Monotonicity and Error Type Differentiability in Performance Measures for Target Detection and Tracking in Video

Ido Leichter, Member, IEEE, and Eyal Krupka, Member, IEEE

Abstract—There exists an abundance of systems and algorithms for multiple target detection and tracking in video, and many measures for evaluating the quality of their output have been proposed. The contribution of this paper lies in the following: first, it argues that such performance measures should have two fundamental properties—monotonicity and error type differentiability; second, it shows that the recently proposed measures do not have either of these properties and are thus less usable; third, it composizes a set of simple measures, partly built on common practice, that does have these properties. The informativeness of the proposed set of performance measures is demonstrated through their application on face detection and tracking results.

Index Terms—Performance evaluation, tracking, multiple targets.

1 INTRODUCTION

Multiple target detection and tracking in video is an important research subject in computer vision, and many systems have been developed for this task. A necessary supplement to these systems is a means for quantitatively evaluating their performance. Such a quantitative evaluation is necessary for two reasons: enabling a comparison between these systems in terms of performance, as well as enabling the tuning of their parameters for performance optimization. The quantitative evaluation of the performance of a system is done by applying the system on a specific dataset, and then calculating a set of performance measures that quantify the quality of the system’s output with respect to the dataset’s ground truth (GT). Repeating this for multiple systems or for different parameter settings using a common dataset provides a means for quantitatively comparing the systems’ performances or finding the optimal parameter setting. Various tools for annotating datasets and for calculating performance measures have been introduced (e.g., [5], [2], [12]).

The problem of defining performance measures in the context of multiple target detection and tracking is ill-posed in the sense that there is no single “correct” or “best” set of measures. Therefore, many sets of measures have been proposed, many times as a result of multi-party efforts attempting to reach an agreed set of measures. One example of such efforts is the First-Thirteenth IEEE International Workshop Series on Performance Evaluation of Tracking and Surveillance (PETS, 2000-2010). Another example is the VACE metrics [7], developed in the course of a series of evaluations conducted in the framework of the US Government’s Video Analysis and Content Extraction (VACE) program. The CLEAR MOT metrics [4] were developed as part of the CLEAR consortium [3], coordinated by the US National Institute of Standards and Technology (NIST) and by Karlsruhe Institute of Technology. A close variant of the latter metrics is the CLEAR metrics described in [7]. Another recently proposed set of performance measures consists of the pair of information theoretic measures in [6]. An elaborated review of earlier performance measures and additional evaluation programs is provided in [7].

As mentioned, the problem of defining performance measures in the context of multiple target detection and tracking is ill-posed. Nevertheless, we claim that any set of measures should have the following two fundamental properties. The first, termed here monotonicity, is that the elimination of an error or the addition of a success should result in each measure in the set being improved or unchanged (an exact definition is provided in Sec. 5.1). The second property, termed here error type differentiability, is that the set of measures should be informative about the system’s performance with respect to each of the different basic error types. Otherwise, as is explained in Sec. 5.2, the performance measures may not be able to tell which system or parameter setting is better for the application at hand. In fact, if the performance measures are not error type differentiating, it may not be clear what application they may be used for. This paper shows that the recently proposed performance measures do not have either of these two properties and are thus less usable. A set of measures that does have these properties is proposed as well.

2 TERMINOLOGY

Throughout the paper the following terminology is used:

- **Truth target** — An instance of a true target object in a frame.
- **System target** — An instance of an object reported by the system in a frame.
- **Target’s location** — The values of variables that define the location of a truth or system target. For example, when the target’s location is approximated by an axes-aligned square, these variables may consist of the square’s center and length of side. It is assumed here that the annotated locations of the truth targets and the estimated locations of the system targets are in the same state space. In case the spaces are different, they should be reduced to a common one.
- **Truth Track** — The sequence of all truth targets corresponding to a true target object. Note that the frame indices in the track are not necessarily consecutive, as the object may be temporarily occluded or outside the camera’s field of view.
- **System Track** — The sequence of all system targets corresponding to an object reported by the system. Similarly to a truth track, the frame indices in a system track are not necessarily consecutive.
• GT annotation – The set consisting of all truth tracks.
• System’s output – The set consisting of all system tracks.

3 BASIC TYPES OF ERROR AND SUCCESS

There are various ways to partition the possible detection and tracking errors into a set of basic error types. The most common basic types of error addressed in the literature are:

1) False negative – No system target is associated with a truth target. The performance is degraded as the number of false negatives increases.
2) False positive – A system target is not associated with a truth target. The performance is degraded as the number of false positives increases.
3) Fragmentation – A truth track contains targets that are associated with different system tracks. The performance is degraded as the number of fragmented truth tracks increases and as these tracks become more severely fragmented.
4) Merger – A pair of truth tracks contain target pairs – one target of each track – that are associated with a common system track. The performance is degraded as the number of merged truth track pairs increases and these track pairs become more severely merged.
5) Deviation – The system target’s location deviates from its associated truth target’s location. The performance is degraded as the deviations become greater.

The first four basic error types are illustrated in Fig. 1. These error types are basic in the sense that other types of error are combinations of these types. For example, a “gap” [6] is a sequence of consecutive false negatives corresponding to one truth track, an “extra” [6] is a sequence of consecutive false positives in one system track, and an “ID swap” [11] is a combination of fragmentation errors and a merger related to two truth tracks and two system tracks.

Complementary to an error, the lack of a potential error is considered as a success. Symmetrically to the detection and tracking error types, the corresponding types of detection and tracking success are:

1) True positive – A system target is associated with a truth target. The performance is improved as the number of true positives increases.
2) True negative – The lack of a system target that would have been a false positive. The performance is improved as the number of true negatives increases.
3) Identification – A truth track contains targets that are associated with the same system track. The performance is improved as the number of identification successes increases and as the truth tracks corresponding to these successes become less fragmented. Note that, unless the truth track is completely fragmented (i.e., each of the system targets associated with it is of a different system track) or wholly identified as a single identity (i.e., all system targets associated with it are of the same system track), both a fragmentation error and an identification success are associated with this truth track.
4) Differentiation – A pair of truth tracks contain target pairs – one target of each track – that are associated with different system tracks. The performance is improved as the number of differentiation successes increases and as the truth track pairs corresponding to these success become less severely merged. As before, both a merger error and a differentiation success may be associated with a pair of truth tracks.
5) Proximity – The system target is in proximity to its associated truth target. The performance is improved as the proximities between the system targets and their associated truth targets become greater. Note that, unless the pair of system and truth targets matches perfectly, both a proximity success and a deviation error are associated with it.

4 RECENTLY PROPOSED PERFORMANCE MEASURES

Several sets of performance measures have been recently proposed: VACE [7], CLEAR [7], CLEAR MOT [4], and the information theoretic measures [6].

The VACE metrics [7] consist of two scores – Sequence Frame Detection Accuracy (SFDA) and Average Tracking Accuracy (ATA). SFDA evaluates the performance with respect to false negatives, false positives and deviation errors; ATA evaluates the performance with respect to all errors.

The CLEAR metrics [7] consist of four scores – Normalized Multiple Object Detection Accuracy (N-MODA), Normalized Multiple Object Detection Precision (N-MODP), Multiple Object Tracking Accuracy (MOTA), and Multiple Object Tracking Precision (MOTP). N-MODP and MOTP evaluate the performance with respect to deviation errors, N-MODA evaluates the performance with respect to false negatives and false positives, and MOTA evaluates the performance with respect to false negatives, false positives, fragmentation, and mergers.

A close variant of the CLEAR metrics is the CLEAR MOT metrics [4]. These metrics consist of the MOTA and MOTP scores. However, in the MOTA score of the CLEAR MOT metrics the penalties for the different errors are fixed, whereas in the CLEAR metrics the penalties are tunable.

In [6], a pair of information theoretic measures was proposed: Truth Information Completeness and False Information Ratio. The former quantifies the percentage of the truth information captured and the latter quantifies the amount of false information generated.
5.1 Monotonicity

Type differentiability

5.1.1 Lack of monotonicity in previous measures.

The elimination or reduction of an error or the addition or increase of a success should never result in a degradation of the measures. More exactly:

If the system’s output or the GT annotation is modified such that an error (success) is eliminated (added) or reduced (increased) with no error (success) introduced (removed) or worsened (moderated), then each measure should not be degraded.

In particular, each measure should not be degraded upon the elimination of a false negative or the addition of a true positive, the elimination of a false positive or the addition of a true negative, the elimination or reduction of a fragmentation error or the addition or increase of an identification success, etc.

Usually, the elimination of an error or the addition of a success should result in the improvement of one or more measures. Allowing that all measures remain unchanged as well accounts for the case where the type of eliminated (added) error (success) is irrelevant to all measures, or in the boundary case where the error (success) is eliminated (added) when there have not been any successes (errors) of the complementary type beforehand.

5.1.1 Lack of monotonicity in previous measures. The aforementioned performance measures (VACE [7], CLEAR [7], CLEAR MOT [4], and the information theoretic measures [6]) are not monotonic. This is shown in Figs. 2-4, which provide simple, concrete examples that the elimination of errors or the addition of successes may result in the degradation of one or more measures. Even worse, for each set of measures, the error eliminations or success additions cause the degradation of one or more of the measures and the improvement of none.

5.2 Error type differentiability

The set of measures should provide a performance evaluation with respect to each of the basic error types alone. The reason for this requirement is that the severity of the different error types is application dependent. One detection and tracking system may be better than another in one context and worse in another context. For example, in applications that require high recall, such as surveillance, a false negative is typically more severe than a false positive or a deviation; in applications that require high precision such as video summarization for entertainment purposes it is the other way around.

5.2.1 Lack of error type differentiability in previous measures. The VACE metrics were recently proposed in [7]. As mentioned, these metrics consist of only two measures, each is a single score that evaluates the performance with respect to multiple error types. Since these metrics lack error type differentiability, they usually cannot provide an indication about how good the system is for the application at hand, or which of several systems is better for it. Furthermore, the collective quantification of the errors of different basic types may render the relation between the measure and the quality of the system’s output inappropriate in many cases. Two examples of such a case are shown in Figs. 5 and 6.

Other recently proposed performance measures are the CLEAR metrics [7]. As mentioned, these metrics consist of four measures. While two of them (N-MODP and MOTP) evaluate the performance with respect to deviation errors alone, each of the other two measures (N-MODA and MOTA) is a single score that evaluates
Fig. 4. The information theoretic measures in [6] are not monotonic. Thick (Thin) lines are truth (system) tracks. Image (a) illustrates a GT annotation consisting of one track, as well as two system tracks. The green one corresponds to a section of the truth track, and the red one is too distant, which makes all red system targets false positives. Image (b) illustrates the same GT annotation and the green system track lengthened and unified with the other system track. Note that the latter 500 system targets (red in (a), now green) remain distant from the truth targets and thus remain being false positives. The modifications in (b) to the system’s output eliminate 50 false negatives without introducing any new errors. However, instead of being improved, both information theoretic measures are degraded: the pair (truth information completeness, false information ratio) changes from (0.4178, 0.6013) to (0.3982, 0.6018). These figures were obtained assuming 1000 frames and 10000 states [6] per frame.

The performance with respect to multiple error types. Therefore, the CLEAR metrics do not differentiate between most of the error types as well. As mentioned, a close variant of these metrics that has this problem as well is the CLEAR MOT metrics [4]. There, the MOTA measure assigns the same penalty for a false negative in a single frame and for the merger error resulting from a mismatch error [4] although the latter error is more severe than the former in most contexts; see Fig. 7.

As mentioned, a pair of information theoretic measures was proposed in [6] – Truth Information Completeness and False Information Ratio. The first measure is affected by false negatives, false positives, and mergers; the second measure is affected by these errors as well as by fragmentation. Therefore, although concise, this pair of measures also lack error type differentiability. A clear illustration of this is provided in Fig. 7 in [6], where it is shown that the pair of measures may be very similar for different basic error types.

6 MONOTONIC AND ERROR TYPE DIFFERENTIATING PERFORMANCE MEASURES

Following are provided five error measures, each one measures the quality of the system’s output with respect to one of the basic error types (and its complementary success type) listed in Sec. 3.

Fig. 5. The frame detection accuracy as defined by the VACE metrics is identical in both cases. Thick (Thin) boxes are truth (system) targets. Image (a) illustrates a frame consisting of two truth targets, each is a 100x100 square. The system detected the left-hand side target as an 87x87 square and did not detect the right-hand side target. Image (b) illustrates the same frame with both targets detected, the left-hand side target is detected exactly as before and the right-hand side target is detected as a 50x50 square. In (a) there is a false negative whereas in (b) the deviations are greater in average. These two opposite factors cannot be reflected by the VACE metrics, which consist of only one frame-based measure (SFDA). In fact, the frame detection accuracy as defined by the VACE metrics is identical in both cases (0.50), although the quality of the system’s output in (b) is considered higher than that in (a) for most applications – in (b) the right-hand side target is inaccurately localized, whereas in (a) it is not detected at all. Any reduction of one the system targets’ sizes in (b) will even make this system’s output less favorable than that in (a).

Each measure is nonnegative, zero if and only if there are no errors of its respective type, as well as monotonic. As shown, four measures out of the five may be interpreted as prior probabilities that a specific error of the corresponding type occurs, and the other measure is simply the mean deviation error. A method for obtaining a smaller set of monotonic and error type differentiating measures is discussed in Sec. 8. Note that the set of measures is partly built on common practice. Specifically, the first, second, and fifth measures have been widely used.

Inevitably, the performance measures depend on the specific matching between the truth and system targets. In the following formulation of the measures it is assumed that such a matching has already been done in each frame, and that each matching is one-to-one. Different possible matching schemes, including the specific one used here, are discussed in Sec. 6.8.

A Matlab function that calculates the performance measures is provided in [8] (download the ZIP compressed file).

6.1 Notation

- \( T \) – the number of frames in the video. The frame indices are 1, 2, \ldots, \( T \).
- \#truth tracks – the number of truth tracks in the GT annotation. Each truth track is associated with a GT ID \( \{1, \ldots, \#truth tracks\} \).
Fig. 6. The VACE metrics favor case (a) over case (b). Thick (Thin) lines are truth (system) tracks. Image (a) illustrates a GT annotation consisting of two tracks, as well as one system track that matches the black truth track perfectly (i.e., overlap ratio [7] equals 1). Image (b) illustrates the same three tracks plus the red system track that corresponds to the blue truth track. The red track is inaccurate, which causes its mean overlap ratio with the corresponding half of the blue track to be 0.25. In (b) there are less false negatives but the deviations are greater in average. This is not reflected in the VACE metrics, both favoring case (a) over case (b) ((SFDA,ATA)=(0.667,0.667) vs (0.646,0.563), respectively). In fact, the preference of case (a) by the VACE metrics contradicts the typical tendency to favor case (b) over case (a) – in (b) the blue target is poorly tracked whereas in (a) it is not tracked at all.

- #system tracks – the number of system tracks in the system’s output. Each system track is associated with a system ID \( i \in \{1, \ldots, \#system tracks\} \).
- \( m^{GT}(i, t) : \{1, \ldots, \#system tracks\} \times \{1, \ldots, T\} \rightarrow \{0, 1, \ldots, \#system tracks, \emptyset\} \) – truth target presence and matching function:
  \[
  m^{GT}(i, t) = \begin{cases} 
  j, & \text{truth target } (i, t) \text{ is matched to system target } (j, t); \\
  0, & \text{truth target } (i, t) \text{ is not matched to any system target although it appears in frame } t; \\
  \emptyset, & \text{no target of truth track } i \text{ appears in frame } t. 
  \end{cases}
  \]

- \( m^{S}(j, t) : \{1, \ldots, \#system tracks\} \times \{1, \ldots, T\} \rightarrow \{0, 1, \ldots, \#system tracks, \emptyset\} \) – system target presence and matching function:
  \[
  m^{S}(j, t) = \begin{cases} 
  i, & \text{system target } (j, t) \text{ is matched to truth target } (i, t); \\
  0, & \text{system target } (j, t) \text{ is not matched to any truth target although it is contained in the system’s output}; \\
  \emptyset, & \text{the system’s output does not contain the system target } (j, t). 
  \end{cases}
  \]

- \( x^{GT}(i, t) \) – the location of truth target \((i, t)\).
- \( x^{S}(j, t) \) – the location of system target \((j, t)\).

6.2 Measuring false negatives

The measure of false negatives is defined to be the ratio between their number and the total number of truth targets in all frames:

\[
\text{False Negative Rate} = \frac{\{ (i, t) : m^{GT}(i, t) = 0 \} \cup \{ (i, t) : m^{GT}(i, t) \neq 0 \}}{\#truth tracks \times \{1, \ldots, T\}}. \quad (3)
\]

The false negative rate (FNR) approximates the prior probability that a given truth target in a frame is unmatched to any system target. It ranges between 0 (i.e., no false negatives) to 1 (i.e., all truth targets are unmatched). In the pathological case where there are no truth targets at all, this measure is undefined. This is in agreement with the fact that a false negative rate can not be calculated when there are zero positives in the dataset.

6.3 Measuring false positives

The measure of false positives is defined to be their number, normalized by the sequence length and image area \(A\):

\[
\text{False Positive Rate} = \frac{1}{T \cdot A} \left| \{ (j, t) : m^{S}(j, t) = 0 \} \right|. \quad (4)
\]

The false positive rate (FPR) estimates the average number of false positives in a frame per unit image area. It is a nonnegative measure that is 0 if and only if there are no false positives. In statistical analysis, the FPR is defined as the ratio between the number of false positives and the total number of false instances. It thus approximates the conditional probability that an instance is erroneously classified as positive given that it is negative. However, in the context of target detection and tracking in video, the total number false instances is not well defined. This is the reason we normalize this measure by the video "volume" \(T \cdot A\), which makes the measure proportional to the prior probability that the system’s output erroneously contains a system target in a specific frame and location where a truth target is not present. A similar approach was taken in [6].

6.4 Measuring fragmentation

Given a particular truth track, let the measure of its fragmentation error be the ratio between the number of unordered pairs of targets of this track that are matched to different system tracks and the total number of unordered pairs of matched truth targets of this track. Symbolically, denote the set of all matched targets
of truth track \(i\) by \(\mathcal{M}_i^{GT} = \{(i, t) : m^{GT}(i, t) \notin \{0, \emptyset\}\}\). Then this measure is

\[
\text{Fragmentation Index}_i = \frac{\left| \left\{ (i_1, t_1), (i_2, t_2) : (i_1, t_1), (i_2, t_2) \in \mathcal{P}_{i}^{\text{same}}, m^{GT}(i_1, t_1) \neq m^{GT}(i_2, t_2) \right\} \right|}{|\mathcal{P}_{i}^{\text{same}}|},
\]

(5)

where \(\mathcal{P}_{i}^{\text{same}} = \{(i_1, t_1), (i_2, t_2) : (i_1, t_1) \in \mathcal{M}_i^{GT}, (i_2, t_2) \in \mathcal{M}_i^{GT}, t_1 \neq t_2\}\) is the set of all unordered pairs of matched truth targets of truth track \(i\).

The fragmentation index of a truth track approximates the prior probability that a specific pair of matched truth targets of this track are erroneously matched to system targets of different tracks. The measure of fragmentation of the system’s output is defined to be a weighted average of the fragmentation indices of all truth tracks.

The weight of each track is the length of its matched part:

\[
\text{Weight of Track } i = m^{GT}(i, t).
\]

The measure of mergers of the system’s output is defined to be a weighted average of the merger indices of all truth tracks.

\[
\text{Merger Index}_i = \frac{\sum_{(i_1, t_1), (i_2, t_2) \in \mathcal{P}_{i}^{\text{diff}}} \left| \mathcal{M}_i^{GT} \right|}{|\mathcal{P}_{i}^{\text{diff}}|}. \quad (6)
\]

6.5 Measuring mergers

Given a particular pair of truth tracks, let the measure of their merger error be the ratio between the number of unordered pairs of targets of the same truth track that are matched to different system tracks and the total number of unordered pairs of matched truth targets of the same track. This would approximate the prior probability that a specific pair of matched truth targets of the same track is erroneously fragmented when the drawing of the pair is performed uniformly over the set of all pairs of matched truth target of the same track. However, this would make the weight of each truth track quadratic in the length of its matched part rather than linear as in (6). In the pathological case where \(\forall i \ P_{i}^{\text{same}} = \emptyset\) (i.e., \(\forall i \ |\mathcal{M}_i^{GT}| \leq 1\)), the measure is undefined. This is in agreement with the fact that, in this case, there are no fragmentation errors neither identification successes.

Note that this measure as well as the former may be efficiently calculated by counting the number of matches per each system-truth ID pair. Therefore, there is no need to explicitly iterate over all target pairs.

6.6 Measuring deviations

The measure of deviations is simply the mean distance between system targets and corresponding truth targets:

\[
\text{Mean Deviation} = \frac{\sum_{(j,t) \in \mathcal{M}^{S}} d(x^{S}(j,t), x^{GT}(m^{S}(j,t), t))}{|\mathcal{M}^{S}|},
\]

(9)

where \(\mathcal{M}^{S} = \{(j,t) : m^{S}(j,t) \notin \{0, \emptyset\}\}\) is the set of all matched system targets in all frames. In the pathological case where \(\mathcal{M}^{S} = \emptyset\) (i.e., there are no matched system targets at all), the measure is undefined. This is in agreement with the fact that no measure of system targets’ deviation can be calculated when there are no matched system targets.

6.7 Fulfillment of monotonicity and error type differentiability

Each basic type of error and success influences its corresponding error measure only. Therefore, the proposed set of performance measures differentiates between errors of the different basic types. The examples in Sec. 5.1.1, which show that the previous measures are not monotonic, exemplify that the proposed set of measures are monotonic: in Fig. 2 the Merger Index reduces from 1 in case (a) to 0 in case (b) and all other measures remain unchanged; in Fig. 3 the False Negative Rate reduces from 0.5 in case (a) to 0 in case (b) and all other measures remain unchanged;
in Fig. 4 the False Negative Rate reduces from 0.55 in case (a) to 0.5 in case (b) and all other measures remain unchanged. The monotonicity with respect to each of the basic types of error and success is proved in the appendix.

6.8 Matching between truth and system targets

The performance measures depend on the matching between the truth targets and the system targets. This matching depends on the conditions under which a specific pair of truth and system targets may be considered to be a legitimate match. Relaxed conditions generally imply lower false negative and false positives rates, but a higher mean deviation. Strict conditions generally imply the opposite. These conditions limit the set of legitimate truth-system target pairs. However, there may be multiple legitimate one-to-one matchings in a frame, and various schemes may be used to choose among them. Each scheme may produce a different matching, which may result in different performance values. Moreover, some matching schemes are biased with respect to one or more performance measures. For example, in the VACE metrics [7], two matchings are produced: a frame-level matching that maximizes the SFDA measure, and a track-level matching that maximizes the ATA measure. In the CLEAR MOT metrics [4], the matching in each frame is generated via a special procedure that favors ID matching consistency with previous frames, few false negatives and false positives, and a small sum of deviation errors.

Here, the matching between the truth targets and the system targets in each frame is generated as follows:

1) All legitimate truth-system target pairs are identified.
2) The distance \(d\) between each legitimate pair is calculated.
3) The maximum bipartite matching \(^1\) between the truth and system targets that has the smallest sum of distances is calculated.

The maximum matching in Step 3 can be calculated efficiently as follows: 1. Augment the smaller of the two sets, that of the truth targets and that of the system targets, with imaginary targets so that the two sets will be of equal size. 2. Assign a very large distance (larger than the sum of all other distances) to each pair of truth and system targets that cannot be matched or that contains an imaginary target. 3. Solve the corresponding linear assignment problem (e.g., by the Hungarian algorithm).

7 EXPERIMENTS

7.1 Tested system and datasets

To test the proposed set of measures, they were employed to measure and compare the quality of the results produced by various operational modes of an in-house developed offline face detection and tracking system. Given a video, this system detects faces in it and tracks them through the video. In its default operational mode (“Mode 0”), the system sequentially executes the following main steps:

1) Shot boundary detection.
2) For each detected shot:
   a) Face detection in all frames. The face detector is an implementation of the Viola-Jones face detector [10] that was trained for detecting faces in poses (yaw rotation) that range between frontal and profile. In order to detect roll rotated faces, the face detector in Mode 0 is applied on images rotated by \(\pm 30\) degrees as well.

   \(^1\) A maximum bipartite matching in a bipartite graph is a matching that consists of the maximum possible number of matches.

b) Agglomerative clustering of the detected faces into system tracks. The clustering is based on spatiotemporal proximity and face size similarity.

c) Tracking the last (first) face of each system track in the forward (backward) temporal direction. In the process of tracking a face, if its location overlaps the location of the first (last) face of another system track, the tracking of this face is terminated, and the extended former track and the latter track are merged as they are likely of the same identity. If the tracking of a specific face terminates before such a merging occurs (due to low confidence or due to reaching a shot boundary), the corresponding system track’s extension is discarded as it may be false. The tracking is accomplished by a color-based, general object tracker [9].

Two datasets were used for testing. Dataset 1 consisted of about 18 minutes from Episode 1 of the first season of the TV series “Coupling,” whose annotation is available at [1]. Dataset 2 consisted of the first 15 minutes of the movie “Up in the Air.” The tested system is expected and attempts to maintain the same track ID after the face was temporarily occluded or outside the camera’s field of view (but in the same video shot). Therefore, in each shot, one truth track was generated per identity. This resulted in 604 truth tracks consisting of 45,502 targets in Dataset 1, and in 397 truth tracks consisting of 26,794 targets in Dataset 2.

The tested system approximates the location of a target (i.e., a face) by a square and thus provides its center coordinates and size. In the annotation of Dataset 1 [1], only the center coordinates of the targets were provided. Therefore, the distance function between a system target’s location and a truth target’s location was the Euclidean distance between their centers, measured as the fraction of the corresponding square’s side length. In Dataset 2, the faces were annotated by squares of variable size. There, the distance function was \(1 - a\) where \(a \in [0,1]\) is the squares’ overlapping area, measured as a fraction of their union area. In Dataset 1 (2), a matching between a truth target and a system target was allowed under the condition that their distance was not greater than 0.5 (0.85).

7.2 Operational modes and results

The performance measures obtained for the two datasets under various operational modes (Modes 1–5) were calculated and compared to those obtained under Mode 0 (described above in Sec. 7.1). In the calculation of FPR (4), the image area in the corresponding dataset is taken as the unit area size (i.e., \(A = 1\)). Thus, for these datasets FPR equals the mean number of false positives per frame. The results are summarized in Table 1 and analyzed in what follows.

In Mode 1, Step 2(c), which contains the tracking by the general object tracker, is skipped over. On one hand, this results in fewer detected truth targets, and the FNR in Mode 1 is indeed degraded as compared to that in Mode 0. On the other hand, this results in fewer false system targets as well, and the FPR in Mode 1 is indeed improved. Skipping all the system track mergers in Step 2(c) results in much higher fragmentation, and the fragmentation index in Mode 1 is indeed significantly degraded as compared to that in Mode 0. As some of the track mergers in Step 2(c) are erroneous, skipping them reduces the number of merger errors. Thus, the merger index in Mode 1 is indeed improved as compared to that in Mode 0. Another consequence of skipping Step 2(c) is that the mean deviation is improved. This results from the lack of the system targets generated by the color-
based general tracker, whose target localization is less accurate than that of the face detector.

In Mode 2, the shot boundary detection (Step 1) is skipped over and the entire video is treated as a single shot. This results in tracks erroneously spanning multiple video shots, and the merger index is indeed significantly degraded as compared to that in Mode 0. As mentioned, if during the tracking of a specific face in Step 2(c) a shot boundary is reached, the corresponding system track extension is discarded as it may be false. In Mode 2, there are no detected shot boundaries that may cause the discard of these false system targets. Thus, the FPR is indeed significantly degraded as compared to that in Mode 0. On the other hand, a fraction of these non-discarded system targets are in fact true, and the FNR in Mode 2 is indeed better than that in Mode 0. All these non-discarded, true system targets were obtained by the general object tracker, which provides less accurate localization than that provided by the face detector. Thus, the mean deviation in Mode 2 is indeed degraded as compared to that in Mode 0.

In Mode 3, in order to reduce the runtime, the face detector in Step 2(a) is fully applied on every tenth frame only. In other frames, it is only applied in locations and scales close to faces detected in nearby frames. Thus, the set of system targets in this mode lacks part of those in Mode 0. Some of the lacking system targets are false and the others are true. This is indeed reflected by the FNR and FPR differences between the two operational modes.

In Mode 4, the face detector is applied on unrotated images only. As in Mode 3, the set of system targets in this mode lacks part of the those in Mode 0, which results in the same consequences as before. Moreover, although in this operational mode the face detector is applied on unrotated images only, part of the rotated faces are still detected. However, the localization of these detected rotated faces is less accurate than that in Mode 0, where the “full” detector is applied. This results in a degradation of the mean deviation.

In Mode 5, the face detector’s thresholds were lowered to minimum. As can be seen, this resulted in a marginal reduction in the number of false negatives, as well as an increase in the number of false positives and, in Dataset 1, an increase in fragmentation errors. The additional true positives are low-confidence face detections returned by the face detector. Such low-confidence detections tend to be less accurate in location, which is reflected in the marginal degradation of the mean deviation in this operational mode as compared to that in Mode 0.

## 8 Conclusion

Two important properties that any set of performance measure should have are monotonicity and error type differentiability. It was shown that the recently proposed measures do not have either of these properties. In addition, a set of five intuitive measures that are applicable in other contexts as well. Two properties that any set of performance measures that have these properties could be defined.

The proposed set of measures was composed for the preceding specific types of error and success. For other sets of error and success types, different measures may be defined. In particular, it might be desired to have a smaller set of measures. This may be accomplished by: 1. unifying subsets of the basic error and success types between which the differentiation is less important for the application at hand; 2. measuring the performance with respect to a unified error type by the (possibly weighted) sum of the corresponding subset of measures. The obtained smaller set of measures will remain monotonic and error type differentiating with respect to the new set of more general error types.

Finally, it is important to clarify that the two properties – monotonicity and error type differentiability – cannot serve as sufficient conditions for appropriateness of a set of performance measures. These two properties, however, are necessary conditions for appropriateness that are applicable in other contexts as well.

## References


