Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2008, Article ID 246309, 10 pages
doi:10.1155/2008/246309

Research Article

Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics

Keni Bernardin and Rainer Stiefelhagen

Interactive Systems Lab, Institut für Theoretische Informatik, Universität Karlsruhe, 76131 Karlsruhe, Germany

Correspondence should be addressed to Keni Bernardin, keni@ira.uka.de

Received 2 November 2007; Accepted 23 April 2008

Recommended by Carlo Regazzoni

Simultaneous tracking of multiple persons in real-world environments is an active research field and several approaches have been proposed, based on a variety of features and algorithms. Recently, there has been a growing interest in organizing systematic evaluations to compare the various techniques. Unfortunately, the lack of common metrics for measuring the performance of multiple object trackers still makes it hard to compare their results. In this work, we introduce two intuitive and general metrics to allow for objective comparison of tracker characteristics, focusing on their precision in estimating object locations, their accuracy in recognizing object configurations and their ability to consistently label objects over time. These metrics have been extensively used in two large-scale international evaluations, the 2006 and 2007 CLEAR evaluations, to measure and compare the performance of multiple object trackers for a wide variety of tracking tasks. Selected performance results are presented and the advantages and drawbacks of the presented metrics are discussed based on the experience gained during the evaluations.

1. INTRODUCTION

The audio-visual tracking of multiple persons is a very active research field with applications in many domains. These range from video surveillance, over automatic indexing, to intelligent interactive environments. Especially in the last case, a robust person tracking module can serve as a powerful building block to support other techniques, such as gesture recognizers, face or speaker identifiers, head pose estimators [1], and scene analysis tools. In the last few years, more and more approaches have been presented to tackle the problems posed by unconstrained, natural environments and bring person trackers out of the laboratory environment and into real-world scenarios.

In recent years, there has also been a growing interest in performing systematic evaluations of such tracking approaches with common databases and metrics. Examples are the CHIL [2] and AMI [3] projects, funded by the EU, the U.S. VACE project [4], the French ETISEO [5] project, the U.K. Home Office *iLIDS* project [6], the CAVIAR [7] and CREDS [8] projects, and a growing number of workshops (e.g., PETS [9], EEMCV [10], and more recently CLEAR [11]). However, although benchmarking is rather straightforward for single object trackers, there is still no general agreement on a principled evaluation procedure using a common set of objective and intuitive metrics for measuring the performance of multiple object trackers.

Li et al. in [12] investigate the problem of evaluating systems for the tracking of football players from multiple camera images. Annotated ground truth for a set of visible players is compared to the tracker output and 3 measures are introduced to evaluate the spatial and temporal accuracy of the result. Two of the measures, however, are rather specific to the football tracking problem, and the more general measure, the "identity tracking performance," does not consider some of the basic types of errors made by multiple target trackers, such as false positive tracks or localization errors in terms of distance or overlap. This limits the application of the presented metric to specific types of trackers or scenarios.

Nghiem et al. in [13] present a more general framework for evaluation, which covers the requirements of a broad range of visual tracking tasks. The presented metrics aim at allowing systematic performance analysis using large

amounts of benchmark data. However, a high number of different metrics (8 in total) are presented to evaluate object detection, localization and tracking performance, with many dependencies between separate metrics, such that one metric can often only be interpreted in combination with one or more others. This is for example the case for the "tracking time" and "object ID persistence/confusion" metrics. Further, many of the proposed metrics are still designed with purely visual tracking tasks in mind.

Because of the lack of commonly agreed on and generally applicable metrics, it is not uncommon to find tracking approaches presented without quantitative evaluation, while many others are evaluated using varying sets of more or less custom measures (e.g., [14–18]). To remedy this, this paper proposes a thorough procedure to detect the basic types of errors produced by multiple object trackers and introduces two novel metrics, the multiple object tracking precision (MOTP), and the multiple object tracking accuracy (MOTA), that intuitively express a tracker's overall strengths and are suitable for use in general performance evaluations.

Perhaps the work that most closely relates to ours is that of Smith et al. in [19], which also attempts to define an objective procedure to measure multiple object tracker performance. However, key differences to our contribution exist: again, a large number of metrics are introduced: 5 for measuring object configuration errors, and 4 for measuring inconsistencies in object labeling over time. Some of the measures are defined in a dual way for trackers and for objects (e.g., MT/MO, FIT/FIO, TP/OP). This makes it difficult to gain a clear and direct understanding of the tracker's overall performance. Moreover, under certain conditions, some of these measures can behave in a nonintuitive fashion (such as the CD, as the authors state, or the $\overline{\text{FP}}$ and $\overline{\text{FN}}$, as we will demonstrate later). In comparison, we introduce just 2 overall performance measures that allow a clear and intuitive insight into the main tracker characteristics: its precision in estimating object positions, its ability to determine the number of objects and their configuration, and its skill at keeping consistent tracks over time.

In addition to the theoretical framework, we present actual results obtained in two international evaluation workshops, which can be seen as field tests of the proposed metrics. These evaluation workshops, the classification of events, activities, and relationships (CLEAR) workshops, were held in spring 2006 and 2007 and featured a variety of tracking tasks, including visual 3D person tracking using multiple camera views, 2D face tracking, 2D person and vehicle tracking, acoustic speaker tracking using microphone arrays, and even audio-visual person tracking. For all these tracking tasks, each with its own specificities and requirements, the here-introduced MOTP and MOTA metrics, or slight variants thereof, were employed. The experiences made during the course of the CLEAR evaluations are presented and discussed as a means to better understand the expressiveness and usefulness, but also the weaknesses of the MOT metrics.

The remainder of the paper is organized as follows. Section 2 presents the new metrics, the MOTP and the MOTA and a detailed procedure for their computation. Section 3 briefly introduces the CLEAR tracking tasks and their various requirements. In Section 4, sample results are shown and the usefulness of the metrics is discussed. Finally, Section 5 gives a summary and a conclusion.

2. PERFORMANCE METRICS FOR MULTIPLE OBJECT TRACKING

To allow a better understanding of the proposed metrics, we first explain what qualities we expect from an ideal multiple object tracker. It should at all points in time find the correct number of objects present; and estimate the position of each object as precisely as possible (note that properties such as the contour, orientation, or speed of objects are not explicitly considered here). It should also keep consistent track of each object over time: each object should be assigned a unique track ID which stays constant throughout the sequence (even after temporary occlusion, etc.). This leads to the following design criteria for performance metrics.

(i) They should allow to judge a tracker's precision in determining exact object locations.

(ii) They should reflect its ability to consistently track object configurations through time, that is, to correctly trace object trajectories, producing exactly one trajectory per object.

Additionally, we expect useful metrics

(i) to have as few free parameters, adjustable thresholds, and so forth, as possible to help making evaluations straightforward and keeping results comparable;

(ii) to be clear, easily understandable, and behave according to human intuition, especially in the occurrence of multiple errors of different types or of uneven repartition of errors throughout the sequence;

(iii) to be general enough to allow comparison of most types of trackers (2D, 3D trackers, object centroid trackers, or object area trackers, etc.);

(iv) to be few in number and yet expressive, so they may be used, for example, in large evaluations where many systems are being compared.

Based on the above criteria, we propose a procedure for the systematic and objective evaluation of a tracker's characteristics. Assuming that for every time frame t, a multiple object tracker outputs a set of hypotheses $\{h_1, \ldots, h_m\}$ for a set of visible objects $\{o_1, \ldots, o_n\}$, the evaluation procedure comprises the following steps.

For each time frame t,

(i) establish the best possible correspondence between hypotheses h_j and objects o_i,

(ii) for each found correspondence, compute the error in the object's position estimation,

(iii) accumulate all correspondence errors:

  (a) count all objects for which no hypothesis was output as misses,

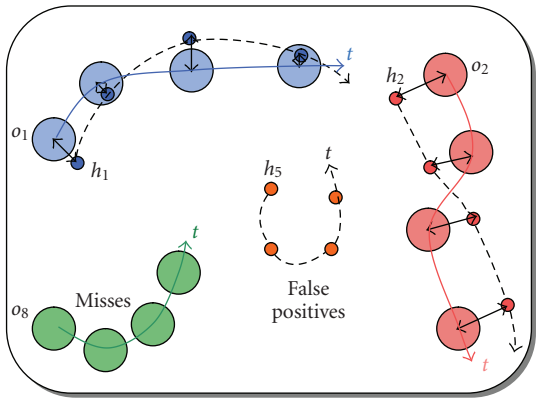

FIGURE 1: Mapping tracker hypotheses to objects. In the easiest case, matching the closest object-hypothesis pairs for each time frame t is sufficient.

(b) count all tracker hypotheses for which no real object exists as false positives,

(c) count all occurrences where the tracking hypothesis for an object changed compared to previous frames as mismatch errors. This could happen, for example, when two or more objects are swapped as they pass close to each other, or when an object track is reinitialized with a different track ID, after it was previously lost because of occlusion.

Then, the tracking performance can be intuitively expressed in two numbers: the "tracking precision" which expresses how well exact positions of persons are estimated, and the "tracking accuracy" which shows how many mistakes the tracker made in terms of misses, false positives, mismatches, failures to recover tracks, and so forth. These measures will be explained in detail in the latter part of this section.

2.1. Establishing correspondences between objects and tracker hypotheses

As explained above, the first step in evaluating the performance of a multiple object tracker is finding a continuous mapping between the sequence of object hypotheses $\{h_1, \ldots, h_m\}$ output by the tracker in each frame and the real objects $\{o_1, \ldots, o_n\}$. This is illustrated in Figure 1. Naively, one would match the closest object-hypothesis pairs and treat all remaining objects as misses and all remaining hypotheses as false positives. A few important points need to be considered, though, which make the procedure less straightforward.

2.1.1. Valid correspondences

First of all, the correspondence between an object o_i and a hypothesis h_j should not be made if their distance $\text{dist}_{i,j}$ exceeds a certain threshold T. There is a certain conceptual boundary beyond which we can no longer speak of an error in position estimation, but should rather argue that the tracker has missed the object and is tracking something else. This is illustrated in Figure 2(a). For object area trackers (i.e., trackers that also estimate the size of objects or the area occupied by them), distance could be expressed in terms of the overlap between object and hypothesis, for example, as in [14], and the threshold T could be set to zero overlap. For object centroid trackers, one could simply use the Euclidian distance, in 2D image coordinates or in real 3D world coordinates, between object centers and hypotheses, and the threshold could be, for example, the average width of a person in pixels or cm. The optimal setting for T therefore depends on the application task, the size of objects involved, and the distance measure used, and cannot be defined for the general case (while a task-specific, data-driven computation of T may be possible in some cases, this was not further investigated here. For the evaluations presented in Sections 3 and 4, empirical determination based on task knowledge proved sufficient). In the following, we refer to correspondences as *valid* if $\text{dist}_{i,j} < T$.

2.1.2. Consistent tracking over time

Second, to measure the tracker's ability to label objects consistently, one has to detect when conflicting correspondences have been made for an object over time. Figure 2(b) illustrates the problem. Here, one track was mistakenly assigned to 3 different objects over the course of time. A mismatch can occur when objects come close to each other and the tracker wrongfully swaps their identities. It can also occur when a track was lost and reinitialized with a different identity. One way to measure such errors could be to decide on a "best" mapping (o_i, h_j) for every object o_i and hypothesis h_j, for example, based on the initial correspondence made for o_i, or the correspondence (o_i, h_j) most frequently made in the whole sequence. One would then count all correspondences where this mapping is violated as errors. In some cases, this kind of measure can however become nonintuitive. As shown in Figure 2(c), if, for example, the identity of object o_i is swapped just once in the course of the tracking sequence, the time frame at which the swap occurs drastically influences the value output by such an error measure.

This is why we follow a different approach: only count mismatch errors once at the time frames where a change in object-hypothesis mappings is made; and consider the correspondences in intermediate segments as correct. Especially in cases where many objects are being tracked and mismatches are frequent, this gives us a more intuitive and expressive error measure. To detect when a mismatch error occurs, a list of object-hypothesis mappings is constructed. Let $M_t = \{(o_i, h_j)\}$ be the set of mappings made up to time t and let $M_0 = \{\cdot\}$. Then, if a new correspondence is made at time $t+1$ between o_i and h_k which contradicts a mapping (o_i, h_j) in M_t, a mismatch error is counted and (o_i, h_j) is replaced by (o_i, h_k) in M_{t+1}.

The so constructed mapping list M_t can now help to establish optimal correspondences between objects and hypotheses at time $t+1$, when multiple valid choices exist. Figure 2(d) shows such a case. When it is not clear, which

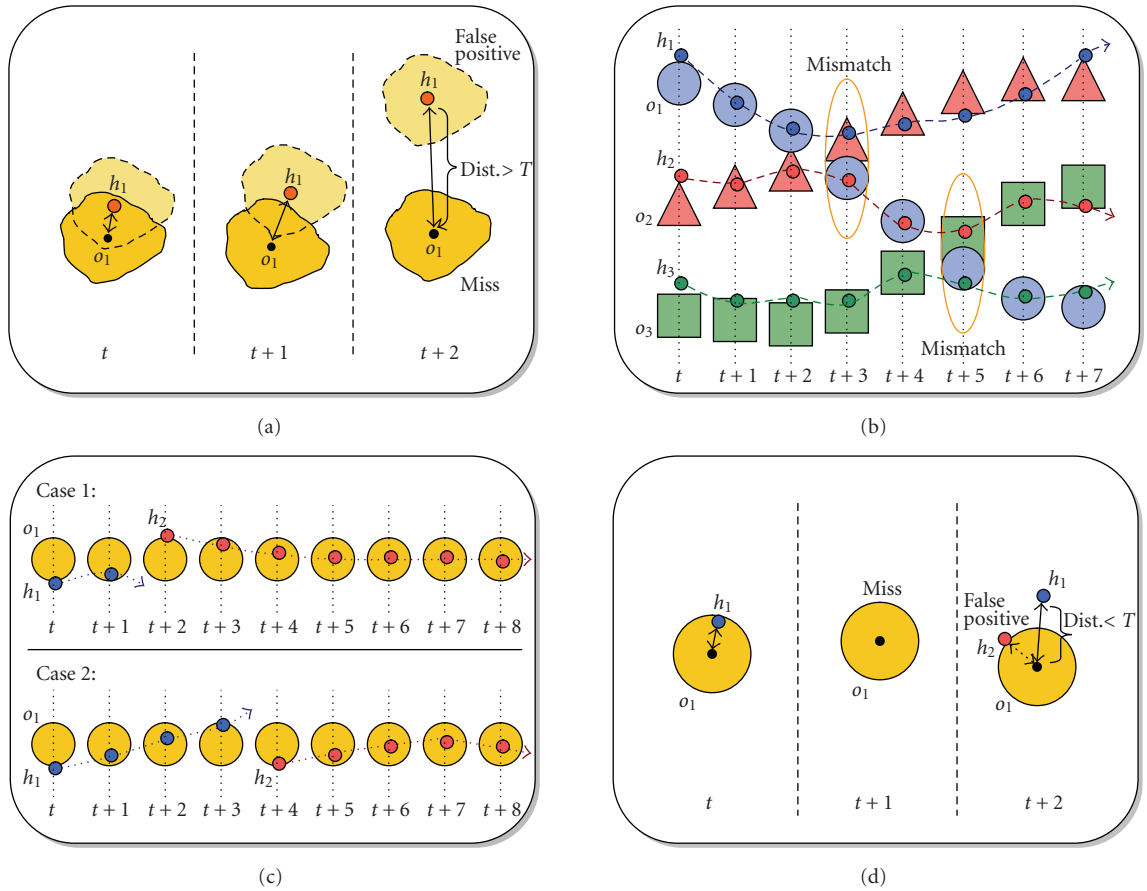

FIGURE 2: Optimal correspondences and error measures. (a) When the distance between o_1 and h_1 exceeds a certain threshold T, one can no longer make a correspondence. Instead, o_1 is considered missed and h_1 becomes a false positive. (b): Mismatched tracks. Here, h_2 is first mapped to o_2. After a few frames, though, o_1 and o_2 cross paths and h_2 follows the wrong object. Later, it wrongfully swaps again to o_3. (c): Problems when using a sequence-level "best" object-hypothesis mapping based on most frequently made correspondences. In the first case, o_1 is tracked just 2 frames by h_1, before the track is taken over by h_2. In the second case, h_1 tracks o_1 for almost half the sequence. In both cases, a "best" mapping would pair h_2 and o_1. This however leads to counting 2 mismatch errors for *case 1*; and 4 errors for *case 2*, although in both cases only one error of the same kind was made. (d): Correct reinitialization of a track. At time t, o_1 is tracked by h_1. At $t+1$, the track is lost. At $t+2$, two valid hypotheses exist. The correspondence is made with h_1 although h_2 is closer to o_1, based on the knowledge of previous mappings up to time $t+1$.

hypothesis to match to an object o_i, priority is given to h_o with $(o_i, h_o) \in M_t$, as this is most likely the correct track. Other hypotheses are considered false positives, and could have occurred because the tracker outputs several hypotheses for o_i, or because a hypothesis that previously tracked another object accidentally crossed over to o_i.

2.1.3. Mapping procedure

Having clarified all the design choices behind our strategy for constructing object-hypothesis correspondences, we summarize the procedure as follows.

Let $M_0 = \{\cdot\}$. For every time frame t, consider the following.

(1) For every mapping (o_i, h_j) in M_{t-1}, verify if it is still valid. If object o_i is still visible and tracker hypothesis h_j still exists at time t, and if their distance does not exceed the threshold T, make the correspondence between o_i and h_j for frame t.

(2) For all objects for which no correspondence was made yet, try to find a matching hypothesis. Allow only one-to-one matches, and pairs for which the distance does not exceed T. The matching should be made in a way that minimizes the total object-hypothesis distance error for the concerned objects. This is a minimum weight assignment problem, and is solved using Munkres' algorithm [20] with polynomial runtime complexity. If a correspondence (o_i, h_k) is made that contradicts a mapping (o_i, h_j) in M_{t-1}, replace (o_i, h_j) with (o_i, h_k) in M_t. Count this as a mismatch error and let mme_t be the number of mismatch errors for frame t.

(3) After the first two steps, a complete set of matching pairs for the current time frame is known. Let c_t be

the number of matches found for time t. For each of theses matches, calculate the distance d_t^i between the object o_i and its corresponding hypothesis.

(4) All remaining hypotheses are considered false positives. Similarly, all remaining objects are considered misses. Let fp_t and m_t be the number of false positives and misses, respectively, for frame t. Let also g_t be the number of objects present at time t.

(5) Repeat the procedure from step 1 for the next time frame. Note that since for the initial frame, the set of mappings M_0 is empty, all correspondences made are initial and no mismatch errors occur.

In this way, a continuous mapping between objects and tracker hypotheses is defined and all tracking errors are accounted for.

2.2. Performance metrics

Based on the matching strategy described above, two very intuitive metrics can be defined.

(1) The multiple object tracking precision (MOPT):

$$\text{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}. \tag{1}$$

It is the total error in estimated position for matched object-hypothesis pairs over all frames, averaged by the total number of matches made. It shows the ability of the tracker to estimate precise object positions, independent of its skill at recognizing object configurations, keeping consistent trajectories, and so forth.

(2) The multiple object tracking accuracy (MOTA):

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}, \tag{2}$$

where m_t, fp_t, and mme_t are the number of misses, of false positives, and of mismatches, respectively, for time t. The MOTA can be seen as derived from 3 error ratios:

$$\overline{m} = \frac{\sum_t m_t}{\sum_t g_t}, \tag{3}$$

the ratio of misses in the sequence, computed over the total number of objects present in all frames,

$$\overline{fp} = \frac{\sum_t fp_t}{\sum_t g_t}, \tag{4}$$

the ratio of false positives, and

$$\overline{mme} = \frac{\sum_t mme_t}{\sum_t g_t}, \tag{5}$$

the ratio of mismatches.

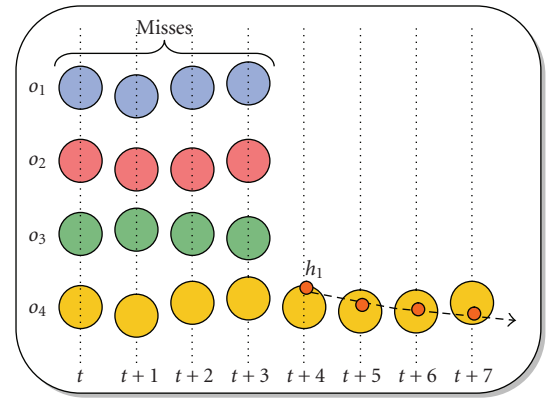

Figure 3: Computing error ratios. Assume a sequence length of 8 frames. For frames t_1 to t_4, 4 objects $o_1, \ldots, o_4$ are visible, but none is being tracked. For frames t_5 to t_8, only o_4 remains visible, and is being consistently tracked by h_1. In each frame $t_1, \ldots, t_4$, 4 objects are missed, resulting in 100% miss rate. In each frame $t_5, \ldots, t_8$, the miss rate is 0%. Averaging these frame level error rates yields a global result of $(1/8)(4.100 + 4.0) = 50\%$ miss rate. On the other hand, summing up all errors first, and computing a global ratio yield a far more intuitive result of 16 misses/20 objects = 80%.

Summing up over the different error ratios gives us the total error rate E_{tot}, and $1 - E_{\text{tot}}$ is the resulting tracking accuracy. The MOTA accounts for all object configuration errors made by the tracker, false positives, misses, mismatches, over all frames. It is similar to metrics widely used in other domains (such as the word error rate (WER), commonly used in speech recognition) and gives a very intuitive measure of the tracker's performance at detecting objects and keeping their trajectories, independent of the precision with which the object locations are estimated.

Remark on computing averages: note that for both MOTP and MOTA, it is important to first sum up all errors across frames before a final average or ratio can be computed. The reason is that computing ratios r_t for each frame t independently before calculating a global average $(1/n)\sum_t r_t$ for all n frames (such as, e.g., for the $\overline{\text{FP}}$ and $\overline{\text{FN}}$ measures in [19]) can lead to nonintuitive results. This is illustrated in Figure 3. Although the tracker consistently missed most objects in the sequence, computing ratios independently per frame and then averaging would still yield only 50% miss rate. Summing up all misses first and computing a single global ratio, on the other hand, produces a more intuitive result of 80% miss rate.

3. TRACKING EVALUATIONS IN CLEAR

The theoretical framework presented here for the evaluation of multiple object trackers was applied in two large evaluation workshops. The classification of events, activities, and relationships (CLEARs) workshops [11] as organized in a collaboration between the European CHIL project, the U.S. VACE project, and the National Institute of Standards and Technology (NIST) [21] (as well as the AMI project, in 2007), and were held in the springs of 2006 and 2007. They

represent the first international evaluations of their kind, using large databases of annotated multimodal data, and aimed to provide a platform for researchers to benchmark systems for acoustic and visual tracking, identification, activity analysis, event recognition, and so forth, using common task definitions, datasets, tools, and metrics. They featured a variety of tasks related to the tracking of humans or other objects in natural, unconstrained indoor and outdoor scenarios, and presented new challenges to systems for the fusion of multimodal and multisensory data. A complete description of the CLEAR evaluation workshops, the participating systems, and the achieved results can be found in [22, 23].

The authors wish to make the point here, that these evaluations represent a systematic, large-scale effort using hours of annotated data, and with a substantial amount of participating systems, and can therefore be seen as a true practical test of the usefulness of the MOT metrics. The experience from these workshops was that the MOT metrics were indeed applicable to a wide range of tracking tasks, made it easy to gain insights into tracker strengths and weaknesses, to compare overall system performances, and helped researchers publish and convey performance results that are objective, intuitive, and easy to interpret. In the following, the various CLEAR tracking tasks are briefly presented, highlighting the differences and specificities that make them interesting from the point of view of the requirements posed to evaluation metrics. While in 2006, there were still some exceptions, in 2007 all tasks related to tracking, for single or multiple objects, and for all modalities, were evaluated using the MOT metrics.

3.1. 3D visual person tracking

The CLEAR 2006 and 2007 evaluations featured a 3D person tracking task, in which the objective was to determine the location on the ground plane of persons in a scene. The scenario was that of small meetings or seminars, and several camera views were available to help determine 3D locations. Both the tracking of single persons (the lecturer in front of an audience) and of multiple persons (all seminar participants) were attempted. The specifications of this task posed quite a challenge for the design of appropriate performance metrics: measures such as track merges and splits, usually found in the field of 2D image-based tracking, had little meaning in the 3D multicamera tracking scenario. On the other hand, errors in location estimation had to be carefully distinguished from false positives and false track associations. Tracker performances were to be intuitively comparable for sequences with large differences in the number of ground truth objects, and thus varying levels of difficulty. In the end, the requirements of the 3D person tracking task drove much of the design choices behind the MOT metrics. For this task, error calculations were made using the Euclidian distance between hypothesized and labeled person positions on the ground plane, and the correspondence threshold was set to 50 cm.

Figure 4 shows examples for the scenes from the seminar database used for 3D person tracking.

3.2. 2D face tracking

The face tracking task was to be evaluated on two different databases: one featuring single views of the scene and one featuring multiple views to help better resolve problems of detection and track verification. In both cases, the objective was to detect and track faces in each separate view, estimating not only their position in the image, but also their extension, that is, the exact area covered by them. Although in the 2006 evaluation, a variety of separate measures were used, in the 2007 evaluation, the same MOT metrics as in the 3D person tracking task, with only slight variations, were successfully applied. In this case, the overlap between hypothesized and labeled face bounding boxes in the image was used as distance measure, and the distance error threshold was set to zero overlap.

Figure 5 shows examples for face tracker outputs on the CLEAR seminar database.

3.3. 2D person and vehicle tracking

Just as in the face tracking task, the 2D view-based tracking of persons and vehicles was also evaluated on different sets of databases representing outdoor traffic scenes, using only slight variants of the MOT metrics. Here also, bounding box overlap was used as the distance measure.

Figure 6 shows a scene from the CLEAR vehicle tracking database.

3.4. 3D acoustic and multimodal person tracking

The task of 3D person tracking in seminar or meeting scenarios also featured an acoustic subtask, where tracking was to be achieved using the information from distributed microphone networks, and a multimodal subtask, where the combination of multiple camera and multiple microphone inputs was available. It is noteworthy here, that the MOT measures could be applied with success to the domain of acoustic source localization, where overall performance is traditionally measured using rather different error metrics, and is decomposed into speech segmentation performance and localization performance. Here, the miss and false positive errors in the MOTA measure accounted for segmentation errors, whereas the MOTP expressed localization precision. As a difference to visual tracking, mismatches were not considered in the MOTA calculation, as acoustic trackers were not expected to distinguish the identities of speakers, and the resulting variant, the A – MOTA, was used for system comparisons. In both, the acoustic and multimodal subtasks, systems were expected to pinpoint the 3D location of active speakers and the distance measure used was the Euclidian distance on the ground plane, with the threshold set to 50 cm.

ITC-irst UKA AIT IBM UPC

Figure 4: Scenes from the CLEAR seminar database used in 3D person tracking.

UKA AIT

Figure 5: Scenes from the CLEAR seminar database used for face detection and tracking.

Figure 6: Sample from the CLEAR vehicle tracking database (i-LIDS dataset [6]).

Site/system	MOTP	Miss rate	False pos. rate	Mismatches	MOTA
System A	92 mm	30.86%	6.99%	1139	59.66%
System B	91 mm	32.78%	5.25%	1103	59.56%
System C	141 mm	20.66%	18.58%	518	59.62%
System D	155 mm	15.09%	14.5%	378	69.58%
System E	222 mm	23.74%	20.24%	490	54.94%
System F	168 mm	27.74%	40.19%	720	30.49%
System G	147 mm	13.07%	7.78%	361	78.36%

Figure 7: Results for the CLEAR'07 3D multiple person tracking visual subtask.

4. EVALUATION RESULTS

This section gives a brief overview of the evaluation results from select CLEAR tracking tasks. The results serve to demonstrate the effectiveness of the proposed MOT metrics and act as a basis for discussion of inherent advantages, drawbacks, and lessons learned during the workshops. For a more detailed presentation, the reader is referred to [22, 23].

Figure 7 shows the results for the CLEAR 2007 Visual 3D person tracking task. A total of seven tracking systems with varying characteristics participated. Looking at the first column, the MOTP scores, one finds that all systems performed fairly well, with average localization errors under 20 cm. This can be seen as quite low, considering the area occupied on average by a person and the fact that the ground truth itself, representing the projections to the ground plane of head centroids, was only labeled to 5–8 cm accuracy. However, one must keep in mind that the fixed threshold of 50 cm, beyond which an object is considered as missed completely by the tracker, prevents the MOTP from rising too high. Even in the case of uniform distribution of localization errors, the MOTP value would be 25 cm. This shows us that, considering the predefined threshold, *System E* is actually not very precise at estimating person coordinates, and that *System B*, on the other hand, is extremely precise, when compared to ground truth uncertainties. More importantly still, it shows us that the correspondence threshold T strongly influences the behavior of the MOTP and MOTA measures. Theoretically, a threshold of $T = \infty$ means that all correspondences stay valid once made, no matter how large the distance between object and track hypothesis becomes.This reduces the impact of the MOTA to measuring the correct detection of the number of objects, and disregards all track swaps, stray track errors, and so forth, resulting in an also unusable MOTP measure. On the other hand, if T approximates 0, all tracked objects will eventually be considered as missed, and the MOTP and MOTA measures lose their meaning. As a consequence, the single correspondence threshold T must be carefully chosen based on the application and evaluation goals at hand. For the CLEAR 3D person tracking task, the margin was intuitively set to 50 cm, which produced reasonable results, but the question of determining the optimal threshold, perhaps automatically in a data driven way, is still left unanswered.

The rightmost column in Figure 7, the MOTA measure, proved somewhat more interesting for overall performance comparisons, at least in the case of 3D person tracking, as it was not bounded to a reasonable range, as the MOTP was. There was far more room for errors in accuracy in the complex multitarget scenarios under evaluation. The

best and worst overall systems, *G* and *F* reached 78% and 30% accuracy, respectively. Systems *A*, *B*, *C*, and *E*, on the other hand, produced very similar numbers, although they used quite different features and algorithms. While the MOTA measure is useful to make such broad high-level comparisons, it was felt that the intermediate miss, false positive and mismatch errors measures, which contribute to the overall score, helped to gain a better understanding of tracker failure modes, and it was decided to publish them alongside the MOTP and MOTA measures. This was useful, for example, for comparing the strengths of systems *B* and *C*, which had a similar overall score.

Notice that in contrast to misses and false positives, for the 2007 CLEAR 3D person tracking task, mismatches were presented as absolute numbers as the total number of errors made in all test sequences.This is due to an imbalance, which was already noted during the 2006 evaluations, and for which no definite solution has been found as of yet: for a fairly reasonable tracking system and the scenarios under consideration, the number of mismatch errors made in a sequence of several minutes labeled at 1 second intervals is in no proportion to the number of ground truth objects, or, for example, to the number of miss errors incurring if only one of many objects is missed for a portion of the sequence.This typically resulted in mismatch error ratios of often less than 2%, in contrast to 20–40% for misses or false positives, which considerably reduced the impact of faulty track labeling on the overall MOTA score. Of course, one could argue that this is an intuitive result because track labeling is a lesser problem compared to the correct detection and tracking of multiple objects, but in the end the relative importance of separate error measures is purely dependent on the application. To keep the presentation of results as objective as possible, absolute mismatch errors were presented here, but the consensus from the evaluation workshops was that according more weight to track labeling errors was desirable, for example, in the form of trajectory-based error measures, which could help move away from frame-based miss and false positive errors, and thus reduce the imbalance.

Figure 8 shows the results for the visual 3D single person tracking task, evaluated in 2006. As the tracking of a single object can be seen as a special case of multiple object tracking, the MOT metrics could be applied in the same way. Again, one can find at a glance the best performing system in terms of tracking accuracy, *System D* with 91% accuracy, by looking at the MOTA values. One can also quickly discern that, overall, systems performed better on the less challenging single person scenario. The MOTP column tells us that *System B* was remarkable, among all others, in that it estimated target locations down to 8.8 cm precision. Just as in the previous case, more detailed components of the tracking error were presented in addition to the MOTA. In contrast to multiple person tracking, mismatch errors play no role (or should not play any) in the single person case. Also, as a lecturer was always present and visible in the considered scenarios, false positives could only come from gross localization errors, which is why only a detailed analysis of the miss errors was given. For better understanding,

Site/system	MOTP	Miss rate (dist > T)	Miss rate (no hypo)	MOTA
System A	246 mm	88.75%	2.28%	−79.78%
System B	88 mm	5.73%	2.57%	85.96%
System C	168 mm	15.29%	3.65%	65.44%
System D	132 mm	4.34%	0.09%	91.23%
System E	127 mm	14.32%	0%	71.36%
System F	161 mm	9.64%	0.04%	80.67%
System G	207 mm	12.21%	0.06%	75.52%

Figure 8: Results for the CLEAR'06 3D Single Person Tracking visual subtask

they were broken down into misses resulting from failures to detect the person of interest (*miss rate (no hypo)*), and misses resulting from localization errors exceeding the 50 cm threshold (*miss rate* (dist $>$ T)). In the latter case, as a consequence of the metric definition, every miss error was automatically accompanied by a false positive error, although these were not presented separately for conciseness.

This effect, which is much more clearly observable in the single object case, can be perceived as penalizing a tracker twice for gross localization errors (one miss penalty, and one false positive penalty). This effect is however intentional and desirable for the following reason: intelligent trackers that use some mechanisms such as track confidence measures, to avoid outputting a track hypothesis when their location estimation is poor, are rewarded compared to trackers which continuously output erroneous hypotheses. It can be argued that a tracker which fails to detect a lecturer for half of a sequence performs better than a tracker which consistently tracks the empty blackboard for the same duration of time. This brings us to the noteworthy point: just as much as the types of tracker errors (misses, false positives, distance errors, etc.) that are used to derive performance measures, precisely "*how*" these errors are counted, the procedure for their computation when it comes to temporal sequences, plays a major role in the behavior and expressiveness of the resulting metric.

Figure 9 shows the results for the 2007 3D person tracking acoustic subtask. According to the task definition, mismatch errors played no role and just as in the visual single person case, components of the MOTA score were broken down into miss and false positive errors resulting from faulty segmentation (*true miss rate, true false pos. rate*), and those resulting from gross localization errors (*loc. error rate*). One can easily make out *System G* as the overall best performing system, both in terms of MOTP and MOTA, with performance varying greatly from system to system. Figure 9 demonstrates the usefulness of having just one or two overall performance measures when large numbers of systems are involved, in order to gain a high-level overview before going into a deeper analysis of their strengths and weaknesses.

Figure 10, finally, shows the results for the 2007 face tracking task on the CLEAR seminar database. The main difference to the previously presented tasks lies in the fact that 2D image tracking of the face area is performed and the distance error between ground truth objects and tracker hypotheses is expressed in terms of overlap of the

Site/system	MOTP	True miss rate	True false pos. rate	Loc. error rate	A-MOTA
System A	257 mm	35.3%	11.06%	26.09%	1.45%
System B	256 mm	0%	22.01%	41.6%	−5.22%
System C	208 mm	11.2%	7.08%	18.27%	45.18%
System D	223 mm	11.17%	7.11%	29.17%	23.39%
System E	210 mm	0.7%	21.04%	23.94%	30.37%
System F	152 mm	0%	22.04%	14.96%	48.04%
System G	140 mm	8.08%	12.26%	12.52%	54.63%
System H	168 mm	25.35%	8.46%	12.51%	41.17%

Figure 9: Results for the CLEAR'07 3D person tracking acoustic subtask.

Site/system	MOTP (overlap)	Miss rate	False pos. rate	Mismatch rate	MOTA
System A	0.66	42.54%	22.1%	2.29%	33.07%
System B	0.68	19.85%	10.31%	1.03%	68.81%

Figure 10: Results for the CLEAR'07 face tracking task.

respective bounding boxes. This is reflected in the MOTP column. As the task required the simultaneous tracking of multiple faces, all types of errors, misses, false positives, and mismatches were of relevance, and were presented along with the overall MOTA score. From the numbers, one can derive that although systems *A* and *B* were fairly equal in estimating face extensions once they were found, *System B* clearly outperformed *System A* when it comes to detecting and keeping track of these faces in the first place. This case again serves to demonstrate how the MOT measures can be applied, with slight modifications but using the same general framework, for the evaluation of various types of trackers with different domain-specific requirements, and operating in a wide range of scenarios.

5. SUMMARY AND CONCLUSION

In order to systematically assess and compare the performance of different systems for multiple object tracking, metrics which reflect the quality and main characteristics of such systems are needed. Unfortunately, no agreement on a set of commonly applicable metrics has yet been reached.

In this paper, we have proposed two novel metrics for the evaluation of multiple object tracking systems. The proposed metrics—the multiple object tracking precision (MOTP) and the multiple object tracking accuracy (MOTA)—are applicable to a wide range of tracking tasks and allow for objective comparison of the main characteristics of tracking systems, such as their precision in localizing objects, their accuracy in recognizing object configurations, and their ability to consistently track objects over time.

We have tested the usefulness and expressiveness of the proposed metrics experimentally, in a series of international evaluation workshops. The 2006 and 2007 CLEAR workshops hosted a variety of tracking tasks for which a large number of systems were benchmarked and compared. The results of the evaluation show that the proposed metrics indeed reflect the strengths and weaknesses of the various used systems in an intuitive and meaningful way, allow for easy comparison of overall performance, and are applicable to a variety of scenarios.

ACKNOWLEDGMENT

The work presented here was partly funded by the *European Union* (EU) under the integrated project CHIL, *Computers in the Human Interaction Loop* (Grant no. IST-506909).

REFERENCES

[1] M. Voit, K. Nickel, and R. Stiefelhagen, "Multi-view head pose estimation using neural networks," in *Proceedings of the 2nd Workshop on Face Processing in Video (FPiV '05), in association with the 2nd IEEE Canadian Conference on Computer and Robot Vision (CRV '05)*, pp. 347–352, Victoria, Canada, May 2005.

[2] CHIL—Computers In the Human Interaction Loop, http://chil.server.de/.

[3] AMI—Augmented Multiparty Interaction, http://www.amiproject.org/.

[4] VACE—Video Analysis and Content Extraction, http://www.informedia.cs.cmu.edu/arda/vaceII.html.

[5] ETISEO—Video Understanding Evaluation, http://www.silogic.fr/etiseo/.

[6] The i-LIDS dataset, http://scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/video-based-detection-systems/i-lids/.

[7] CAVIAR—Context Aware Vision using Image-based Active Recognition, http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[8] F. Ziliani, S. Velastin, F. Porikli, et al., "Performance evaluation of event detection solutions: the CREDS experience," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '05)*, pp. 201–206, Como, Italy, September 2005.

[9] PETS—Performance Evaluation of Tracking and Surveillance, http://www.cbsr.ia.ac.cn/conferences/VS-PETS-2005/.

[10] EEMCV—Empirical Evaluation Methods in Computer Vision, http://www.cs.colostate.edu/eemcv2005/.

[11] CLEAR—Classification of Events, Activities and Relationships, http://www.clear-evaluation.org/.

[12] Y. Li, A. Dore, and J. Orwell, "Evaluating the performance of systems for tracking football players and ball," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '05)*, pp. 632–637, Como, Italy, September 2005.

[13] A. T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin, "ETISEO, performance evaluation for video surveillance systems," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '07)*, pp. 476–481, London, UK, September 2007.

[14] R. Y. Khalaf and S. S. Intille, "Improving multiple people tracking using temporal consistency," MIT Department of Architecture House_n Project Technical Report, Massachusetts Institute of Technology, Cambridge, Mass, USA, 2001.

[15] A. Mittal and L. S. Davis, "M2Tracker: a multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo," in *Proceedings of the 7th European Conference on Computer Vision (ECCV '02)*, vol. 2350 of *Lecture Notes in Computer Science*, pp. 18–33, Copenhagen, Denmark, May 2002.

[16] N. Checka, K. Wilson, V. Rangarajan, and T. Darrell, "A probabilistic framework for multi-modal multi-person tracking,"

in *Proceedings of the IEEE Workshop on Multi-Object Tracking (WOMOT '03)*, Madison, Wis, USA, June 2003.

[17] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proceedings of the 7th International Conference on Multimodal Interfaces (ICMI '05)*, pp. 61–68, Torento, Italy, October 2005.

[18] H. Tao, H. Sawhney, and R. Kumar, "A sampling algorithm for tracking multiple objects," in *Proceedings of the International Workshop on Vision Algorithms (ICCV '99)*, pp. 53–68, Corfu, Greece, September 1999.

[19] K. Smith, D. Gatica-Perez, J. Odobez, and S. Ba, "Evaluating multi-object tracking," in *Proceedings of the IEEE Workshop on Empirical Evaluation Methods in Computer Vision (EEMCV '05)*, vol. 3, p. 36, San Diego, Calif, USA, June 2005.

[20] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society of Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[21] NIST—National Institute of Standards and Technology, http://www.nist.gov/.

[22] R. Stiefelhagen and J. Garofolo, Eds., *Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006*, vol. 4122 of *Lecture Notes in Computer Science*, Springer, Berlin, Germany.

[23] R. Stiefelhagen, J. Fiscus, and R. Bowers, Eds., *Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops*, vol. 4625 of *Lecture Notes in Computer Science*, Springer, Berlin, Germany.