A Protocol for Evaluating Video Trackers Under Real-World Conditions

Tahir Nawaz and Andrea Cavallaro

***Abstract*—The absence of a commonly adopted performance evaluation framework is hampering advances in the design of effective video trackers. In this paper, we present a single-score evaluation measure and a protocol to objectively compare trackers. The proposed measure evaluates tracking *accuracy* and *failure*, and combines them for both summative and formative performance assessment. The proposed protocol is composed of a set of trials that evaluate the *robustness* of trackers on a range of test scenarios representing several real-world conditions. The protocol is validated on a set of sequences with a diversity of targets (head, vehicle, person) and challenges (occlusions, background clutter, pose changes, scale changes) using six state-of-the-art trackers, highlighting their strengths and weaknesses on more than 187000 frames. The software implementing the protocol and the evaluation results are made available online and new results can be included, thus facilitating the comparison of trackers.**

***Index Terms*—Performance evaluation, video trackers, evaluation measure, protocol, trials.**

I. Introduction

UNLIKE OTHER areas of image processing and computer vision such as disparity estimation [1], optical flow computation [2] and video coding [3] that consistently use commonly accepted evaluation procedures, video tracking still lacks a standard way to evaluate and compare algorithmic performance.

Although a number of efforts have been made toward performance evaluation of trackers in the form of evaluation campaigns (ETISEO, CLEAR, PETS, i-LIDS, CAVIAR) and small-scale evaluation frameworks ([4], [5], [6], [7], [8]), the performance of trackers is still tested using different evaluation criteria and varying datasets, thus hindering an effective evaluation and comparison. Moreover, because of the complexity of the evaluation task, many performance criteria contain multiple measures [4], [6], [7], which are difficult to combine in order to rank various algorithms. A single-score evaluation criterion that can comprehensively encapsulate the overall tracking performance would be desirable to simplify the performance comparison task.

Performance evaluation may involve the computation of the discrepancy between the estimated and the ground-truth position and size of the target [4], [6], [9], [10]. The discrepancy is computed based on a distance-based criterion [9], [10], [11] or an overlap-based criterion [4], [7], [12]. *Distance-based criteria* use the concept of distance minimization between estimation and ground truth to evaluate performance. Naive distance-based evaluation may not include target size variations in the evaluation procedure [8] and may not effectively reflect instances of tracking failure [13] that refers to the case of no-overlap between estimated and ground-truth states. *Overlap-based criteria* compute the amount of overlap between estimation and ground truth. Overlap-based evaluation mostly takes into account target size variations (with some exceptions [4], [5], [6]) and can therefore detect instances of tracking failure. However, existing overlap-based criteria [5], [7], [8], [14] use hard thresholds or fixed parameters that restrict their use to application-specific tracking performance assessment.

In this paper, we propose a threshold-independent overlap-based criterion that summarizes tracking performance based on a new evaluation measure, which takes into account target size variations. The proposed measure quantifies how *accurately* and how *long* a target is tracked across a sequence. Moreover, we propose a protocol with a comprehensive set of trials that evaluate trackers on a wide range of test scenarios representing several real-world operational conditions. The trials quantify the *robustness* of a tracker to noisy inputs, processing and communication delays, video compression and varying scene conditions such as illumination changes. To the best of our knowledge, this is the first initiative that enables evaluating the robustness of the performance of tracking algorithms under such a wide variety of real-world conditions. The resulting performance evaluation tool is made available online as an open source software[1].

The organization of the paper is as follows. The proposed evaluation criterion and the protocol are discussed in Sec. II and Sec. III, respectively. This is followed by the experimental validation in Sec. IV. Section V concludes the paper.

II. Combined Tracking Performance Score

Let an estimated trajectory R be represented as:

$$R = \{(x_k, y_k, A_k)\}_{k=1}^{K^R}, \tag{1}$$

where (x_k, y_k) is the estimated target position (e.g. its centroid), A_k is the information about the estimated target area in the kth frame and K^R is the total number of frames for which the tracker generated an output. Let the corresponding ground-truth trajectory be:

$$G = \{(\hat{x}_k, \hat{y}_k, \hat{A}_k)\}_{k=1}^{K^G}, \tag{2}$$

Manuscript received February 29, 2012; revised June 14, 2012; accepted October 14, 2012. This work was supported in part by the EU under the FP7 project APIDIS (ICT-216023). Tahir Nawaz was supported by the Erasmus Mundus Joint Doctorate in Interactive and Cognitive Environments, which is funded by the Education, Audiovisual & Culture Executive Agency (FPA n° 2010-0012). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jenq-Neng Hwang.

T. Nawaz & A. Cavallaro are with the Centre for Intelligent Sensing, Queen Mary University of London, London E1 4NS, UK (e-mail: {tahir.nawaz,andrea.cavallaro}@eecs.qmul.ac.uk).

[1] http://www.eecs.qmul.ac.uk/~andrea/pft2

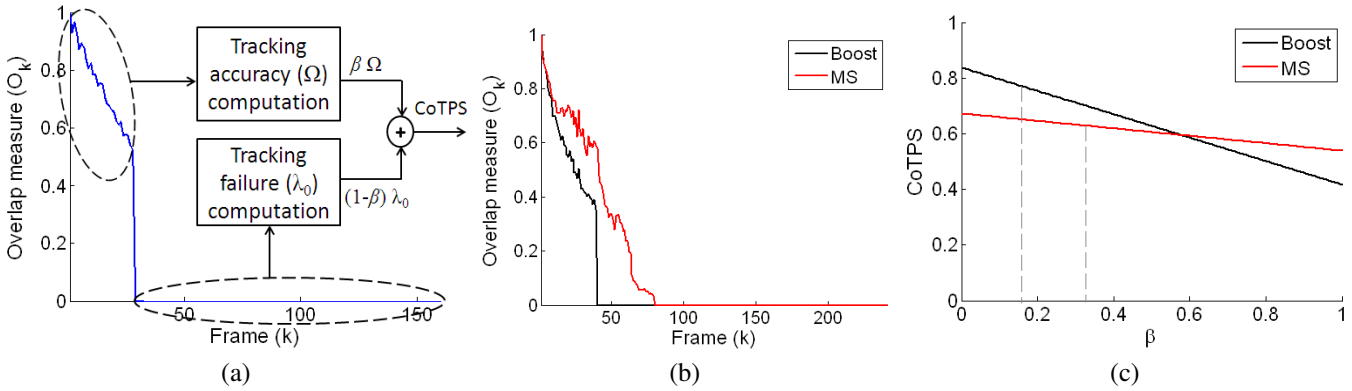


Fig. 1. (a) Schematic diagram of the proposed evaluation measure formulated by combining contributions that quantify tracking accuracy and tracking failure. (b) The result of Mean-Shift (MS) [16] (in red) and of the online boosting-based tracker (Boost) [17] (in black) on the AVSS 2007 sequence containing target H_4 (described in Sec. III). The $CoTPS$ values are 0.628 for MS and 0.770 for Boost. β for MS and Boost is computed using Eq. 8. (c) Comparison between values of β computed adaptively. $CoTPS$ plotted for a range of β values for the tracking results of Boost and MS from the example in (b). The interpretation of the performance can significantly change depending on the value of β. A preset value could lead to an incorrect evaluation.

where $(\hat{x}_k, \hat{y}_k)$ is the ground-truth target position, $\hat{A}_k$ is the ground-truth target area in the kth frame and K^G is the total number of frames in which the target exists. A_k and $\hat{A}_k$ may be in the form of a bounding box, a bounding ellipse or a bounding contour. Without loss of generality, let A_k and $\hat{A}_k$ be bounding boxes that define the width and the height of the target.

We firstly compute the amount of overlap, O_k, across R as follows [15]:

$$O_k = \frac{|TP_k|}{|TP_k| + |FP_k| + |FN_k|}, \tag{3}$$

where $O_k \in [0,1]$ and $|\cdot|$ represents the cardinality of a set. TP_k, FP_k and FN_k are the sets of true positive (correctly estimated), false positive (incorrectly estimated) and false negative (missed) pixels of a target at frame k. Note that $O_k = 0$ if the tracker does not produce a bounding box when the target is present or if a bounding box is produced when no target is present.

The *tracking accuracy* quantifies the extent to which the estimated trajectory overlaps the ground-truth trajectory, considering only frames with $O_k \neq 0$ (Fig. 1(a)) and is computed as [15]:

$$\hat{\lambda} = \frac{\hat{N}_l}{\hat{N}}, \tag{4}$$

where $\hat{N}_l = |\hat{F}_l|$ and $\hat{F}_l = \{f_k : O_k \in (0, \hat{\tau}),\ \hat{\tau} \in (0,1],\ \forall k\}$; and $\hat{N} = |\hat{F}|$, with $\hat{F} = \{f_k : O_k \neq 0,\ \forall k\}$, is the number of frames with $O_k \neq 0$.

Computing $\hat{\lambda}$ for a fixed value of $\hat{\tau}$ necessitates an application-dependent decision, since different values of $\hat{\tau}$ may be appropriate for different tracking tasks. To overcome this limitation, instead of computing $\hat{\lambda}$ for a fixed value of $\hat{\tau}$, we accumulate its value over the full range of $\hat{\tau}$ values. In particular, we use an increment of $\Delta\hat{\tau} = 0.01$ to obtain $\hat{\lambda}(\hat{\tau})$ and therefore, the score that quantifies tracking accuracy across the sequence, Ω, is computed as

$$\Omega = \Delta\hat{\tau} \sum_{\hat{\tau} \in (0,1]} \hat{\lambda}(\hat{\tau}), \tag{5}$$

where $\Omega \in [0,1]$. The smaller Ω, the higher the tracking accuracy. Ω can be regarded as an approximation of the area under the curve of $\hat{\lambda}(\hat{\tau})$.

Tracking failures correspond to instances of target loss. The tracking failure score, λ_0 ($\lambda_0 \in [0,1]$), is defined as

$$\lambda_0 = \frac{N^0}{N}, \tag{6}$$

where $N^0 = |F^0|$, with $F^0 = \{f_k : O_k = 0,\ \forall k\}$, and $N = |F|$, with $F = \{f_k : \forall k\}$. The smaller λ_0, the smaller the tracking failure score.

We combine the information on tracking accuracy and tracking failure in a single score to facilitate performance ranking. The proposed *Combined Tracking Performance Score*, $CoTPS$ ($CoTPS \in [0,1]$), is computed as follows:

$$CoTPS = \beta\Omega + (1-\beta)\lambda_0, \tag{7}$$

where β is a *penalty*, with $\beta \in [0,1]$. The smaller $CoTPS$, the better the tracking performance. Figure 1(b) plots O_k for two tracking results whose comparison is shown using $CoTPS$.

Note that a preset value of β may lead to incorrect performance evaluation (see Fig. 1(c)). β is computed adaptively:

$$\beta = \frac{\hat{N}}{N}, \tag{8}$$

where $\hat{N}$ is the number of frames in which the tracker has partially or completely tracked the target ($O_k > 0$), thus restricting the inclusion of any extra influence of Ω in the computation of $CoTPS$. Similarly, $(1-\beta)$ applied to λ_0 is proportional to $(N - \hat{N})$, i.e. the number of frames in which the tracker has failed ($O_k = 0$), which are also the same frames used in the estimation of λ_0, thus restricting the inclusion of any extra influence of λ_0 in the computation of $CoTPS$.

Let us consider the result of the Mean-Shift tracker (MS) [16] in Fig. 1(b). In this example, a penalty of $\beta = 0.328$ (computed using Eq. (8)) is applied to Ω since the tracker is successful ($O_k > 0$) in 32.8% frames ($\hat{N} = 79$ and $N = 241$). Similarly, a penalty of $(1-\beta) = 0.672$ is applied to λ_0 since the tracker has failed ($O_k = 0$) in 67.2% frames ($N - \hat{N} = 162$

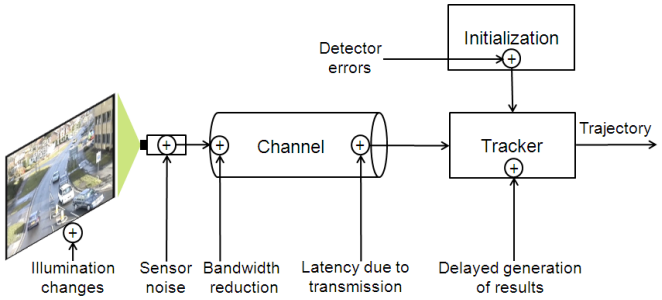


Fig. 2. Conceptual illustration of various distortions that may affect the performance of a tracker in a real-world application. These distortions include initialization errors caused by detector, sensor noise, latency due to transmission over the channel or due to the delayed generation of results by the tracker, illumination changes in the scene, and video compression.

and $N = 241$). The adaptive computation of β allows us to include accurate contributions of Ω and λ_0 in the estimation of $CoTPS$.

III. Evaluation Protocol

Trackers operate under various challenges in real-world applications and therefore these challenges should be considered when evaluating and comparing performance. Tracking challenges include initialization errors caused by a detector; sensor noise; latency due to the transmission of video data over a channel or due to the delayed generation of results by the tracker; changing illumination in the scene; and compression of the video data (Fig. 2). The proposed evaluation protocol enables evaluation under these challenges to quantify the robustness of trackers.

Given a set of M trackers[2] $T = \{T_j\}_{j=1}^{M}$, we aim to evaluate tracker T_j on a set of trials $P = \{P_i\}_{i=1}^{Z}$, where each trial simulates a specific real-world operational condition.

Trials 1, 2, 3 (P_1, P_2, P_3) evaluate the robustness of trackers to *initialization errors* possibly introduced by a detector. These errors are simulated by perturbing the position of the initializing bounding box in *Trial 1* (P_1), the size (width and height) of the bounding box in *Trial 2* (P_2), and both the position and size in *Trial 3* (P_3). The amount of perturbation is added while ensuring at least an overlap of $\hat{O}\%$ between the bounding boxes of the original (ground-truth) initialization and the perturbed initialization. The number of perturbed initializations generated on P_1, P_2 and P_3 are n_1, n_2 and n_3, respectively.

Trial 4 (P_4) evaluates robustness to *noisy video data* generated by low-cost sensors. On P_4, a set of n_4 test sequences are generated by adding to the original sequence $\hat{l}\times$ (the estimated variance of) the zero-mean Gaussian noise of a low-quality webcamera (Creative webcam VF0330). The estimated standard deviations of its red, green and blue channels are $\sigma_1 = 8.59$, $\sigma_2 = 8.40$ and $\sigma_3 = 11.96$, respectively.

Trial 5 (P_5) evaluates robustness to *latency* due to transmission over a channel or due to the delayed generation of the results by the tracker. On P_5, the protocol generates a set of n_5 test sequences by periodically dropping $m-1$ frames from the original sequence.

Trial 6 (P_6) evaluates robustness to changing illumination in the scene. On P_6, a set of n_6 test sequences are generated by synthetically increasing ($+\Delta L$) or decreasing ($-\Delta L$) illumination over time (in the original sequence) with saturation by adding (subtracting) $\Delta L = 0, 1, \ldots, L$ to (from) the pixel values of frames $k = 1, 2, \ldots, K$, respectively. If the number of frames in the sequence is $K > (L+1)$, a value of $\Delta L = L$ is maintained for the remaining frames.

Trials 7, 8 (P_7, P_8) evaluates robustness to bandwidth reduction of the video data. On P_7, test sequences are generated by gradually increasing the compression ratio of the original sequence. We chose Motion JPEG compression because of its suitability for video tracking applications. In Motion JPEG, the extent of compression ratio depends on a quality parameter ζ. The higher ζ, the better the visual quality and the lower the compression ratio, where $\zeta \in [0, 100]$. To ensure evaluation under strong compression ratios, a set of n_7 test sequences are generated on this trial by gradually reducing ζ. On P_8, a set of n_8 test sequences are generated by reducing the resolution of the original video frames by $\rho\%$.

On each trial P_i, T_j is tested with the original (ground truth) initialization I_t and the original video sequence V_t which contains a target H_t, where $H = \{H_t\}_{t=1}^{J}$ is a set of targets. To study its variation in performance, each tracker T_j is tested with the initialization $I_{t,i}$ and test sequence $V_{t,i}$ which are generated on trial P_i by modifying I_t or V_t in a pre-defined manner such that the applied modification simulates a specific real-world scenario: $I_{t,i} = P_i(I_t)$, and $V_{t,i} = P_i(V_t)$.

Let $R_{t,i}^{j}$ be the trajectory of target H_t estimated by testing tracker T_j on trial P_i with $V_{t,i}$ and $I_{t,i}$: $R_{t,i}^{j} = T_j(V_{t,i}, I_{t,i})$. The performance of tracker T_j is computed by evaluating the estimated trajectory $R_{t,i}^{j}$ of the target with respect to its ground-truth trajectory G_t using the proposed evaluation criterion (Sec. II) thus obtaining the performance score: $CoTPS_{t,i}^{j} = \Psi(R_{t,i}^{j}, G_t)$, where $\Psi(\cdot)$ represents the procedure involved in the evaluation criterion (see Sec. II). Based on $CoTPS_{t,i}^{j}$, we compare the performance of the trackers under consideration.

Table I summarizes the trials and the values of the corresponding parameters (these parameters accomplished statistically significant results, as discussed at the end of Sec. IV). Using the proposed protocol, a tracker is tested on each sequence of the dataset in original form and in its 24 variations generated on different trials. Each tracker is tested with 60 perturbations of the initialization on the original video sequence. A deterministic tracker is therefore run 85 times, whereas a probabilistic tracker is run $85 \times n$ times for its evaluation using the protocol, where n denotes the number of runs for each test of a trial.

We selected the *dataset* by taking into account the diversity of targets and test scenarios, their availability and the challenges involved. The dataset contains three target classes, namely *head*, *vehicle* and *person*. The sequences are chosen from PETS, CAVIAR, AVSS and SPEVI datasets, which are publicly available. A range of tracking challenges are present in the dataset such as partial/total occlusions, pose

[2]Please note that the index j refers to different trackers or to different parameter settings for the same tracker.

TABLE I
DESCRIPTION OF EIGHT TRIALS COVERING VARIOUS REAL-WORLD CHALLENGES ('RWC') AS ILLUSTRATED IN FIG. 2. THE PROTOCOL GENERATES 60 INITIALIZATIONS BY ADDING PERTURBATIONS TO THE ORIGINAL (GROUND-TRUTH) TARGET INITIALIZATION AND 24 TEST SEQUENCES BY MODIFYING THE ORIGINAL VIDEO.

RWC	Trial	Description	Parameters
Initialization errors	P_1	Position	$n_1 = 20$, $\hat{O} = 50$
	P_2	Size	$n_2 = 20$, $\bar{O} = 50$
	P_3	Position and size	$n_3 = 20$, $\tilde{O} = 50$
Sensor noise	P_4	Noisy video	$n_4 = 6$, $\hat{l} = 1, 2, \ldots, 6$
Latency	P_5	Frame dropping	$n_5 = 4$, $m = 2, 4, 6, 8$
Illumination	P_6	Changing illumination	$n_6 = 2$, $L = 200$
Bandwidth reduction	P_7	Video compression	$n_7 = 4$, $\zeta = 75, 50, 25, 0$
	P_8	Resolution reduction	$n_8 = 8$, $\rho = 10, 20, \ldots, 80$

TABLE II
DESCRIPTION OF THE DATASET. $C^H_{ini}, C^H_{min}, C^H_{max}, K$ AND C^f DENOTE, IN PIXELS, THE INITIAL TARGET SIZE, THE MINIMUM TARGET SIZE, MAXIMUM TARGET SIZE, THE NUMBER OF FRAMES IN THE SEQUENCE AND THE FRAME SIZE, RESPECTIVELY. KEY. PC: POSE CHANGES; SC: SCALE CHANGES; SSC: SMALL SC; PO: PARTIAL OCCLUSIONS; TO: TOTAL OCCLUSIONS; BC: BACKGROUND CLUTTER.

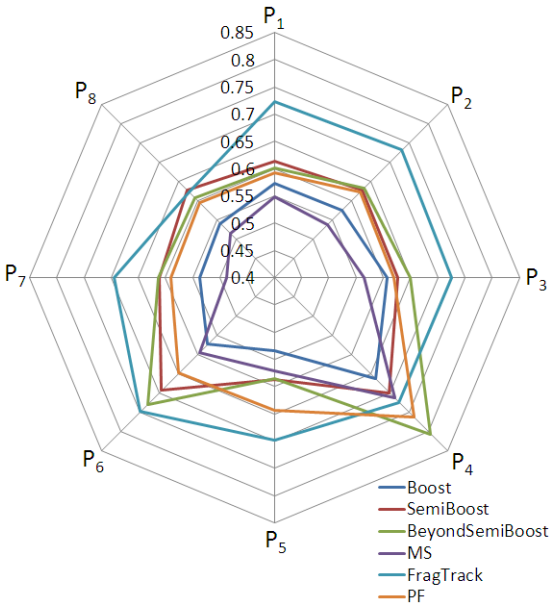
Target	Class	C^H_{ini}	C^H_{min}	C^H_{max}	K	C^f	Challenges
H_1	Head	139 × 91	7488	15965	430	576 × 720	PC, SSC
H_2	Head	62 × 66	370	40128	550	240 × 320	PC, SC, PO
H_3	Vehicle	227 × 108	2067	24516	160	576 × 768	SC, PC
H_4	Vehicle	99 × 103	870	10197	241	576 × 720	BC, SC, PC
H_5	Person	30 × 87	180	3444	150	576 × 768	PO, TO, SC, PC
H_6	Person	73 × 28	638	4410	750	288 × 384	PO, PC, SC, BC

changes, background clutter and small/large scale changes. The selected sequences include two *head* targets H_1 and H_2 from SPEVI [18], two *vehicle* targets H_3 and H_4 from PETS 2000 [19] and AVSS 2007 [20], respectively, and two *person* targets H_5 and H_6 from PETS 2010 [21] and CAVIAR [22], respectively.

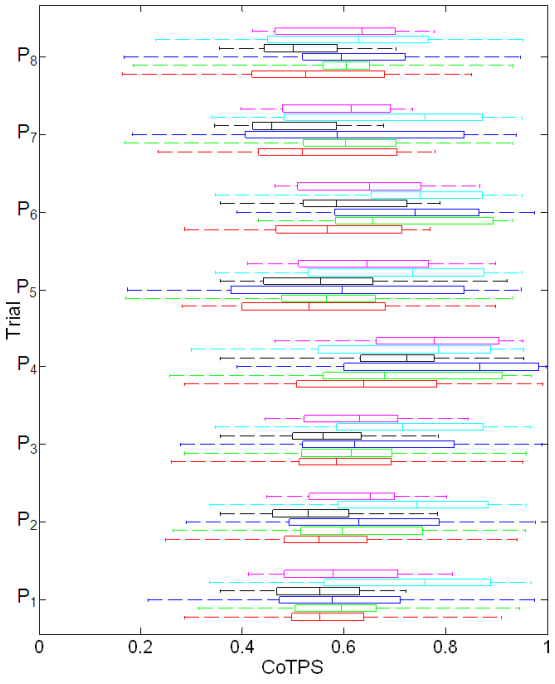
Table II summarizes the dataset in terms of initial target size (C^H_{ini}), minimum and maximum sizes of the visible part of target (C^H_{min} and C^H_{max}), number of frames (K), frame size (C^f) and the challenges present in the sequence.

IV. EXPERIMENTAL ANALYSIS AND VALIDATION

We demonstrate the effectiveness of the proposed score and protocol by evaluating and comparing six state-of-the-art trackers. The selected trackers can be divided into two categories: standard trackers and boosting-based trackers. Standard trackers are Mean Shift (MS) [16], the fragments-based tracker (FragTrack) [23], and Particle Filter (PF) [24]. Boosting-based trackers are Boost [17], the semi-supervised on-line boosting-based tracker (SemiBoost) [25], and beyond semi-supervised boost (BeyondSemiBoost) [26]. *The parameters of all trackers are fixed throughout the experiments*. We discuss the performance comparison on each trial P_i, on each target H_j and on each target class, and verify the statistical significance of the obtained results. Each tracker is tested on a total of 187144 frames. PF, being a probabilistic tracker, is run $n = 10$ times on each test of each trial and the mean value of its $CoTPS$ on the n runs is considered. The choice of $n = 10$ is made based on the analysis of the behavior of the mean $CoTPS$ of PF computed by running it with each of the six targets $H_1, H_2, ..., H_6$ for a variation of n: the fluctuation in the mean $CoTPS$ tends to decrease after $n = 5$ and becomes stable for $n \to 10$.



(a)



(b)

Fig. 3. Performance comparison of trackers on each trial. (a) Mean $CoTPS$ (μ_C) of trackers on each trial ($P_1, P_2, ..., P_8$) with all targets. (b) $CoTPS$ of trackers on each trial computed with all targets; Boost (red), SemiBoost (green), BeyondSemiBoost (blue), MS (black), FragTrack (cyan) and PF (magenta). The dispersion value (d_C) for a tracker is computed as difference between its maximum and minimum $CoTPS$ values on a trial.

A. Trial-wise comparison

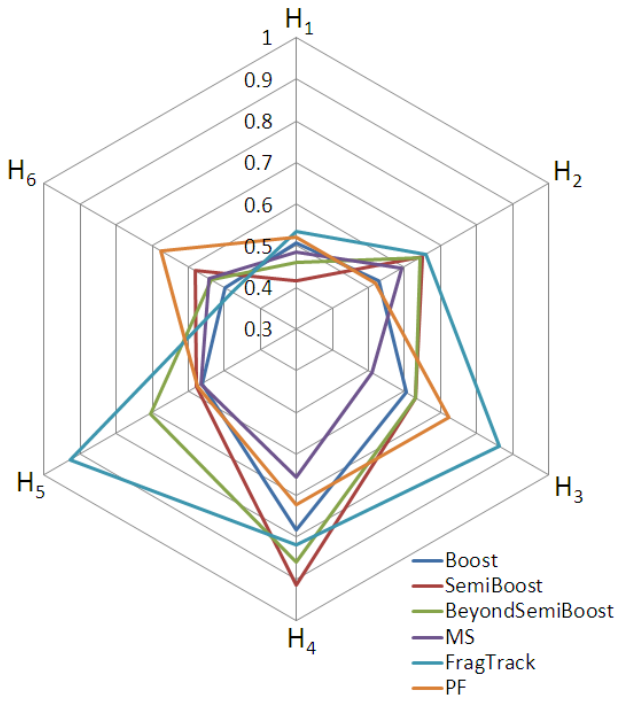
Figure 3 shows the mean $CoTPS$ (μ_C) of trackers on each P_i computed with all targets and their robustness in terms of the dispersion of their $CoTPS$ (d_C) computed with all targets as $d_C = CoTPS_{max} - CoTPS_{min}$, where $CoTPS_{max}$ and $CoTPS_{min}$ are the maximum and the minimum values of $CoTPS$ of a tracker on a trial, respectively.

MS consistently tracks more accurately in the presence

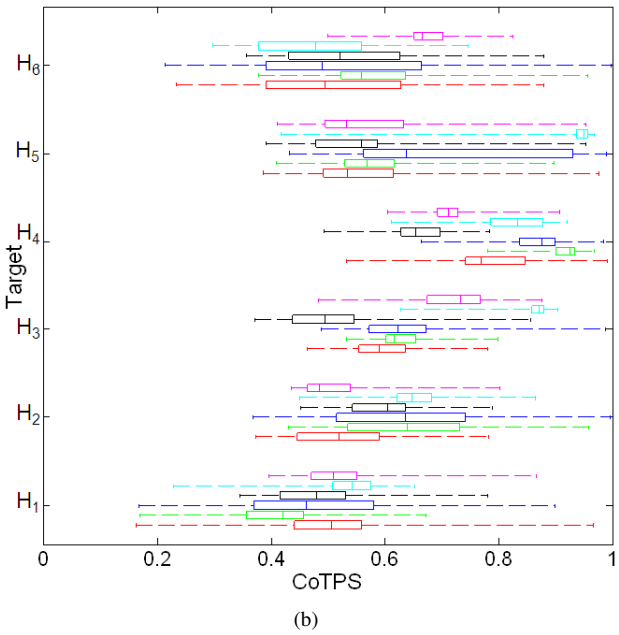
of initialization errors than other trackers (smaller μ_C on P_1, P_2, P_3 in Fig. 3(a)); whereas FragTrack shows inferior performance than other trackers in the presence of initialization errors. In fact, unlike MS, FragTrack uses a fragment-based representation of the target [23] and a perturbation in its initialization can lead to the inclusion of non-target patches in the target model thus resulting in the accumulation of tracking errors over time. Among the remaining trackers, Boost shows closer performance to MS on these trials (Fig. 3(a)) followed by PF and the other two boosting-based trackers. Additionally, in terms of robustness to initialization errors, MS and PF outperform the boosting-based trackers and FragTrack (smaller d_C of MS and PF in Fig. 3(b)). The reason of the increased sensitivity of the boosting-based trackers is that any perturbation to initialization may affect their learning process. The performance of the BeyondSemiBoost decreased the most with noisy video data (highest μ_C on P_4). The online adaptation model of Boost enables it to cope with noisy video data (smallest μ_C on P_4). PF is more robust to deal with noise (smaller d_C than the remaining trackers). On P_5, Boost shows the best performance (smallest μ_C) followed by MS, BeyondSemiBoost, SemiBoost, PF and FragTrack, respectively. Frame dropping may result in abrupt movements of target: standard trackers are more robust to increasing levels of frame dropping than boosting-based trackers. PF is the most robust tracker (smallest d_C on P_5) and BeyondSemiBoost is the least robust. Boost has the best performance under changing illumination (smallest μ_C on P_6), because of its ability to adapt to appearance changes [17]. PF is the most robust with changing illumination (smallest d_C on P_6). The d_C of MS is the closest to that of PF. An interesting observation regarding the performance of boosting-based trackers on P_6 is that both μ_C and d_C increase from Boost to SemiBoost and from SemiBoost to BeyondSemiBoost, which suggests that the evolution of the boosting-based trackers has resulted in a decreased ability to cope with changing illumination. The results also highlight the sensitivity of FragTrack to deal with changing illumination (the highest μ_C and the highest d_C on P_6). MS has the best performance on P_7 both in terms of μ_C and the robustness (d_C) to cope with the compressed video data. In terms of μ_C, Boost shows the closest performance to MS followed by PF, SemiBoost, BeyondSemiBoost and FragTrack, respectively; and in terms of d_C, PF has the closest performance to MS followed by Boost, FragTrack, BeyondSemiBoost and SemiBoost, respectively. Finally, on P_8, MS again outperforms other trackers in terms of μ_C and the robustness (d_C) to deal with resolution changes. The performance of Boost is closer to that of MS in terms of μ_C as compared to remaining trackers. Moreover, the performance of PF is closer to that of MS in terms of d_C than the other trackers. Based on the performance analysis on P_7 and P_8, the performance of MS is the least affected by compression or reduction in the resolution.

B. Target-wise comparison

Figure 4 shows the mean $CoTPS$ of trackers (μ_C) on each target ($H_1, H_2, ..., H_6$) and their robustness in terms of dispersion of their $CoTPS$ (d_C) computed in all trials.



(a)



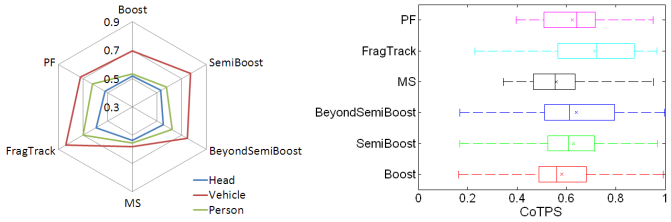
(b)

Fig. 4. Performance comparison of trackers on each target. (a) Mean $CoTPS$ (μ_C) of trackers on each target ($H_1, H_2, ..., H_6$) with all trials. (b) $CoTPS$ of trackers on each target computed with all trials; Boost (red), SemiBoost (green), BeyondSemiBoost (blue), MS (black), FragTrack (cyan) and PF (magenta). The dispersion value (d_C) for a tracker is computed as difference between its maximum and minimum $CoTPS$ values on a target.

The performance of SemiBoost is the best on H_1 in terms of μ_C, followed by BeyondSemiBoost, MS, Boost, PF and FragTrack (Fig. 4(a)). In terms of d_C, the results show a smaller variation in performance of the standard trackers compared to the boosting-based trackers (Fig. 4(b)). There is a pose change of the target (H_1) around frame 107 of the sequence (Fig. 5(a)), where the boosting-based trackers lose the target (Boost only tracks a very small part of target in this frame). H_2

(a) H_1 (b) H_2 (c) H_3 (d) H_4 (e) H_5 (f) H_6

Fig. 5. Sample tracking results generated by the trackers using the trials of the proposed evaluation protocol. Boost: red; SemiBoost: green; BeyondSemiBoost: blue; MS: black; FragTrack: cyan; PF: magenta. For the complete set of results, please see http://www.eecs.qmul.ac.uk/~andrea/pft2.



(a) (b)

Fig. 6. (a) Performance comparison of trackers on each target class (*head, vehicle, person*) based on mean $CoTPS$. (b) Cumulative performance of trackers: the mean $CoTPS$ of trackers computed on all trials with all targets are shown with a '×' in the corresponding boxplots. The dispersion value (d_C) for a tracker is computed as difference between maximum and minimum $CoTPS$ values in its boxplot.

presents challenges such as partial occlusions, pose changes and scale changes. PF has the best performance in terms of μ_C. The μ_C of Boost is the closest to PF. Moreover, MS and PF have a smaller variation (d_C) in their performance on H_2 compared to the remaining trackers. There is a significant pose change of target (360° turning) around frame 145 (Fig. 5(b)), where only MS, PF and Boost have tracked. On H_3, MS outperforms the other trackers as shown by its smallest μ_C. This is because the appearance of target H_3 is very bright and well-distinguished from the background, and the use of color distribution enables MS to track well on the various generated test sequences containing H_3. SemiBoost has the smallest variation in performance on H_3 (smallest d_C). H_3 undergoes gradual change in its scale and pose across the sequence. MS deals with these challenges and tracks consistently well, followed by PF (Fig. 5(c)). H_4 is challenging due to the presence of background clutter, similar objects (vehicles) and scale changes, and all trackers have obtained high μ_C on it. MS has the best performance on H_4 in terms of μ_C followed by PF, Boost, FragTrack, BeyondSemiBoost and SemiBoost, respectively. In terms of variation in performance, although d_C for SemiBoost is smaller than for the other trackers, this is less important as its $CoTPS$ is mostly very high. Among the remaining trackers, d_C of standard trackers is smaller than Boost and BeyondSemiBoost. The appearance of H_4 is very similar to that of the road making it challenging to track. All trackers have struggled to track this target with MS showing the best tracking followed by PF (Fig. 5(d)). Boost and MS have similar performance on H_5 in terms of μ_C (μ_C of PF and SemiBoost are also comparable to them). SemiBoost shows a smaller variation (d_C) in its performance on H_5 as compared to other trackers. H_5 faces a severe occlusion around frame 51 where only PF can track the target after the occlusion (Fig. 5(e)). H_6 has challenges such as the presence of targets of the same class (person), partial occlusions and small pose changes. FragTrack outperforms the other trackers in terms of μ_C as it can deal well with pose changes and partial occlusions [23]. The sequences containing H_2 and H_5 also involve pose changes and partial occlusions but FragTrack has not performed as well on them (Fig. 4(a)). H_2 involves significant pose changes and H_5 involves severe occlusions, suggesting that FragTrack can cope better with small pose changes and partial occlusions. Figure 5(f) shows frame 359 involving partial occlusion where FragTrack performs well (PF also tracks a small part of the target).

C. Discussion

Figure 6(a) shows the performance of trackers for each target class (*head, vehicle, person*). Each tracker has its best performance on *head*, followed by *person* and *vehicle*. The overall best performance on *head* and *person* tracking is by Boost. The performance of PF is closer to Boost on *head* tracking. The overall best performance on *vehicle* tracking is by MS. There is an inconsistency in the performance of FragTrack on *person* tracking: while it has achieved the best performance on H_6, its performance reduces significantly on H_5 (Fig. 4(a)), as discussed earlier.

Figure 6(b) shows the *cumulative performance* of trackers on all trials and all targets. MS has the best perfomance in terms of μ_C followed by Boost, PF, SemiBoost, Beyond-SemiBoost and FragTrack, respectively. PF is more robust than the remaining trackers as shown by its smaller d_C. Finally, overall, the standard trackers are more robust when dealing with various test scenarios than the boosting-based trackers (smaller d_C of the former set of trackers in Fig. 6(b)). MS handles better initialization errors and outperforms other trackers with compressed videos and resolution reductions. Boost copes well with noise, with frame dropping and with changing illumination. Among standard trackers, MS and PF can handle small as well as large pose changes; whereas FragTrack can only deal with small pose changes. Among boosting-based trackers, Boost outperforms SemiBoost and BeyondSemiBoost in handling pose changes. PF can handle partial and total occlusions better than all the other trackers.

To conclude, we tested the *statistical significance* of $CoTPS$ using the Welch ANOVA test [27], a modified version of the one-way ANalysis Of VAriance (ANOVA) test [28], commonly employed to test statistical significance of multiple groups of data (in our case, there are six groups each

containing a set of $CoTPS$ of a tracker) whose variances are unequal [29]. Statistical significance was achieved on each trial, on each target and on each target class at the standard significance level $\alpha = 0.05$.

V. Conclusion

We introduced a new overlap-based criterion for the performance evaluation of video trackers on extended targets. The proposed criterion quantifies performance by combining tracking accuracy and tracking failure scores. We also presented a new evaluation protocol that quantifies the robustness of trackers under various real-world conditions, which are encapsulated in a series of trials. An extensive experimental analysis and validation is presented in the form of a statistically significant performance comparison of six state-of-the-art trackers. The implementation of the protocol is available online to provide the research community with a platform to present and compare the performance of their trackers.

Our future work involves extending the proposed evaluation criterion to multi-target tracking performance evaluation. Moreover, as the proposed trials are generic and not designed specifically for a ground-truth-based evaluation, we aim to use them in combination with standalone evaluation criteria.

References

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. of Comput. Vis.*, vol. 47, no. 1/2/3, pp. 7–42, 2002.

[2] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. of Comput. Vis.*, vol. 92, no. 1, pp. 1–31, March 2011.

[3] *Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference*, Int. Telecom. Union, 2011. [Online]. Available: http://www.itu.int/rec/T-REC-J.341-201101-I/en. Last accessed on 12/6/2012.

[4] J. Black, T. Ellis, and P. Rosin, "A novel method for video tracking performance evaluation," in *Proc. of IEEE Int. Work. on Vis. Surveill.-Perform. Evaluat. of Tracking and Surveill.*, 2003, pp. 125–132.

[5] L. M. Brown, A. W. Senior, Y.-L. Tian, J. Connell, A. Hampapur, C. f. Shu, H. Merkl, and M. Lu, "Performance evaluation of surveillance systems under varying conditions," in *Proc. of IEEE Int. Work. on Vis. Surveill.-Perform. Evaluat. of Tracking and Surveill.*, 2005, pp. 1–8.

[6] F. Bashir and F. Porikli, "Performance evaluation of object detection and tracking systems," in *Proc. of IEEE Int. Work. on Perform. Evaluat. of Tracking and Surveill.*, 2006, pp. 7–14.

[7] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 31, no. 2, pp. 319–336, February 2009.

[8] H. Li, C. Shen, and Q. Shi, "Real-time visual tracking using compressive sensing," in *Proc. of IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2011, pp. 1305–1312.

[9] A. Senior, A. Hampapur, Y. Tian, L. Brown, S. Pankanti, and R. Bolle, "Appearance models for occlusion handling," in *Proc. of Int. Work. on Perform. Evaluat. of Tracking and Surveill.*, Hawaii, December 2001.

[10] B. Ristic, B.-N. Vo, D. Clark, and B.-T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *IEEE Trans. on Signal Process.*, vol. 59, no. 7, pp. 3452–3457, July 2011.

[11] X. Mei, H. Ling, Y. Wu, E. Blasch, and L. Bai, "Minimum error bounded efficient l_1 tracker with occlusion detection," in *Proc. of IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2011.

[12] T. Nawaz and A. Cavallaro, "PFT: A protocol for evaluating video trackers," in *Proc. of IEEE Int. Conf. on Image Process.*, Brussels, September 2011.

[13] P. Pan, F. Porikli, and D. Schonfeld, "Recurrent tracking using multifold consistency," in *Proc. of IEEE Int. Work. on Perform. Evaluat. of Tracking and Surveill., in conjunction with IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2009.

[14] F. Yin, D. Makris, and S. A. Velastin, "Performance evaluation of object tracking algorithms," in *Proc. of IEEE Int. Work. on Perform. Evaluat. of Tracking and Surveill.*, Rio de Janeiro, Brazil, 2007.

[15] E. Maggio and A. Cavallaro, *Video tracking: theory and practice*. Wiley, 2011.

[16] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 25, no. 5, pp. 564–577, May 2003.

[17] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proc. of IEEE Conf. on Comput. Vis. and Pattern Recognit.*, USA, 2006, pp. 260–267.

[18] http://www.eecs.qmul.ac.uk/~andrea/spevi.html. Last accessed on 12/6/2012.

[19] ftp://ftp.cs.rdg.ac.uk/pub/PETS2000/. Last accessed on 12/6/2012.

[20] http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html. Last accessed on 12/6/2012.

[21] http://www.cvg.rdg.ac.uk/PETS2010/a.html#s2l1. Last accessed on 12/6/2012.

[22] http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/. Last accessed on 12/6/2012.

[23] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. of IEEE Conf. on Comput. Vis. and Pattern Recognit.*, USA, 2006, pp. 798–805.

[24] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Proc. of European Conf. on Comput. Vis.*, 2002, pp. 661–675.

[25] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proc. of European Conf. on Comput. Vis.*, Heidelberg, 2008, pp. 234–247.

[26] 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 *Proc. of Int. Conf. on Comput. Vis.*, USA, 2009.

[27] B. L. Welch, "On the comparison of several mean values: An alternative approach," *Biometrika*, vol. 38, no. 3-4, pp. 330–336, 1951.

[28] R. A. Fisher, "The logic of inductive inference," *Journal of the Royal Statist. Society*, vol. 98, pp. 39–82, 1935.

[29] K. Moder, "Alternatives to f-test in one way anova in case of heterogeneity of variances (a simulation study)," *Psycholog. Test and Assess. Modeling*, vol. 52, no. 4, pp. 343–353, 2010.

Tahir Nawaz received the Bachelor of Mechatronics Engineering degree from the National University of Sciences and Technology (NUST) in 2005, the M.Sc. degree in vision and robotics (VIBOT), a joint Masters program in three European universities: Heriot-Watt University, Edinburgh, U.K., University of Girona, Girona, Spain, and the University of Burgundy, Dijon cedex, France, in 2009.

He worked for four months in 2010 with Medicsight PLC, London, U.K., as Scientific R&D Intern on the analysis of haustral folds (part of human colon anatomy) imaged with Computed Tomography. Since 2010, he has been with Queen Mary University, London, under the supervision of Prof. A. Cavallaro, first as a Research Assistant from September to December 2010 and then as a Research Student since January 2011. His current research interests include performance evaluation of video tracking, environment learning and shape analysis.

Andrea Cavallaro received the Laurea (summa cum laude) degree from the University of Trieste, Trieste, Italy, in 1996, and the Ph.D. degree from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 2002, both in electrical engineering. He served as a Research Consultant with the Image Processing Laboratory, University of Trieste, from 1996 to 1998, working on compression algorithms for very low bitrate video coding. He was a Research Assistant with the Signal Processing Laboratory, EPFL, from 1998 to 2003. He has been with Queen Mary University of London, London, U.K., since 2003, where he is a Professor of multimedia signal processing. He has authored more than 130 papers, including ten book chapters and two books, Multi-Camera Networks (Elsevier, 2009) and Video Tracking (Wiley, 2011).

Dr. Cavallaro was the recipient of a Research Fellowship with British Telecommunications from 2004 to 2005, three Student Paper Awards at IEEE ICASSP in 2005, 2007, and 2009, respectively, and the Best Paper Award at IEEE AVSS 2009. He is an elected member of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee. He served as the Technical Chair for IEEE AVSS 2011, WIAMIS 2010, and EUSIPCO 2008; and as a General Chair for IEEE/ACM ICDSC 2009, BMVC 2009, and IEEE AVSS 2007. He is the Area Editor for the IEEE Signal Processing Magazine and an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING and the IEEE TRANSACTIONS ON SIGNAL PROCESSING.