Mr. Lin's main research is to apply deep neural networks on computer vision tasks such as object recognition.

Yi Wu received the B.S. degree in automation from Wuhan University of Technology, Wuhan, China, in 2004 and the Ph.D. degree in pattern recognition and intelligent systems from Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2009. Since 2009, he has been a Lecturer with the School of Information and Control Engineering, Nanjing University of Information Science and Technology, Nanjing, China. From May 2010 to June 2012, he was a Post-Doctoral Fellow with Temple University, Philadelphia, PA, USA. From July 2012 to April 2014, he was a Post-Doctoral Fellow with the University of California, Merced, CA, USA. His research interests include computer vision, medical image analysis, multimedia analysis, and machine learning.

Ming-Hsuan Yang is an associate professor in Electrical Engineering and Computer Science at University of California, Merced. He received the PhD degree in computer science from the University of Illinois at Urbana-Champaign in 2000. Prior to joining UC Merced in 2008, he was a senior research scientist at the Honda Research Institute working on vision problems related to humanoid robots. He coauthored the book Face Detection and Gesture Recognition for Human-Computer Interaction (Kluwer Academic 2001) and edited special issue on face recognition for Computer Vision and Image Understanding in 2003, and a special issue on real world face recognition for IEEE Transactions on Pattern Analysis and Machine Intelligence. Yang served as an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence from 2007 to 2011, and is an associate editor of the International Journal of Computer Vision, Image and Vision Computing and Journal of Artificial Intelligence Research. He received the NSF CAREER award in 2012, the Senate Award for Distinguished Early Career Research at UC Merced in 2011, and the Google Faculty Award in 2009. He is a senior member of the IEEE and the ACM.



Shuicheng Yan is currently an Associate Professor at the Department of Electrical and Computer Engineering at National University of Singapore, and the founding lead of the Learning and Vision Research Group (http://www.lv-nus.org). Dr. Yan's research areas include machine learning, computer vision and multimedia, and he has authored/co-authored hundreds of technical papers over a wide range of research topics, with Google Scholar citation $>$13,000 times and H-index 50. He has been serving as an associate editor of IEEE TKDE, TCSVT and ACM Transactions on Intelligent Systems and Technology (ACM TIST). He received the Best Paper Awards from ACM MM'13 (Best Paper and Best Student Paper), ACM MM12 (Best Demo), PCM'11, ACM MM10, ICME10 and ICIMCS'09, the runner-up prize of ILSVRC'13, the winner prizes of the classification task in PASCAL VOC 2010-2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honorable mention prize of the detection task in PASCAL VOC'10, 2010 TCSVT Best Associate Editor (BAE) Award, 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, and 2012 NUS Young Researcher Award.