Intelligent Video Systems: A Review of Performance Evaluation Metrics that use Mapping Procedures

X. Desurmont
Xavier@desurmont.eu

C. Carincotte
Multitel
2, rue Pierre et Marie Curie,
B7000 Mons,
Belgium

F. Brémond
INRIA
2004, route des Lucioles, BP93,
06902 Sophia Antipolis Cedex,
France

Abstract

In Intelligent Video Systems, most of the recent advanced performance evaluation metrics perform a stage of mapping data between the system results and ground truth. This paper aims to review these metrics using a proposed framework. It will focus on metrics for events detection, objects detection and objects tracking systems.

1. Introduction

Performance evaluation has become an increasingly important topic when dealing with video intelligent systems. However, while many concurrent metrics exist, they are not formalised in the same way which make it difficult to compare them in a fair manner. For some applications, metrics need to perform a mapping (an assignment) between Result and Ground truth data. This paper proposes, in a first step, to introduce that class of metrics. In a second step, it reviews successively existing metrics for event detection, object detection and object tracking systems. Finally it concludes with a summary of the review and proposes some guidelines for designing new metrics.

2. Evaluation metrics

Usually, during evaluation procedure of an Intelligent Video System (IVS), a metric calculates scores (SC) by comparing the IVS' Result (RS) with the Ground Truth (GT) which is the expected correct result (often made by manual annotation).

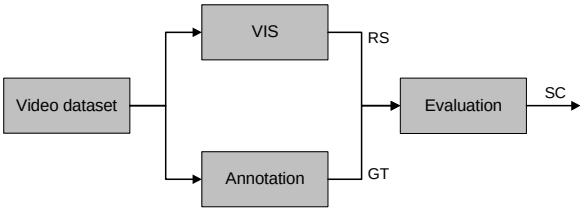

Figure 1. Performance evaluation framework.

In most cases, RS is a set of several entities (e.g. events) that could differ (e.g. time, location, etc.) from the GT because of possible errors of the RS (e.g. non detection). These errors could be qualitative or/and quantitative. In the following section, we describe these two types of errors and some related scoring techniques.

1) Qualitative errors come from qualitative processes such as classification (e.g. misclassification). Every process leads to practical decisions that can be evaluated ("is this pixel part of the background or the foreground?"). A GT lets you classify RS decisions as correct or incorrect. When it is correct it is called "true" and when it is incorrect it is called "false". The comparison of RS and GT values entails four possible issues:

- •True positives (TPs): RS confirmed by GT,
- •False positives (FPs): RS not matched in GT,
- •True negatives(TNs):RS rejected and not part of GT,
- •False negatives (FNs): RS rejected but part of GT.

Note that in a detection problem, which is slightly different from binary classification, "true positives" are typically named "correct detections", "false positives" are "false detections/alarms" and "false negatives" are "non detections". Some useful metrics derived from TPs, FPs, FNs and TNs are important for gathering information about the performances of a detection system [Altman94]:

General name	***Function***
Detection Rate (DR) or Sensitivity	$N_{tp}/(N_{tp}+N_{fn})$
Classification:False Positive Rate (FPR)	$N_{fp}/(N_{fp}+N_{tn})$
Detection: False Alarm Rate (FAR)	N_{fp} per time units

Table 1. Derived values from the contingency table.

2) Quantitative errors are made by quantitative processes. Typical errors affect position, the object's shape, the object's speed or the delay/advance in a time stamp. To quantify these errors we use scoring techniques that quantify the accuracy of the detection or the tracking algorithms. Examples are "average number

of observations before tracking is initiated"; "average number of frames before tracking is terminated", Euclidean distance between the RS position and the GT position of an object or the distance from nearest segments in the two bounding boxes [Brémond97], etc. We formalise these scoring techniques as Entity Precision Score (EPS) that evaluate a quantitative result of features representing an entity such as time, position, size, colour and shape, track and speed. EPS is usually specific to an entity such as an object or event. It could be assimilated as a result of similarity distance. Table 2 shows some generic examples.

Features	***Entity Precision Score (EPS)***
Position *(x,y)*	$\sqrt{(x_{rs}-x_{gt})^2+(y_{rs}-y_{gt})^2}$
Bounding box [Kasturi09]	$\frac{B_{gt}\cap B_{rs}}{B_{gt}\cup B_{rs}}$ Intersection/overlap ratio
Time *t*	$\lvert t_{rt}-t_{gt}\rvert$

Table 2. Examples of entity precision scores for different features.

3) Dual qualitative/quantitative scoring techniques exist because IVS algorithms often make composite qualitative and quantitative errors. Take the application of face detection in images. In that case, errors can involve the detection of the face but also the precision of the face's position.

In the particular case of object tracking, the output of a tracking system is the set of trajectories of objects in the scene. As described by Smith *et al.* [Smith05], there are key properties for a good tracker, such as (i) tracking objects well; placing the correct number of trackers at the correct locations for each frame, (ii) identifying objects well; tracking individual objects consistently over a long period of time. Typical errors are thus about locations and identities. When the tracking system mismatches two objects because of an inversion, this can be seen as an identification error. On the other hand, when the position given by the system differs slightly from the ground truth, it is considered a location error. Thus, visual tracking evaluation should be made for both qualitative and quantitative errors. Figure 2 and Figure 3 show the two possible types of error. Sometimes the error cannot be classified clearly in a misidentification or location drift; it really depends on the interpretation.

In dual qualitative/quantitative scoring techniques, an important issue is the mapping of RS and GT. It consists of choosing matches between entities of the RS and the GT. Since manual mapping of all the data would take a lot of time, this should be handled automatically by the evaluation metric. Indeed, several approaches have been proposed [Senior01, Bruneaut05, Brown05, Smith05, Etiseo06, Manohar06, Desurmont06, Bernardin08 and Kasturi09] to tackle the mapping issue in different ways. In the following section we review some of these metrics.

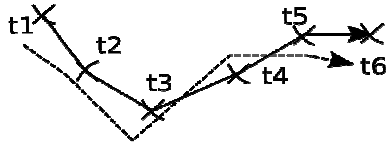

Figure 2. Location error between GT (continuous line) and RS (dashed line).

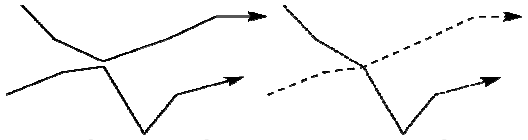

Figure 3. Identity error between GT (continuous lines) and RS (dashed lines).

3. Review of metrics with mapping

In order to review the metrics of the literature, we need to formalize one concept: the "system behaviour model" (SBM). Indeed a metric evaluates the result of a system according to an implicit model of that system's behaviour. In that scope, the SBM describes the possible errors the metric is able to cope with. For example, in detection systems, a common SBM defines that the system can produce the following errors: false detection, non detection and jitter in the time stamp of the detection.

We contend that evaluation problems are ill-posed as they usually do not define the SBM clearly. This sometimes induces inconsistence in the metric with the respect to the SBM. For example, with detection systems, if the metric is not able to take account of "time-stamp jitter", the metric will not be able to count most correct detections because of impossibility of matching events that are not exactly placed in GT and RS at the time-stamp.

In this section we try to highlight some problems raised by various metric strategies in the literature, with focus on qualitative and quantitative errors: event detection (time), object detection (space) and object tracking (time, space and identities). We describe these metrics by reviewing their SBMs, one-to-one EPS processes and mapping processes and then highlight possible drawbacks. Some metrics may have been mis-interpreted due to their complexity and to the lack of details provided by the available documents describing the metrics.

3.1. Event detection metrics

We propose investigating the different event detection metrics using the toy example of Figure 4 representing a case in which there are three GT events and four RS events. Note that in this example, events have a temporal duration and are represented as a time interval with a beginning and ending time. In that example, experts usually consider RS1, RS2 and RS4 to be correct detections and RS3 as a false detection.

Figure 4. Example of result events *vs* ground truth events.

1) Bruneaut *et al.* [Bruneaut05] proposed a metric in the framework of Challenge of Real-time Event Detection Solutions (CREDS) in 2005. The SBM handles temporal shifts. The EPS is a function of the delay/anticipation and of duration ratio of RS and GT events. The CREDS metric defines how to compute correct detections and false positive and false negative detections. Then a weight is assigned to each of these detections. The overall score for a given scenario is the sum of all the correct, false and non detection scores. The metric matches events from ground truth and result with the handling of time shifts. A match is defined as the first occurrence of a result event that overlaps a ground truth event in time; it is considered a correct detection. If multiple result events overlap with the same ground truth event then only the first (in time) result event is matched, while the others are classified as false detections.

The major drawback of this method is when events are so frequent that the possibility of early and delayed events entails the overlapping of several events, thus resulting in wrong matches. Figure 5 shows assignments from the toy example where the GT3 event overlaps with both the RS3 and then the RS4 event. The evaluation matches RS3-GT3, i.e., events that do not correspond! Moreover, the metric does not state what is happening if a unique result event overlaps with two ground truth events. It is not clear which ground truth event should be matched.

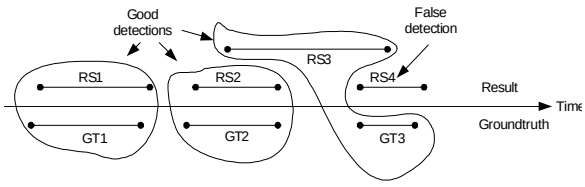

Figure 5. Example of CREDS mapping between RS and GT.

2) The "Text REtrieval Conference" sponsors a video "track" devoted to research in automatic segmentation, indexing and content-based retrieval of digital video dubbed Trecvid. It proposes a metric for evaluating event detection [Trecvid08] that is an improvement of Bruneaut *et al.*'s proposal. This metric works with GT and RS events defined with Viper XML format with a start and end time. RS events should also provide a decision confidence for the event. The SBM handles temporal shifts, correct, false and non detections. The EPS between two events is more or less proportional to the sum of the intersection of the time interval and the decision confidence. Equations (1) show a simplified version of the correspondence matrix computation procedure. The latter uses an event alignment procedure with a one-to-one mapping with GT and RS using the Hungarian solution [Munkres57] to the bipartite graph matching problem by modelling event observations as nodes in the bipartite graph. The toy example with mapping is shown in Figure 6 and Figure 7.

$$Kernel(RSi,GTj)=\begin{cases}\phi & if \quad Mid(RSi) > End(GTj)+\Delta_T \\ \phi & if \quad Mid(RSi) < Beg(GTj)-\Delta_T \\ 1 & + \quad TC(RSi,GTj)\end{cases}$$

$$TC(RSi,GTj)=\frac{Min(End(GTj),End(RSi))-Max(Beg(GTj),Beg(RSi))}{End(GTj)-Beg(GTj)} \quad \textbf{(1)}$$

$Beg(\,)$= The beginning of event's time span
$Mid(\,)$= The midpoint of event's time span
$End(\,)$= The end of event's time span
$\Delta_T = 0.5(s)$; a constant differentiating the mappable and unmappable events

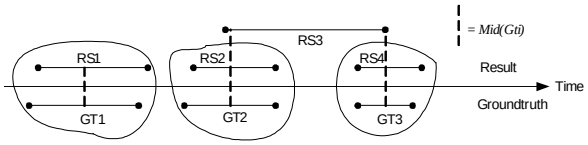

Figure 6. Example of TRECVID mapping between RS and GT.

$$Kernel(RSi,GTj)=\begin{bmatrix} 11/12 & \phi & \phi \\ \phi & 9/10 & \phi \\ \phi & 5.5/10 & 3/6 \\ \phi & \phi & 6/6 \end{bmatrix} \Rightarrow Mapping=\begin{bmatrix} M & \phi & \phi \\ \phi & M & \phi \\ \phi & \phi & \phi \\ \phi & \phi & M \end{bmatrix}$$

Figure 7. Kernel values and best mapping for Figure 6 example.

3) We note that, because of the definition of the events' score, the Trecvid and the CREDS metrics are not able to evaluate events with no duration such as systems that simply trigger off alarms. Desurmont *et al* [Desurmont06] propose a metric to handle evaluation of

these duration-less events. We use the toy example shown in Figure 8: When looking at all the events on the same timeline to analyse the system, one will probably match them as A-α, B-β, C-δ and thus conclude that there are three good detections, one false alarm (D) and a miss-detection (γ).

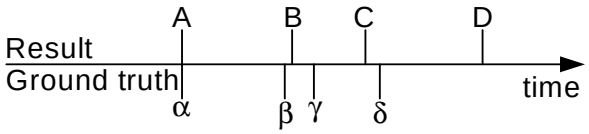

Figure 8. Representation of events on a timeline.

The SBM assumes that the possible deviations in the event detection system are a set of false positives, false negatives, delays and advances for each potential individual event. In practical terms, this means that it sometimes allows no match between events of GT and RS and matches can also be made with events having different time stamps. The aim of the approach is to process a dynamic re-alignment of the system's RS according to the GT in order to find the best mapping by minimising an overall cost. The EPS is a cost defined as an absolute difference of time between events. Costs are also set to false positives (FPDist) and false negatives (FNDist). The global cost minimisation can be optimised with a dynamic programming approach based on "dynamic time warping" and "sequence alignment" (Needleman/Wunsch techniques). It then becomes a straightforward matter of counting the number of matches, the number of false detections and the number of non-detections. Computational complexity analysis: Let N be the number of events in the ground truth and M the number of events in the result. The algorithm of dynamic programming used in this proposal is a complexity of $O(N \times M)$. It is lower than the Trecvid proposal, which uses the Hungarian algorithm with a $O(\max(N,M)^3)$ complexity.

Summary of event detection metrics review: The Trecvid approach is fully consistent. It proposes an SBM and then the procedure chooses the mapping that maximises an overall score. For duration-less events, the problem formalization of Desurmont *et al* is similar as Trecvid but the implementation uses a faster algorithm.

3.2. Object detection metrics

We propose investigating the different object detection metrics using the toy example of Figure 9 representing three cases of result objects given a ground truth of three objects GT1, GT2 and GT3. An object is defined by a Bounding Box (BBOX) region. On the top a) we can consider that there are three correct detections and three result objects (RS1, RS2 and RS3) that intersect respectively with ground truth objects GT1, GT2 and GT3 only. In the middle b) we can also consider that there are three correct detections but some minor overlap problems, for results RS1 and RS2 overlap with several ground truth objects. On the bottom c) we can consider that GT1 is detected with a fragmentation problem (RS1 and RS2), GT2 is not detected and GT3 is partially detected by RS3.

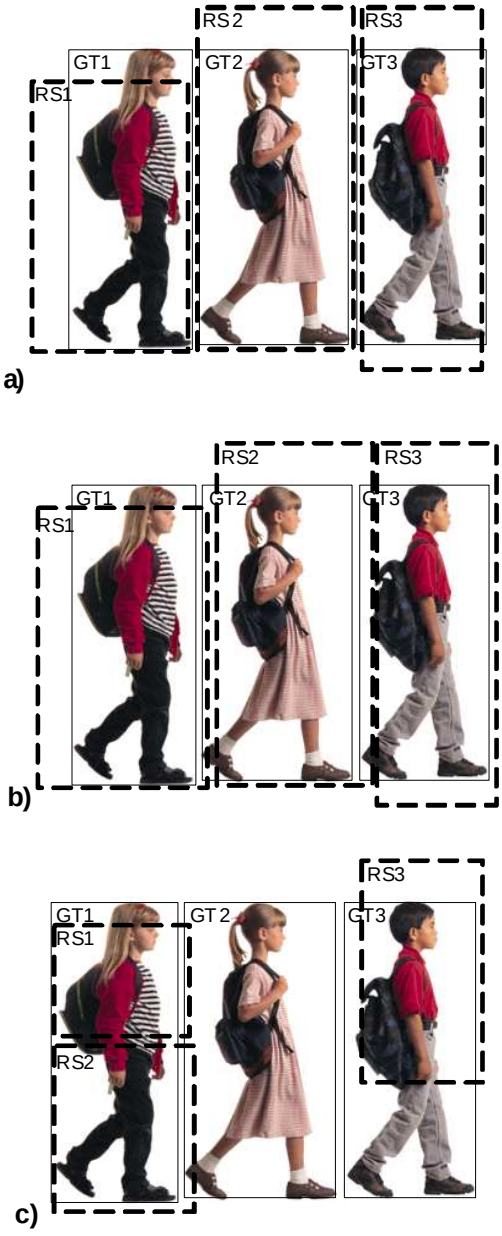

Figure 9. Three examples of RS BBOXs *vs.* a given GT.

1) Nascimento *et al*. [Nascimento04] suggest a method for object detection evaluation. The SBM includes correct detection, false detection, non detection, merge, split and split-merge. The EPS between one GT region and one RS region is binary: 1 if there is a spatial intersection, 0 if not. The method accounts for a correct detection when the RS region matches one and only one GT region, false detection when the RS region has no correspondence with the GT, non detection when the GT region has no correspondence with the RS, merge region when the RS region is associated with several GT regions, split region when the GT region is associated with several RS regions and finally split-merge region when the region is at the same time a split and a merged region.
The drawback of the mapping procedure is that the spatial noise in RS regions entails inconsistent metric scores. In the example of Figure 9, b) should be considered to be three correct detections with small spatial deviation while the proposed metric considers it to be two splits and two merges while c) should be considered a miss-detection of GT2 and over-segmentation of GT1 and GT3 but the proposed metric considers it two splits and two merges. Thus b) and c) are scored the same by the proposed metric while the RS in c) should be considered worse than the RS of c).

2) The Etiseo project [Etiseo06] proposes a detection metric that counts correctly detected, misdetected and falsely-detected objects. EPS between objects should be chosen between several ones like the overlapping ratio or the maximum deviation. The issue of matching pairs of RS and GT data is done by first computing a one-to-one EPS. Second some matches are done when the measurements are above a threshold. Thus this matching is neither unique nor optimal. In Figure 9, c) RS1 and RS2 are matched with GT1 and GT2 (RS1-GT1, RS1-GT2, RS2-GT1, RS2-GT2) and thus no misdetection is detected.

3) Manohar *et al*. [Manohar06] propose a frame-level measurement of object detection (FDA) that accounts for the objects correctly detected, miss-detected and falsely-detected. The EPS between objects consist of computing the spatial overlap (see overlap ratio in Table 2) between ground truth and result BBOX of two objects matched by a mapping procedure. Then the sum of the overlaps of objects is normalised over the average of the number of ground truth and result objects in order to build the FDA.

The mapping of object pairs is built using the "Hungarian algorithm" [Munkres57] with the criteria of FDA maximisation. The mapping procedure entails a unique comprehensive score. However, one disadvantage is that there is no minimum for the overlapping ratio between matched objects, so that objects are sometimes matches despite having only a very narrow intersection. In the toy example of Figure 9, a) the mapping will be GT1-RS1, GT2-RS2 and GT3-RS3, b) the mapping will be also GT1-RS1, GT2-RS2 and GT3-RS3, and c) GT1-RS1, GT2-RS2 and GT3-RS3. Thus GT2-RS2 is wrong.

Summary of object detection metrics review: Manohar's approach seems interesting because it is consistent with its defined SBM. Others methods can produce inconsistent scores.

3.3. Object tracking metrics

We propose investigating some object tracking metrics using the toy example of Figure 10: Three objects appear at time t1 and are correctly detected (GT1-RS1, GT2-RS2, GT3-RS3) and tracked until time t3, when GT2 and GT1 cross each other's path, causing a tracking error, and then RS3 is wrongly attached to GT1. At time t4, GT2 and GT3 are near, which causes some position errors for RS2 and RS3. Then there are no more errors until the end of the sequence at time t6.

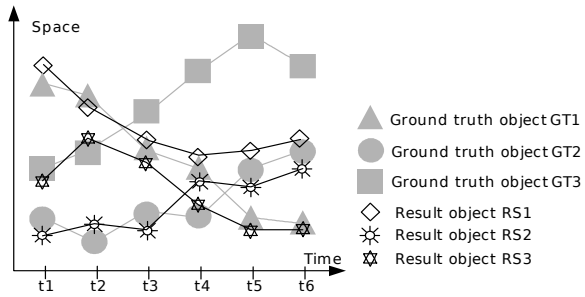

Figure 10. Toy example of object tracking for investigating tracking evaluations.

We can divide the methods of object tracking evaluation roughly into two groups. The first group proposes to map tracks from GT and RS in a sharing strategy, that is, each track in the ground truth can be assigned to one or more tracks from the results and vice versa. In our example this means that GT1 can be matched with RS1 and RS3 at the same time (at times t3 and t4).

1) In Senior *et al*.'s proposal [Senior01], the metric that matches system tracks to ground truth tracks first computes the EPS from the distance (based on spatial proximity and the overlap duration) between each possible pair of tracks from GT and RS, then a correspondence matrix is constructed using the track distance measure and finally track correspondence

mapping is established by thresholding this matrix. Each track in the ground truth can be assigned one or more tracks from the results but not vice versa. This accommodates fragmented tracks but then the method is not able to state anything about some problems encountered by tracking algorithms (e.g. in Figure 10 at time t4, t5 and t6 the method may interpret RS1-GT2 as a fragmentation of RS2-GT2 whereas it is clearly a "merge error").

2) Brown *et al*. [Brown05] propose to enhance Senior's proposal with a two pass match between results tracks and ground truth tracks in a "system-track-matching" and a "GT-track-matching" but made with local criteria with possibilities of multiple matching for a unique track (see Figure 11). However, this method has the same drawback as Senior's proposal when it comes to misinterpreting some split/merge problems.

```
1. System-Track-Matching – for every system
track find all "GT-matches"
     "GT-match"    =    Temporal-Overlap    AND
Spatial-Overlap
 Temporal-Overlap = overlap/(system duration)
       Spatial-Overlap = GT centroid inside
E1% enlarged system bounding box
   If cumulative temporal/spatial overlap <
T1, then system track has
    insufficient matches and is labelled as
FP.
   If multiple GT-matches, then this system
track has merge error = # matched GT tracks

2. GT-Track-Matching – for every GT track find
all "system-matches"
     "System-match"  =  Temporal-Overlap  AND
Spatial-Overlap
       Temporal-Overlap      =      overlap/(GT
duration)
       Spatial-Overlap   =   system   centroid
inside E2% enlarged GT bounding box
   If cumulative temporal/spatial overlap <
T2, then GT track has
    insufficient matches and is labelled as
FN.
   If multiple system-matches, then this GT
track has fragmentation error =
    # matched Sys tracks
```

Figure 11. GT/RS Matching procedure for tracking proposed by Brown *et al*.

The second group of methods for object tracking evaluation proposes mapping between tracks from GT and RS that are chosen over a large set of matching possibilities using the maximisation of a criterion.

3) Manohar *et al*. [Manohar06] propose a tracking metric similar to their object detection evaluation scheme but in which "objects" are changed by "tracks". They try to match tracks from results and ground truth in order to maximise the spatial overlap (which is the EPS) as a whole, again using the "Hungarian algorithm". However, the underlying SBM does not integrate the notion of misidentification of the tracking algorithm. (e.g., in Figure 10 the mapping will be GT1-RS1, GT2-RS2 and GT3-RS3 and the misidentification that starts at time t3 is not detected and just seen as a location error). Furthermore, as shown by Bernardin *et al*, this kind of approach can become non intuitive [Bernardin08], e.g. Figure 12: For Case 1, RS1 matched with GT1, causing 2 mismatches at time t1 and t2; for Case 2, RS1 matched with GT1, yielding 1 mismatch at time t1; and for Case 3, RS2 matched with GT1, causing 2 mismatches at times t5 and t6. The three error's cases are semantically similar but a metric that matched a unique GT track to a unique RS track would give a better score for Case 2.

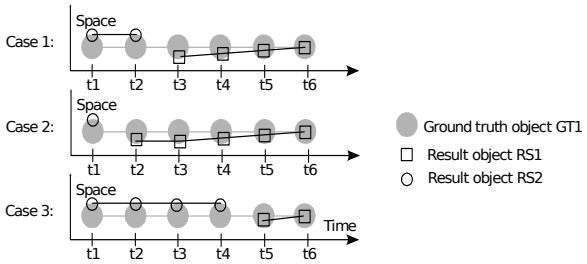

Figure 12. Three different cases of a similar identity swap error during tracking.

4) Bernardin *et al*. [Bernardin08] propose the CLEAR MOT Metrics based on two values: MOTP (Multiple Object Tracking Precision) (2), which measures, as EPS, the error of positions of tracked objects, and MOTA (Multiple Object Tracking Accuracy), which measures the number of occurrences of errors such as loss of tracks and mismatches. They count mismatches errors only once at the frame where a change in GT-RS mapping is made (when each mismatch starts). The SBM thus handles problems such as misidentification that can occur at anytime. d_t^i is the distance between a result object and its corresponding ground truth object at time t. c_t is the number of matches found for time t. The matching is driven by chronological order. When new tracks start they are mapped with the "Hungarian algorithm" but only for this first frame of the tracks. Thus it is not the best possible matching in terms of MOTP maximisation on the overall time sequence. Figure 13 shows an example: Following the mapping procedure we have *MOTP=10.83*. Now, if we reverse all the positions of objects between time t3 and t1, we have *MOTP=15.16* (the computation can be found in [desurmont09].) Thus we have two different MOTP for two semantically similar cases.

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \quad (2)$$

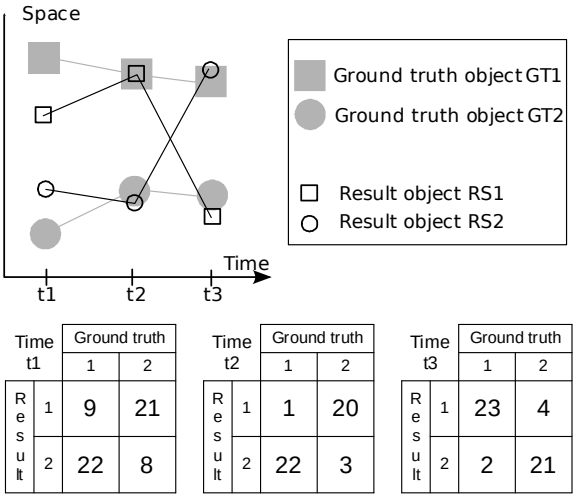

Figure 13. Example of tracking that shows a problem for the MOTP metric.

Summary of object tracking metrics: Manohar's approach achieves the best matching of RS and GT but the underlying SBM output does not allow any statements about misidentification. In contrast, Bernardin's approach takes account of misidentification, but the procedure for finding the "best" mapping is not optimal: the score is maximised but only frame-by-frame, not for the whole sequence. Such state-of-the-art approaches highlight the fact that there is no consistent object tracking evaluation metric.

4. Conclusion of metrics proposals' review

We reviewed and analysed some evaluation metrics highlighting some underlying concepts that we tried to formalise. Indeed the analysis was conducted with regard to four important aspects, namely, the metric type, the EPS (entity precision score), the SBM (system behaviour model) and the procedure to choose a mapping for GT and RS.

We summarise our review in Table 3. Three reviewed metrics produce consistent scores in any case (with respect to their SBM): Trecvid08's proposal for event detection with interval duration, Desurmont06's proposal for duration-less event detection and Manohar *et al.*'s proposal for object detection. We haven't found any object tracking metric that is fully consistent.

We don't claim that any metric is good or bad. Indeed for some reviewed metrics, the SBM is very complex and thus it is difficult to build an evaluation algorithm that avoids any inconsistency. It may be why there is no consistent metrics for object tracking systems. Moreover some reviewed metrics may have been mis-interpreted due to their complexity and to the lack of details provided by the available documents describing the metrics.

Based on our review, we propose some guidelines when designing a new metric:

- Define clearly the SBM with the real possible errors of any system that could be evaluated,
- Define clearly the rules of possible match between entities of RS and GT,
- Define a deterministic score for each possible local error (EPS and qualitative score),
- (Optional but useful for practical reason) find an optimised fast way to browse all possible mappings and related global scores to find the optimal solution (e.g. with dynamic programming algorithm.)

Next generation metrics may include all ideas of the reviewed metrics in order to cope with the complexity of all applications as well as the need of objective and consistent evaluation.

References

[Altman94] D.G. Altman, M. Bland, "Diagnostic tests 2: predictive values", BMJ 1994.

[Bernardin08] K. Bernardin, R. Stiefelhagen," Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics", EURASIP Journal on Image and Video Processing, EURASIP Journal on Image and Video Processing, Special Issue on Video Tracking in Complex Scenes for Surveillance Applications, Volume 2008, Article ID 246309, May 2008.

[Brown05] M. Brown, A.W. Senior, Y. Tian, J. Connell, A. Hampapur, C. Shu, H. Merkl, and M. Lu, "Performance Evaluation of Surveillance Systems Under Varying Conditions", Proceedings IEEE Workshop on PETS, pp 1-8, Breckenridge, CO USA, 7 January 2005.

[Bruneaut05] P. Bruneaut, A. Cavallaro, T. Kelliher, Lucio Marcenaro, F. Porikli, S. Velastin, F. Ziliani, "Performance Evaluation Of Event Detection Solutions: the CREDS experience", CREDS - Special session, AVSS, pp. 201-206, September 2005.

[Desurmont06] X. Desurmont, R. Sebbe, F. Martin, C. Machy and J.-F. Delaigle. Performance Evaluation of Frequent Events Detection Systems. IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS). Pp 15-21, New York, USA. June 2006.

[Desurmont09] X. Desurmont, “Objective Performance Evaluation and Optimal Allocation Framework for Video Analysis Methods”, Phd Thesis, Ecole Polytechnique de Louvain, Louvain-la-neuve, Belgium, June 2009.

[Etiseo06] http://www-sop.inria.fr/orion/ETISEO/iso_album/etimetrics_definition-v2.pdf, Etiseo Internal Technical Note, "Metrics Definition", version 2.0 – Approved, 6th June 2006.

[Kasturi09] 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 Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 2, Pages: 319-336, Feb 2009.

[Manohar06] V. Manohar, P. Soundararajan, H. Raju, D. Goldgof, R. Kasturi, and J. Garofolo, "Performance Evaluation of Object Detection and Tracking in Video", In the Seventh Asian Conference on Computer Vision, LNCS 3852, pp. 151-161, Jan 2006

[Munkres57] 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.

[Nascimento04] J. Nascimento and J.S. Marques, “New performance evaluation metrics for object detection algorithms”, 6th International Workshop on Performance Evaluation for Tracking and Surveillance (PETS 2004), ECCV, Prague, Czech Republic, May 2004.

[Senior01] A. Senior *et al*., “Appearance Models for Occlusion Handling", 3rd International Workshop on Performance Evaluation of Tracking and Surveillance, PETS2001, Hawaii, 2001.

[Trecvid08] http://www.nist.gov/speech/tests/trecvid/2008/doc/EventDet08-EvalPlan-v07.htm updated on the 21st August 2008.

Reference	Metric type	Entity Precision score (EPS)	System Behaviour Model (SBM)	Comments about the mapping procedure
CREDS, Bruneaut05	Event detection	Function of the delay/anticipation and of duration ratio of RS and GT events.	Temporal shift, correct, false and non detection	Not fully defined, can produce inconsistent scores
Trecvid08	Event detection	Sum of the intersection of time interval and decision confidence	Temporal shift, correct, false and non detection	Uses optimal one-to-one matching [Munkres57]: **consistent**
Desurmont06	Duration-less event detection	Absolute time difference	Temporal shift, correct, false and non detection	Uses dynamic time warping optimal one-to-one matching: **consistent**
Nascimento04	Object detection	Binary: 1 if there is a spatial intersection, 0 if not	Spatial shift, correct, false and non detection, merge, split and split-merge	Can produce inconsistent scores
Etiseo06	Object detection	Several possibilities of EPS: overlapping ratio, maximum deviation, etc.	Spatial shift, correct, false and non detection	Can produce not optimal scores
Manohar06	Object detection	Spatial overlap ratio	Spatial shift, correct, false and non detection	Uses optimal one-to-one matching [Munkres57]: **consistent**
Senior01	Object tracking	Spatial proximity and the overlap duration between each possible pair of tracks	Spatial shift, correct, false and non detection, split of tracks	Can produce inconsistent scores
Brown05	Object tracking	Spatial proximity and the overlap duration between each possible pair of tracks	Spatial shift, correct, false and non detection, split and merge of tracks	does misinterpret some split/merge problems, can produce inconsistent scores
Manohar06	Object tracking	Spatial overlap ratio for the objects along the track	Spatial shift, correct, false and non detection	Can produce inconsistent scores
CLEAR MOT, Bernardin08	Object tracking	MOTP: Multiple Object Tracking Precision (2)	Spatial shift, correct, false and non detection, misidentification	Munkres57 one-to-one matching but only on the first frame, so not optimal for the whole track, can produce inconsistent scores

Table 3. Summary of reviewed metrics according to metric type, EPS and SBM.