ETISEO

Internal Technical Note

Metrics Definition

version 2.0 – Approved

IN_ETI_1_004

Document Title :	Metrics Definition
Document version :	version 2.0
Document status :	Approved
Date :	2006-01-06
Availability :	
Authors :	Silogic - Inria

Abstract

This document presents criteria & metrics used for the ETISEO evaluation.

Keyword List

Metrics, Ground-truth, annotation, evaluation, tracking, detection, classification, event recognition.

DOCUMENT CHANGE LOG

Document Issue.	Date	Reasons for change
Version 1- draft 4	16 Nov 2005	Metrics detail.
Version 2 - 0	6 Jan 2006	Apply changes after seminar feedback.

APPLICABLE AND REFERENCE DOCUMENTS (A/R)

A/R	Reference	Title
1	E01	Introduction to evaluation and metrics
2		Video Performance Evaluation Resource (VIPER) – Performance Evaluation Manual
3		Performance Evaluation Metrics and Statistics for Positional Tracker Evaluation. C.J. Needham, R.D. Boyle

Table of contents

1. GENERAL REMARKS ON METRICS

1.1. RESULT OF THE METRICS

When measuring the ability of an algorithm to face a certain problem, we will use metrics that can output:

- A defined value for example 2D localisation error, time interval between object apparition in the ground truth and its detection in the system... In these cases the metrics can compute statistics on these values (mean, deviation, min & max values…), the statistics being for example computed along the video sequence.

- A number of successes or failures of the algorithm faces to situations. In that case, we use the standard definitions:
  - The **True Positive (TP)**: the system has detected a real situation (exists in reference data and algorithm results).
  - The **False Negative (FN)**: a real situation has been missed by the system (exists only in reference data).
  - The **False Positive (FP)**: the system has detected a situation that is not real (exists only in algorithm results).
  - The **True Negative (TN)**: entity that does not fit neither with a reference data, nor an algorithm result.

These data may provide also more information that will characterize algorithms as:

- The **precision** = TP / (TP + FP),
- The **sensitivity** = TP / (TP+FN)
- The **specificity** = TN / (FP + TN),
- The **F-score** = 2*precision*sensitivity / (precision + sensitivity): harmonic mean of the precision and sensitivity.

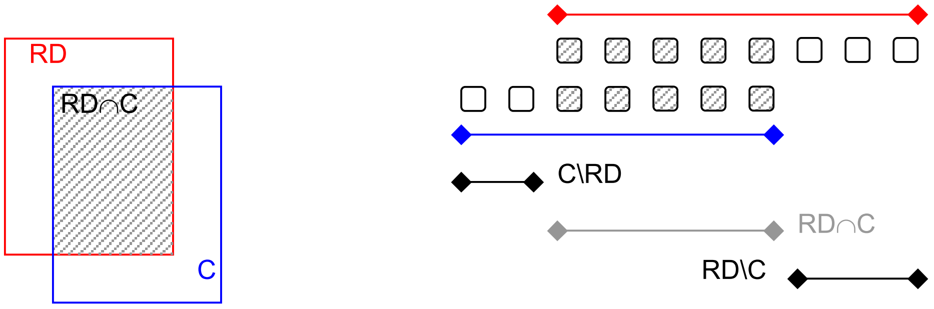
1.2. MATCHING BETWEEN REFERENCE DATA AND OBSERVATION

In order to compute some metrics, we will sometime need to match objects from the reference data to objects belonging to the observation (algorithm results).

We define hereafter two majors sets used for matching purpose, areas and time intervals.

We call:

- **RD**: time interval or area of interest of a Reference Data
- **C**: time interval or area of interest of a Candidate data (result of a system).



(a) (b)

Figure 1. (a) area of interests (b) time intervals- of a reference data and a candidate for distance comparison

To evaluate the matching between a candidate result and a reference data, we may use following distances:

- **D1-The Dice coefficient**: Twice the shared, divided by the sum of both intervals: 2*card(RD∩C) / (card(RD) + card(C)).
- **D2-The overlapping**: card(RD∩C) / card(RD).
- **D3-Bertozzi** and al. metric: (card(RD∩C))^2 / (card(RD) * card(C)).
- **D4-The maximum deviation** of the candidate object or target according to the shared frame span: Max { card(C\RD) / card(C), card(RD\C) / card(RD) }.

These distances will provide values that may be interpreted "as it" but also permit, after thresholding, to validate the matching criteria.

1.3. TASKS

Criteria and metrics are organised considering the following tasks:

- ❑ T1) Detection of physical objects of interest,
- ❑ T2) Localisation of physical objects of interest,
- ❑ T3) Tracking of physical objects of interest,
- ❑ T4) Classification of physical objects of interest,
- ❑ T5) Event recognition.

1.4. PRECISION – SENSITIVITY OR STATISTIC

Some of the following metrics computes a unique statistic value. In some cases equivalent GD good detection & FD false detection number can be determined. Precision and sensitivity can be calculated. The corresponding metrics are pointed with an index (*).

2. T1- DETECTION OF PHYSICAL OBJECTS OF INTEREST

2.1. C1.1 NUMBER OF PHYSICAL OBJECTS

The first criteria focus on objects of interest being detected in each frame.

M1.1.1: Number of detected objects compared to reference data according only to their presence.

Per frame, the metric computes:

- Good Detection, GD = MIN(number of objects detected, number of objects in Ref)
- False Detection, FD = number of objects detected - GD.
- Miss Detection MD, = number of objects in Ref - GD.
- OD = GD + FD: all object detections.
- RD = GD + MD: all reference data.
- Precision: number of GD / number of OD.
- Sensitivity: number of GD / number of RD.

The global precision and sensitivity are computed by meaning per frame precision and sensitivity for each frame that contain at least one object.

2.2. C1.2 NUMBER OF PHYSICAL OBJECTS USING THEIR BOUNDING BOX

This second criterion also measures the number of object being observed. This time not only the presence of the object is used but also its bounding box to generate Good detection, False detection…

M1.2.1: Number of detected objects compared to reference data using their bounding box. The matching between an observation and a reference data is done using one of the distances D1 to D4 using the bounding box being threshold.

A good detection corresponds to a reference data overlapping an observation. When there are several overlapping, the best overlap is kept as the good correspondence, and removed for further association.

Per frame, the metric computes:

- GD = number of observed object that match with reference data.
- FD = number of observed object that do not match with reference data.
- MD = number of reference data having no match with any observed object.
- OD = GD + FD: all object detections.
- RD = GD + MD: all reference data.
- Precision: number of GD / number of OD.
- Sensitivity: number of GD / number of RD.

3. T2- LOCALISATION OF PHYSICAL OBJECTS OF INTEREST

This evaluation concerns the evaluation of the 2D position of objects in each frame. For this part, we use bounding boxes or object centres, representing the object 2D localization (reference data).

3.1. C2.1 PHYSICAL OBJECTS AREA

This criterion of evaluation qualifies the area corresponding to physical objects of interest, globally in an image.

M2.1.1: Computation between bounding boxes area of detected blobs and references data regions, for each image. The metric computes the following quantities based and the bounding box:

- Good Localisation, GL = pixels belonging to both the reference data set and the blob set
- False Localisation, FL = pixels belonging to the blob set but not to the reference data set.
- Miss Localisation ML, = pixels belonging to the reference data set but not to the blob set.
- FLR = pixels belonging neither to the reference data set nor the blob set.
- OL = GL + FL: all detections.
- RL = GL + ML: all reference data.
- Precision: number of GL / number of OL.
- Sensitivity: number of GL / number of RL.
- Specificity: number of FLR / number of N.
- F-score: 2*precision*sensitivity / (precision + sensitivity)

Percentages for a video clip are computed as the sum of percentages per image divided by the number of images containing at least one reference data.

3.2. C2.2 PHYSICAL OBJECT AREA FRAGMENTATION (SPLITTING)

This criterion qualifies the fragmentation of object in the observation. One reference data has several corresponding observed data.

M2.2.1(*): number of detected blobs per reference data per image, using their bounding boxes and the thresholding of one of the distances D1-D4. To obtain a percentage between [0;100], we compute the inverse. Percentages for a video clip are computed as the sum of percentages per image divided by the number of images containing at least one reference data.

$$Split = \frac{1}{NB_{image}} \sum_{image} \left(\frac{1}{NB_{\mathrm{Re}\,fData}} \sum_{\mathrm{Re}\,fData} \frac{1}{NBBlob_{/\mathrm{Re}\,fData}} \right)$$

Notes: Having a split less than 100% suppose that reference data have been split into two different blobs.

3.3. C2.3 PHYSICAL OBJECT AREA INTEGRATION (MERGE)

This criterion qualifies the integration of object in the observation. One observed data has several corresponding reference data.

M2.3.1(*): number of reference bounding boxes per detected object, per image, using their bounding boxes and the thresholding of one of the distances D1-D4. Percentages for a video clip are computed as the sum of percentages per image divided by the number of images containing at least one reference data.

$$Merge = \frac{1}{NB_{image}} \sum_{image} \left(\frac{1}{NB_{Blob}} \sum_{Blob} \frac{1}{NBR\grave{e}fData_{/Blob}} \right)$$

Notes: A Merge less than 100% suppose that reference data have been merged into one blob.

3.4. C2.4 PHYSICAL OBJECTS CENTROID LOCALISATION

This criterion qualifies the precision of the localisation of object centroids being measured has centre of bounding box.

M2.4.1: average of the distance 2D between centres of gravity of physical objects and references data, matching considering D1-D4.

M2.4.2: average of the distance 3D between centres of gravity of physical objects and references data, matching considering D1-D4.

Other statistics can be computed like standard deviation, maximum and minimum values…

4. T3- TRACKING OF PHYSICAL OBJECTS OF INTEREST

4.1. C3.1 FRAME-TO-FRAME TRACKING: LINK BETWEEN TWO FRAMES

This part of the evaluation estimates if the link between two physical objects detected at two consecutive time instants is correctly computed or not.

M3.1.1: Number of links between physical objects compared to reference data links. The metric uses following comparison information:

a. Detected object bounding box at time t and t+1 are related to the same reference data using one the distance D1 to D4 and a threshold.

b. A link exists between detected objects at time t+1 and t and a link exist also in reference data.

If there are several links between detected objects related to the same reference data, the one which maximize the overlap with reference data is kept as the good link, and is removed for further association.

Using this we will compute:

- Good tracking, GT = reference data link matching a link between two physical objects.

- False tracking, FT = a link between two physical objects not matching any reference data.

- Miss tracking, MT1 = reference data link not found due to frame-to-frame tracking shortcomings (rejects of case b.).

- MT2 = reference data link not found due to detection or frame-to-frame tracking shortcomings (rejects of case a. and b.).

- OT = GT + FT: all detected links.

- RT1 = GT + MT1: reference data links between correctly classified physical objects.

- RT2 = GT + MT2: all reference data links.

- Precision: number of GT / number of OT.

- Sensitivity1: number of GT / number of RT1.

- Sensitivity2: number of GT / number of RT2.

- F-score1: 2*precision*sensitivity1 / (precision + sensitivity1).

- F-score2: 2*precision*sensitivity2 / (precision + sensitivity2).

4.2. C3.2 NUMBER OF OBJECT BEING TRACKED DURING TIME

This criterion measures the global ability of the system to detect and follow an object during time.

M3.2.1: Number of detected reference data according to their presence intervals. Defining a threshold for a given distance ($\subset$ [D1;D4]) between both presence intervals (of detected object and reference data) permits extracting:

- GT = reference data matching detected object.
- FT = detected object having no sufficient matching with any reference data.
- MT = reference data having no sufficient matching with detected object.
- OT = GT + FT: all object detections.
- RT = GT + MT: all reference data.
- Precision: number of GT / number of OT.
- Sensitivity: number of GT / number of RT.
- F-score: 2*precision*sensitivity / (precision + sensitivity)

This criterion put in evidence:

- Object occurrence for which we do not need spatial information
- Objects persistence along time.

4.3. C3.3 TRACKING TIME EVALUATION

This criterion measures the percent of time an object present in the reference data has been observed and tracked with a consistent ID over tracking period.

M3.3.1: percentage of time during which a reference data is detected and tracked.

The match between a reference data (RD) and a physical object (C) is done with the bounding box (distance D1 to D4 and threshold) and with the constraint that object ID is constant over the time.

With this first match, we may obtain multiple candidates. We choose one candidate with one of these following criterions (to define):

- The candidate is the first physical object that match the reference data,
- The candidate has the greatest intersection time interval with the reference data.

Then we are able to compute the mean time during which a reference data is well tracked:

$$T_{Tracked} = \frac{1}{NB_{\mathrm{Re}fData}} \sum_{\mathrm{Re}fData} \frac{card(RD \cap C)}{card(RD)}$$

4.4. C3.4 PHYSICAL OBJECT ID FRAGMENTATION

Tracking results (M3.1) can be efficient without ID persistence. For the evaluation of the ID persistence, we will add the following metric to measure over the time how many objects are associated to one reference object (ID fragmentation).

M3.4.1: Number of different ID that can take a reference data ($NumObjectID_{/RefData}$). The different detected objects ID are identified with their bounding box and with the fact that their intersection intervals with the reference data are disjointed.

$$Persistence = \frac{1}{NB_{\mathrm{Re}fData}} \sum_{\mathrm{Re}fData} \frac{1}{NumObjectID_{/\mathrm{Re}fData}}$$

Notes: The higher the persistence is (best is 100%), the better the persistence of the ID is.

DG? FD? Cf M1.1 Number / precision

4.5. C3.5 PHYSICAL OBJECT ID CONFUSION CRITERION

We will compute in a similar way the criterion of confusion. An example of confusion may be 2 peoples that meet each other and that permute their ID.

M3.5.1(*): Number of reference data ID that can take a detected object ($NumRefID_{/DetectedObject}$). The different reference data ID are identified with their bounding box and with the fact that their intersection intervals with the physical object are disjointed. To transform the result in a percentage, we will compute the inverse:

$$Confusion = \frac{1}{NB_{DetectedObjectMatch\mathrm{Re}fData}} \sum_{DetectedObjectMatch\mathrm{Re}fData} \frac{1}{NumR\grave{e}fID_{/DetectedObject}}$$

Notes: The more the confusion is close to 100%, the more the tracking algorithm is robust to confusion along the time.

4.6. C3.6 PHYSICAL OBJECT 2D TRAJECTORIES

The first criterion estimates whether 2D trajectories of physical objects are correctly detected over the duration of their presence in the scene or not.

M3.6.1: Number of detected trajectories compared to reference data trajectories. The match between two trajectories is done by computing the maximum distance between them and threshold this result.

If there are several trajectories of tracked objects related to the same reference data, the one which maximizes the sum over time of the spatial overlap with reference data is kept as the good trajectory, and is removed for further association.

- GT = reference data trajectories matching physical object trajectories.
- FT = physical object trajectories not matching any reference data trajectories.
- MT = reference data trajectories not found.
- OT = GT + FT: all detected trajectories.
- RT = GT + MT: reference data trajectories.
- Precision: number of GT / number of OT.
- Sensitivity: number of GT / number of RT.
- F-score: 2*precision*sensitivity / (precision + sensitivity).

4.7. C3.7 3D LOCALISATION EVALUATION

For certain sequence, we could be able to have the 3D objects trajectories in the scene. This ground truth information may be useful to evaluate the 3D reconstruction system capabilities.

M3.7.1 Trajectories may be a line or a series of 3D points. In both case, we will be able to extract:

1. The average of the distances between detected object points and its trajectory,
2. The variance of the distances between detected object points and its trajectory,
3. The minimum and maximum distance between detected object points and its trajectory.

5. T4- CLASSIFICATION OF PHYSICAL OBJECTS OF INTEREST

5.1. C4.1 OBJECT TYPE OVER THE SEQUENCE

M4.1.1 Number of correctly classified physical objects of interest in each frame.

For each frame we compute:

- Good classification, GC = for all the class, Sum of MIN (number of occurrence of the class in the results , number of occurrence of the class in Ref) .
- False classification, FC = for all the class, Sum of (number of occurrence of the class in the results – GD).
- Miss classification, MC = for all the class, Sum of number of Reference object unclassified.

- OC = GC + FC: all detections.

- RC = GC + MC: reference data.

- Precision: number of GC / number of OC.

- Sensitivity: number of GC / number of RC.

Percentages for a video clip are computed as the sum of percentages per image divided by the number of images containing at least one reference data.

M4.1.2: Number of correctly classified physical objects of interest in each frame using also their bounding box.

This metrics uses the same criterion as above but using also bounding boxes areas for object association between observation and reference data. We can refine the following definitions:

- Miss classification, MC1 = Physical objects not classified due to classification shortcomings (e.g., unknown).

- Miss classification, MC2 = Physical objects not classified due to classification shortcomings (e.g., unknown) or due to error in detection.

Then we have:

- RC1 = GC + MC1: correctly detected reference data.

- RC2 = GC + MC2: reference data.

- Precision: number of GC / number of OC.

- Sensitivity1: number of GC / number of RC1.

- Sensitivity2: number of GC / number of RC2.

- F-score1: 2*precision*sensitivity1 / (precision + sensitivity1).

- F-score2: 2*precision*sensitivity2 / (precision + sensitivity2).

Percentages for a video clip are computed as the sum of percentages per image divided by the number of images containing at least one reference data.

M4.1.3: Number of correctly classified physical objects of interest in each frame using also their bounding box and ID persistence.

Same criterion as above but using also bounding boxes areas and ID persistence for object association between observation and reference data.

5.2. C4.2 OBJECT CLASSIFICATION PER TYPE

This criterion computes the Precision and Sensibility per type of object. The output of this criterion is as many metrics as number of object types.

M4.2.1: Number of correctly classified physical objects of interest, in each frame per type, using their bounding boxes.

- Good classification, GC = physical objects of interest correctly classified.
- False classification, FC = physical objects of interest wrongly classified.
- Miss classification, MC1 = physical objects not classified due to classification shortcomings (e.g., unknown).
- MC2 = physical objects not classified due to detection or classification shortcomings (e.g., lack of contrast).
- OC = GC + FC: all detections.
- RC1 = GC + MC1: correctly detected reference data.
- RC2 = GC + MC2: reference data.
- Precision: number of GC / number of OC.
- Sensitivity1: number of GC / number of RC1.
- Sensitivity2: number of GC / number of RC2.
- F-score1: 2*precision*sensitivity1 / (precision + sensitivity1).
- F-score2: 2*precision*sensitivity2 / (precision + sensitivity2).

Percentages for a video clip are computed as the sum of percentages per image divided by the number of images containing at least one reference data.

5.3. C4.3 TIME PERCENTAGE GOOD CLASSIFICATION

In this evaluation part, we suppose that a correspondence between reference data and one or more detected objects have been done, according to their bounding boxes and the ID persistence.

M4.3.1: On a common frame span of a candidate-referenceData pair, we deduce the percentage of frames were the object is well categorized: $card\{ RD \cap C, Type(C) = Type(RD) \} / card(RD \cap C)$. We also deduce the percentage of categorization achievement per object type.

Notes: we may evaluate other similar object attributes (subtype…) as the same way.

6. T5- EVENT RECOGNITION

6.1. C5.1 NUMBER OF EVENTS RECOGNISED OVER THE SEQUENCE

M5.1.1 Number of correctly detected events for the sequence.

Here events in reference data and observation are only compared by their “names. For the global sequence we compute:

- Good recognition, GR = for all the events, Sum of MIN (number of occurrence of the event in the results, number of occurrence of the event in the Ref)
- False recognition, FR= for all the events, Sum of (number of occurrence of the event in the results – GD).
- Miss recognition, MR = for all the events, Sum of number of Reference event un-recognized..
  - OR = GR + FR: all detections.
  - RR = GR + MR: reference data.
  - Precision: number of GR / number of OR.
  - Sensitivity: number of GR / number of RR.

M5.1.2 Number of recognised scenario that are correctly recognised in time, per scenario name. This second metric adds a time constraint to M5.1.1. A detected scenario is consider equals to a reference one if they have the same name, plus a time constraint which is:

- For events:
  i. Their mean times are not very far from each other,
  ii. The detected event interval is included in the reference event interval (depends how we do the ground truth)
- For states: we use thresholding for a given distance (⊂ [D1;D4]) according to the frame spans of the reference and detected scenario (same as in part 1.2).

This metric permits extracting per scenario name:

- Good recognition, GR = reference data scenarios matching recognized scenarios.
- False recognition, FR = recognized scenarios not matching reference data scenarios.
- Miss recognition, MR = reference data scenarios not found.
- ER = GR + FR: all recognized scenarios.
- RR = GR + MR: all reference data scenarios.
- Precision: number of GR / number of ER.
- Sensitivity: number of GR / number of RR.
- F-score: 2*precision*sensitivity / (precision + sensitivity).

We then deduce the precision and sensitivity of the sequences has the average on all scenario name detected almost once (that the system is able to recognize).

6.2. C5.3-SCENARIO PARAMETERS

M5.3.1: Number of recognised scenario that are correctly recognised in time and with the right parameters, per scenario name. For a candidate-reference data pair extracted from the metric 5.2 we compare, in a second time, the objects involved in these scenarios. A scenario is counted as correct if its parameters "match" the reference data ones.

The equivalence between scenario parameters is computed as follow (AND):

- If one of the parameter is a physical object present in the scene and is represented by a bounding box, we compare both bounding boxes with one if the methods D1-D4?
- If the one of the parameter has a specific class, we compare its class with the reference data.
- If one of the parameter is a contextual one (zone equipment…): the name should be the same.

This metric permit extracting per scenario name:

- GR = reference data scenarios matching recognized scenarios.
- FR = recognized scenarios not matching reference data scenarios.
- MR = reference data scenarios not found.
- ER = GR + FR: all recognized scenarios.
- RR = GR + MR: all reference data scenarios.
- Precision: number of GR / number of ER.
- Sensitivity: number of GR / number of RR.
- F-score: 2*precision*sensitivity / (precision + sensitivity).

We then deduce the precision and sensitivity of the sequences has the average on all scenario name detected almost once (that the system is able to recognize).

Notes:

In the case of scenario recognition exploiting 2D context and 2D objects location, we can compute the results and reference data matching.

In the case of scenario recognition exploiting 3D context and 3D objects location, we need the 2D information localisation results for object involved in a scenario for this part of the evaluation. If only 3D information is available, we re-project in 2D the centre of gravity and verify that it is contained in the reference data bounding box.

Others criterions of evaluation could be:

- Average of the initial and ending time differences between detected state and reference data state.
- Average of the mean time difference between detected event and reference data event.