AMI COMMUNICATION

AUGMENTED
MULTI-PARTY
INTERACTION
http://www.amiproject.org/

Evaluation scheme for Tracking in AMI

S. Schreiber [a]
D. Gatica-Perez [b]

AMI WP4 Tracking: Evaluation scheme 1.0

10 January 2006

AUGMENTED
MULTI-PARTY
INTERACTION
http://www.amiproject.org/

[a] Technische Universität München, Germany
[b] IDIAP, Switzerland

1 Introduction

Since a number of tracking algorithms for AMI meeting scenarios is developed at several institutes, there is a certain necessity to agree on a common scheme to evaluate the performance of the different approaches. In the following paragraph a fundamental concept based on [1] for such a scheme is introduced, defining how to evaluate multiple object tracking for unknown configurations.

2 Coverage test

In order to determine the quality of a tracking result for a single object, we introduce two shape-independent measures, indicating if a ground truth object is being tracked and which $\mathcal{E}_i$ is connected to which $\mathcal{GT}_j$:

$$\begin{aligned} &\text{Recall} && \alpha_{i,j} = \frac{|\mathcal{E}_i \cap \mathcal{GT}_j|}{|\mathcal{GT}_j|} \\ &\text{Precision} && \beta_{i,j} = \frac{|\mathcal{E}_i \cap \mathcal{GT}_j|}{|\mathcal{E}_i|} \end{aligned}$$

While the first measurement (recall) represents the ratio of the ground truth area, which is covered by the estimate, the precision embodies the ratio of the estimate area covered by the ground truth. As it can be shown very easily, both α and β must be high to obtain good tracking results. For this reason, a coverage test using the F-measure [2]

$$F_{i,j} = \frac{2\alpha_{i,j}\beta_{i,j}}{\alpha_{i,j} + \beta_{i,j}} \tag{1}$$

has to be passed, returning only a high value if $\alpha_{i,j}$ and $\beta_{i,j}$ are high. This test is considered to be passed, if $F_{i,j}$ exceeds a fixed threshold t_c and thus determines, that $\mathcal{GT}_j$ is being tracked by $\mathcal{E}_i$.

3 Configuration test

To facilitate the explanations in the following sections some definitions will be introduced at first. In this document labeled tracking targets are denoted as ground truth objects $\mathcal{GT}$, tracker outputs are referred to as estimates $\mathcal{E}$. The output of a tracking approach is considered to be correct, if and only if one $\mathcal{GT}$ (resp. $\mathcal{E}$) is tracking exactly one $\mathcal{GT}$ (resp. $\mathcal{E}$). In the following sections there will be defined what kind of errors arise and how they can be detected.

3.1 Configuration error measures

In this context, configuration means the number, the location and the size of all objects in a frame of the scenario. According to the above definition of a correct tracker output, a configuration error occurs if the size or the location of a certain $\mathcal{E}_i$ and its related $\mathcal{GT}_j$ do not match. To identify all types of errors that may occur, 4 configuration measures are introduced:

a) **Measure** FP - False positive. There is an $\mathcal{E}$ indicating an object, where no $\mathcal{GT}$ is.

b) **Measure** FN - False negative. A $\mathcal{GT}$ is not tracked by an $\mathcal{E}$.

c) **Measure** MT - Multiple trackers. More than one $\mathcal{E}$ is associated with only one $\mathcal{GT}$. In order to obtain the subjective impression of a human spectator each excess $\mathcal{E}$ is counted as a MT error.

d) **Measure** MO - Multiple objects. More than one $\mathcal{GT}$ is associated with only one $\mathcal{GT}$. Again a MO error is assigned for each excess $\mathcal{E}$.

For each of these errors above an example is depicted in Fig. 1, where the ground truth is marked with green, the estimates with red resp. blue colored boxes.

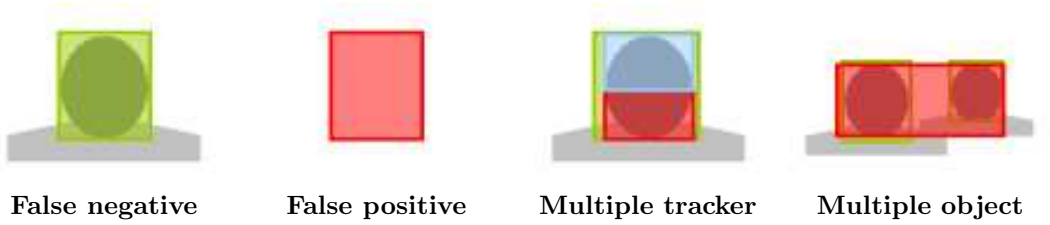


Figure 1: Example for the configuration errors

3.2 Occlusion handling

Situations with occlusion will be treated in a special manner, since MO or MT errors might occur although the estimates are correctly placed. For this reason ground truth labels are enlarged by an additional flag occ_j indicating an occlusion in the image data. This flag is defined for each object and is set to one, if the ratio of the ground truth area from object j, which is covered by the ground truth object k, exceeds a certain threshold t_o.

$$occ_j = \begin{cases} 1, & \exists \mathcal{GT}_k s.t. |\mathcal{GT}_j \cap \mathcal{GT}_k| > t_o \\ 0, & otherwise \end{cases} \tag{2}$$

For all situations with a set occlusion flag there will be no evaluation of any error, i.e. none of the error measurement scores introduced above is increased and thus no ground truth data has to be available for these frames.

3.3 Configuration evaluation procedure

To enable a performance evaluation of different tracking approaches evaluated on diverse data sets, all those measurements presented above have to be normalized by both the number of ground truth objects $N^t_{\mathcal{GT}}$ per frame and the number of frames n as listed in the structure chart below. Since there may occur frames with no $\mathcal{GT}$ labeled at all, normalizing by simply $N^t_{\mathcal{GT}}$ would fail and thus the denominator was chosen to $max(N^t_{\mathcal{GT}}, 1)$ to avoid a division by zero for $N^t_{\mathcal{GT}} = 0$.
For an easy comparison of tracking algorithms a quality measure $\overline{ME}$ is computed from the error measurements. Since the human impression does not consider one of the error types much more severe than other ones, again the F-measure is used to compute the quality measure.

Structure chart for the configuration evaluation procedure

- calculate $F_{i,j}$ for each $\mathcal{E}_i$ combined with each $\mathcal{GT}_j$
- if $F_{i,j} > t_c$
  - if $\mathcal{GT}_j$ not already mapped: map $\mathcal{GT}_j \rightarrow \mathcal{E}_i$
    else increment MO

  else increment FP
- if $F_{i,j} > t_c$
  - if $\mathcal{E}_i$ not already mapped: map $\mathcal{E}_i \rightarrow \mathcal{GT}_j$
    else increment MT

  else increment FN

- report $\overline{FP}$, $\overline{FN}$, $\overline{MT}$ and $\overline{MO}$

$$\overline{FP} = \frac{FP}{n}\sum_{t=0}^{n}\frac{1}{max(N_{\mathcal{GT}}^t,1)} \quad , \quad \overline{FN} = \frac{FN}{n}\sum_{t=0}^{n}\frac{1}{max(N_{\mathcal{GT}}^t,1)}$$
$$\overline{MT} = \frac{MT}{n}\sum_{t=0}^{n}\frac{1}{max(N_{\mathcal{GT}}^t,1)} \quad , \quad \overline{MO} = \frac{MO}{n}\sum_{t=0}^{n}\frac{1}{max(N_{\mathcal{GT}}^t,1)}$$

- compute $\overline{ME} = \frac{4\overline{FN}\,\overline{FP}\,\overline{MT}\,\overline{MO}}{\overline{FN}+\overline{FP}+\overline{MT}+\overline{MO}}$

---

4 Identification test

In the field of tracking, identification means that a particular $\mathcal{E}$ tracks exactly one $\mathcal{GT}$ over its entire lifetime and thus correctly identifies this ground truth object. Among several methods to associate identities that could be considered, each with its assets and drawbacks, an approach based on a "majority rule" was chosen to represent the identification associations. Thus a $\mathcal{GT}_j$ is said to be identified by that $\mathcal{E}_i$ which tracks object j most of the time, and vice versa $\mathcal{E}_i$ identifies that $\mathcal{GT}_j$ where it spent most of the time.

4.1 Identification error measures

Examining tracking scenarios there arise two different types of identification failures. The first type occurs, when one estimate i suddenly stops tracking ground truth object j and another estimate k continues tracking this ground truth object. The second error type results from swapping the ground truth paths, i.e. an estimate i initially tracks $\mathcal{GT}_j$ and after a while changes to track $\mathcal{GT}_k$. To detect all these identification errors, the measures listed below are introduced:

a) **Measure** FIT - Falsely identified tracker. A $\mathcal{E}_i$ which passed the coverage test for $\mathcal{GT}_j$ is different to that identifying this ground truth object before.

b) **Measure** FIO - Falsely identified object. A $\mathcal{GT}_j$ which passed the coverage test for $\mathcal{E}_i$ has not been the identified object in the frame before.

Since these measurements only report changes in associations of $\mathcal{E}$s and $\mathcal{GT}$s, a purity measure is introduced to evaluate the degree of consistency to associations between a $\mathcal{E}$ and a $\mathcal{GT}$.

a) **Measure** OP - Object purity. If $\mathcal{GT}_j$ is the ground truth object which has been identified by $\mathcal{E}_i$ for most of the time, then OP is the ratio of frames that $\mathcal{GT}_j$ is correctly identified by $\mathcal{E}_i$ ($n_{i,j}$) to the overall number of frames (n_j) $\mathcal{GT}_j$ exists.

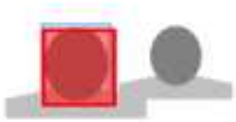
Again the errors mentioned above are visualized in the example (Fig. 2) below, where the each box describes an estimate.

4.2 Identification evaluation procedure

Similar to the configuration evaluation procedure again all measurements have to be normalized by the number of ground truth objects $N_{\mathcal{GT}}^t$ per frame and the number of frames n as listed in the

Situation at time step t-1

Situation at time step t : FIT



Situation at time step t : FIO

Figure 2: Example for the identification errors

structure chart below. For the identification task it is difficult to create only one value indicating the performance of the algorithm, thus all three measures should be reported to get an idea of the quality of the identification capability of an approach.

Structure chart for the identification evaluation procedure

- if ${\mathcal{GT}}_{j,t} \rightarrow \mathcal{E}_{i,t}$
  - if ${\mathcal{GT}}_{j,t-1} \rightarrow \mathcal{E}_{k,t-1}$ increment FIT
  - if ${\mathcal{GT}}_{j,t-1}$ not mapped before increment FIO
- report $\overline{FIT}$, $\overline{FIO}$, $\overline{OP}$

$$\begin{aligned}
\overline{FIT} &= \frac{FIT}{n} \sum_{t=0}^{n} \frac{1}{max(N^t_{\mathcal{GT}}, 1)}, \\
\overline{FIO} &= \frac{FIO}{n} \sum_{t=0}^{n} \frac{1}{max(N^t_{\mathcal{GT}}, 1)}, \\
\overline{OP} &= \frac{1}{N_{\mathcal{GT}}} \sum_{j=0}^{N_{\mathcal{GT}}} \frac{n_{i,j}}{n_j}
\end{aligned}$$

5 Training and Evaluation Video Set

To get comparable evaluation results for the tracking algorithms developed by the different partners in AMI we will define a common video set for the evaluation. This video set should contain as much of the challenges which have led to the acquisition of the special side-corpus AV16.7-ami, thus the following sets have been defined for the evaluation, which may only be used for the evaluation task itself and not e.g. for tuning parameters:

Eval I : Sequences from the side corpus AV16.7ami (2, 3, 9, 12, 14)
Eval II : Sequence from the AMI core corpus (1008b)
Eval III : Sequences from the side corpus AV16.7ami (1, 8, 13, 16)

Since each of the specified sequences consists of three avi-files (left, right and central camera view) on which our algorithms will be evaluated, this material offers a total amount of approximately 1.5 h of video data for the evaluation of tracking modules. For the deliverable only results for Eval I and Eval II have to be reported. Below you can find the weblinks to get the video sequences:

AMI core corpus : http://mmm.idiap.ch/private/AMIzone/idiapHub.html
AV16.7ami : ftp://mmm.idiap.ch/private/ami/906401383/

All measurement errors introduced above will be reported according to this video evaluation set. The video material is fully annotated using different annotation rates depending on the level of dynamics of the person in the sequence. The advantage of this proceeding is a reduction of the effort in annotating parts (especially easy parts like seated people) while giving more "annotation resolution" on parts that are more interesting for tracking (e.g. somebody leaving). For this reason videos will be annotated based on three different levels of accuracy:

- Slow (1 frame/ 5 seconds) - people seated or standing for several minutes
- Middle (1 frame/ 1 second) - people standing for one minute or so max
- Fast (2 frames/ 1 second) - people entering/seating/standing up/moving to white board

The annotation data can be found at ftp://mmm.idiap.ch/private/ami/906401383/. To derive the annotation resolution please refer to the frame number explicitely given in the files. All other video material from the AMI corpus (both main and side corpus) - except the evaluation test set mentioned above - is free to be used for training the detectors and modules of the invented tracking algorithms.

6 Data storage format

In order to facilitate a joint evaluation in the scope of AMI tracking technologies, a common evaluation tool has been developed and spread among all partners (also downloadable at http://www.idiap.ch/ smith/AMItrack.html). For simplifying the usage of this tool each tracking algorithm has to provide the output in the same way, i.e. a head bounding box is generated enclosing each tracked object. This result has to be stored for the evaluation tool in a simple ASCII-file according to the following file format:

frame [frame number]
    object [identifier] <tab> [head bounding box]
    object [identifier] <tab> [head bounding box]

In this file format description all expressions in brackets have to be replaced by the real numbers. For each frame, first provide the frame number (the results and ground truths must cover the same set of frame numbers), followed by the object parameters. Object parameters include a unique identifier and the location of the object in the image. The identifiers need not (and should not necessarily) match between the ground truth and tracking results, but they should be consistent within each. For each frame, provide the object parameters of every object present (in the results or the ground truth). If there are no ground truths or estimates present, just provide the frame number. Objects must be represented by bounding boxes (in both tracking and ground truth). The bounding boxes are defined by four numbers, (x,y,w/2,h/2). The point (x,y) indicates the location of the center of the bounding box, w/2 is the distance from the center to one of the vertical edges (or half-width), and h/2 is the distance from the center to one of the horizontal edges (or half-height). All coordinates have to be referenced to the top left image origin.

References

[1] K. Smith, S. Ba, J. Odobez, and D. Gatica-Perez, "Evaluating multi-object tracking," San Diego, CA, USA, June 2005, vol. Workshop on Empirical Evaluation Methods in Computer Vision (EEMCV).

[2] C. J. Van Rijsbergen, *Information Retrieval*, Butterworth-Heinemann, Newton, MA, USA, 1979.