ASSESSING TRACKING ASSESSMENT MEASURES

Tahir Nawaz *Fabio Poiesi* *Andrea Cavallaro*

Centre for Intelligent Sensing, Queen Mary University of London, United Kingdom
{tahir.nawaz,fabio.poiesi,andrea.cavallaro}@eecs.qmul.ac.uk

ABSTRACT

We propose a methodology to quantitatively compare the relative performance of tracking evaluation measures. The proposed methodology is based on determining the probabilistic agreement between tracking result decisions made by measures and those made by humans. We use tracking results on publicly available datasets with different target types and varying challenges, and collect the judgments of 90 skilled, semi-skilled and unskilled human subjects using a web-based performance assessment test. The analysis of the agreements allows us to highlight the variation in performance of the different measures and the most appropriate ones for the various stages of tracking performance evaluation.

***Index Terms*—** Video tracking, evaluation measures, subjective assessment.

1. INTRODUCTION

Several performance evaluation measures have been introduced to measure the quality of video tracking results [1–6]. These evaluation measures, in turn, need to be assessed to understand their relative performance. Discrepancy-based empirical measures evaluate performance by quantifying the deviation of tracking results from a ground truth over time at frame level [7] or at sequence level [8]. The measures may evaluate tracking performance based, for example, on the extent of spatial match between the tracked region and the ground-truth target region over time. The spatial match may be determined in the form of the number of common pixels [7] or coincidence between the tracked and ground-truth regions [1]. Coincidence is defined as the existence of the centroid of one region within the other region.

While efforts have been made to empirically assess measures in other research areas, including information retrieval [9], data clustering [10] and image compression [11], to the best of our knowledge no attempt has yet been made at a direct quantitative assessment of measures in the area of video tracking. The comparison of measures was indirectly performed by considering the performance of algorithms [12] and by studying the inter-measure correlation [13] without explicitly analyzing the performance of the measures. Moreover, a previous study [14] analyzed the agreement among the ground-truth labelings (for different tasks including tracking) by humans to examine the possible variations in their annotations without aiming at assessing the measures.

In this paper we propose a methodology for the quantitative assessment of discrepancy-based evaluation measures with respect to human judgement. The comparison and analysis are based on determining the probabilistic agreement between the decisions made by measures and those made by humans on tracking results (Fig. 1). We assess seven measures on tracking results generated on ten publicly available datasets with three target types (head, full body, vehicle).

This paper is organized as follows. Sec. 2 formulates the problem and explains the statistical significance test used in the analysis. Sec. 3 describes the assessed evaluation measures. Sec. 4 describes the subjective evaluation procedure with respect to which the measures will be assessed in Sec. 5. Conclusions are drawn in Sec. 6.

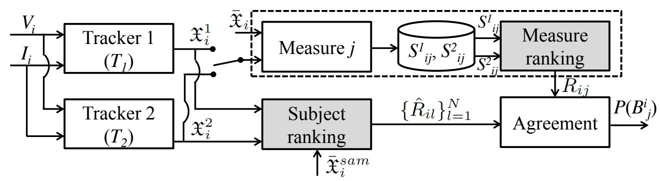


Fig. 1. Empirical assessment of measures with respect to human judgement. T_1 and T_2 are tested on video clip (V_i) with initialization (I_i). S^1_{ij} and S^2_{ij} are performance scores computed using the measure j by evaluating $\mathfrak{X}^1_i$ and $\mathfrak{X}^2_i$, the estimated trajectories of T_1 and T_2 on V_i, with respect to ground-truth trajectory $\bar{\mathfrak{X}}_i$. R_{ij} is the decision of the measure j based on S^1_{ij} and S^2_{ij}. $\hat{R}_{il}$ is the decision of the human subject l: $l = 1, \ldots, N$, based on $\mathfrak{X}^1_i$ and $\mathfrak{X}^2_i$ while also using the available ground-truth samples, $\bar{\mathfrak{X}}^{sam}_i$. $P(B^i_j)$ denotes the amount of agreement on V_i between R_{ij} and the set of human judgements, $\{\hat{R}_{il}\}^N_{l=1}$.

2. PRELIMINARIES

2.1. Problem formulation

Let us consider two trackers, T_1 and T_2, run on the ith video clip, $V_i : i = 1, ..., M$, with target initialization, I_i. The trackers generate the respective trajectories, $\mathfrak{X}^1_i$ and $\mathfrak{X}^2_i$, in each clip i. $\mathfrak{X}^1_i$ and $\mathfrak{X}^2_i$ are a sequence of states over frames: $\mathfrak{X}^1_i$=$\{X^1_{ik}\}^{K^1_i}_{k=1}$, where X^1_{ik} is the estimated state of T_1 at frame k of V_i, and K^1_i is the number of frames where $\mathfrak{X}^1_i$ exists. X^1_{ik} may contain information about the target position (x^1_{ik}, y^1_{ik}) and the occupied region A^1_{ik}: X^1_{ik}=$\{(x^1_{ik}, y^1_{ik}), A^1_{ik}\}$. Let $\bar{\mathfrak{X}}_i$, $\bar{X}_{ik}$, $\bar{K}_i$, $(\hat{x}_{ik}, \hat{y}_{ik})$ and $\hat{A}_{ik}$ represent the corresponding ground-truth of the quantities defined above. $\mathfrak{X}^1_i$ and $\mathfrak{X}^2_i$ are evaluated with respect to $\bar{\mathfrak{X}}_i$ using one out of J measures ($j = 1, ..., J$) to obtain their evaluation scores, S^1_{ij} and S^2_{ij}, respectively.

Based on the comparison between S^1_{ij} and S^2_{ij} we define the rank R_{ij} as: R_{ij}=$(1, 2)$ if S^1_{ij} is better than S^2_{ij}; R_{ij}=$(2, 1)$ if S^2_{ij} is better than S^1_{ij}; or R_{ij}=$(1.5, 1.5)$ if S^1_{ij}=S^2_{ij}. R_{ij}=$(1.5, 1.5)$ defines a tie between T_1 and T_2 [15]. Similarly, let $\hat{R}_{il}$ be the judgement (decision) of the lth human subject (*s.t.* $l = 1, ..., N$) in ranking $\mathfrak{X}^1_i$

This work was supported in part by the Artemis JU and in part by the UK Technology Strategy Board through COPCAMS Project under Grant 332913. 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 no 2010-0012).

and $\mathfrak{X}_i^2$. $\hat{R}_{il}$ is defined as R_{ij}, where j in R_{ij} is replaced by l.

2.2. Statistical significance test

This section discusses the statistical significance test to check the intra-subject agreement. To test the statistical significance for decisions of a sample of judges (subjects), we define two hypotheses, the *null hypothesis* (H_0) and *alternate hypothesis* (H_a), which are defined as follows. H_0: a set of judges cannot distinguish the performance of two trackers on a video; H_a: a set of judges can distinguish the performance of two trackers on a video.

We aim to statistically check whether H_a is valid by rejecting H_0 according to a level of significance, α, which indicates the probability of rejecting a true null hypothesis and is often set to 0.05 [15]. We choose a test that can be applied for ranked data and account for ties, namely the Friedman's Two-Way ANOVA test (the Friedman's test) [15]. The Friedman's test, χ^2, for a video is computed as

$$\chi^2 = \frac{12}{NF(F+1)} \sum_{f=1}^{F} \left(\sum_{l=1}^{N} \hat{R}_{il}(f) \right)^2 - 3N(F+1), \qquad (1)$$

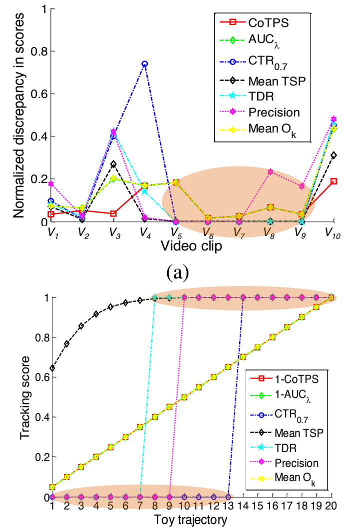
where N is the number of judges, $\hat{R}_{il}(f)$ is the rank assigned to tracker T_f on V_i by subject l such that f=$\{1,2\}$ because we consider a pair of trackers (F=2). To test the statistical significance at α=0.05, the χ^2 value is compared to the value corresponding to F-1 degrees of freedom in the χ^2 table of critical values [15] that is equal to 3.841. If $\chi^2 > 3.841$, the statistical significance is achieved and H_0 is rejected. $\chi^2 \in [0, N]$. Let us consider an example with $N = 50$: if $\hat{R}_{il} = (1,2)$ for 50% of the subjects and $\hat{R}_{il} = (2,1)$ for the remaing subjects, $\chi^2 = 0$; if $\hat{R}_{il} = (1,2)$ for 62% of the subjects and $\hat{R}_{il} = (2,1)$ for the remaining subjects, $\chi^2 = 2.880$; if $\hat{R}_{il} = (1,2)$ for 63% of the subjects and $\hat{R}_{il} = (2,1)$ for the remaining subjects, $\chi^2 = 3.920$; if $\hat{R}_{il} = (1,2)$ for 75% of the subjects and $\hat{R}_{il} = (2,1)$ for the remaining subjects, $\chi^2 = 13.520$; if $\hat{R}_{il} = (1,2)$ for 75% of the subjects and $\hat{R}_{il} = (1.5, 1.5)$ for the remaining subjects, $\chi^2 = 28.880$; if $\hat{R}_{il} = (1,2)$ for 100% of the subjects, $\chi^2 = 50$.

3. MEASURES

We consider the following state-of-the-art evaluation measures: Mean Overlap ($\overline{O}$) [16], Precision ($\hat{P}$), Track Detection Rate (TDR) [1], Area Under lost-track ratio Curve (AUC_λ) [5], Combined Tracking Performance Score (CoTPS) [6], Tracking Success Probability (TSP) [7] and Mean Dice (MD) *vs.* Correct Track Ratio (CTR) curve [8]. AUC_λ and CoTPS quantify performance based on the *lost-track ratio*. TSP, MD-vs-CTR and $\hat{P}$ need presetting of parameters, whereas TDR, AUC_λ, CoTPS and $\overline{O}$ do not require presetting of parameters. All the measures are bounded in $[0, 1]$. We use the symbol ($\uparrow$) to indicate that the higher the score, the better the result, whereas ($\downarrow$) indicates that the lower the score, the better the result.

Mean Overlap: The overlap, O_k ($\uparrow$), between $\hat{A}_{ik}$ and A_{ik} is defined as O_k=$\frac{|\hat{A}_{ik} \cap A_{ik}|}{|\hat{A}_{ik} \cup A_{ik}|}$. The Mean Overlap ($\overline{O}$) is computed as the average of O_k across the frames where the target exists.

Precision, $\hat{P}$ ($\uparrow$), is defined as $\hat{P}$=$\frac{|TP|}{|TP|+|FP|}$, where $|TP|$ and $|FP|$ are the number of true and false positives across the sequence, respectively. An estimation is a true positive if the overlap $O_k \geq \tau_3$ and a false positive if $O_k < \tau_3$. We use $\tau_3 = 0.25$ for head targets and $\tau_3 = 0.50$ for person and vehicle targets as done in [17].



(a)

(b)

Fig. 2. Ability of the measures to distinguish different tracking results. (a) Normalized discrepancy in the evaluation scores of each measure for the tracking pair on each clip, $V_1, \ldots, V_{10}$. (b) Evaluation scores computed using measures for 20 toy trajectories.

The *Track Detection Rate*, TDR ($\uparrow$), is defined as ratio of the number of true positive coincidences ($|TC|$) across $\mathfrak{X}_i^1$ or $\mathfrak{X}_i^2$ and the number of ground-truth points across $\bar{\mathfrak{X}}_i$, ($\bar{K}_i$), i.e. TDR $= \frac{|TC|}{\bar{K}_i}$. A true positive coincidence occurs when the ground-truth position of a target in a frame coincides with the estimated target area.

The *Area under lost-track ratio curve*, AUC_λ ($\downarrow$), is defined as AUC_λ=$\Delta\tau_2 \sum_{\tau_2=0}^{1} \lambda(\tau_2)$ and quantifies the tracking performance based on the area under the lost-track ratio curve, $\lambda(\tau_2)$, which represents the percentage of lost tracks for a given threshold τ_2. A track is considered *lost* in a frame if $O_k < \tau_2$. $\lambda(\tau_2)$ is generated for a variation of τ_2 with an increment of $\Delta\tau_2$.

The *Combined Tracking Performance Score*, CoTSP ($\downarrow$), is computed as CoTPS=$\beta\Omega + (1-\beta)\lambda_0$, where β is a weighting factor computed adaptively and is proportional to the number of frames with $O_k > 0$. The tracking accuracy Ω is computed similarly to AUC_λ, but using only the frames with $O_k > 0$. The tracking failure, λ_0, is the percentage of failed frames ($O_k = 0$).

The *Tracking Success Probability*, TSP$_k$ ($\uparrow$), is defined at frame k as: TSP$_k$=$\frac{\exp(\nu \cdot a(\hat{A}_{ik}, A_{ik}))}{1+\exp(\nu \cdot a(\hat{A}_{ik}, A_{ik}))}$, where $a(\hat{A}_{ik}, A_{ik})$ quantifies the overlap between $\hat{A}_{ik}$ and A_{ik} [18]. We use the mean TSP score ($\overline{\text{TSP}}$) across the trajectory and the fixed parameter ν=11.8 [7].

Mean Dice vs. Correct Track Ratio curve, MD-vs-CTR ($\uparrow$). Let the Dice score D_k be defined as D_k=$\frac{2|\hat{A}_{ik} \cap A_{ik}|}{|\hat{A}_{ik}|+|A_{ik}|}$, where $0 \leq D_k \leq 1$. The Correct Track Ratio (CTR) is the percentage of frames where D_k is greater than a threshold. Mean Dice (MD) is the average of the D_k scores that are greater than this threshold. The MD-vs-CTR curve plots MD against CTR, computed for the full range of possible thresholds. To quantify the tracking performance we use the CTR value corresponding to MD of at least 0.7, i.e. $\min\{\text{MD}\}_{MD \geq 0.7}$, denoted as $CTR_{0.7}$. A Dice score ≥ 0.7 is considered to be a satisfactory tracking result [8]; the threshold of 0.7 is used for $CTR_{0.7}$, thus showing the long-term tracking ability as the percentage of the sequence where the target is tracked with MD of at least 70%.

We are interested in analyzing the ability of measures to distinguish (slightly) different tracking results. Fig. 2(a) shows the nor-

Table 1. Summary of the dataset. Key: FS: Frame Size; K: number of frames in V_i; t: duration of the clip.

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
	Clemson head tracking				SPEVI		PETS2000	AVSS2007	PETS2010	CAVIAR
K	51	83	50	50	100	29	30	30	30	100
***t* (sec)**	7	11	6	7	7	4	4	4	4	11
FS	96×128	96 × 128	96 × 128	96 × 128	576 × 720	240 × 320	576 × 768	576 × 720	576 × 768	288 × 384
Target	Head						Vehicle		Person	

malized discrepancy between evaluation scores of each measure for tracker pairs on M video clips (where M=10 as discussed in Sec. 4), which is the absolute difference between the evaluation scores of tracker pairs computed using the measure divided by its range. $\overline{O}$, AUC_λ and CoTPS consistently distinguish tracker pairs on all clips (normalized discrepancy > 0), whereas the remaining measures are unable to distinguish results (i.e. normalized discrepancy=0) from V_5 to V_9 as highlighted in Fig. 2(a), except $\hat{P}$ that could distinguish performance on V_8 and V_9.

We show the variation of the scores of the measures using 20 toy trajectories, each having a constant overlap (for the whole sequence) of $0.05, 0.10, \ldots, 1$, respectively. The overlap is as $a(\cdot)$ for $\overline{\text{TSP}}$, as O_k for AUC_λ, CoTPS and $\hat{P}$, and as D_k for $CTR_{0.7}$. For TDR, coincidence is achieved throughout a trajectory when $O_k \geq 0.4$ (i.e. for trajectory 8 to trajectory 20). In Fig. 2(b) we can clearly see two groups of measures. The first group includes (1-CoTPS), (1-AUC_λ) and $\overline{O}$, which can each discriminate the results throughout overlap variations. The second group includes $\overline{\text{TSP}}$, $\hat{P}$, $CTR_{0.7}$ and TDR, which are often not able to distinguish variations in results (as highlighted in Fig. 2(b)) due to the hard decisions caused by their preset thresholds on the overlap or coincidence.

4. SUBJECTIVE EVALUATION

We use ten test videos (V_1 to V_{10}) with different target types (head, vehicle, person), challenges (scale change, pose change, occlusion, clutter) and scenarios (indoor, outdoor). The video clips are from publicly available datasets including AVSS 2007 challenge [19], CAVIAR [20], Clemson head tracking [21], PETS 2000 [22], PETS 2010 [23], SPEVI [24] (Tab. 1, Fig. 3). As trackers we use the mean-shift tracker [25], a particle filter-based tracker [26], the fragments-based tracker [27], the online boosting tracker [28], the semi-supervised online boosting tracker [29] and the beyond semi-supervised boosting tracker [30].

(a) V_1 (b) V_2 (c) V_3 (d) V_4 (e) V_5 (f) V_6 (g) V_7 (h) V_8 (i) V_9 (j) V_{10}

Fig. 3. Visualization of the first frame of video clips with targets indicated in green bounding boxes. Datasets: (a-d) Clemson head tracking, (e-f) SPEVI, (g) PETS 2000, (h) AVSS 2007 challenge, (i) PETS 2010 and (j) CAVIAR.

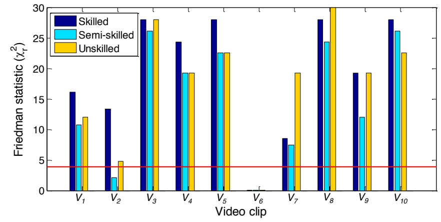


Fig. 4. Statistical significance using the Friedman test (χ^2) on each V_i for the skilled, semi-skilled and unskilled (subject) samples. The red line indicates the critical value corresponding to the standard significance level, $\alpha = 0.05$.

We asked subjects to rank the results of tracker pairs $(\mathfrak{X}_i^1, \mathfrak{X}_i^2)$ on all V_i. For each V_i, the tracking results are shown with $\mathfrak{X}_i^1$ and $\mathfrak{X}_i^2$ superimposed as a sequence of bounding boxes over time. Three samples of subjects are distinguished as *skilled*, *semi-skilled* and *unskilled* in target tracking. N_1, N_2 and N_3 denote the size of the skilled, semi-skilled and unskilled samples ($N_1 = N_2 = N_3 = 30$). None of the subjects was involved in this work [14].

The subjective evaluation tests were performed using a website [31] that, after providing the instructions, shows the tracking results of tracker pairs (T_1, T_2) side-by-side. The gray color of the background (red=green=blue=130) of the webpage follows the recommendation by ITU for relaxing human eyes [32]. For each clip the ground-truth tracking samples are also provided as a reference for the first, middle and last frames. We show short clips to help subjects remember the tracking results, thereby aiming to minimize the uncertainty in their judgment. The clips are played in a loop and can be viewed multiple times. Each subject chooses the tracker, 'Left' or 'Right', which is deemed to be the best or chooses 'Same' if the result of each tracker in the pair appears indistinguishable.

We perform the Friedman's test on each V_i for skilled (N=N_1), semi-skilled (N=N_2) and unskilled (N=N_3) samples separately (Sec. 2.2). Fig. 4 shows the results for the statistical significance: for skilled and unskilled subjects the statistical significance is achieved for all V_i except for V_6; for semi-skilled subjects the statistical significance is achieved for all V_i except for V_2 and V_6. The reason for the statistical insignificance on V_6 is that the subjects could not distinguish tracking results (Fig. 5(a)). In fact, the results in V_6 seem comparable (Fig. 6(f)).

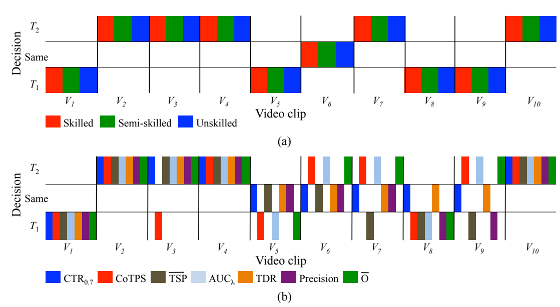


Fig. 5. Decision (ranking) for each video sequence (V_i). The ranking between the tracker pair (T_1, T_2) given on each V_i by (a) (most of) the skilled, semi-skilled and unskilled subjects, (b) the evaluation measures. 'T_1', 'T_2' and 'Same' on the vertical axis show T_1 considered the best, T_2 considered the best and both trackers considered the same, respectively.

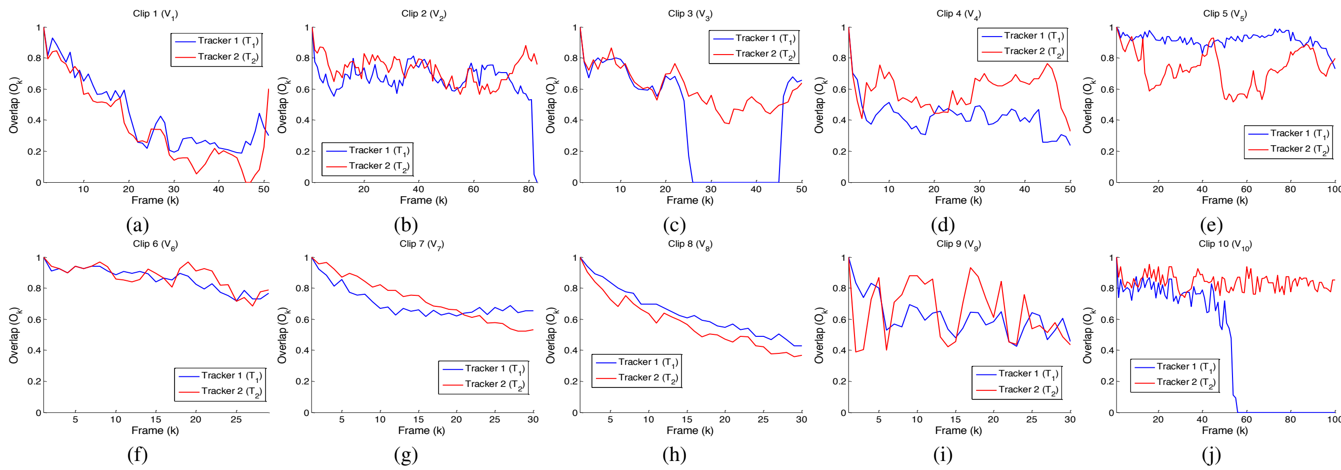


Fig. 6. Amount of overlap (common pixels), O_k, between estimated and ground-truth results for T_1 and T_2 on V_1 (a) to V_{10} (j).

5. MEASURE-SUBJECT AGREEMENT

We devise a probabilistic criterion for computing the measure-subject agreement. Let us consider a set of events of a sample of subjects (skilled, semi-skilled or unskilled) in a probability space for each V_i, which is defined as follows: $\mathbf{E}^i$=$\{E_1^i, E_2^i, E_3^i\}$: E_1^i= $\{T_1(V_i) \succ T_2(V_i)\}$; E_2^i= $\{T_2(V_i) \succ T_1(V_i)\}$; E_3^i= $\{T_1(V_i) \equiv T_2(V_i)\}$. $T_1(V_i)$ is the result of tracker T_1 on V_i; the symbol $\succ$ indicates preference and the symbol $\equiv$ means that two results are indistinguishable.

We can compute the probability of each event occurring as $P(E_r^i)$=$\frac{n_{E_r^i}}{n_{E_1^i}+n_{E_2^i}+n_{E_3^i}}$ $\forall r = 1, 2, 3$, where $n_{E_r^i}$ denotes the number of times E_r^i occurs for each V_i and for each sample. We find the probability, $P(B_j)$, of the jth measure (B_j^i has the same probability space as E_r^i) by calculating the total probabilities for M independent sets of events computed from each sample of subjects: $P(B_j)$=$\frac{1}{M}\sum_{i=1}^{M}\sum_{r=1}^{3} P(B_j^i|E_r^i)P(E_r^i)$, where M is the normalization factor. We use $P(B_j)$ to quantify the agreement between the jth measure and each sample of subjects (i.e. skilled, semi-skilled and unskilled) (Tab. 2).

The measures with the overall highest agreement with the three subject samples are $\hat{P}$ and $\overline{\text{TSP}}$ (Tab. 2). AUC_λ and $\overline{O}$ also consistently achieve high $P(B_j)$. CoTPS has a lower $P(B_j)$ due to an inappropriate decision on V_3 (Fig. 5(b)). $CTR_{0.7}$ and TDR show the lowest $P(B_j)$ for the three subject samples. Moreover, each measure has the highest $P(B_j)$ for skilled subjects followed by unskilled and semi-skilled subjects.

CoTPS, AUC_λ and $\overline{O}$ are mostly in agreement (Fig. 5(b)) and can capture slight changes in tracking results even when humans show uncertainty in distinguishing them. The ability to capture these changes is useful in accurately ranking the tracking results. For example, these three measures can distinguish the trackers on V_6 by judging T_2 as better (Fig. 5(b)), despite the fact that the majority of skilled (97%), semi-skilled (90%) and unskilled (90%) subjects judge them indistinguishable. A limitation in CoTPS can be seen on V_3 where T_1 is judged to be better than T_2, which is opposite to the judgement of the remaining measures and subjects as well. This limitation is due to the non-linear (quadratic) behavior of CoTPS due to its failure term, $\lambda_0 = 1 - \beta$. $\overline{\text{TSP}}$ and $\hat{P}$ are mostly in agreement (Fig. 5(b)) and also with respect to subjects (Tab. 2). $\overline{\text{TSP}}$ and $\hat{P}$ indeed penalize bad tracking results and poorly discriminate between good results (Fig. 2(b)). TDR and $CTR_{0.7}$ have the lowest agreement ($P(B_j)$) with subjects and have a limited ability to distinguish tracking results. Fig. 5(b) shows that 50% of video clips are judged 'Same' and this does not correspond to the judgment of subjects (Fig. 5(a)). Additionally, the smallest $P(B_j)$ of TDR indicates that tracking evaluation based on the coincidence criterion is not reflecting human judgment.

Overall, $\hat{P}$ and $\overline{\text{TSP}}$ generally show the highest agreement with human judgment, whereas CoTPS, AUC_λ and $\overline{O}$ have a better ability to distinguish similar tracking results. This confirms that a two-stage procedure for the evaluation and comparison of trackers is desirable [33]. First $\hat{P}$ should be used to group trackers in performance classes, where each class contains trackers with comparable results. Next the evaluation should be further refined within each class using, for example, $\overline{O}$.

Table 2. Assessment in terms of the measure agreement ($P(B_j)$) with the skilled, semi-skilled and unskilled subject samples. The brighter the cell, the better (higher) the agreement.

Measure	$\overline{\text{TSP}}$	$\hat{P}$	$CTR_{0.7}$	CoTPS	AUC_λ	$\overline{O}$	TDR
Skilled	0.74	0.74	0.58	0.61	0.71	0.71	0.58
Semi-skilled	0.68	0.67	0.52	0.57	0.66	0.66	0.52
Unskilled	0.70	0.71	0.53	0.61	0.70	0.70	0.53

6. CONCLUSIONS

We proposed a methodology to empirically assess tracking measures based on the law of total probability that quantifies the agreement between their decisions and those of human subjects in terms of ranking trackers' results. The results unveiled interesting aspects of the assessed measures. While $\hat{P}$ and $\overline{\text{TSP}}$ exhibit the highest agreement with humans, both have a limited ability to distinguish tracking results. $CTR_{0.7}$ and TDR showed the lowest agreement. AUC_λ and $\overline{O}$ are parameter independent, have a better ability to distinguish results and show a substantially higher agreement with humans (although lower than $\hat{P}$ and $\overline{\text{TSP}}$). Moreover, we observed that $\hat{P}$ and $\overline{O}$ should be used jointly for a thorough performance evaluation and comparison of trackers. Future work will involve assessing the reliability and stability of the measures, and performing the analysis on a larger video set.

7. REFERENCES

[1] J. Black, T. Ellis, and P. Rosin, "A novel method for video tracking performance evaluation," in *Proc. of VS-PETS Work.*, 2003.

[2] F. Bashir and F. Porikli, "Performance evaluation of object detection and tracking systems," in *Proc. of PETS Work.*, 2006.

[3] 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 PAMI*, vol. 31, no. 2, pp. 319–336, 2009.

[4] I. Leichter and E. Krupka, "Monotonicity and error type differentiability in performance measures for target detection and tracking in video," in *Proc. of CVPR*, 2012.

[5] T. Nawaz and A. Cavallaro, "PFT: a protocol for evaluating video trackers," in *Proc. of IEEE ICIP*, 2011.

[6] T. Nawaz and A. Cavallaro, "A protocol for evaluating video trackers under real-world conditions," *IEEE Trans. on IP*, vol. 22, no. 4, pp. 1354–1361, 2013.

[7] H. Li, C. Shen, and Q. Shi, "Real-time visual tracking using compressive sensing," in *Proc. of CVPR*, 2011.

[8] S. Salti, A. Cavallaro, and L. D. Stefano, "Adaptive appearance modeling for video tracking: Survey and evaluation," *IEEE Trans. on IP*, vol. 21, no. 4334-4348, pp. 10, 2012.

[9] C. Buckley and E. M. Voorhees, "Evaluating evaluation measure stability," in *Proc. of ACM SIGIR Conf. on Res. and Dev. in Inf. Ret.*, 2000.

[10] P. Kranen, H. Kremer, T. Jansen, T. Seidl, A. Bifet, G. Holmes, and B. Pfahringer, "Clustering performance on evolving data streams: Assessing algorithms and evaluation measures within moa," in *Proc. of IEEE ICDM Work.*, 2010.

[11] A. Mayache, T. Eude, and H. Cherifi, "A comparison of image quality models and metrics based on human visual sensitivity," in *Proc. of ICIP*, 1998.

[12] A. Baumann, M. Boltz, J. Ebling, M. Koenig, H. S. Loos, M. Merkel, W. Niem, J. K. Warzelhan, and J. Yu, "A review and comparison of measures for automatic video surveillance systems," *EURASIP JIVP*, 2008.

[13] R. Martin and J. M. Martinez, "Correlation study of video object trackers evaluation metrics," *IET EL*, vol. 50, no. 5, pp. 361–363, 2014.

[14] T. List, J. Bins, J. Vazquez, and R. B. Fisher, "Performance evaluating the evaluator," in *Proc. of PETS Work*, 2005.

[15] D. Israel, *Data Analysis in Business Research: A Step-By-Step Nonparametric Approach*, SAGE Pub. Pvt. Ltd., 2008.

[16] Visual Object Tracking Challenge (VOT2013), "http://votchallenge.net/index.html," last accessed on January 2014.

[17] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. of CVPR*, 2011.

[18] Hanxi Li, Chunhua Shen, and Qinfeng Shi, "Real-time visual tracking using sparse representation," *http://arxiv.org/abs/1012.2603*, 2010.

[19] "http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html," last accessed on February 2014.

[20] "http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/," last accessed on February 2014.

[21] "http://www.ces.clemson.edu/~stb/research/headtracker/seq," last accessed on February 2014.

[22] "ftp://ftp.cs.rdg.ac.uk/pub/PETS2000/," last accessed on February 2014.

[23] "http://www.cvg.rdg.ac.uk/PETS2009/a.html#s2," last accessed on February 2014.

[24] "http://www.eecs.qmul.ac.uk/~andrea/spevi.html," last accessed on February 2014.

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

[26] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Proc. of ECCV*, 2002.

[27] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. of CVPR*, 2006.

[28] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proc. of CVPR*, 2006.

[29] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proc. of ECCV*, 2008.

[30] 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 ICCV Work.*, 2009.

[31] "Subjective evaluation webpage," http://www.eecs.qmul.ac.uk/~andrea/subjeval.html. Last accessed on June 2014.

[32] Subjective video quality assessment methods for multimedia applications, "http://videoclarity.com/pdf/t-rec-p.910-199909-ipdf-e1.pdf," last accessed on October 2013.

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