CLEAR 2007 Evaluation Plan

3D Person Tracking Task

Editor:
Keni Bernardin (keni@ira.uka.de)

Version 1.0

0 Introduction

The goal of the 3D person tracking task is to track multiple people in 3D using the recordings from a wide variety of sensors. The current scenario is that of a small interactive seminar with 4-7 participants mostly located around a meeting table. For tracking, the synchronized data from 5 video cameras with overlapping field of view and from at least 80 microphones, grouped into various arrays is available. Thus, the task is not determine person positions in one image, or with respect to a microphone pair, but rather to find the true person positions in the scene.

Depending on the input streams to be used, the task is broken down into 3 subtasks, visual, acoustic, and multimodal person tracking, with slightly varying constraints, goals and metrics.

The challenge in the 3D person tracking task comes from the natural, unscripted and highly variable nature of the recordings, made in several sites under different conditions.

1 Label file description

1.1 3D label files

For the evaluation of the 3D person tracking task, the reference positions of persons in the room must be known. To make the definition of a person position unambiguous, and to ease the annotation effort, the centroid of a person's head is taken as reference point.
High labeling accuracy is reached by annotating the seminar participants' head centroids in all room camera views. These 2D positions are then used together with camera calibration information to generate the actual 3D positions. The 3D positions are expressed in mm, relative to the room coordinate frame.

Along with the room positions, a unique person ID is assigned to every person, which remains constant for the duration of the sequence. This is to evaluate tracking consistency over time.

The format of the 3D label file is as follows:

For each second of recording, a line is generated containing the actual timestamp followed by a list of IDs and positions of the persons currently visible. A person is considered visible if its head centroid can be seen in at least 2 corner camera views. No fields are written for persons not visible in the actual timeframe.

The structure of a label line is

<TimeStamp> {<ID_i> <$HeadCentroid_i$ x y z>}

And an example line could be:

<Timestamp> <ID1> <x> <y> <z> <ID4> <x> <y> <z> <ID5> <x> <y> <z> …

Note that even if for one of the frames to be labeled there is no person visible in at least 2 camera views, a line is generated in the label file anyway, containing nothing but the timestamp. This is to ease up file consistency checks, scoring, etc…

1.2 Additional label files for the acoustic and multimodal subtasks

For ease of scoring of the acoustic and multimodal person tracking subtasks, slightly modified versions of the above described 3D labels were created, using speaker activity information gained from manual speech transcriptions. The format of these files will be described in the appropriate sections for the concerned subtasks.

1.3 Notes on general labeling rate:

For most tasks using the CLEAR database, manual annotations were produced only every second, i.e. every 15th, 25th or 30th video frame, depending on the frame rate. This is to reduce the effort needed to label longer sequences of video with multiple persons. In most tasks, tracking systems, detection systems, etc may still output hypotheses at a much higher rate (e.g. 15 fps for a visual person tracker, or every 100ms for an acoustic tracker), but the performance of the systems will be evaluated only once per second on the key time frames that have been labeled.

2 Task and Metric Description

2.1 Visual Person Tracking

Leading partner	**UKA-ISL**
Contact	**Keni Bernardin (keni@ira.uka.de)**

2.1.1 Description of the Task

The specification of this task has not changed, compared to the first evaluation run, with the exception that single person tracking on lecture-like data is no longer an objective.

For the current evaluation, only Interactive Seminars are considered and the goal is to continuously track all persons visible in the recording sequence. This requires to simultaneously track multiple persons in 3D in each frame. Although no person identification is aimed at, track persistency will be measured, which means tracking systems should be capable of keeping a consistent ID for each person for the whole length of a sequence, even through lengthy visual gaps.

Although the objective is to track people as a whole, for ease of annotation and evaluation person positions are defined as the 2D projections of their head centers to the ground. The projections to the ground plane will also be extracted from the tracker output and will be used in computing accuracy measures.

All views from the four corner cameras and from the ceiling camera may be used. Intrinsic and extrinsic camera calibration information and a set of images depicting the empty recording room (for easier initial background subtraction) are provided for each seminar recording.

For the CLEAR 2007 evaluation, two data sets will be distributed: The development set, which can be used to train or tune tracking systems, and the evaluation set, which will be used to measure system performance. While all information which can be gathered from the development set is legitimate to tune systems, no information gained manually from the evaluation set (such as manually segmented table or chair positions, person colors, manual person count, selected cameras for a specific seminar, etc) is admissible. Here, only information derived automatically (and therefore part of the tracking algorithm) may be used to adapt systems and enhance performance.

Person Tracking systems will be evaluated in terms of localization precision, as well as their ability to estimate the number of persons, and to keep consistent track of them.

For the current evaluation, the speed of tracking systems, measured as realtime factor, and the nature of the algortihms (online-capable, fully or semi-automatic, requiring pre- or post-processing, with automatic or manual chaining of processing steps, one- two- or multiple-pass, etc.) are also evaluation criteria and should be reported.

2.1.2 Database

The available data for the visual person tracking task includes, for each seminar:

- The video streams from the four fixed corner cameras and from the top camera.
- A set of images showing the empty recording room (for eventual background subtraction)
- The intrinsic and extrinsic calibration information for each camera.
- A description of the room layout showing e.g. the room dimensions and the position and orientation of the room coordinate frame.

2.1.3 Format of label files for scoring

See section 1.1 for the 3D label file format.

2.1.4 Output file format

To ease up the evaluation process, the same format as for the reference 3D label files should also be used for the output hypothesis files of a person tracker. One output file should be created per seminar. It should ideally contain one line for each processed frame (usually 15/25/30 per second), with the following format:

<TimeStamp> {<ID_i> <$HeadCentroid_i$ x y z>}

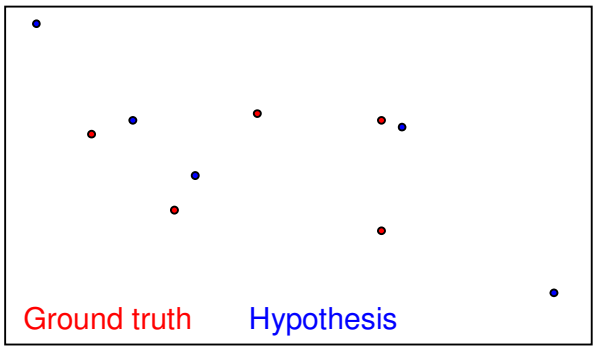
The scoring tool will then evaluate the tracker performance only on time frames for which manual labels exist by comparing each reference label with the closest tracker hypothesis in time.

2.1.5 Metrics

Evaluating the performance of systems for the tracking of multiple people requires a set of metrics that intuitively reflect a tracker's performance in determining the number of people present, their exact position, and in keeping consistent track over time.

An evaluation script fed with the output of a multi-person tracker and with the reference person positions must do the following things:

1) Find a mapping between the persons indicated by the ground truth and the hypotheses of the tracker (The correspondence problem).
2) For each individual mapping, determine the precision with which the person's position was estimated.
3) Count all ground truth persons for which no tracker hypothesis was output as misses.

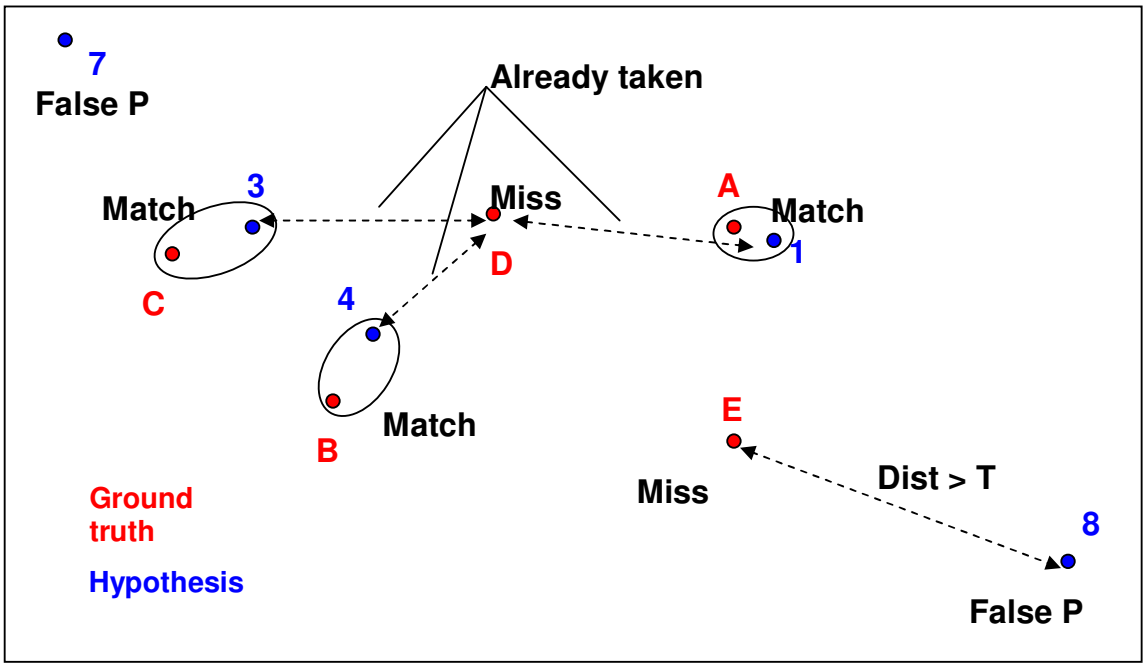


4) Count all hypotheses for which no ground truth exists as false positives
5) Make sure that the persons were tracked correctly over time. This includes:
   - Checking that persons were not substituted for each other, for example when they walk close to each other.
   - Checking that a track is correctly recovered after it was lost, for example when a person was occluded, re-entered the room, etc

The performance of the tracker can then be intuitively expressed in two numbers: The "tracking precision" which expresses how good the tracker is at estimating the exact positions of persons, and the "tracking accuracy" which shows how well the system keeps track of persons, how many mistakes were made in terms of misses, false positives, mismatches, failures to recover tracks, etc. A brief introduction to these metrics will be given in the following. For a more detailed explanation, the reader is referred to [1].

2.1.5.1 Finding a mapping between ground truth and tracker hypothesis

As explained before, the first step in evaluating the performance of a multi-person tracker is finding a mapping between the sequence of person positions $h_1 \ldots h_m$ output by the tracker and the positions $o_1 \ldots o_n$ marked in the ground truth labels. Keeping in mind that only the (2D) projections on the floor of those 3D positions are considered for distance calculations, the mapping is done in the following way:



Let $M_0 = \{\ \}$ be the initial mapping between tracker hypotheses and ground truth points. For every time frame *t*,

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

2) For all ground truth persons for which no correspondence was made yet, try to find a matching hypothesis. Allow only one to one matches. To find optimal correspondences that minimize the overall distance error, Munkre's algorithm is used. Only pairs for which the distance does not exceed the threshold *T* are valid. 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 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 ground truth person o_i and its corresponding hypothesis.

4) All remaining hypotheses are considered false positives. Similarly, all remaining ground truth persons 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 persons 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.

The one free parameter, the threshold T, must be carefully chosen, as it defines the error beyond which a tracking hypothesis should reasonably be considered a miss. If T is chosen too small, all error measures become meaningless. If it is chosen too large, the number of mismatches and misses can not be correctly estimated. Based on the experience from the previous evaluation, it has been agreed to use a threshold of $T = 500$ mm.

In the above described way, a continuous mapping between the ground truth and the tracker hypotheses is defined and all occurring tracking errors are accounted for. Based on this, two very intuitive metrics can be defined.

2.1.5.2 Definition of metrics

Metrics	
Number of metrics	2
List metrics	• Multiple Object Tracking Precision (MOTP) • Multiple Object Tracking Accuracy (MOTA)

Metrics Name	**Multiple Object Tracking Precision (MOTP) [mm]**
Description	This is the precision of the tracker when it comes to determining the exact position of a tracked person in the room. It is calculated as follows: $$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}$$ It is the total Euclidian distance error for matched ground truth – hypothesis pairs over all frames, averaged by the total number of matches made. It shows the ability of the tracker to find correct positions, and is independent of its errors in keeping tracks over time, estimating the numbers of persons, etc.

Metrics Name	**Multiple Object Tracking Accuracy (MOTA) [%]**
Description	This is the accuracy of the tracker when it comes to keeping correct correspondences over time, estimating the number of persons, recovering tracks, etc. $$MOTA = 1 - \frac{\sum_t m_t + fp_t + mme_t}{\sum_t g_t}$$ It is the sum of all errors made by the tracker, false positives, misses, mismatches, over all frames, averaged by the total number of ground truth positions. It is similar to accuracy metrics widely used in other domains and gives a very intuitive measure of the tracker's performance independent of its ability to determine exact person locations. In the definition of the *MOTA* given here, the three different types of errors, misses, false positives and mismatches, are equally weighted. But other weighting or cost functions are

also thinkable. Therefore, to ensure clarity of the results, the individual error ratios $\frac{\sum_t m_t}{\sum_t g_t}$, $\frac{\sum_t fp_t}{\sum_t g_t}$ and $\frac{\sum_t mme_t}{\sum_t g_t}$ are also to be reported.

Note that for the acoustic person tracking task (where identity mismatches are not evaluated), a slight variant of the *MOTA* metric, in which all mismatch errors are ignored, will be used to measure tracker performance only for the active speaker at each point in time. This will be done using manual speaker segmentation labels and only on segments where at most one person is speaking. This separate measure, the $A-MOTA = 1 - \frac{\sum_t m_t + fp_t}{\sum_t g_t}$, will be reported for the Acoustic Person Tracking Task. The original *MOTA* will be computed for the Visual and the Multimodal Tasks.

2.2 Acoustic Person Tracking

Leading partner	**ITC-IRST**
Contact	**Maurizio Omologo (omologo@itc.it)**

2.2.1 Introduction

The goal of the acoustic person tracking task is to localize and track one or more speakers within a room. The target scenario is that of an Interactive Seminar, with a small number of participants, discussing a topic of common interest. As the participants take turns speaking, tracking algorithms are expected to determine their positions, but are not required to identify the speakers or to detect speaker switches. As a result, for this evaluation, track persistency will not be evaluated, i.e. identity mismatches will not be penalized.

As opposed to the previous evaluation, the current task definition places greater emphasis on the speech activity detection (SAD). While segments containing speech from multiple speakers will still be ignored in scoring, this time around silence segments will not be excluded and tracking algorithms must detect when speech is present, in addition to localizing the sound source, in order to avoid being penalized for missed speech frames or for false positive detections.

The output of the tracker should be, for every time frame, at most the 3D position of one person (the current speaker), with a generic track ID.

Since the localization error along the scene's *z*-coordinate is less critical for person tracking and more difficult to derive in an accurate way, the tracking system performance will be evaluated by considering 2 dimensions (*x,y*). Person positions are defined as the 2D projections of their head centers to the ground and only the (*x,y*)-components of the tracker output will be used in computing accuracy measures.

The speaker localization algorithms can be applied to all the available far-field microphones. Exact geometry and positioning information for all microphone arrays usable for source localization is available. The far-field microphones include at least the 3 T-shaped microphone arrays (with 4 mics each) of the minimal sensor setup, the 64 channel NIST

MarkIII microphone array, and the table top microphones (although no position information is given for these).

2.2.2 Temporal axis for evaluation

In the CLEAR database, microphone signals are sampled at 44.1 kHz sampling frequency.

In principle, a speaker localization system may produce a set of coordinates at a very high rate, such as the typical rate (100 Hz, i.e. every 10 ms) of an Automatic Speech Recognition (ASR) front-end or more, but in the given scenarios the adoption of a reduced rate in the range of, for instance, 1-10 Hz (which means that a set of coordinates will be produced every 100-1000 ms) is more adequate. This choice is more consistent with the rate of ground truth position labeling and more appropriate for scoring purposes. In the following of this document, a rate of 10 Hz is assumed.

Given 10 Hz rate, since the speaker tracking system provides coordinates with a faster rate than the manual labels, the evaluation tool will use the track estimates closest in time to the reference positions for scoring. If the system output is subject to heavy fluctuations over a 1s window (for ex. with individual occurrences of silence between valid position hypotheses, due to SAD errors), it is therefore advisable to average or interpolate the system output (in whichever way one may choose) to avoid penalties (misses, etc). If the speaker localization system produces data with a slower rate, or is not able to produce a set of coordinates for some labeled frames, the evaluation tool will classify those missing data as misses.

2.2.3 Speech Activity Detection constraints

Although it may be considered as a minor comment, it is worth noting that, introducing a SAD pre-processor, one can also expect to have some false alarms due to it. However, we expect that some automatic localization systems (but not all of them) will be conceived to derive an estimate of the speaker coordinates only when the SAD preprocessing module has detected a certain speech activity. Other localization systems may work in a different way, for instance producing a set of coordinates according to a given confidence measure of the localization reliability (and to a related thresholding). The latter situation would lead to other possible false alarms of the speaker localization system, not corresponding to frames detected as speech by the SAD module. Here, the combined SAD and speaker localization systems are evaluated in terms of a single false positive rate, no matter which approach was taken.

2.2.4 Database

The available data for the acoustic person tracking task includes, for each seminar:

- The audio recordings from all far-field microphones
- The position and array geometry for each microphone
- A description of the room layout showing e.g. the room dimensions and the position and orientation of the room coordinate frame.

2.2.5 Format of label files for scoring

The label file format for the acoustic person tracking task is essentially the same as described in section 1.

The difference lies in the fact that only one person position is labeled per time frame, namely that of the active speaker, and that time intervals containing more than one active speaker are not labeled at all (not even a timestamp). Therefore, the label files may contain longer gaps in time, compared to the visual 3D label files, which will be ignored by the scoring tool.

2.2.6 Output file format

- The format is essentially the same as in section 2.1.4, with the difference that only one track hypothesis should be output per frame (that of the active speaker). So the format of one output line is:

  <TimeStamp> <ID_i> <$HeadCentroid_i$ x y z>

  or just

  <TimeStamp>

  if no active speaker is being tracked.

2.2.7 Metrics

Metrics	
Number of metrics	2
List metrics	• Multiple Object Tracking Precision (MOTP) • (Audio-) Multiple Object Tracking Accuracy (A-MOTA) See Section 2.1.5 for details about the MOTP and A-MOTA metrics

For the CLEAR evaluation, acoustic source localization accuracy will be evaluated with the same metrics used for visual person tracking as described in Section 2.1.5 with the exception that identity mismatches will not be counted. For this, the original MOTA metric was modified to the **Audio-MOTA** version (**A-MOTA**) that excludes mismatch errors (*mme*) from the MOTA formula. We do not expect audio trackers to distinguish between speakers yet (this may be done in the future, though). The **MOTP** metric is identical to the one in visual person tracking. Evaluating vision- and audio-based technologies with the same set of metrics in this way will enable their performance to be better compared.

2.3 Multimodal Person Tracking

Leading partner	**UKA-ISL**
Contact	**Keni Bernardin** keni@ira.uka.de

2.3.1 Description of the Task

The goal of the multimodal tracking task is to continuously track the last known active speaker in an Interactive Seminar setting. Algorithms should be able to detect speech

automatically, to determine which of the attendees is speaking and recognize speaker switches, and to continuously output the position of the last known speaker even through periods of silence. As for the acoustic person tracking task, only segments in which no more than one speaker is active will be considered in scoring. These valid segments, containing non-overlapping speech or silence, are determined on the basis of manual transcriptions.

The output of the tracker should be, for every time frame, at most the 3D position of one person (the last active speaker), along with the corresponding track ID. The track ID itself should remain constant for each speaker throughout the sequence.

Although the objective is to track people as a whole, for ease of annotation and evaluation person positions are defined as the 2D projections of their head centers to the ground. The projections to the ground plane will also be extracted from the tracker output and will be used in computing accuracy measures.

All views from the four corner cameras and from the ceiling camera, and all far-field microphones may be used for tracking. Exact microphone positions and geometry, intrinsic and extrinsic camera calibration information and a set of images depicting the empty recording room (for easier initial background subtraction) are provided for each seminar recording.

For the CLEAR 2007 evaluation, two data sets will be distributed: The development set, which can be used to train or tune tracking systems, and the evaluation set, which will be used to measure system performance. While all information which can be gathered from the development set is legitimate to tune systems, no information gained manually from the evaluation set (such as manually segmented table or chair positions, person colors, manual person count, selected cameras for a specific seminar, etc) is admissible. Here, only information derived automatically (and therefore part of the tracking algorithm) may be used to adapt systems and enhance performance.

Multimodal person tracking systems will be evaluated in terms of localization precision, as well as their ability to determine and keep consistent track of the correct speaker.

For the current evaluation, the speed of tracking systems, measured as realtime factor, and the nature of the algortihms (online-capable, fully or semi-automatic, requiring pre- or post-processing, with automatic or manual chaining of processing steps, one- two- or multiple-pass, etc.) are also evaluation criteria and should be reported.

2.3.2 Database

The available data for the multimodal person tracking task includes, for each seminar:

- The video streams from the four fixed corner cameras and from the top camera.
- A set of images showing the empty seminar room (for background subtraction).
- The intrinsic and extrinsic calibration information for each camera.
- The audio recordings from all far-field microphones
- The position and array geometry for each microphone
- A description of the room layout showing e.g. the room dimensions and the position and orientation of the room coordinate frame.

2.3.3 Format of label files for scoring

The label file format for the multimodal person tracking task is essentially the same as described in section 1.

The difference lies in the fact that only one person position is labeled per time frame, namely that of the last known active speaker, and that time intervals containing more than one active speaker are not labeled at all (not even a timestamp). Therefore, the label files may contain longer gaps in time, compared to the visual 3D label files, which will be ignored by the scoring tool.

2.3.4 Output file format

- The format is essentially the same as in section 2.1.4, with the difference that only one track hypothesis should be output per frame (that of the last active speaker). So the format of one output line is:

  <TimeStamp> <ID_i> <$HeadCentroid_i$ x y z>

  or just

  <TimeStamp>

  if no active speaker is being tracked.

2.3.5 Metrics

Metrics	
Number of metrics	2
List metrics	• Multiple Object Tracking Precision (MOTP) • Multiple Object Tracking Accuracy (MOTA) See Section 2.1.5 for details about the MOTP and MOTA metrics

3 Submission protocol

3.1 Result submission

3.1.1 File naming convention

A system is identified by the following name:

EXP-ID: = <SITE>_<TASK>_<DATA>_<SYSTEM>

Where:

<SITE> = <Abbreviation for the institution making the submission>

<TASK> = VPT | APT | MPT

<DATA> = EVAL07

<SYSTEM> = PRIMARY | CONTRAST-XXX | CONTRAST-YYY

The list of tasks is :

- ☐ VPT = Visual Person Tracking
- ☐ APT = Acoustic Person Tracking
- ☐ MPT = Multimodal Person Tracking

3.1.2 Structure of the archive

One result file has to be submitted per seminar.

For each system, the submitted archive should contain the following structure

<EXP-ID>/<EXP-ID>.TXT

<EXP-ID>/<RESULT_FILE_1>.PT

<EXP-ID>/<RESULT_FILE_2>.PT

.

.

.

<EXP-ID>/<RESULT_FILE_N>.PT

Each result file <RESUL_FILE_N> is a UNIX text file with the name of the task (PT) as extension.

The archive should be a tgz, tar or a zip file.

3.1.2.1 Format and content of the result file for each seminar segment

For each segment of a seminar a result file should be produced. The format of the file is described in section 1.

The name of the result file is:

<SEMINAR>.PT

where:

<SEMINAR> = AIT_20051010 | AIT_20051011_B | … | UPC_20050727

Example:

A primary submission from XXX on the evaluation set for Visual Person Tracking should be an archive

XXX_VPT_EVAL07_PRIMARY.tar.gz

The archive should contain the following files :

XXX_VPT_EVAL07_PRIMARY/XXX_VPT_EVAL07_PRIMARY.TXT

XXX_VPT_EVAL07_PRIMARY/AIT_20051010.PT

XXX_VPT_EVAL07_PRIMARY/AIT_20051011_B.PT

.

.

XXX_VPT_EVAL07_PRIMARY/IBM_20050827B_A.PT

.

.

XXX_VPT_EVAL07_PRIMARY/UPC_20050725B_B.PT

3.1.3 System description

For each system, a one page system description <EXP-ID>.TXT must be provided describing the data used, the approaches (algorithms), the configuration, and, more importantly:

- The processing time expressed as realtime factor (number of seconds needed for processing / number of seconds in the sequence). For the processing time, only the time consumed by the actual algorithm should be counted, specifically excluding the time for loading and uncompressing jpeg images or audio files, for eventual display, or the time consumed by manual operation in between a multi-pass or multi-stage algorithm. If an algorithm runs on several machines in parallel, the true parallel processing times may be counted together. Processing times needed for sequential steps must be counted separately, even if they are performed on different machines.
- Along with the processing speed, the number of machines (if distributed), of processors, the CPU types and speed should be stated. As an example statement:

  " The algorithm runs at realtime factor 1.5 on 3 Pentium 3GHz Dual Core machines."

  or

  "Realtime factor: 1.5, (1x Pentium 2.4GHz + 2x Pentium 3GHz Dual core) "

  The statement should not be significantly longer. No elaborate descriptions of hardware, operating systems or libraries are required.
- The nature of the algorithm, namely if it is a one-pass or multiple-pass algorithm; if it is fully automatic, or needs human intervention to chain pre- or post-processing steps; if it works in a batch processing manner, or if it is online-capable, etc.

These properties will be made official and will serve to judge the quality of a system, just as much as the precision and accuracy metrics.

3.1.4 Submission procedure

The system results must be uploaded to the following FTP server:

ftp://ftp-clear.ilkd.uni-karlsruhe.de/

The deadline for submissions is:

- □ March 28th 23h59 CET for Visual Person Tracking, Acoustic Person Tracking and Multimodal Person Tracking

4 References

[1] Keni Bernardin, A. Elbs, Rainer Stiefelhagen, **"Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment".** The Sixth IEEE International Workshop on Visual Surveillance, VS 2006, Graz, Austria, 2006-05-13