EVALUATION OF ON-LINE QUALITY ESTIMATORS FOR OBJECT TRACKING*

Juan C. SanMiguel[1], *Andrea Cavallaro*[2] *and Jose M. Martinez*[1]

[1]Video Processing and Understanding Lab - Universidad Autónoma of Madrid (Spain)
[2]Multimedia and Vision Group – Queen Mary University of London (United Kingdom)

ABSTRACT

Failure of tracking algorithms is inevitable in real and on-line tracking systems. The online estimation of the track quality is therefore desirable for detecting tracking failures while the algorithm is operating. In this paper, we propose a taxonomy and present a comparative evaluation of online quality estimators for video object tracking. The measures are compared over a heterogeneous video dataset with standard sequences. Among other results, the experiments show, that the Observation Likelihood (OL) measure is an appropriate quality measure for overall tracking performance evaluation, while the Template Inverse Matching (TIM) measure is appropriate to detect the start and the end instants of tracking failures.

***Index Terms*—** performance evaluation without ground-truth, visual tracking quality, video surveillance

1. INTRODUCTION

Tracking objects in video sequences is a critical task in many computer vision applications. No tracking algorithm can perfectly perform in all conditions due to the variability and uncertainty of the data generated for example by crowded environments, clutter, changing illumination and occlusions. Consequently, failure of tracking algorithms is inevitable in real tracking systems. Ground-truth information is typically used for the evaluation of a tracking algorithm [1]. However, ground-truth annotations are very expensive to produce and therefore they usually cover only a small portion of video sequences and therefore a small percentage of data variability. This limitation makes it difficult to extrapolate the results of the tracking performance evaluation to new sequences. Moreover, ground-truth annotations are not available when a tracker has to operate while the video stream is being captured. Online performance evaluation not based on ground-truth is therefore a desirable evaluation alternative. The idea is to analyze online the intermediate results of the tracker (e.g., target position estimation or observation likelihood) at each time instance to determine the accuracy of the tracker.

In this paper, we propose a taxonomy for online quality estimators and we present a comparative evaluation by objectively evaluating representative measures to understand their advantages and drawbacks on commonly used datasets. The result of this evaluation is a recommendation on which measure to use to detect specific performance characteristics of a tracker.

This paper is organized as follows: Section 2 describes the proposed classification and representative state-of-art measures; Section 3 discusses experimental results and finally Section 4 describes the conclusions and future work.

2. ONLINE TRACK QUALITY ESTIMATION

Online track quality measures that do not make use of ground-truth can be classified into three main categories, namely trajectory-based, feature-based, and hybrid-based.

2.1. Trajectory-based measures

Trajectory-based measures use the information from the generated trajectories (the time-series representing the estimated position of the target over time) to measure the quality of a track. We have identified three sub-categories: *Forward-based measures, Model-based measures* and *Forward-backward measures.*

Forward-based quality measures analyze short trajectory segments. For example, Motion Smoothness (MS) [2] analyzes the trajectory increment for two adjacent frames:

$$MS(t) = L_2(T_t,\ T_{t-1}) \quad , \qquad (1)$$

where L_2 is the Euclidean distance and T_t represents the target position at time t. This measure is not robust against fast position changes in case of successful tracking.

Model-based quality measures use partial trajectories and online acquired (trajectory) models to evaluate the track

* Work partially supported by the Spanish Government (TEC2007- 65400 SemanticVideo), Cátedra Infoglobal-UAM for "Nuevas Tecnologías de video aplicadas a la seguridad", Consejería de Educación of the Comunidad de Madrid and European Social Fund. Part of the work reported in this paper was done during a research stay of the first author under a research grant (funded by UAM) at Queen Mary University of London (UK).

DATASET	TARGETS	SIZE	CHARACTERISTICS
D1	H1	320x240	Scale change, clutter, occlusion
D2	H2,H3, H4,H5	128x96	Scale/appearance change, clutter, occlusions
PETS01	P1,P2,P3,P4	768x576	Scale change, occlusions
CAVIAR	P5	384x288	Clutter
VISOR	P6, P7	352x288	Scale change, occlusions

Table 1: Description of the evaluation dataset

quality. For example, [3] uses the similarity between the acquired trajectories and the dynamically computed trajectory clusters as a measure of tracking quality.

Finally, *Forward-backward* quality measures apply an additional backward tracking analysis and the overall quality is derived from similarities between the forward and the backward trajectories. For example, [4] uses the Template Inverse Matching (TIM) and it is defined as:

$$TIM(t) = \sqrt{\left(\frac{C_x(T_{t-1}) - C_x(T'_{t-1})}{W(T_{t-1})}\right)^2 + \left(\frac{C_y(T_{t-1}) - C_y(T'_{t-1})}{H(T_{t-1})}\right)^2}\ , \quad (2)$$

where C_x, C_y, H and W, are the center coordinate, height and width of the target, respectively. Additionally, T_t and T'_{t-1} are the targets estimated positions at time *t-1* for forward and backward tracking (using as template the target estimated at time *t*). Unlike MS, TIM is robust against fast position changes in case of successful tracking because the change can be recovered by the backward tracking.

2.2. Feature-based measures

Feature-based measures are based on the analysis of the internal stages or output of the tracking algorithm. We have identified two sub-categories: *Model-dependent* and *Model-independent approaches*.

Model-dependent approaches use complementary features of the target model to measure the track quality. For example, [5] uses motion and color edges to measure the track quality of a contour tracking algorithm.

Model-independent approaches use measures that are independent of the target model and consider the output (or internal stages) of the tracking algorithm. Different types of measures are proposed for deterministic tracking approaches (e.g., weak similarity criterion [4]) or probabilistic tracking approaches (e.g., Observation Likelihood [6] or State Covariance Analysis [7]). For example, [6] computes the negative log-likelihood of the current observation and it is approximated as follows:

$$OL(t) \approx -\log \frac{1}{N} \sum_{i=1}^{N} \omega_i(t)\,, \quad (3)$$

where $\omega_i(t)$ is the sample i of the observation likelihood distribution at time t and N is the number of samples. OL is useful for measuring quick changes in the observation likelihood of the tracked target. However, its performance

Figure 1: Target initialization for the evaluation dataset. (From top-left to bottom-right) faces: H1, H2, H3, H4 and H5; pedestrians: P1, P2, P3, P4, P5, P6 and P7.

decreases for slow changes. Moreover, [7] analyzes the covariance of the target state distribution before and after weighting by the observation likelihood. It is defined as:

$$COV(t) = \det[C_b]/\det[C_a]\,, \quad (4)$$

where C_b and C_a are, respectively, the covariance matrix of the state distribution before and after weighting by the observation likelihood (assuming equal weights for the computation of C_b). The value of COV increases when the distribution uncertainty grows (in terms of variance) and tends to zero as the state distribution becomes more peaked.

2.3. Hybrid-based measures

Hybrid-based measures include methods that combine several approaches. Temporal and non-temporal analyses are combined to increase the performance of the track quality estimation. For example, [8] proposes a weighted sum between motion smoothness, temporal length and size/color similarity. Its advantage is the performance improvement for complex scenes that contain different tracking failures. However, a priori information (the weights of the measures) has to be determined for each scene.

3. COMPARATIVE EVALUATION

3.1. Experimental setup

The measures under analysis are the following: Motion Smoothness (MC) [2], Template Inverse Matching (TIM) [4], Observation Likelihood (OL) [6] and State Covariance analysis (COV) [7]. The evaluation dataset is composed of sequences from the D1[1] and D2[2] public datasets for face tracking, the PETS2001 dataset[3], the CAVIAR dataset[4] and the VISOR dataset[5]. Their characteristics are summarized in Table 1 and the target initialization is shown in Fig. 1. As tracking algorithm we used a color-based particle filter [9].

[1] ftp://motinas.elec.qmul.ac.uk/pub/single_face
[2] http://www.ces.clemson.edu/~stb/research/headtracker
[3] http://www.cvg.rdg.ac.uk/PETS2001/
[4] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/
[5] http://imagelab.ing.unimore.it/visor/

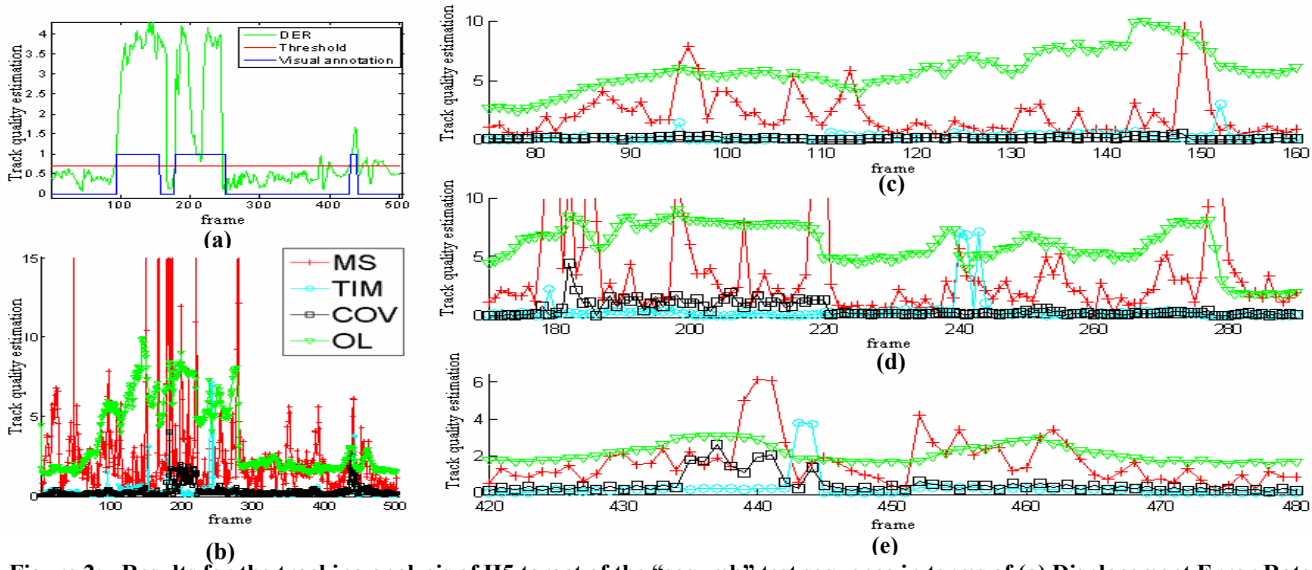

Figure 2: Results for the tracking analysis of H5 target of the "seq_mb" test sequence in terms of (a) Displacement Error Rate (DER) and (b) Track quality measures. Additionally, a zoom of the track quality estimation measurements is show for the (c) Clutter (frames 90-155), (d) Appearance change (frames 170-240) and (e) Occlusion (frames 435-445) failures.

In order to evaluate the measures, we analyze the overall quality and the quality at the start/end of a failure (that is useful for failure detection). To compare the overall quality, we use ROC analysis evaluating the discrimination of their values between two classes: successful and unsuccessful tracks. An unsuccessful track is determined if the Displacement Error Rate (DER) [10] is above a certain value that depends on the minimum allowed overlap between the ground-truth and detected areas (e.g., a value of $\sqrt{2}/2$ indicates a minimum overlap of 50%). DER computes the distance between the target estimation, T_e, and the ground-truth information, T_{GT}, as follows:

$$DER = \frac{d_E(T_e, T_{GT})}{\sqrt{\mathrm{A}_{\mathrm{GT}}}} = \frac{\sqrt{(x_e - x_{GT})^2 + (y_e - y_{GT})^2}}{\sqrt{\mathrm{A}_{\mathrm{GT}}}} \quad , \quad (5)$$

where $(x,y)_e$, $(x,y)_{GT}$ and A_{GT} are, the center location of the target estimated/annotated and the area of the ground-truth annotation, respectively. Additionally, the accuracy of the DER measure has been verified by visually defining the instants when the tracker fails.

3.2. Overall quality comparison

Due to the statistically nature of the selected tracking algorithm, several runs have been performed to get meaningful results. A summary of the experiments is shown in Table 2. Feature-based measures (OL and COV) perform better than trajectory-based measures (MS and TIM), because the tested sequences contained failures that produced a change in the state or observation likelihood and sometimes a fast position change of the target. Additionally, feature-based measures presented high standard deviation showing their dependency on the probabilistic analyzed data. Among the trajectory-based measures, TIM presents the highest AUC value due to the robustness introduced by the backward tracking stage. MS has low accuracy and depends on the movement of the target. Among the feature-based measures, OL outperforms COV as the majority of the evaluated failure types implied a change in the target observation likelihood (e.g., particle weights in a particle filter framework) instead a variation of the state variance.

MEASURE	Area Under Curve (AUC)	False Positive rate	True Positive rate
MS [2]	0.55 ± 0.0599	0.43 ± 0.0795	0.53 ± 0.0727
TIM[4]	0.69 ± 0.0358	0.37 ± 0.0481	0.60 ± 0.0651
OL [6]	0.78 ± 0.0887	0.20 ± 0.0554	0.65 ± 0.1133
COV [7]	0.70 ± 0.0675	0.35 ± 0.0619	0.72 ± 0.0986

Table 2: Comparative results for the ROC analysis using 10 runs (in terms of average ± standard deviation)

3.3. Analysis of the performance of the quality measures

A detailed analysis of the results in the test sequences has been performed to evaluate the DER and the selected quality measures. Fig. 2 shows an example of the DER and the track quality measurements for the H5 target of the "seq_mb" sequence. Fig. 2a shows that the DER produced a good measurement of the track quality useful for the overall quality comparison.

Regarding the quality of the measures, TIM provided high values at the starting/ending failure frames when the position change was due to model target dissimilarities (Fig. 2d and 2e) and similarities with clutter (Fig. 2c). During the failures, TIM did not provide high values because there was not a position change, whilst outside the failures its lower values indicated high-quality measurements. MS provided

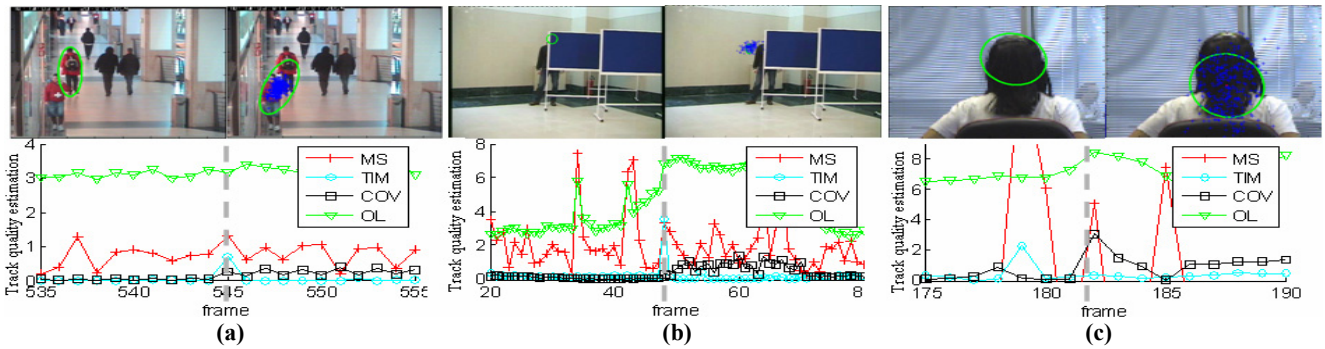

Figure 3: Examples of track quality measures for (a) clutter (target P5, frame 545 of the sequence "ThreePastShop2cor"), (b) an occlusion (target P6, frame 45 of the sequence "Visor_1") and (c) appearance change (target H5, frame 182 of the sequence "seq_mb"). Results correspond to ground-truth (top-left), tracking results (particles) (top-right) and the measures (bottom).

high values during the failures (starting/ending and duration) but also presented high values outside them. In general, MS obtained low performance to measure track quality. OL provided high values during the tracking failures because there was a change in observation likelihood in all the failures. At the starting/ending frames, its results depended on the rate of the observation likelihood change as demonstrated for the slow (Fig. 2e) and medium (Fig. 2c) changes. COV obtained good quality measurements in case of occlusion and appearance changes because the state distribution variance is increased to search the lost target. Moreover, the performance of COV is reduced when the target model is adapted to the wrong track (frames 220-260 of Fig. 2d).

Illustrative examples of the track quality measures are shown in Fig. 3. An example of similarities with the target model is shown in Fig. 3a (clutter). The small change in the observation likelihood and state distribution variance produced a low quality measurement of track quality by OL and COV. TIM obtained good results because there was a change in position (to the wrong estimated target). MS obtained low quality measurement because the target movement was small. Fig. 3b and 3c show an example of dissimilarities with the target model (an occlusion and an appearance change). The observation likelihood and state distribution variance changes were measured, respectively, by OL and COV obtaining high quality measurements. TIM obtained poor results because the backward tracking used the wrong tracked target as template being only adequate its use for the instant when the dissimilarities were produced. MS also obtained high quality measures because the quick changes of target movement

4. CONCLUSIONS AND FUTURE WORK

This paper introduced a taxonomy and presented a comparative evaluation of online quality estimation measures for video object tracking. Existing measures have been classified into three main categories and representative measures for each category have been described. Experimental results using a heterogeneous dataset showed that different measures should be applied to evaluate the overall performance or the start/end time of failure. For measuring the start and the end of a tracking failure the TIM measure is outperforming the other measures both for target model similarities (e.g., clutter) and dissimilarities (e.g., appearance changes and occlusions); whereas when considering the overall quality, the OL measure obtained the best results.

As future work, we will analyze the impact of the various thresholds applied on the quality measures (their values and variability) to detect the tracking failure.

5. REFERENCES

[1] Nascimento, J.C.; Marques, J.S. "Performance Evaluation of Object Detection Algorithms for Video Surveillance", IEEE Transactions on Multimedia, 8(4): 761-774, 2006.

[2] Wu H. and Zheng, Q.; "Self-evaluation of visual tracking systems", Proc of ASC, Orlando (FL, USA), 29 Nov.-2 Dec. 2004.

[3] Piciarelli, C.; Foresti, G.L.; Snidaro, L., "Trajectory clustering and its applications for video surveillance", Proc. of AVSS'05, Como (Italy), 15-16 Sept. 2005.

[4] Liu, R.; Li, S.; Yuan, X.; He, R.; "Online Determination of Track Loss Using Template Inverse Matching", Proc. of VS2008, Marseille (France), 17 Oct. 2008.

[5] Erdem, C.; Sankur & Tekalp, A. "Performance measures for video object segmentation and tracking", IEEE Transactions on Image Processing, 13(7):937-951, 2004.

[6] N. Vaswani, "Additive change detection in nonlinear systems with unknown change parameters", IEEE Transactions on Signal Processing, 55(3):859-872, 2007.

[7] Badrinarayanan, V.; Perez, P.; Le Clerc, F., Oisel, L.; "Probabilistic Color and Adaptive Multi-Feature Tracking with Dynamically Switched Priority Between Cues", Proc of ICCV'07, Rio de Janeiro (Brasil), 14-21 Oct. 2007.

[8] Chau, D.; Bremond, F.; Thonnat, M.; "Online evaluation of tracking algorithm performance", Proc. of ICDP'09, Kingston (UK), 3 Dec. 2009.

[9] Nummiaro, K.; Koller-Meier, E.; Van Gool, E., "An adaptive colour-based particle filter", Image and Vision Computing, 21(1): 99-110, 2003.

[10] Han, Z.J.; Ye, Q.X.; Jiao, J.B., "Online feature evaluation for object tracking using Kalman Filter", Proc. of ICPR'08, Tampa (FL, USA), 8-11 Dec. 2008.