Measures of effective video tracking

Tahir Nawaz, Fabio Poiesi, Andrea Cavallaro

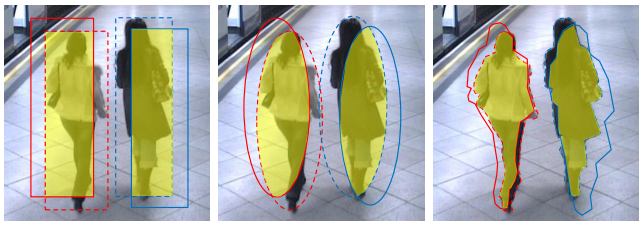
***Abstract*—To evaluate multi-target video tracking results, one needs to quantify the accuracy of the estimated target-size and the cardinality error as well as measure the frequency of occurrence of ID changes. In this paper we survey existing multi-target tracking performance scores and, after discussing their limitations, we propose three parameter-independent measures for evaluating multi-target video tracking. The measures take into account target-size variations, combine accuracy and cardinality errors, quantify long-term tracking accuracy at different accuracy levels, and evaluate ID changes relative to the duration of the track in which they occur. We conduct an extensive experimental validation of the proposed measures by comparing them with existing ones and by evaluating four state-of-the-art trackers on challenging real-world publicly-available datasets. The software implementing the proposed measures is made available online to facilitate their use by the research community.**

***Index Terms*—Multi-target video tracking, evaluation measure, accuracy, cardinality error, ID changes.**

I. Introduction

VIDEO tracking is a widely researched topic with applications in event detection, surveillance and behavior analysis. These applications may involve simultaneous tracking of multiple moving targets using a point-target representation (e.g. feature-point tracking) or an extended-target representation (e.g. in face or person tracking) [1]–[6]. Point-target representations use target position information, whereas extended-target representations also include information about the region covered by the target in the image plane [6], [7]. A tracking error is generally quantified by computing the discrepancy between estimated and ground-truth target regions [8], [9] (Fig. 1). Ground-truth-free tracking evaluation frameworks also exist that provide performance assessment by enforcing constraints such as time reversibility [10], [11] and feature consistency [12], [13] of the estimated tracks.

Unlike single-target tracking evaluation [9], [14], [15], multi-target tracking evaluation requires solving the assignment problem in order to establish associations between estimated and ground-truth states [8], [16]–[19]. The association may be determined using position only (point-based assignment) or region information as well (region-based assignment). *Point-based assignment* determines associations based on distance minimization between estimated and ground-truth tracks [16], [17]. *Region-based assignment* determines associations based on the overlap between estimated and ground-truth target regions [8], [18] or by establishing their *coincidence* [19]. Coincidence means that the centroid of one (e.g. estimated target) lies within the region of the other. Finally, the assignment may be solved at frame level [4] or at sequence level [16].



(a) (b) (c)

Fig. 1. Accuracy error for extended targets: overlap between estimated (solid line) and corresponding ground-truth (dotted line) regions defined as (a) bounding box, (b) bounding ellipse or (c) contour. Image from the iLids Easy sequence [20].

Three important aspects to be evaluated for multi-target tracking are accuracy, cardinality and number of ID changes [8], [16]. The *accuracy* quantifies the closeness of agreement between estimated and ground-truth states [21], and it can be calculated as an error score (i.e. distance [16], overlap [8], [17]) (Fig. 1) or based on true positives (correct estimations), false positives (incorrect estimations) and false negatives (missed estimations) [7]. The *cardinality error* is the difference between the number of estimated and ground-truth targets. *ID changes* are the incorrect associations between estimated and ground-truth targets.

Evaluation measures can be categorized into distance-based [16], [22]–[24] and overlap-based measures [8], [9], [17], [18]. *Distance-based measures* may not be suitable to evaluate changes in target size [16], [23], [24] or their values may not explicitly detect instances of tracking failure, which is defined as a zero-overlap between estimated and corresponding ground-truth states [22]–[24]. *Overlap-based measures* [8], [9], [18] generally consider the estimated target-size variations and can detect instances of tracking failure. Both distance-based and overlap-based measures may need presetting of parameters [8], [16]. For example, a cut-off parameter is used to define an upper bound [16]. Then the number of false positive (i.e. its spatial overlap with the ground truth is insufficient) and false negative estimations (i.e. missed estimation having spatial overlap with the ground truth to be zero) is determined by comparing their spatial overlaps with a pre-defined threshold [8]. Moreover, some existing measures are numerically unbounded [8], [25] and not well defined for the worst tracking case.

To address the above-mentioned limitations, we propose three overlap-based measures for multiple extended-target

Manuscript received November 12, 2012; revised May 18, 2013 and August 24, 2013; accepted October 11, 2013. This work has been partially funded by the Artemis JU and TSB as part of the COPCAMS project under GA number 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 n° 2010-0012). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hassan Foroosh.

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

video trackers that evaluate performance at frame level taking into account (i) the accuracy and the cardinality errors; (ii) long-term tracking accuracy using lost-track-ratio information; and (iii) ID changes in a parameter-independent manner. We provide an extensive experimental validation of the proposed measures in terms of comparison with existing measures and in the form of evaluation of four state-of-the-art multi-target trackers on challenging real-world datasets. The software implementation for the measures is made available online at http://www.eecs.qmul.ac.uk/~andrea/mtte.html.

This paper is organized as follows. A detailed review of the existing measures is presented in Sec. II. Sec. III describes the proposed measures, followed by their experimental validation in Sec. IV. Sec. V concludes the paper.

II. Related work: survey

Discrepancy-based performance assessment operates at frame level [16] or at sequence level by considering either individual tracks [17] or all the tracks [8]. We can identify three categories of multi-target tracking evaluation measures: Point-based Assignment and Position-based (PAP) evaluation, Region-based Assignment and Position-based (RAP) evaluation, and Region-based Assignment and Size-based (RAS) evaluation. These three categories are discussed after the definition of the notation we will use in this paper.

A. Notation

Let $X_{k,j}$ be the state of target j estimated by a tracker at frame k and defined as

$$X_{k,j} = (x_{k,j}, y_{k,j}, A_{k,j}, l_j), \tag{1}$$

where $(x_{k,j}, y_{k,j})$ define the position of the target, $A_{k,j}$ is its region information on the image plane and l_j is the target ID, and $k = 1, \ldots, K$. K is the number of frames in the video sequence. $A_{k,j}$ may be represented in the form of a bounding box [5], a bounding ellipse [6] or a bounding contour [26]. In the case of *point* targets, the estimated state of the target j at frame k does not contain $A_{k,j}$ and it is denoted as $X'_{k,j}$. $\mathbf{X}_k$ is the set of estimated states of multiple targets:

$$\mathbf{X}_k = \{X_{k,1}, \ldots, X_{k,j}, \ldots, X_{k,u_k}\}, \tag{2}$$

where $u_k = |\mathbf{X}_k|$ is the number of estimated targets at frame k (i.e. the cardinality of $\mathbf{X}_k$). The track $\mathfrak{X}_j$ of target j is defined as a sequence of states over time:

$$\mathfrak{X}_j = \{X_{k,j}\}_{k=k^j_{ini}}^{k^j_{end}}, \tag{3}$$

where k^j_{ini} and k^j_{end} denote the initial and final frame numbers of $\mathfrak{X}_j$, respectively, and K_j is the number of frames spanned by $\mathfrak{X}_j$. The set containing all the estimated tracks $\mathcal{X}$ in the sequence is

$$\mathcal{X} = \{\mathfrak{X}_j\}_{j=1}^{U}, \tag{4}$$

where U denotes the number of estimated tracks. Similarly, $\bar{X}_{k,i}$, $(\bar{x}_{k,i}, \bar{y}_{k,i}, \bar{A}_{k,i}, \bar{l}_i)$, $\bar{\mathbf{X}}_k$, v_k, $\bar{X}'_{k,i}$, $\bar{\mathfrak{X}}_i$, $\bar{k}^i_{ini}$, $\bar{k}^i_{end}$, $\bar{K}_i$, $\bar{\mathcal{X}}$ and V are the corresponding ground-truth notations for $X_{k,i}$, $(x_{k,i}, y_{k,i}, A_{k,i}, l_i)$, $\mathbf{X}_k$, u_k, $X'_{k,j}$, $\mathfrak{X}_j$, k^j_{ini}, k^j_{end}, K_j, $\mathcal{X}$ and U, respectively.

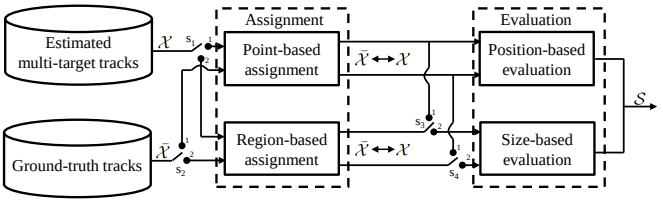


Fig. 2. General procedure for the computation of the evaluation score $\mathcal{S}$ for multiple extended-target tracking. Three different modalities are possible: using a point-based solution for the assignment problem and for evaluation ($s_1 = s_2 = 1, s_3 = s_4 = 2$); using a region-based solution for the assignment problem and a point-based solution for evaluation ($s_1 = s_2 = 2, s_3 = s_4 = 1$); using a region-based solution for the assignment problem and information about target position and size for evaluation ($s_1 = s_2 = 2, s_3 = s_4 = 2$).

B. PAP evaluation

PAP measures use a point-based assignment and evaluate target position only, without considering temporal size-changes (Fig. 2). Examples of PAP measures include Object Tracking Error (OTE), the Wasserstein's distance-based metric, the Optimal Sub-Pattern Assignment (OSPA) metric, Tracker Detection Rate (TRDR), False Alarm Rate (FAR), Track Detection Rate (TDR) and Track Fragmentation (TF).

OTE [17] computes the average positional distance between ground-truth and estimated track pairs. The assignment associates an estimated track with the ground-truth track that minimizes the average Euclidean distance across their common frames [22]. For each associated pair t, their OTE_t is calculated as

$$\text{OTE}_t = \frac{1}{\hat{K}_t} \sum_{k=\hat{k}^t_{ini}}^{\hat{k}^t_{end}} \sqrt{(\bar{x}_{k,t} - x_{k,t})^2 + (\bar{y}_{k,t} - y_{k,t})^2}, \tag{5}$$

where $\hat{K}_t = \hat{k}^t_{end} - \hat{k}^t_{ini}$ is the number of frames that are common in both ground-truth and estimated tracks and $\hat{k}^t_{ini}$ and $\hat{k}^t_{end}$ denote the initial and final frame numbers, respectively, of the pair t.

The Wasserstein's distance-based metric [25], $W_p(\bar{\mathbf{X}}_k, \mathbf{X}_k)$, computes the multi-target tracking accuracy as

$$W_p(\bar{\mathbf{X}}_k, \mathbf{X}_k) = \min_{\mathbf{C}} \left(\sum_{j=1}^{u_k} \sum_{i=1}^{v_k} C^k_{j,i} d(X'_{k,j}, \bar{X}'_{k,i})^p \right)^{1/p}, \tag{6}$$

where $d(\cdot)^p$ denotes the *p-norm* with $p \in [1, \infty)$, u_k is the number of estimated targets, v_k is the number of ground-truth targets, and $\mathbf{C}$ is the transportation matrix defining the association costs among all possible pairs of estimated and ground-truth tracks at frame k. The associations that minimize the overall cost are determined using the Hungarian or Munkres algorithms [27], [28].

The OSPA metric [16], [29] defines the tracking error as

$$\mathcal{D}_{p,c}(\bar{\mathbf{X}}_k, \mathbf{X}_k) = \left[\frac{1}{max(u_k, v_k)} \left(\min_{\pi \in \Pi_{u_k}} \sum_{i=1}^{v_k} \left(D_c(\bar{X}'_{k,i}, X'_{k,\pi(i)}) \right)^p + |u_k - v_k| \cdot c^p \right) \right]^{1/p}, \tag{7}$$

where Π_{u_k} represents the set of permutations each containing v_k elements taken from $\{1, 2, \ldots, u_k\}$, $D_c(\bar{X}', X') =$

$\min(c, D(\bar{X}', X'))$ is the cut-off distance between the two states with $c>0$ representing the cut-off parameter and $p \in [1, \infty)$ is the order parameter of the OSPA-based metric. $D(\bar{X}', X')$ denotes the base distance that quantifies the discrepancy between estimated and ground-truth states, and includes localization and labeling errors [16]:

$$D(\bar{X}', X') = \left(||\bar{X}' - X'||_{p'} + \alpha^{p'} \bar{\delta}[\bar{l}, l] \right)^{1/p'}, \quad (8)$$

where $\bar{\delta}[\bar{l}, l]$ is the complement of the Kronecker delta such that $\bar{\delta}[\bar{l}, l] = 0$ if $\bar{l} = l$ and $\bar{\delta}[\bar{l}, l] = 1$ if $\bar{l} \neq l$, and $\alpha \in [0, c]$ is the penalty applied to the labeling error if the frame-level assignment (determined as a result of the minimization in Eq. 7) does not correspond to the global assignment of tracks computed *a priori*. The global assignment is determined based on the minimization of the average distance between estimated and ground-truth tracks [16], [30]. $p' \in [1, \infty)$ denotes the order parameter of the base distance. Typically, $p = p' = 1$ [16]. Unlike OTE and the Wasserstein's distance-based metric, OSPA incorporates the cardinality error in the evaluation procedure, which would be otherwise not be taken into account by the minimization term of the distance error in Eq. 7. In particular, when the assignment is performed, the unassociated targets do not contribute to the accuracy error term and the inclusion of the cardinality error accounts for them. Additionally, the combination of the accuracy and the cardinality error terms yields a single score that facilitates performance comparisons.

TRDR, FAR and TDR [17] evaluate the accuracy using true positives and false positives determined with the coincidence criterion. Although these measures use target-size information in the evaluation, they are PAP measures because they do not evaluate target-size changes over time. For TRDR, FAR and TDR, the assignment between estimated and ground-truth tracks is solved as for OTE.

TRDR quantifies the overall performance at frame k as the ratio of the number of correctly-tracked targets (true positives), $|\widehat{TP}_k|$, to the number of ground-truth targets v_k:

$$\text{TRDR}_k = \frac{|\widehat{TP}_k|}{v_k}. \quad (9)$$

An estimation is considered a true positive if the centroid of the ground-truth bounding box lies (coincides) within the estimated bounding box. If no centroid of ground-truth bounding boxes coincide with an estimated bounding box, the estimation is considered a false positive.

FAR quantifies tracking performance at frame k as the ratio of the number of incorrectly-tracked targets (false positives), $|\widehat{FP}_k|$, to the sum of correctly- and incorrectly-tracked targets, $|\widehat{TP}_k| + |\widehat{FP}_k|$:

$$\text{FAR}_k = \frac{|\widehat{FP}_k|}{|\widehat{TP}_k| + |\widehat{FP}_k|}. \quad (10)$$

TDR quantifies the tracking performance at track level as the ratio between the number of true positive targets in the estimated track $\mathfrak{X}_j$, $|\widehat{TP}_j|$ and the number of frames where the corresponding ground-truth track $\mathfrak{X}_i$ exists, $\bar{K}_i$:

$$\text{TDR}_i = \frac{|\widehat{TP}_j|}{\bar{K}_i}. \quad (11)$$

The evaluation of the consistency of the IDs of targets is provided in the form of TF [17]:

$$\text{TF}_i = |IDC_i|, \quad (12)$$

where $|IDC_i|$ is the number of ID changes with respect to the ground-truth track i, measured as the number of times a ground-truth track i is associated with different estimated tracks. The association between estimated and ground-truth tracks is determined as for OTE.

C. RAP evaluation

RAP measures use a region-based assignment and provide a position-based evaluation. Examples of RAP measures include $\widehat{TP}$ track matches, $\widehat{FP}$ track matches and $\widehat{FN}$ track matches.

The computation of $\widehat{TP}$ track matches, $\widehat{FP}$ track matches and $\widehat{FN}$ track matches [19] is based on the spatial and temporal overlaps between estimated and ground-truth tracks and involves performing the assignment implicitly. If the estimated track j overlaps any ground-truth track i both spatially and temporally, the estimation is considered a $\widehat{TP}$ track match. A spatial overlap is achieved in a frame when the centroid of the estimated track j coincides with the corresponding bounding box of the ground-truth track i. At track level, it is measured for each ground-truth track as the percentage of frames having coincidence between estimated and ground-truth bounding boxes. For a $\widehat{TP}$ match, the temporal overlap, $\bar{O}_{tp}$, between the estimated track j and the corresponding ground-truth track i is defined as

$$\bar{O}_{tp} = \frac{\mathcal{N}_{i,j}^{ov}}{K_i}, \quad (13)$$

where $\mathcal{N}_{i,j}^{ov}$ is the number of concurrent frames between ground-truth track i and estimated track j, i.e. the frames where i and j exist. If the spatial or temporal overlap of the estimated track j with any ground-truth track i is smaller than a threshold τ_2, the estimation is considered to be a $\widehat{FP}$ track match. For a $\widehat{FP}$ match, the temporal overlap, $\bar{O}_{fp}$, between the estimated track j and the corresponding ground-truth track i is defined as

$$\bar{O}_{fp} = \frac{\mathcal{N}_{i,j}^{ov}}{K_j}. \quad (14)$$

Given all estimated tracks, if the spatial or temporal overlap of the ground-truth track i with any estimated track j is smaller than a threshold $\bar{\tau}_2$, the estimation is considered a $\widehat{FN}$ match. For a $\widehat{FN}$ match, the temporal overlap $\bar{O}_{fn}$ is computed as for $\bar{O}_{tp}$ (Eq. 13), i.e. $\bar{O}_{fn} = \bar{O}_{tp}$.

D. RAS evaluation

RAS measures use a region-based assignment and provide tracking evaluation that also takes into account target-size changes over time (size-based evaluation). Examples of RAS measures include Correct Detected Track (CDT), False Alarm Track (FAT), Track Detection Failure (TDF), Multiple Object

Tracking Precision (MOTP), Multiple Object Detection Accuracy MODA, Normalized MODA, Multiple Object Tracking Accuracy (MOTA) and ID changes (IDC).

CDT, FAT and TDF [18] are conceptually similar to $\widehat{TP}$, $\widehat{FP}$ and $\widehat{FN}$ tracks, respectively, as defined in [19]. However, unlike [19], in which the spatial overlap is based on the coincidence of bounding boxes, spatial overlap is defined using the number of common pixels between estimated and ground-truth bounding boxes. This implies that CDT, FAT and TDF include also the variations of target sizes in the evaluation. However, they do not individually evaluate the cardinality error.

For MOTP, MODA, Normalized MODA and MOTA, a one-to-one assignment is achieved at frame level between estimated and ground-truth tracks based on the maximization of spatial overlap values (computed as for CDT, FAT and TDF) between pairs using the Hungarian algorithm [8], [27].

MOTP [8] is a spatio-temporal measure that computes the amount of overlap between estimated and ground-truth tracks:

$$\text{MOTP} = \frac{\sum_{t=1}^{n_m} \sum_{k=\hat{k}_{ini}^t}^{\hat{k}_{end}^t} \frac{|\bar{A}_k^t \cap A_k^t|}{|\bar{A}_k^t \cup A_k^t|}}{\sum_{k=1}^{K} n_m^k}, \tag{15}$$

where n_m is the number of associated estimated and ground-truth track pairs in the sequence, $|\bar{A}_k^t \cap A_k^t|$ is the number of common pixels in $\bar{A}_k^t$ and A_k^t, $|\bar{A}_k^t \cup A_k^t|$ is the number of pixels in $\bar{A}_k^t \cup A_k^t$, and n_m^k is the number of associated estimated and ground-truth target pairs at frame k. The pairs with an overlap greater than a fixed threshold value τ_o are considered in the evaluation procedure.

MODA_k [8] computes tracking performance at frame k by combining the information about the number of false negative estimations $|\widehat{FN}_k|$ and the number of false positive estimations $|\widehat{FP}_k|$:

$$\text{MODA}_k = 1 - \frac{c_1|\widehat{FN}_k| + c_2|\widehat{FP}_k|}{v_k}, \tag{16}$$

where c_1 and c_2 are fixed *a priori*. $\widehat{FP}_k$ and $\widehat{FN}_k$ are determined by comparing the amount of overlap between estimated and corresponding ground-truth targets with the threshold τ_o. Note that MODA is not numerically lower bounded. For example, let $c_1 = c_2 = 1$, $|\widehat{FN}_k| = 2$, $|\widehat{FP}_k| = 6$ and $v_k = 6$; hence $\text{MODA}_k = -0.33$. As $|\widehat{FN}_k|$ and/or $|\widehat{FP}_k|$ increase, MODA_k keeps decreasing without lower bound (Fig. 3). A sequence-level formulation of MODA, the Normalized MODA (N-MODA) [8], is defined as

$$\text{N-MODA} = 1 - \frac{\sum_{k=1}^{K}(c_1|\widehat{FN}_k| + c_2|\widehat{FP}_k|)}{\sum_{k=1}^{K} v_k}. \tag{17}$$

Unlike MODA, MOTA [8] is a sequence-level measure that evaluates tracking performance by including also the information about the number of ID switches ($|IDS_k|$) in each frame, in addition to $|\widehat{FN}_k|$ and $|\widehat{FP}_k|$. The contributions of $|\widehat{FN}_k|$, $|\widehat{FP}_k|$ and $|IDS_k|$ are determined by manually setting the corresponding three application-dependent parameters, c_1, c_2 and c_3, respectively. The contributions are accumulated across the sequence and normalized as follows:

$$\text{MOTA} = 1 - \frac{\sum_{k=1}^{K}(c_1|\widehat{FN}_k| + c_2|\widehat{FP}_k| + c_3|IDS_k|)}{\sum_{k=1}^{K} v_k}, \tag{18}$$

where $\widehat{FP}_k$ and $\widehat{FN}_k$ are determined as in MODA and, as with MODA, it is not numerically lower bounded. For example, let $c_1 = c_2 = c_3 = 1$ and $k = 1, 2$; at $k = 1$, $|\widehat{FN}_1| = 0, |\widehat{FP}_1| = 2, |IDS_1| = 0$, $v_1 = 3$; at $k = 2$, $|\widehat{FN}_2| = 0$, $|\widehat{FP}_2| = 5$, $|IDS_2| = 2$, $v_2 = 3$; hence, $\text{MOTA} = -0.50$.

IDC [18] counts the number of ID changes corresponding to all ground-truth tracks. At each frame, each estimated bounding box is assigned to the ground-truth bounding box with an overlap larger than a predefined threshold. When the amount of overlap for an estimated track and ground-truth track pair falls below the threshold, an ID change is considered to have occurred.

E. Discussion

Table I compares the state-of-the-art multi-target tracking evaluation measures. Existing frame-level measures do not take into account the evaluation of target-size changes [16], [17], [25] and require presetting application-dependent parameters [8], [16]. Additionally, frame-level measures ignore the cardinality error [17], [25]. Sequence-level measures do not evaluate target-size changes (e.g. OTE, TDR [17] and the measures presented in [19]) and use application-dependent thresholds (e.g. MOTA, MOTP [8] and the measures presented in [18]). These measures aim to evaluate the accuracy only while not considering the cardinality error [8], [17]–[19]. Existing sequence-level measures are generally not employed to analyze tracking at varying accuracy levels, which would be desirable and useful to determine the suitability of trackers for different applications or scenarios. ID-change evaluation measures simply incorporate the information about the total number of ID changes or switches in the sequence [8], [17], [18]; however it would be desirable to evaluate ID changes relative to the track duration.

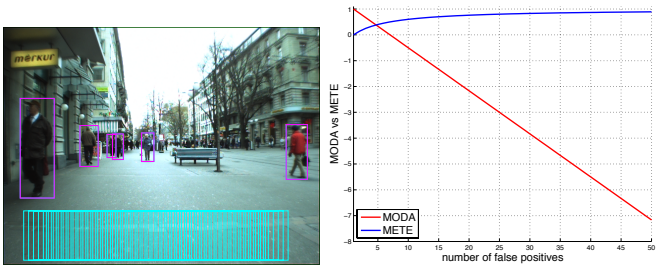
We address the drawbacks of current evaluations and propose three measures, namely the Multiple Extended-target Tracking Error (METE), the Multiple Extended-target Lost-Track ratio (MELT) and the Normalized ID Changes (NIDC).

III. Tracking error measures

A. Multiple extended-target tracking error

The proposed overlap-based Multiple Extended-target Tracking Error (METE) measure combines accuracy and cardinality errors in a parameter-independent manner. The use of spatial overlap information in METE eliminates the need to include the OSPA parameters (Eq. 7), namely the penalty for the estimated states located far away from any of the ground-truth states (p) and the cut-off parameter defining the upper bound (c).

The accuracy error, $\mathcal{A}_k$, represents the extent of the mismatch between estimated and ground-truth states at frame k



(a) (b)

Fig. 3. Unbounded nature of MODA. (a) Sample frame from ETH Bahnhof [31] (six targets). Ground truth and tracker's estimates are shown as magenta and cyan bounding boxes, respectively. The estimated bounding boxes are perfectly overlapping with corresponding ground-truth bounding boxes. By gradually increasing false positives (placed at the bottom of the frame) from the perfectly overlapping scenario, we compute the corresponding MODA and METE values as shown in (b). MODA continues decreasing without lower bound (Eq. 16), whereas METE $\in [0,1]$.

and is defined as

$$\mathcal{A}_k = \min_{\pi \in \Pi_{max(v_k,u_k)}} \sum_{i=1}^{min(v_k,u_k)} \left(1 - O(\bar{A}_{k,i}, A_{k,\pi(i)})\right), \quad (19)$$

where $O(\bar{A}_{k,i}, A_{k,\pi(i)}) = \frac{|\bar{A}_{k,i} \cap A_{k,\pi(i)}|}{|\bar{A}_{k,i} \cup A_{k,\pi(i)}|}$ computes the amount of spatial overlap between $\bar{A}_{k,i}$ and $A_{k,\pi(i)}$; and $O(\cdot) \in [0,1]$ [7], like in Eq. 15. Without loss of generality, we consider here $\bar{A}_{k,i}$ and $A_{k,\pi(i)}$ to be bounding boxes. $\Pi_{max(v_k,u_k)}$ is the set of permutations, each containing $min(v_k,u_k)$ elements, drawn from $\{1,2,\ldots,max(v_k,u_k)\}$. The permutation minimizing the summation term in Eq. 19 establishes the association between estimated and ground-truth states and contributes to the computation of the accuracy error at frame k. This minimization is performed by the Hungarian algorithm [27]. $\mathcal{A}_k \in [0, u_k = v_k]$ when $u_k = v_k$; $\mathcal{A}_k \in [0, v_k]$ when $u_k > v_k$ (i.e. the association is performed only for the v_k terms); and $\mathcal{A}_k \in [0, u_k]$ when $u_k < v_k$ (i.e. the association is performed only for the u_k terms). For the cases when $u_k > v_k$ and $u_k < v_k$, $\mathcal{A}_k$ does not take into account the discrepancy between u_k and v_k (i.e. the unassociated targets), and the accuracy error is in fact computed for the associated pairs only. This justifies the computation of the cardinality error, $\mathcal{C}_k$, namely the discrepancy in estimating the number of targets:

$$\mathcal{C}_k = |u_k - v_k|. \quad (20)$$

We combine $\mathcal{C}_k$ with $\mathcal{A}_k$ to account for the unassociated targets in the evaluation procedure (in OSPA [16], [29]) and to provide a single-score performance evaluation at frame level. METE is therefore computed as:

$$\text{METE}_k = \frac{\mathcal{A}_k + \mathcal{C}_k}{max(v_k,u_k)}, \quad (21)$$

$\text{METE}_k \in [0,1]$: the lower METE_k, the better the tracking result. We explain below the bounds of the measure, where $\text{METE}_k = 0$ for the best tracking case and $\text{METE}_k = 1$ for the worst tracking case.

Best tracking case: $\mathcal{A}_k = 0$: $O(\cdot) = 1$ for all the associated pairs (Eq. 19), and $\mathcal{C}_k = 0$ since $u_k = v_k$ (Eq. 20). This implies $\text{METE}_k = 0$, using Eq. 21.

TABLE I
COMPARISON OF MULTI-TARGET TRACKING EVALUATION MEASURES. KEY: PI: PARAMETER INDEPENDENCE; SE: SIZE-CHANGE EVALUATION; APS: ASSIGNMENT PROBLEM SOLUTION; P: POINT-BASED; R: REGION-BASED; F: FRAME-LEVEL MEASURE; S: SEQUENCE-LEVEL MEASURE; AE: ACCURACY ERROR; CE: CARDINALITY ERROR; PROP.: PROPOSED; $\widehat{TP}_m$: $\widehat{TP}$ MATCHES; $\widehat{FP}_m$: $\widehat{FP}$ MATCHES; $\widehat{FN}_m$: $\widehat{FN}$ MATCHES.

Measure	Ref.	PI	SE	APS	Type	AE	CE
OSPA	[16]			P	F	✓	✓
$W_p(\cdot)$	[25]	✓		P	F	✓	
OTE	[17]	✓		P	S	✓	
TRDR	[17]	✓		P	F	✓	
FAR	[17]	✓		P	F	✓	
TDR	[17]	✓		P	S	✓	
TF	[17]	✓		P	S		
$\widehat{TP}_m$	[19]			R	S	✓	
$\widehat{FP}_m$	[19]			R	S	✓	
$\widehat{FN}_m$	[19]			R	S	✓	
CDT	[18]		✓	R	S	✓	
FAT	[18]		✓	R	S	✓	
TDF	[18]		✓	R	S	✓	
IDC	[18]		✓	R	S		
MODA	[8]		✓	R	F	✓	✓
N-MODA	[8]		✓	R	S	✓	✓
MOTA	[8]		✓	R	S	✓	✓
MOTP	[8]		✓	R	S	✓	
METE	Prop.	✓	✓	R	F	✓	✓
MELT	Prop.	✓	✓	R	S	✓	
NIDC	Prop.	✓	✓	R	S		

Worst tracking case: $\mathcal{A}_k$ has its maximum value, i.e. $\mathcal{A}_k = u_k = v_k$ when $u_k = v_k$, $\mathcal{A}_k = v_k$ when $u_k > v_k$ (the association is performed only for the v_k terms) and $\mathcal{A}_k = u_k$ when $u_k < v_k$ (the association is performed only for the u_k terms). Thus the numerator of Eq. 21 becomes $\mathcal{A}_k + \mathcal{C}_k = v_k = u_k : u_k = v_k$ meaning $\mathcal{C}_k = 0$; $\mathcal{A}_k + \mathcal{C}_k = v_k + |u_k - v_k| = u_k : u_k > v_k$; $\mathcal{A}_k + \mathcal{C}_k = u_k + |u_k - v_k| = v_k : u_k < v_k$. Therefore, $\mathcal{A}_k + \mathcal{C}_k = max(v_k,u_k)$, which implies $\text{METE}_k = 1$, using Eq. 21. The other tracking cases lie within the two bounds of METE as shown in Fig. 3.

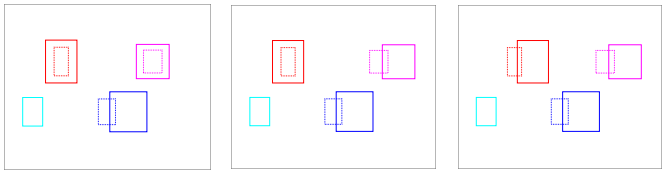
As the same METE values for two trackers may be caused by different accuracy and cardinality error combinations, it may be useful to analyze these errors separately in order to determine their individual influence in the estimation of METE. To this end, we use two error rates, the Accuracy Error Rate (AER):

$$\text{AER} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{A}_k \quad (22)$$

and the Cardinality Error Rate (CER):

$$\text{CER} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{C}_k. \quad (23)$$

Unlike OSPA, METE evaluates changes in the size of extended targets; and unlike MODA, METE is numerically bounded between 0 and 1 (Fig. 3) and parameter-independent. Indeed, the parameter dependence of MODA may not always enable it to distinguish different tracking results (Fig. 4).



MODA = -1.33
METE = 0.81
(a)

MODA = -1.33
METE = 0.88
(b)

MODA = -1.33
METE = 0.94
(c)

Fig. 4. Example of limitations of Multiple Object Detection Accuracy (MODA) [8]: although clearly different, the three cases are not distinguished by MODA [4]. Ground-truth and estimated boxes are shown as dotted and solid lines, respectively. The proposed measure METE can instead distinguish the three cases.

B. Multiple extended-target lost-track ratio

The proposed Multiple Extended-target Lost-Track ratio (MELT) evaluates tracking accuracy across the sequence in a parameter-independent manner and enables analysis at different levels of accuracy. Given $\bar{\mathcal{X}}$ and $\mathcal{X}$, the association is first performed at each frame based on the minimization of the cost $(1-O(\cdot))$ computed for all pairs of estimated and ground-truth targets. Similarly to Eq. 19, the minimization process uses the Hungarian algorithm. The procedure yields a unique assignment at frame level, whereas at track level a ground-truth track may be associated with more than one estimated track due to fragmentations and/or ID changes.

We evaluate accuracy at track level by computing the lost-track ratio (λ_i^τ) for each associated pair of ground-truth track i and estimated track(s) as follows [7]:

$$\lambda_i^\tau = \frac{N_i^\tau}{N_i}, \tag{24}$$

where N_i^τ is the number of frames with spatial overlap $O(\cdot) \leq \tau : \tau \in \mathbb{R}_{(0,1]}$ between the associated pair and N_i is the total number of frames in the ground-truth track i. $\lambda_i^\tau \in [0,1]$; the lower λ_i^τ, the better the performance. We compute the lost-track ratio for a range of a finite number of τ values and obtain $\lambda_i(\tau) = \{\lambda_i^\tau\}_{\tau \in \mathbb{R}_{(0,1]}}$ such that the total number of sampled τ values is S_τ (required for numerical approximation). We compute $\lambda_i(\tau)$ for all V ground-truth tracks to generate the matrix Λ:

$$\Lambda = [\lambda_i^\tau]_{V \times S_\tau}, \tag{25}$$

where V and S_τ are the number of rows and columns of the matrix, respectively. We quantify tracking performance by defining the Multiple Extended-target Lost-Track ratio (MELT$_\tau$):

$$\text{MELT}_\tau = \frac{1}{V} \sum_{i=1}^{V} \lambda_i^\tau, \tag{26}$$

which provides tracking performance at τ $s.t.$ MELT$_\tau \in [0,1]$. The lower MELT$_\tau$, the better the performance. In order to enable the analysis of tracking performance at different accuracy levels, we compute MELT$_\tau$ for different τ values (Fig. 5(c), 5(d)). While the computation of MELT$_\tau$ may be useful from an application viewpoint, the performance comparison among trackers can be facilitated by providing the single-score average tracking performance which is generated as

$$\text{MELT} = \frac{1}{S_\tau} \sum_{\tau \in \mathbb{R}_{(0,1]}} \text{MELT}_\tau. \tag{27}$$

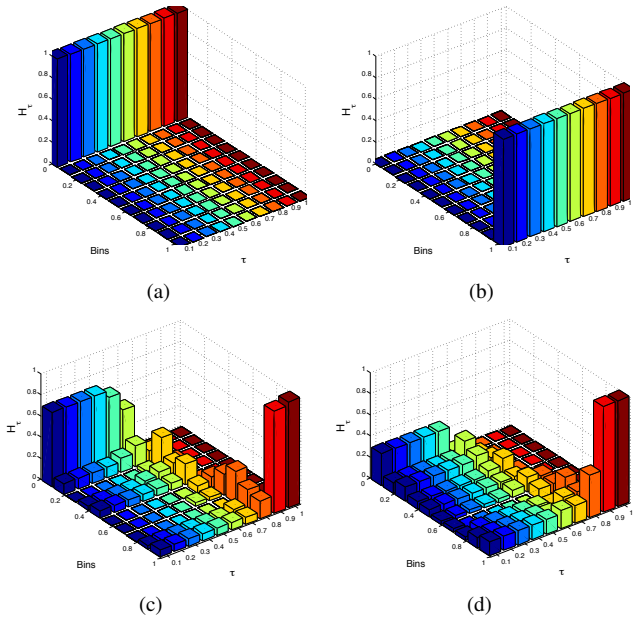


Fig. 5. The probability density function H_τ for a variation of τ values. (a) Ideal tracking result: the lost-track ratio is zero for all tracks at all the values of τ; hence, MELT $= 0$. (b) Worst tracking result: the lost-track ratio is 1 for all tracks at all values of τ; hence, MELT $= 1$. (c-d) MELT and MOTP of the Conditional Random Field based tracker (CRFBT) [6] and the Dynamic Programming-Non-Maxima Suppression based tracker (DP-NMS) [32] on ETH Sunnyday [31]; (c) CRFBT: MELT=0.39, MOTP=0.75; (d) DP-NMS: MELT=0.56, MOTP=0.77.

The performance of a tracker at a particular accuracy level, τ, can be presented by plotting the probability density function, H_τ, of the corresponding lost-track-ratio values (i.e. the values in the column τ of the Λ-matrix (Eq. 25)). Each sample of H_τ represents the percentage of tracks with a particular lost-track-ratio (bin) at a specific value of τ. Bins are the equal-width intervals created by dividing the range of λ_i^τ, where $\lambda_i^\tau \in [0,1]$. Fig. 5 shows examples of H_τ plotted while varying the τ values. The higher the concentration of λ_i^τ values towards bin zero, the better the corresponding tracking performance at τ. Fig. 5(a) shows an ideal tracking result with zero lost-track ratio value for all $\bar{\mathfrak{X}}_i$ at all τ. Similarly, Fig. 5(b) is the worst tracking result: the lost-track ratio is 1 for all $\bar{\mathfrak{X}}_i$ at all τ. Figures 5(c), 5(d) show the results of the Conditional Random Field based tracker (CRFBT) [6] and the Dynamic Programming-Non-Maxima Suppression based tracker (DP-NMS) [32] on ETH Sunnyday [31] using MELT and MOTP. MELT considers CRFBT to be better than DP-NMS and this can be seen from the highest concentration of values of CRFBT in the bins towards zero in Fig. 5(c). Consequently, MELT$_\tau$ values of CRFBT and DP-NMS computed for the variation of τ (Fig. 6(c)) show that CRFBT outperforms DP-NMS. The values of MELT$_\tau$ of CRFBT are lower for all τ than those of DP-NMS, meaning lower lost-track-ratio values and better tracking accuracy. On the other hand, MOTP ranks the performance of two trackers differently (i.e. opposite)

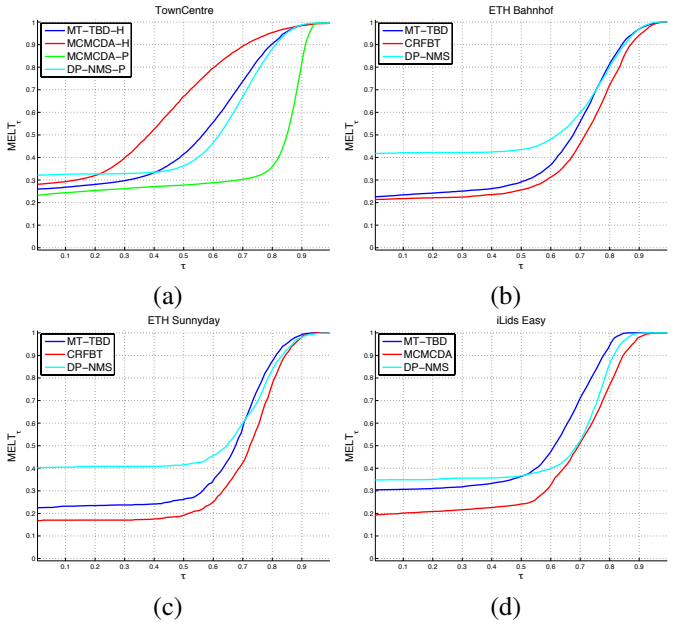
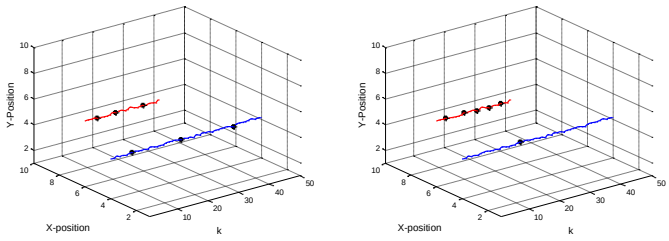


Fig. 6. Evaluation of trackers' results at varying levels of accuracy (τ) using MELT_τ on all sequences. (a) MELT_τ of trackers on TownCentre sequence. 'H' and 'P' in the legend indicate the use of a tracker for head or person tracking, respectively; (b) MELT_τ of trackers on ETH Bahnhof sequence; (c) MELT_τ of trackers on ETH Sunnyday; and (d) MELT_τ of trackers on iLids Easy sequence.



$\text{NIDC}_1 = 0.12$, $\text{TF}_1 = 3$
$\text{NIDC}_2 = 0.06$, $\text{TF}_2 = 3$
$\text{NIDC} = 0.09$, $\text{IDC} = 6$
(a)

$\text{NIDC}_1 = 0.20$, $\text{TF}_1 = 5$
$\text{NIDC}_2 = 0.02$, $\text{TF}_2 = 1$
$\text{NIDC} = 0.11$, $\text{IDC} = 6$
(b)

Fig. 7. Comparison of the proposed Normalized ID Changes (NIDC) measure with Track Fragmentation (TF) [17] and ID Changes (IDC) [18]. (a) and (b) present results of two different trackers in terms of ID changes on the same sequence. Each example shows two ground-truth tracks; ID=1: red ground-truth track; ID=2: blue ground-truth track. ID changes are shown as black dots. (a) The length of the red track ($\text{IDC}_1^{max} = 25$) is shorter than that of the blue track ($\text{IDC}_2^{max} = 50$) and $|IDC_1| = |IDC_2| = 3$. Thus, $\text{NIDC}_1 = 0.12$ and $\text{NIDC}_2 = 0.06$ penalize the red track (shorter) more for the occurrence of the same number of ID changes as the blue track. However, $\text{TF}_1 = \text{TF}_2 = 3$ considers both the cases to be the same. (b) NIDC and TF can distinguish the different ID changes of the two tracks. The IDC measure considers (a) and (b) as the same cases since $\text{IDC} = 6$ for both, whereas NIDC can distinguish (a) and (b).

because it does not take into account the overlap values of the estimated and ground-truth track pairs that are smaller than τ_o, thereby not including the complete tracking accuracy in the assessment. MELT provides a holistic performance assessment taking into account all of the tracking information.

MELT also summarizes tracking performance at different accuracy levels and provides an insight for analysis. For example, consider the MELT_τ plots of DP-NMS and the multi-target track-before-detect (MT-TBD) tracker [33] shown in Fig. 6(c). MELT_τ shows that for $\tau < 0.72$ (approx.), MT-TBD outperforms DP-NMS, after which DP-NMS outperforms MT-TBD. This analysis can be useful in selecting between these two trackers for an application that requires tracking with average overlap (accuracy) of e.g. 80%: DP-NMS would be a more suitable choice than MT-TBD.

C. Normalized ID changes

The proposed Normalized ID Changes (NIDC) measure evaluates the ID changes taking into account the track duration in which they occur. In the case of a comparison of trackers producing tracks of different lengths, the normalization of ID changes is preferable to simply counting the ID changes. Such quantification emphasizes the long-term tracking ability with unique IDs of trackers. Moreover, since the score is normalized it can be more useful than the number of ID changes to compare trackers across different datasets. Unlike IDC [18] and MOTA [8], NIDC is parameter independent since its assignment solution used for detecting ID changes is based on that used in Sec. III-A (Eq. 19).

Let V_{IDC} be the number of ground-truth tracks with at least one ID change and

$$\text{NIDC}_i = \frac{|IDC_i|}{\text{IDC}_i^{max}} \tag{28}$$

be the NIDC value for ground-truth track i and IDC_i^{max} the maximum number of ID changes that can occur for ground-truth track i (i.e. the length of track i). NIDC_i includes a contribution of ID changes for track i that is scaled by IDC_i^{max}, which is proportional to the duration of track i. This penalizes the ID changes by the length of the track in the estimation of NIDC, instead of simply relying on counting ID changes [8], [17], [18]. NIDC quantifies the number of ID changes corresponding to all ground-truth tracks of the sequence:

$$\text{NIDC} = \frac{1}{V_{IDC}} \sum_{i=1}^{V} \text{NIDC}_i, \tag{29}$$

where $\text{NIDC} \in [0, 1]$. The lower NIDC, the better the performance in terms of ID maintenance.

Figure 7 shows two examples that compare NIDC with TF [17] and IDC [18]. The red ground-truth track (ID=1) and the blue ground-truth track (ID=2) shown in Fig. 7(a) have different lengths ($\text{IDC}_1^{max} = 25$ and $\text{IDC}_2^{max} = 50$) but have the same number of ID changes ($|IDC_1| = |IDC_2| = 3$). $\text{NIDC}_1 = 0.12$ is larger than $\text{NIDC}_2 = 0.06$ since the measure penalizes the red track (shorter length) for the occurrence of the same number of ID changes. Unlike NIDC, TF does not distinguish these two cases as $\text{TF}_1 = 3$ and $\text{TF}_2 = 3$, as it does not consider track length. Both NIDC and TF are able to distinguish ID changes of two tracks in Fig. 7(b) as is shown in their listed values. Moreover, the ID changes of two different trackers are shown for the same sequence in Fig. 7(a) and Fig. 7(b), respectively. While IDC does not distinguish the results of the two trackers as $\text{IDC} = 6$ for both of them,

NIDC differentiates between them (NIDC $= 0.09$ for (a) and NIDC $= 0.11$ for (b)).

IV. Experimental validation and analysis

We validate the effectiveness of the proposed measures by comparing them with state-of-the-art measures and by evaluating the performance of recently proposed trackers on real-world publicly-available datasets.

A. Experimental setup

We use four real-world datasets, namely TownCenter [4], ETH Bahnhof [31], ETH Sunnyday [31] and iLids Easy [20]. The datasets contain a high density of targets with occlusions. *TownCentre*, recorded from an overhead static camera, is composed of 4491 frames of size 1920×1080 pixels recorded at 25 fps. The ground truth has 231 head/person-tracks with an average of 16 people per frame. ETH *Bahnhof* and *Sunnyday*, recorded from a human-height moving camera, are composed of 999 and 354 frames, respectively, with a frame size of 640×480 recorded at 14 fps. The ground truth of Bahnhof has 95 person-tracks with an average of eight people per frame, while that of Sunnyday has 30 person-tracks with an average of five people per frame. *iLids Easy* is composed of 5220 frames of size 720×576 pixels recorded at Westminster subway station (London, UK) at 25 fps. The ground truth has 17 person-tracks with an average of 1.9 people per frame.

We use four state-of-the-art trackers in the experimental validation including a combination of the Kanade-Lucas-Tomasi tracker [34] with Markov-Chain Monte-Carlo Data Association (MCMCDA) algorithm [4], a data association algorithm with the online learned Conditional Random Field Based Tracker (CRFBT) [6], a Multi-Target Track-Before-Detect (MT-TBD) with a post-processing stage [33], and the Dynamic Programming Non-Maxima Suppression based tracker (DP-NMS) [32]. Tracking includes head and person (full-body) tracks from both static and moving cameras. DP-NMS is tested on TownCentre, ETH Bahnhof and Sunnyday, and iLids Easy sequences for person tracking. MT-TBD is used for head tracking on the TownCentre sequence and for person tracking on the ETH Bahnhof and Sunnyday, and iLids Easy sequences. MCMCDA is used for head tracking on TownCentre and for person tracking on TownCentre and iLids Easy sequences. CRFBT is tested on ETH Bahnhof and Sunnyday sequences for person tracking. Table II summarizes the datasets and lists the trackers used on the respective sequences. The parameter values of all trackers are those used in the original papers. For the computation of N-MODA, we use $\tau_o = 0.50$ in the case of person tracking and $\tau_o = 0.25$ in the case of head tracking, as done in [4].

TABLE II
Summary of the datasets used. Key: FS: Frame Size; NF: Number of Frames; MC: Moving Camera; NT: Number of Tracks; ANPPF: Average Number of People Per Frame; TM: Tracking Methods tested; C: Crowdness; O: Occlusions; VS: Variable Speed; IC: Illumination Changes; SC: Scale Changes.

Dataset	FS	NF	Challenges	MC	NT	ANPPF	TM
TownCentre	1920×1080	4491	C,O,VS,SC		231	16	[4], [32], [33]
ETH Bahnhof	640×480	999	C,O,IC,SC	✓	95	8	[6], [32], [33]
ETH Sunnyday	640×480	354	C,O,IC,SC	✓	30	5	[6], [32], [33]
iLids Easy	720×576	5220	O,VS,IC,SC		17	1.9	[4], [32], [33]

B. Comparison of measures

We compare the proposed METE, MELT and NIDC measures with relevant state-of-the-art measures, namely N-MODA, MOTP and IDC. Table III shows the scores of all measures obtained for all trackers.

The evaluation results using METE and N-MODA on TownCentre with head tracking (TownCentre-H) and with person tracking (TownCentre-P), and on Sunnyday show an agreement between both measures in terms of the relative ranking of trackers. However, there are disagreements on Bahnhof and iLids Easy. On Bahnhof, N-MODA of DP-NMS and MT-TBD are the same. This is because the normalization in N-MODA formulation (Eq. 17) is with respect to the number of false positives and false negatives of tracking only and it does not consider the number of true positives. Since the total number of false positives and false negatives for DP-NMS (3525) and MT-TBD (3514) is comparable, their N-MODA is comparable. Interestingly, the number of true positives for DP-NMS and MT-TBD are 5030 and 6222, respectively. On the other hand, METE ranks MT-TBD higher than DP-NMS since it implicitly takes into account true positives, false positives and false negatives. On iLids Easy, N-MODA ranks MT-TBD as the best tracker, which is not consistent with METE that ranks MCMCDA as the best. N-MODA shows the best performance for MT-TBD because the total number of its false positives and false negatives (3639) is smaller than that of DP-NMS (3843) and MCMCDA (3698). METE, as discussed above, ranks their performance effectively by considering also true positives (in addition to false positives and false negatives) that are 6632, 6705 and 7969 for DP-NMS, MT-TBD and MCMCDA, respectively.

While MELT and MOTP agree on their relative ranking of trackers on TownCentre-H and TownCentre-P, they disagree on the remaining sequences (Tab. III). In the case of Bahnhof, MOTP of MT-TBD and DP-NMS are the same, whereas MELT ranks MT-TBD higher than DP-NMS. The MELT$_\tau$ plots also show a better performance of MT-TBD for most of the variations of τ than DP-NMS (Fig. 6(b)). The disagreement of MOTP is due to its dependence on the threshold value τ_o. MOTP considers only the overlap values of pairs greater than τ_o, which may lead to the exclusion of some tracking information in the performance assessment. On the other hand, MELT uses all of the tracking information in the performance assessment to present a comprehensive performance evaluation that can more effectively reflect the trackers' comparison. In the case of Sunnyday, there is a disagreement between MELT and MOTP in selecting the best tracker, as already discussed in Sec. III-B. In the case of iLids Easy, MOTP of DP-NMS and MCMCDA are comparable; however, based on their MELT scores and MELT$_\tau$ plots (Fig. 6(d)), the difference in their performance is clear. The inconsistencies of MOTP in Sunnyday and iLids Easy are due to its parameter dependency.

NIDC and IDC agree in their relative evaluation of trackers on TownCentre, Bahnhof and iLids Easy (Tab. III). The effectiveness of NIDC can be noticed in the case of Sunnyday.

TABLE III

OVERALL COMPARISON OF TRACKERS ON DIFFERENT DATASETS WITH DIFFERENT EVALUATION MEASURES. THE COLORED CELLS INDICATE THE TRACKER'S PERFORMANCE: THE DARKER THE COLOR, THE BETTER THE PERFORMANCE. KEY: TOWNCENTRE-H: HEAD TRACKING PERFORMED ON TOWNCENTRE SEQUENCE; TOWNCENTRE-P: PERSON TRACKING PERFORMED ON TOWNCENTRE SEQUENCE; METE: MULTIPLE EXTENDED-TARGET TRACKING ERROR; MELT: MULTIPLE EXTENDED-TARGET LOST TRACK RATIO; NIDC: NORMALIZED ID CHANGES; AER: ACCURACY ERROR RATE; CER: CARDINALITY ERROR RATE; MLT: MEAN LENGTH OF GROUND-TRUTH TRACKS HAVING ID CHANGE(S); N-MODA: NORMALIZED MULTIPLE OBJECT DETECTION ACCURACY; MOTP: MULTIPLE OBJECT TRACKING PRECISION; IDC: ID CHANGES; μ: MEAN VALUE OVER THE SEQUENCE; σ: STANDARD DEVIATION OF VALUES OVER THE SEQUENCE IN THE CASE OF METE, AND STANDARD DEVIATION OF VALUES OF ACCURACY ERROR ($\mathcal{A}$) AND CARDINALITY ERROR ($\mathcal{C}$) OVER THE SEQUENCE FOR AER AND CER, RESPECTIVELY.

Tracker	Dataset	METE $\mu(\sigma)$	MELT	NIDC	AER (σ)	CER (σ)	N-MODA	MOTP	IDC	MLT
MT-TBD [33]	TownCentre-H	0.53 (0.08)	0.54	0.031	6.82 (2.54)	2.14 (1.92)	0.55	0.64	1798	320.00
MCMCDA [4]		0.62 (0.07)	0.65	0.038	8.48 (2.74)	1.82 (1.62)	0.46	0.51	1913	330.12
DP-NMS [32]	TownCentre-P	0.48 (0.08)	0.53	0.043	5.06 (1.52)	2.67 (2.02)	0.58	0.71	2637	321.61
MCMCDA [4]		0.33 (0.09)	0.37	0.030	3.64 (1.54)	1.81 (1.62)	0.62	0.80	1519	336.44
DP-NMS [32]	ETH Bahnhof	0.53 (0.13)	0.57	0.039	1.45 (0.69)	3.07 (1.85)	0.58	0.75	229	109.92
MT-TBD [33]		0.44 (0.12)	0.46	0.050	2.42 (1.19)	1.56 (1.34)	0.58	0.75	307	103.51
CRFBT [6]		0.39 (0.12)	0.42	0.035	1.99 (0.86)	1.49 (1.26)	0.68	0.77	158	124.91
DP-NMS [32]	ETH Sunnyday	0.44 (0.11)	0.56	0.042	1.16 (0.55)	1.34 (0.93)	0.66	0.77	43	68.68
MT-TBD [33]		0.47 (0.11)	0.46	0.041	1.60 (0.57)	1.09 (0.84)	0.61	0.73	56	91.50
CRFBT [6]		0.46 (0.12)	0.39	0.028	1.46 (0.52)	1.06 (0.78)	0.63	0.75	31	82.20
DP-NMS [32]	iLids Easy	0.40 (0.26)	0.52	0.011	0.40 (0.36)	0.65 (0.86)	0.60	0.74	104	632.87
MT-TBD [33]		0.53 (0.22)	0.54	0.007	0.50 (0.36)	0.96 (1.10)	0.63	0.70	54	632.87
MCMCDA [4]		0.36 (0.26)	0.43	0.029	0.51 (0.45)	0.51 (0.76)	0.62	0.75	227	605.06

IDC considers the performance of DP-NMS to be better than MT-TBD. NIDC shows a slightly better performance for MT-TBD than DP-NMS despite the fact that the former has produced more ID changes than the latter. This is because NIDC provides ID evaluation while considering also the track length. Since MLT (mean length of ground-truth tracks having ID change(s)) of the MT-TBD is much higher than DP-NMS, NIDC penalizes less the ID changes of the former.

To conclude, unlike METE, the dependence of MODA on the preset overlap threshold limits its ability to clearly distinguish different tracking results (Fig. 8(a-c)); and unlike MELT, the threshold dependency of MOTP may result in an inaccurate evaluation of tracking performance (Fig. 9).

C. Evaluation of trackers

We now discuss the effectiveness of the proposed measures by highlighting strengths and weaknesses of selected trackers.

On TownCentre-H, MT-TBD outperforms MCMCDA using mean METE, MELT, NIDC and AER (see r.[1] 1, 2 in Tab. III), which is also confirmed in the MELT_τ plots (Fig. 6(a)). MT-TBD has a better NIDC than MCMCDA because of its better ID management mechanism, which involves minimizing the mixing of target particles in the Bayesian state estimation [33]. Interestingly, CER differs from the remaining measures and shows better performance for MCMCDA compared to MT-TBD. The higher CER of MT-TBD is due to a greater number of tracking failures or missed targets. Since AER is lower for MT-TBD, this points to fewer occurrences of tracking failures than missed targets.

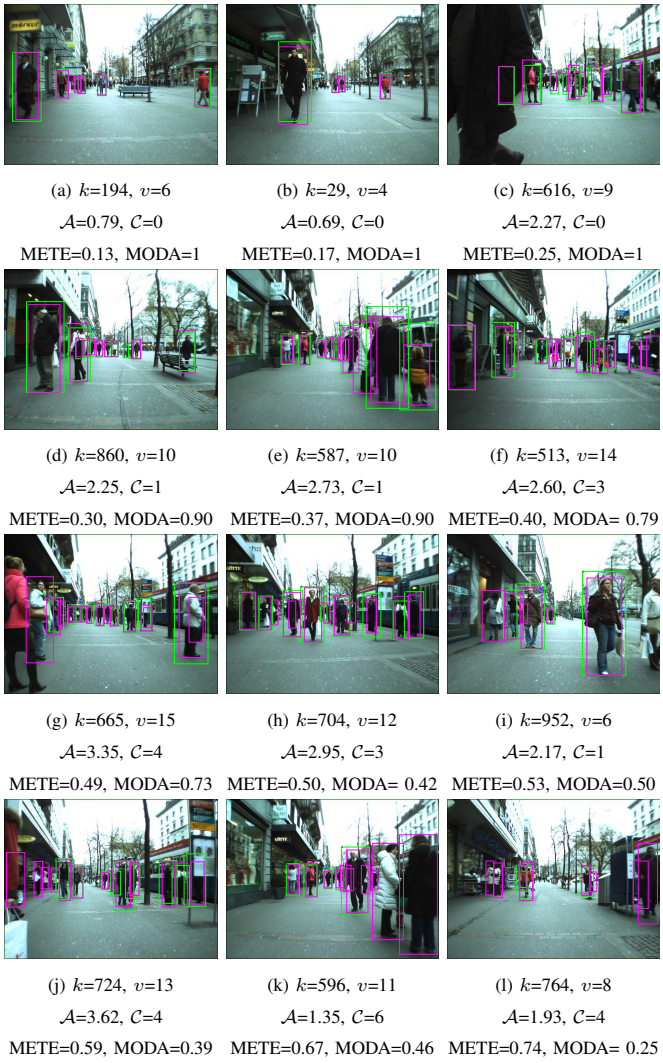
On TownCentre-P, MCMCDA outperforms DP-NMS based on mean METE, MELT, NIDC, AER and CER (see r. 3, 4 in Tab. III). It is also interesting to highlight the clear improvement in the evaluation results of MCMCDA using the proposed measures on TownCentre-P compared to TownCentre-H, which is inline with the results of the original paper [4].

On Bahnhof, mean METE, MELT, NIDC and CER rank CRFBT as the best tracker compared to DP-NMS and MT-TBD (see r. 5, 6, 7 in Tab. III). This is also visible in the MELT_τ plots (Fig. 6(b)). The reason for the best NIDC of CRFBT is its capability in addressing ID changes using motion and appearance 'affinities' [6], enabling it to distinguish and separate nearby targets. There is an inconsistency in the case of AER that ranks CRFBT as second-best tracker after DP-NMS. Furthermore, the CER of DP-NMS is almost twice that of MT-TBD and CRFBT. This is due to the limited capability of DP-NMS, unlike MT-TBD and CRFBT, to link fragmented tracks that increases the cardinality error. The fragmentations in the case of DP-NMS are caused by a worse handling of long-term occlusions compared to MT-TBD and CRFBT (Fig. 10).

On Sunnyday, we tested the same trackers (DP-NMS, MT-TBD and CRFBT) as used on Bahnhof. Some inconsistencies can be noticed in the evaluation results on Sunnyday compared to those on Bahnhof. Firstly, unlike on Bahnhof, the evaluation based on mean METE on Sunnyday shows a better performance of DP-NMS compared to MT-TBD and CRFBT (see r. 8, 9, 10 in Tab. III). This is probably because the person detector [35] used with DP-NMS can better deal with the higher scene brightness in Sunnyday than the detector [36] used with MT-TBD and CRFBT, which results in the improved tracking performance of DP-NMS. Secondly, unlike on Bahnhof, NIDC of MT-TBD is better than DP-NMS on Sunnyday despite the fact that IDC of the former is higher than the latter in both sequences due to the reason discussed in Sec. IV-B.

On iLids Easy, the evaluation of trackers using mean METE and MELT shows the superior performance of MCMCDA compared to DP-NMS and MT-TBD (see r. 11, 12, 13 in Tab. III). The superior mean METE and MELT of MCMCDA over DP-NMS is consistent with their mean METE and

[1] 'r' refers to the row number in Tab. III not considering the row with titles.



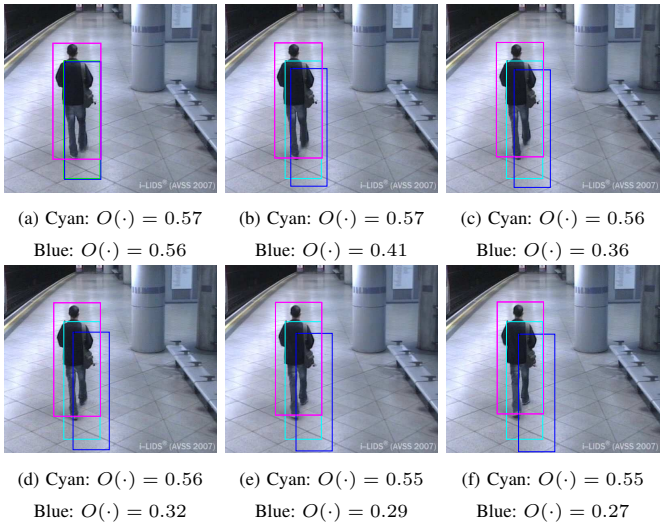
(a) k=194, v=6 $\mathcal{A}$=0.79, $\mathcal{C}$=0 METE=0.13, MODA=1
(b) k=29, v=4 $\mathcal{A}$=0.69, $\mathcal{C}$=0 METE=0.17, MODA=1
(c) k=616, v=9 $\mathcal{A}$=2.27, $\mathcal{C}$=0 METE=0.25, MODA=1
(d) k=860, v=10 $\mathcal{A}$=2.25, $\mathcal{C}$=1 METE=0.30, MODA=0.90
(e) k=587, v=10 $\mathcal{A}$=2.73, $\mathcal{C}$=1 METE=0.37, MODA=0.90
(f) k=513, v=14 $\mathcal{A}$=2.60, $\mathcal{C}$=3 METE=0.40, MODA= 0.79
(g) k=665, v=15 $\mathcal{A}$=3.35, $\mathcal{C}$=4 METE=0.49, MODA=0.73
(h) k=704, v=12 $\mathcal{A}$=2.95, $\mathcal{C}$=3 METE=0.50, MODA= 0.42
(i) k=952, v=6 $\mathcal{A}$=2.17, $\mathcal{C}$=1 METE=0.53, MODA=0.50
(j) k=724, v=13 $\mathcal{A}$=3.62, $\mathcal{C}$=4 METE=0.59, MODA=0.39
(k) k=596, v=11 $\mathcal{A}$=1.35, $\mathcal{C}$=6 METE=0.67, MODA=0.46
(l) k=764, v=8 $\mathcal{A}$=1.93, $\mathcal{C}$=4 METE=0.74, MODA= 0.25

Fig. 8. Evaluation of the results of CRFBT on Bahnhof sequence using METE and MODA. Subscript k is removed from the variables for simplicity in the notation. Ground truth and tracker's estimates are shown as magenta and green bounding boxes, respectively. Results are ordered in terms of ascending METE values.

MELT on TownCentre-P. Moreover, the analysis of MELT$_\tau$ plots (Fig. 6(d)) provides an interesting insight about the performance of MT-TBD and DP-NMS, revealing that MELT$_\tau$ of MT-TBD is better than DP-NMS for $\tau < 0.5$ and the reverse is true thereafter. This suggests that DP-NMS is a more suitable choice for tracking with higher accuracy and MT-TBD should be preferred with lower accuracy since its lost-track-ratio values are smaller at lower τ. Additionally, while CER of DP-NMS is the highest on the rest of the sequences, MT-TBD has the highest CER on iLids Easy. Furthermore, the best NIDC of MT-TBD on iLids Easy, as discussed earlier, is due to its better ID management ability. Interestingly, although MLT of MT-TBD and DP-NMS is the same[2] (see r. 11, 12 in Tab. III), the higher IDC of the latter leads to its inferior NIDC.

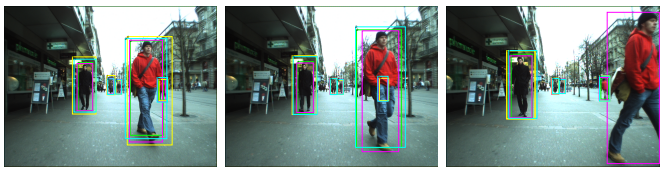
Table III also presents the variation in the performance

[2]The same MLT is because ID change(s) for both trackers have occurred in the same tracks.



(a) Cyan: $O(\cdot) = 0.57$ Blue: $O(\cdot) = 0.56$
(b) Cyan: $O(\cdot) = 0.57$ Blue: $O(\cdot) = 0.41$
(c) Cyan: $O(\cdot) = 0.56$ Blue: $O(\cdot) = 0.36$
(d) Cyan: $O(\cdot) = 0.56$ Blue: $O(\cdot) = 0.32$
(e) Cyan: $O(\cdot) = 0.55$ Blue: $O(\cdot) = 0.29$
(f) Cyan: $O(\cdot) = 0.55$ Blue: $O(\cdot) = 0.27$

Fig. 9. Limitation of Multiple Object Tracking Precision (MOTP) [8]. Cyan tracker: MOTP=0.56, MELT=0.45; Blue tracker: MOTP=0.56, MELT=0.64. Unlike the proposed measure MELT, MOTP does not distinguish two tracking results due to its parameter dependence. Magenta: ground truth.



(a) $k = 16$ (b) $k = 19$ (c) $k = 24$

Fig. 10. Example of target occlusion in Bahnhof sequence. DP-NMS (green bounding box) loses the target due to occlusion in (b); however, the other trackers successfully handle it. Yellow: MT-TBD; cyan: CRFBT; magenta: ground truth.

of the trackers in terms of standard deviation (σ) values for different measures. In the case of METE, σ is comparable on all sequences. As for AER, while MCMCDA has the highest σ on TownCentre and iLids Easy (hence, the highest performance variation over time), MT-TBD has the highest performance variation over time on Bahnhof and Sunnyday. As for CER, the trend of the σ values of trackers on each dataset is the same as the trend of the corresponding CER values.

Figure 8 shows the evaluation results of CRFBT on key frames of Bahnhof using METE and MODA. All the targets are tracked in the results shown in Fig. 8(a), (b) and (c). The value of METE increases from (a) to (c) because of the decrease in the amount of overlap (lower accuracy) among the associated pairs of estimated and ground-truth bounding boxes. The cases shown in Fig. 8(d) and (e) have $\mathcal{C} = 1$; however, METE in (e) is higher than that in (d). In Fig. 8(f), 79% of targets are correctly tracked (11 out of 14), hence its METE value (0.400) is higher than that in (e) where 90% of targets are correctly tracked with a METE value of 0.373. In Fig. 8(g), the percentage of tracked targets reduces further to 73%, hence its METE value is higher than (f). Although the percentage of tracked targets in the case of Fig. 8(h) (75%) is higher than (g), METE is slightly higher in the case of the former because of a more inaccurate overlap

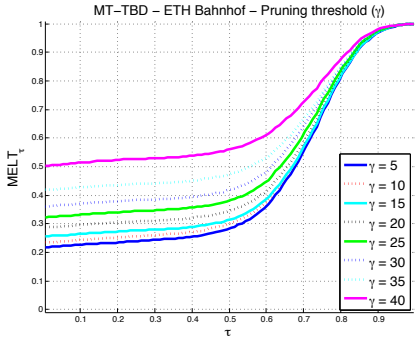
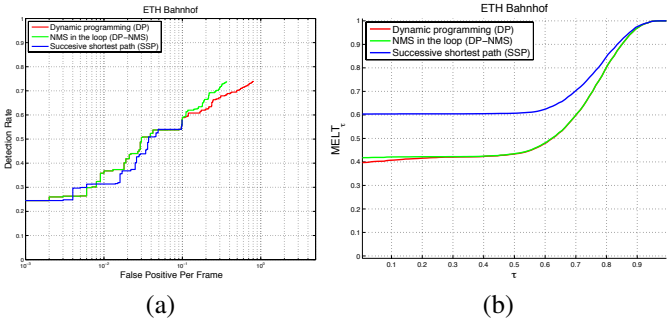


Fig. 11. Effect of the variation in performance of MT-TBD using MELT_τ on Bahnhof by varying the threshold value (γ) used to prune spurious tracks at the track-linking stage.



(a) (b)

Fig. 12. Tracking results obtained using the three algorithms proposed in [32]. (a) Detection rate versus false positives per frame; (b) MELT_τ plots.

in the case of (h). Likewise, METE for the cases shown in Fig. 8(i-l) is influenced by the corresponding accuracy and cardinality errors. MODA does not distinguish among the cases in Fig. 8(a-c) (MODA=1) despite the difference in their respective overlaps. This insensitivity of MODA is due to the threshold (τ_o) used to determine false negatives and false positives (Eq. 18). Another point to highlight is the disagreement between METE and MODA in the cases shown in Fig. 8(h) and (i). Unlike METE, MODA considers the case in (i) to be better than (h). This is because in the case of (h), MODA considers 58% (7 out of 12) of estimated bounding boxes to be correctly associated to those of the ground truth, excluding the third and the sixth pairs (starting from the right) that are not considered to be valid associations since their overlap is below τ_o. Differently, METE, being independent of thresholds, considers these two pairs in the evaluation of the score and penalizes them appropriately. In the case of Fig. 8(i), 66% of the ground-truth targets are correctly associated and there is a presence of a false positive, hence the MODA value is higher for this case.

To summarize, MCMCDA performs better as a person tracker than a head tracker. While DP-NMS has mostly reported the lowest accuracy error, its cardinality error has mostly been the highest. DP-NMS is not able to handle occlusions. Finally, CRFBT is the best tracker based on the evaluation of ID changes, followed by MT-TBD.

D. Performance analysis by varying the parameters of the trackers

We vary a key parameter in MT-TBD framework (Fig. 11) and consider DP-NMS with different methods (Fig. 12), and analyze the variation in their performance based on MELT.

For MT-TBD, we vary γ, which defines the minimum allowed track length [33] such that shorter estimated tracks are pruned. In Fig. 11, MELT_τ values are shown for variations of τ with different γ values. We can notice a gradual deterioration in performance while increasing γ because of an increase in lost-track ratio values.

For DP-NMS, we generate tracking results on Bahnhof with three different algorithms proposed by the authors of [32]. The first method is DP-NMS that is used in Sec. IV-B and IV-C; the second is based on dynamic programming (DP) but without including the non-maximal suppression (NMS) stage in the iterative process; and the third is based on successive shortest path (SSP) [37]. Figure 12(a) shows detection rate versus false positives per frame (FPPF) for DP, DP-NMS and SSP algorithms, while Fig. 12(b) shows their MELT_τ plots. The results in Fig. 12(a) show mostly a lower performance (smaller detection rate) for SSP. This trend is also confirmed for the MELT_τ plots (Fig. 12(b)) since MELT_τ of SSP is consistently higher. This high MELT_τ for SSP is expected because, as discussed in [32], the SSP algorithm generates shorter tracks (higher MELT_τ), while the DP-based algorithms generate longer tracks (lower MELT_τ).

V. Conclusions

We proposed three measures (METE, MELT, NIDC) that quantify key factors in extended multi-target tracking: accuracy, cardinality and ID changes. These measures are parameter independent, numerically bounded and account for target-size changes. METE provides a holistic error assessment using an effective trade-off between accuracy and cardinality errors. MELT enables the analysis of tracking performance at varying accuracy levels that can facilitate the selection of trackers for specific applications. NIDC penalizes ID changes as a function of the length of the track in which they occur. We presented an extensive experimental validation and comparison of these measures with the state-of-the-art measures on recent multi-target trackers using challenging real-world sequences.

The proposed measures are suitable for targets that are modeled in terms of their position and 2D image-plane-occupied area, as commonly considered in the literature [8], [9], [15]. The proposed measures can also be applied to other sensing modalities when a 2D target model is used. Other target models for 2.5D and 3D tracking also exist for different sensing modalities [38]–[40] and future research could investigate the extension of the proposed tracking evaluation approaches for these higher-dimensional target models.

References

[1] X. Wang, X. Ma, and W. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian model," *IEEE Trans. Patt. Ana. Mach. Int.*, vol. 31, no. 3, pp. 539–555, Mar. 2009.

[2] M. Taj and A. Cavallaro, *Recognizing interactions in video*. Intellig. Multimed. Analys. for Secur. Applic, Springer, 2010, vol. 282/2010.

[3] H.-I. Suk, A. Jain, and S.-W. Lee, "A network of dynamic probabilistic models for human interaction analysis," *IEEE Trans. Cir. Sys. Vid. Tech.*, vol. 21, pp. 932–945, 2011.

[4] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. of CVPR*, Colorado Springs, USA, Jun. 2011, pp. 3457–3464.
[5] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Trans. Patt. Ana. Mach. Int.*, vol. 33, no. 9, pp. 1820–1833, Sep. 2011.
[6] B. Yang and R. Nevatia, "An online learned CRF model for multi-target tracking," in *Proc. of CVPR*, Providence, Rhode Island, USA, Jun. 2012, pp. 2034–2041.
[7] E. Maggio and A. Cavallaro, *Video tracking: theory and practice*. Wiley, 2011.
[8] 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. Patt. Ana. Mach. Int.*, vol. 31, no. 2, pp. 319–336, February 2009.
[9] T. Nawaz and A. Cavallaro, "A protocol for evaluating video trackers under real-world conditions," *IEEE Trans. Ima. Proc.*, vol. 22, no. 4, pp. 1354–1361, April 2013.
[10] H. Wu, A. C. Sankaranarayanan, and R. Chellappa, "Online empirical evaluation of tracking algorithms," *IEEE Trans. Patt. Ana. Mach. Int.*, vol. 32, no. 8, pp. 1443–1458, August 2010.
[11] J. SanMiguel, A. Cavallaro, and J. Martinez, "Adaptive on-line performance evaluation of video trackers," *IEEE Trans. Ima. Proc.*, vol. 21, no. 5, pp. 2812–2823, May 2012.
[12] R. T. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Trans. Patt. Ana. Mach. Int.*, vol. 27, no. 10, pp. 1631–1643, October 2005.
[13] D. Chau, F. Bremond, and M. Thonnat, "Online evaluation of tracking algorithm performance," in *Proc. of Int. Conf. on Imag. for Crime Prev. and Detec.*, London, UK, December 2009.
[14] T. Nawaz and A. Cavallaro, "PFT: A protocol for evaluating video trackers," in *Proc. of ICIP*, Brussels, September 2011.
[15] H. Li, C. Shen, and Q. Shi, "Real-time visual tracking using compressive sensing," in *Proc. of CVPR*, 2011, pp. 1305–1312.
[16] B. Ristic, B.-N. Vo, D. Clark, and B.-T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," *IEEE Trans. Sig. Proc.*, vol. 59, no. 7, pp. 3452–3457, July 2011.
[17] J. Black, T. Ellis, and P. Rosin, "A novel method for video tracking performance evaluation," in *Proc. of WPETS*, 2003, pp. 125–132.
[18] F. Yin, D. Makris, and S. A. Velastin, "Performance evaluation of object tracking algorithms," in *Proc. of WPETS*, Rio de Janeiro, Brazil, 2007.
[19] 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 WPETS*, 2005, pp. 1–8.
[20] http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html. Last accessed on 03 october 2012.
[21] *Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions*, International Organization for Standardization Std. ISO 5725-1, Dec. 1994.
[22] A. Senior, A. Hampapur, Y. Tian, L. Brown, S. Pankanti, and R. Bolle, "Appearance models for occlusion handling," in *Proc. of WPETS*, Hawaii, December 2001.
[23] C. Needham and R. Boyle, "Performance evaluation metrics and statistics for positional tracker evaluation," in *Proc. of ICVS*, 2003.
[24] X. Mei, H. Ling, Y. Wu, E. Blasch, and L. Bai, "Minimum error bounded efficient l_1 tracker with occlusion detection," in *Proc. of CVPR*, 2011.
[25] J. R. Hoffman and R. P. S. Mahler, "Multitarget miss distance via optimal assignment," *IEEE Trans. Sys. Man Cyb. Part A: Sys. Hum.*, vol. 34, no. 3, pp. 327 – 336, May 2004.
[26] G. Sundaramoorthi, A. Mennucci, S. Soatto, and A. Yezzi, "Tracking deforming objects by filtering and prediction in the space of curves," in *Proc. of DC*, Shanghai, China, Dec. 2009, pp. 2395–2401.
[27] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
[28] J. Munkres, "Algorithms for assignment and transportation problems," *Journal of the Soc. for Ind. and App. Math.*, vol. 5, no. 1, pp. 32–38, March 1957.
[29] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Trans. Sig. Proc.*, vol. 56, p. 8, 2008.
[30] B. E. Fridling and O. E. Drummond, "Performance evaluation methods for multiple-target-tracking algorithms," in *Proc. SPIE, Signal Data Proces. of Small Targets*, vol. 1481, 1991, pp. 371 – 383.
[31] http://www.vision.ee.ethz.ch/?aess/dataset/. Last accessed: August 2012.
[32] H. Pirsiavash, D. Ramanan, and C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. of CVPR*, Colorado Springs, USA, Jun. 2011, pp. 1201–1208.
[33] F. Poiesi, R. Mazzon, and A. Cavallaro, "Multi-target tracking on confidence maps: an application to people tracking," *Comp. Vision Image Under.*, vol. 117, no. 10, pp. 1257–1272, Oct. 2013.
[34] C. Tomasi and T. Kanade, "Detection and tracking of point features," Carnegie Mellon University, Tech. Rep. CMU-CS-91-132, Apr. 1991.
[35] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Patt. Ana. Mach. Int.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[36] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *Int. Jou. Com. Vis.*, vol. 75, no. 2, pp. 247–266, Nov. 2007.
[37] R. Ahuja, T. Magnati, and J. Orlin, *Network flow: theory, algorithms, and applications*. Prentice Hall, 2008.
[38] X. Suau, J. Ruiz-Hidalgo, and J. R. Casas, "Real-time head and hand tracking based on 2.5D data," *IEEE Trans. Mult.*, vol. 14, no. 3, pp. 575–585, June 2012.
[39] D. Poullin and M. Flecheux, "Passive 3D tracking of low altitude targets using DVB (SFN Broadcasters)," *IEEE Aero. Ele. Sys. Mag.*, vol. 27, no. 11, pp. 36–41, November 2012.
[40] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele, "Monocular visual scene understanding: Understanding multi-object traffic scenes," *IEEE Trans. Patt. Ana. Mach. Int.*, vol. 35, no. 4, pp. 882–897, April 2013.

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. Andrea Cavallaro, first as a Research Assistant from September to December 2010 and then as a PhD Student since January 2011. His research interests include performance evaluation of tracking algorithms, environment learning and shape analysis.

Fabio Poiesi received BSc and MSc degrees in Telecommunication Engineering from the University of Brescia in 2007 and 2010, respectively. He completed the Master thesis at Queen Mary University of London on the estimation of the ball position in basketball matches by looking at the behaviour of players. Since April 2010 he has been a PhD student in the School of Electronic Engineering and Computer Science under the supervision of Prof. Andrea Cavallaro. His research field involves multi-target tracking in highly-populated scenes, behaviour understanding and performance evaluation of tracking algorithms.

Andrea Cavallaro received his Ph.D. in Electrical Engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, in 2002 and the Laurea (Summa cum Laude) in Electrical Engineering from the University of Trieste in 1996. He is Professor of Multimedia Signal Processing and Director of the Centre for Intelligent Sensing at Queen Mary University of London, UK. He was a Research Fellow with British Telecommunications (BT) in 2004/2005 and was awarded the Royal Academy of Engineering teaching Prize in 2007; three student paper awards on target tracking and perceptually sensitive coding at IEEE ICASSP in 2005, 2007 and 2009; and the best paper award at IEEE AVSS 2009. Prof. Cavallaro is Area Editor for the IEEE Signal Processing Magazine and Associate Editor for the IEEE Transactions on Image Processing. He is an elected member of the IEEE Signal Processing Society, Image, Video, and Multidimensional Signal Processing Technical Committee, and chair of its Awards committee. He served as an elected member of the IEEE Signal Processing Society, Multimedia Signal Processing Technical Committee, as Associate Editor for the IEEE Transactions on Multimedia and the IEEE Transactions on Signal Processing, and as Guest Editor for seven international journals. He was General Chair for IEEE/ACM ICDSC 2009, BMVC 2009, M2SFA2 2008, SSPE 2007, and IEEE AVSS 2007. Prof. Cavallaro was Technical Program chair of IEEE AVSS 2011, the European Signal Processing Conference (EUSIPCO 2008) and of WIAMIS 2010. He has published more than 130 journal and conference papers, one monograph on Video tracking (2011, Wiley) and three edited books: Multi-camera networks (2009, Elsevier); Analysis, retrieval and delivery of multimedia content (2012, Springer); and Intelligent multimedia surveillance (2013, Springer to appear).