The CLEAR 2007 Evaluation

Rainer Stiefelhagen[1], Keni Bernardin[1], Rachel Bowers[2], R. Travis Rose[3], Martial Michel[3,4], and John Garofolo[3]

[1] Interactive Systems Lab, Universität Karlsruhe, 76131 Karlsruhe, Germany
{stiefel, keni}@ira.uka.de
[2] Naval Research Laboratory, 4555 Overlook Ave. S.W., Washington, DC 20375
[3] NIST, 100 Bureau Dr - MS 8940, Gaithersburg, MD 20899, USA
[4] Systems Plus, Inc., 1370 Piccard Drive - Suite 270, Rockville, MD 20850, USA

Abstract. This paper is a summary of the 2007 CLEAR Evaluation on the Classification of Events, Activities, and Relationships which took place in early 2007 and culminated with a two-day workshop held in May 2007. CLEAR is an international effort to evaluate systems for the perception of people, their activities, and interactions. In its second year, CLEAR has developed a following from the computer vision and speech communities, spawning a more multimodal perspective of research evaluation. This paper describes the evaluation tasks, including metrics and databases used, and discusses the results achieved. The CLEAR 2007 tasks comprise person, face, and vehicle tracking, head pose estimation, as well as acoustic scene analysis. These include subtasks performed in the visual, acoustic and audio-visual domains for meeting room and surveillance data.

1 Introduction

CLassification of Events, Activities and Relationships (CLEAR) is an international effort to evaluate systems that are designed for perceiving people's identities, activities, interactions and relationships in human-human interaction scenarios, and related scenarios. The first CLEAR evaluation workshop was held in spring 2006 (see [23] for a complete description of CLEAR'06). It hosted a variety of tasks, evaluated on challenging, realistic scenarios, and brought together a number of research institutions from around the world. Prompted by the success of the first evaluation, another round was conducted from January through April 2007, culminating with a 2-day workshop in Baltimore, MD, where system details and results were presented and discussed. The CLEAR 2007 workshop was colocated with the 2007 Rich Transcription (RT) workshop to provide an opportunity for members of both the vision and speech research communities to participate in discussions related to multimedia based evaluations.

1.1 Motivation

Many researchers, research labs and in particular a number of major research projects worldwide – including the European projects CHIL, Computers in the

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2007	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The CLEAR 2007 Evaluation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **National Institute of Standards and Technology (NIST) Gaithersburg, MD 20899 8940**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**34**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Human Interaction Loop [1], and AMI, "Augmented Multi-party Interaction" [2], as well as the US programs VACE, "Video Analysis and Content Extraction" [3], and CALO, "Cognitive Assistant that Learns and Organizes" [4] – are working on technologies to analyze people, their activities, and their interaction. However, common benchmarks for such technologies are usually not available. Most researchers and research projects use their own data sets, annotations, task definitions, metrics and evaluation procedures. As a consequence, comparing the advantages of research algorithms and systems is virtually impossible. Furthermore, this leads to a costly multiplication of data production and evaluation efforts for the research community as a whole.

CLEAR was created to address this problem. Its goal is to provide a common international evaluation framework for such technologies, and to serve as a forum for the discussion and definition of related common benchmarks, including the definition of tasks, annotations, metrics and evaluation procedures. The expected outcomes for the research community from such a common evaluation forum are:

- the definition of widely adopted metrics and tasks
- greater availability of resources achieved by sharing the data collection and annotation burdens
- the provision of challenging multimodal data sets for the development of robust perceptual technologies
- comparability of systems and approaches
- faster progress in developing improved and robust technologies

1.2 Background

CLEAR is a collaborative effort between the US Government funded Video Analysis and Content Extraction (VACE), and the European Commission funded, Computers in the Human Interactive Loop (CHIL) programs, but 2007 has expanded this collaboration to include an evaluation task from the Augmented Multiparty Interaction (AMI) program. As in 2006, this new round of evaluations targeted technologies for tracking, identification, and analysis of human-centered activities, on challenging multimodal databases from various meeting and surveillance domains. As before, the evaluations were open and interested sites not part of the initiating projects were invited to participate.

1.3 Scope and Evaluation Tasks in 2007

The CLEAR 2007 evaluation was organized in conjunction with the Rich Transcription (RT) 2007 evaluation [5], their deadlines were harmonized and this year the workshops were colocated. While the evaluations conducted in RT focus on content-related technologies, such as speech and text recognition, CLEAR is more about context-related multimodal technologies such as person tracking, person identification, head pose estimation, analyzing focus of attention, interaction, activities and events.

The evaluation tasks in CLEAR 2007 can be broken down into four categories:

- tracking (faces/persons/vehicles, 2D/3D, acoustic/visual/audio-visual)
- person identification (acoustic, visual, audio-visual)
- head pose estimation (single view data, multi-view data)
- acoustic scene analysis

These tasks and their various subtasks are described in Section 4. As in the 2006 evaluations, part of the tasks were organized by CHIL and others by VACE, depending on the partner that originally defined them, and on the datasets used. The tasks were run independently in parallel, although care was taken to harmonize task definitions, annotations and metrics wherever possible. In contrast to 2006, the face detection and tracking task was run using the same annotations and metrics for both the CHIL and VACE related subtasks. In addition, the multiple object tracking metrics (see section 3), which were first agreed on in 2006, were further harmonized, and used without exception in all 2007 tracking tracking tasks and subtasks.

1.4 Contributors

As in the previous year, many people and institutions worldwide contributed to the success of CLEAR 2007. Again, the organizers were the Interactive Systems Labs of the Universität Karlsruhe, Germany (UKA) and the US National Institute of Standards and Technology (NIST). The participants and contributors included: the Research and Education Society in Information Technologies at Athens Information Technology, Athens, Greece, (AIT), the Interactive Systems Labs at Carnegie Mellon University, Pittsburgh, PA, USA, (CMU) the Evaluations and Language resources Distribution Agency, Paris, France (ELDA), the IBM T.J. Watson Research Center, RTE 134, Yorktown Heights, USA (IBM), the Centro per la ricerca scientifica e tecnologica at the Fundacione Bruno Kessler, Trento, Italy (FBK-IRST), the Universitat Politécnica de Catalunya, Barcelona, Spain (UPC), the Laboratoire d'Informatique pour la mécanique et les sciences de l'ingénieur at the Centre national de la recherche scientifique, Paris, France (LIMSI), Pittsburgh Pattern Recognition, Inc., Pittsburgh, PA, USA (PittPatt), the department of Electronic Engineering of the Queen Mary University of London, UK (QMUL), the Computer Science and Technology Department of Tsinghua University, Beijing, China (Tsinghua), the Department of Computer Science of the University of Maryland, MD, USA (UMD), the University of Central Florida, USA (UCF), the Institute of Signal Processing of the Technical University of Tampere, Finland (TUT), the Breckman Institute for Advanced Science and Tech. at the University of Illinois Urbana Champaign, USA (UIUC), the IDIAP Research Institute, Martigny, Switzerland (IDIAP), the MIT Lincoln Laboratory, Lexington, MA, USA (MIT), the Institute for Robotics and Intelligent Systems of the University of Southern California, USA (USC), the Institute for Infocomm Research, Singapore (IIR).

UKA, FBK-IRST, AIT, IBM and UPC provided several recordings of "interactive" seminars, which were used for the 3D person tracking tasks, for face

detection, for the person identification tasks and for acoustic event detection. UKA and IDIAP provided several annotated recordings for the head pose estimation task. UPC and FBK-IRST provided different databases with annotated acoustic events used for acoustic event recognition.

Visual and acoustic annotations of the CHIL Interactive Seminar data were mainly done by ELDA, in collaboration with UKA, CMU, AIT, IBM, FBK-IRST and UPC. Packaging and distribution of data coming from CHIL was handled by UKA. The data coming from VACE was derived from a single source for the surveillance data - the Imagery Library for Intelligent Detection Systems (i-LIDS) [6]. The meeting room data was a collection derived from data collected at CMU, the University of Edinburgh (EDI), NIST, the Netherlands Organisation for Applied Scientific Research (TNO), and Virginia Tech (VT). The evaluation scoring software for VACE tasks was contributed by the University of South Florida (USF).

The discussion and definition of the invidual tasks and evaluation procedures were moderated by so-called "task-leaders". These were Keni Bernardin (UKA, 3D person tracking), Ramon Morros (UPC, CHIL-related 2D Face tracking), Rachel Bowers, Martial Michel and Travis Rose (NIST, VACE-related 2D face tracking, 2D person tracking, 2D vehicle tracking), Hazim Ekenel (UKA, visual person identification), Djamel Mostefa (ELDA, acoustic identification), Aristodemos Pnevmatikakis (AIT, audio-visual identification), Michael Voit and Jean-Marc Odobez (UKA and IDIAP, head pose estimation), Andrey Temko (UPC, acoustic event recognition). The tasks leaders were responsible for scoring the evaluation submissions. For CHIL tasks, they were also centrally scored by ELDA.

Note that original plans called for the inclusion of a person detection and tracking task in the unmanned aerial vehicle (UAV) domain using data contributed by the Defense Advanced Research Projects Agency (DARPA) Video Verification of Identity (VIVID) [7] program. Unfortunately, the annotation of this data proved to be too difficult to perform with the sufficient level of consistency required for the purposes of this evaluation. Therefore, the UAV Person Detection and Tracking task was eliminated from the evaluation.

The remainder of this paper is organized as follows: Section 2 first gives a brief overview of the used data sets and annotations, followed by an introduction to the evaluation metrics in Section 3. Section 4 then presents the various evaluation tasks with an overview of the achieved results and discusses some of the outcomes and potential implications for further evaluations. Finally, Section 5 summarizes the experiences gained from the CLEAR'07 evaluation.

Further details on the tasks definitions and data sets can be found in the evaluation plans available on the CLEAR webpage [8].

2 Evaluation Corpora

2.1 CHIL Interactive Seminars

The CHIL-sponsored evaluation tasks of 3D person detection, person identification, face detection and tracking, and acoustic event recognition were carried out using the CHIL Interactive Seminar database. This database features recordings of small seminars with 3 to 8 participants, recorded at 5 different CHIL sites with greatly varying room characteristics. The "lecture-type" Seminar database still used in CLEAR'06 [23], figuring recordings of a lecturer in front of an audience, and focused towards single person analysis were dropped completely in favor of the multiple person scenario. A minimum common sensor setup in the recording rooms guaranteed a certain level of standardization to ease algorithm development and testing. The visual sensor setup includes 4 fixed cameras with overlapping views installed in the room corners and one fisheye ceiling camera. The audio setup includes at least three 4-channel T-shaped microphone arrays and at least one MarkIII 64-channel linear microphone array on the room walls, as well as several close-talking and table top microphones. All data is synchronized, with highest priority on the audio channels which can be used for acoustic source localization and beamforming. A detailed description of the recording rooms, sensors, scenarios and procedures is given in [19]. A total of 25 seminars were recorded in 2006, which were separated into 100 minutes of development and 200 minutes of evaluation data (see Table 1).

Table 1. CHIL Interactive Seminar data used in CLEAR'07

Site	Development	Evaluation
AIT	1 Seminar (20m segment)	4 Seminars (2x 5m segments each)
IBM	1 Seminar (20m segment)	4 Seminars (2x 5m segments each)
IRST	1 Seminar (20m segment)	4 Seminars (2x 5m segments each)
UKA	1 Seminar (20m segment)	4 Seminars (2x 5m segments each)
UPC	1 Seminar (20m segment)	4 Seminars (2x 5m segments each)

For the person identification task, the same development and evaluation seminars were used, but the training and test segments were chosen from different time points to better suit the requirements of the task, as explained in Section 4.5. All video recordings are provided as sequences of single JPEG images at 640x480, 768x576, 800x600 or 1024x768 pixels resolution and at 15, 25 or 30fps, depending on the recording site and camera. The audio recordings are provided as single channels sampled at 44.1kHz, 24 bits per sample, in the WAV or SPHERE formats, depending on the recording sensor. In addition, information about the calibration of every camera, the location of every sensor, the recording room dimensions, and a few empty room images for background modeling are supplied for each seminar. The development and evaluation segments are

(a) ***AIT*** (b) ***UKA*** (c) ***FBK-IRST***

(d) ***IBM*** (e) ***UPC***

Fig. 1. Scenes from the 2007 CHIL Interactive Seminar database

annotated, providing 3D and 2D head centroid locations, face bounding boxes, facial features such as the eyes and nose bridge, and audio transcriptions of speech and other acoustic events. Fig. 1 shows example scenes from the 2007 Interactive Seminar database.

2.2 VACE Related Datasets

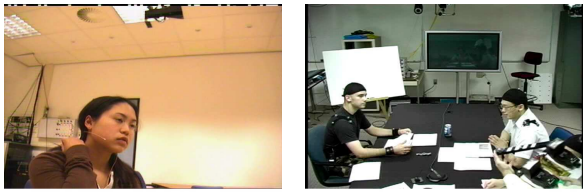
Table 2. Evaluation data

Data	Raw data	Training	Evaluation
Multi-Site Meetings	160GB	50 Clips (Face)	45 Clips (Face)
i-LIDS Surveillance	38GB	50 Clips (Person)	50 Clips (Person)
i-LIDS Surveillance	38GB	50 Clips (Moving Vehicle)	50 Clips (Moving Vehicle)

The evaluation data were assembled using two databases, multi-site meetings and surveillance data (Table 2). The surveillance data originate from the 2006 Imagery Library for Intelligent Detection Systems (i-LIDS) [6], distributed by the United Kingdom's Home Office via collaboration with NIST. All videos are in MPEG-2 format using either 12 or 15 I-frame rate encoding. The annotations are provided in ViPER (the Video Performance Evaluation Resource tool) format [16, 18]. The Multi-Site Meetings are composed of datasets from different sites, samples of which are shown in Fig. 2:

1. CMU (10 Clips)
2. EDI (10 Clips)
3. NIST (10 Clips)
4. TNO (5 Clips)
5. VT (10 Clips)

(a) ***CMU*** **(b)** ***EDI*** **(c)** ***NIST***



(d) ***TNO*** **(e)** ***VT***

Fig. 2. Scenes from Multi–Site Meetings

Sample annotations for the moving vehicle and the person tracking in surveillance tasks are shown in Figs. 3 and 4.

2.3 Other Datasets

In addition to the above described databases, some tasks were carried out using other datasets more suited to their requirements. The Head Pose Estimation task was performed on two databases: One recorded at UKA, using 4 corner cameras with overlapping views of the room, and one extracted from the AMI Meeting database, featuring single views of a meeting table. These databases and their annotations are explained further in Section 4.6. For the Acoustic Event Recognition task, although development and evaluation was mostly based on the CHIL Interactive Seminar database, 2 databases of isolated acoustic events, recorded at UPC and ITC, which were also used in the CLEAR 2006 evaluation, were included in the development set. More details are given in Section 4.7.

3 About Tracking Metrics

The reason tracking metrics are specifically presented here is because these same metrics were used in many of the CLEAR tasks, including 3D visual, acousic and

Fig. 3. Sample annotation for vehicle. MOBILE objects are marked by black boxes. STATIONARY objects are marked by white boxes. The shaded region indicates where mobile vs. stationary is ambiguous.

audio-visual person tracking, face tracking, and 2D person and vehicle tracking. As opposed to other tasks, such as face identification, for which well known and widely accepted metrics exist, there is yet no common standard in the tracking community for the evaluation of multiple object trackers. Most measures are designed with the characteristic of a specific domain in mind (e.g. merges and splits in 2D visual tracking, coming from the tradition of 2D foreground blob analysis), and are not suited for application to other domains (such as e.g. acoustic tracking, 3D tracking, etc). For the first CLEAR evaluation in 2006, an effort was undertaken to harmonize the metrics used in the different tracking tasks under consideration in the CHIL and VACE communitites. The resulting metrics, the Multiple Object Tracking Precision ($MOTP$) and the Multiple Object Tracking Accuracy ($MOTA$), should for the first time offer a general framework for the evaluation of multibody trackers in all domains and for all modalities. The MOT metrics are only briefly sketched in the following. For a detailed explanation, the reader is referred to [11, 14, 22]. The metrics used in the person identification, head pose estimation and acoustic event recognition tasks are described together with the respective task descriptions in Section 4.

3.1 The MOT Tracking Metrics

The Multiple Object Tracking (MOT) metrics build upon a well defined procedure to calculate the basic types of errors made by multiple object trackers over a tracking sequence: Imprecisions in the estimated object locations, failures to estimate the right number of objects, and failures to keep a consistent labeling of these objects in time. Given that for every time frame t a multiple object tracker

Fig. 4. Sample annotation for a person in surveillance.

outputs a set of hypotheses $\{h_1 \dots h_m\}$ for a set of visible objects $\{o_1 \dots o_n\}$, let c_t be the number of object-hypothesis correspondences made for frame t and d_t^i be the distance between object o_i and its corresponding hypothesis. Let further g_t be the number of objects and fp_t, m_t and mme_t be the number of false positives, misses, and track ID mismatch errors made for frame t. Then the $MOTP$ is defined as:

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

and the $MOTA$ as:

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

For the distance d_t^i between an object and a tracker hypothesis, various measures can be used without changing the general framework. For the CLEAR 3D person tracking tasks, e.g., the euclidian distance on the ground plane between annotated and tracked object centroids was used, whereas for the 2D face, person and vehicle tracking tasks, the spatial overlap between annotated and tracked bounding boxes, G_t^i and D_t^i, was used.

$$d_t^i = \frac{|G_t^i \cap D_t^i|}{|G_t^i \cup D_t^i|} \tag{3}$$

4 CLEAR 2007 - Evaluation Tasks and Results

The CLEAR tasks can be broken down into four main categories: tracking tasks, identification tasks, head pose estimation and acoustic event recognition. Table 3 shows the different CLEAR 2007 tasks.

Table 3. CLEAR'07 tasks

Task name	Organizer	Section	Database
Tracking			
3D Person Tracking (A,V,AV)	CHIL	4.1	Interactive Seminars
2D Face Det. & Tracking (V)	CHIL/VACE	4.2	Int. Sem./Multi-Site Meetings
2D Person Tracking (V)	VACE	4.3	Surveillance Data
2D Vehicle Tracking (V)	VACE	4.4	Surveillance Data
Person Identification (A,V,AV)	CHIL	4.5	Interactive Seminars
Head Pose Estimation (V)	CHIL/AMI	4.6	Seminars[1], AMI Meetings
Acoustic Event Recognition	CHIL	4.7	Int. Sem., Isolated Events

4.1 3D Person Tracking

The objective of the 3D person tracking task is to estimate the trajectories on the ground plane of the participants in CHIL Interactive Seminar recordings (see Fig. 5). As in the previous evaluation, it is broken down into 3 subtasks: Visual, acoustic and multimodal tracking. For all subtasks, the $MOTP$ and $MOTA$ metrics described in Section 3 are applied, evaluating both localization precision and tracking accuracy. The database for evaluation consisted of 200 minutes of recordings from 5 different CHIL sites and included the streams from 4 corner cameras and a panoramic ceiling camera, from at least 12 audio channels coming from 3 T-shaped microphone arrays, and from at least 64 more audio channels captured by a MarkIII microphone array. The scenes figured 3 to 8 seminar participants engaged in natural interaction, and were cut out as 5 minute segments from various points inside the seminars, such that they did often not include the starting phase, where persons enter the room. Trackers therefore had to be capable of acquiring person tracks at any point in the sequence, of adapting their person models, and had to automatically cope with the variability of all CHIL rooms without room specific tuning.

Some notable changes to the CLEAR'06 tracking task should be mentioned here:

[1] For this task, a number of interactive seminars were recorded and annotated in 2006. These seminars, however, were not part of the dataset used for the tracking and identification tasks.

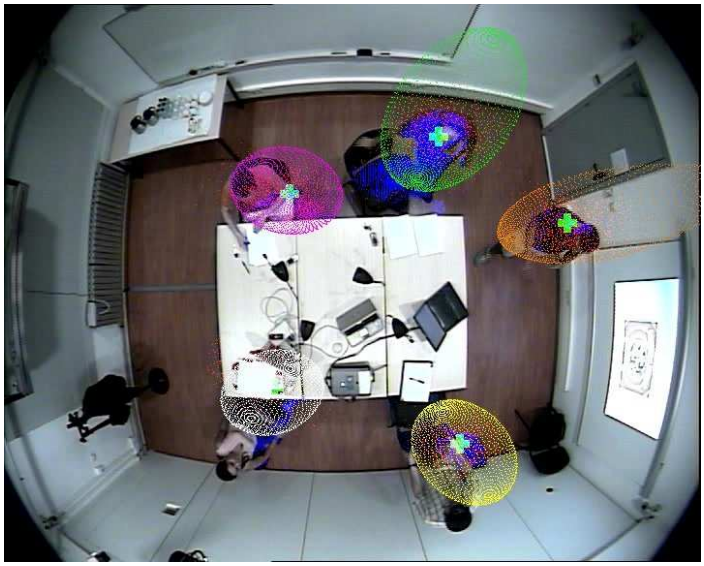


Fig. 5. Example screenshot of a 3D person tracking system running on Interactive Seminar data (Image taken from [15]).

- First of all, the single person tracking scenarios (lecture scenarios) were dropped completely. Only scenarios involving the tracking of multiple persons were considered.
- The acoustic subtask was extended and required trackers to automatically detect segments of speech in addition to performing localization. This means that segments of silence or noise were now included in the evaluation data. Segments containing cross-talk, though, were still considered as "don't care" segments.
- The multimodal subtask was redefined and the conditions *A* and *B* from CLEAR'06 were dropped. The goal in this evaluation was to audio-visually track the last known speaker. This implies that the tracking target has to be determined acoustically, tracked audio-visually, segments of silence have to be bridged using only the visual modality, and the target has to be switched automatically when a new speaker becomes active. The defined task cannot be solved well using monomodal trackers. This change in the task definition was made to achieve a better balance of the modalities and to better show the advantages of multimodal fusion.

Fig. 6 shows the results for the visual subtask. A total of 7 systems from 4 sites participated. Various approaches, such as particle filters, Kalman filters, and heuristic-based trackers were represented and a variety of features, gained from the multiple views, were used. These include foreground segmentation support maps, person colors, body or face detections, edge contours, etc. The best performing system in terms of accuracy (78.36%) was a particle filter based tracker

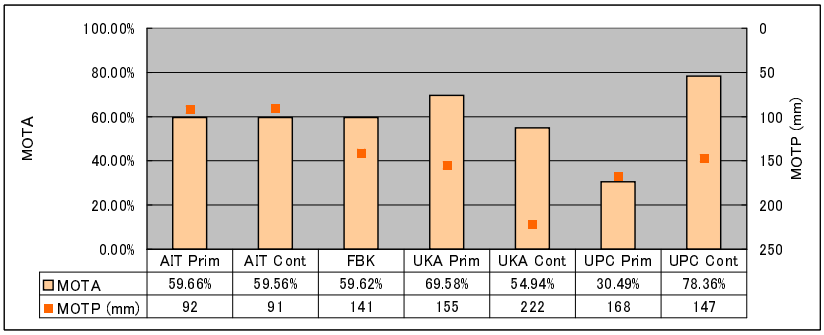


Fig. 6. 3D Person Tracking – Visual subtask. The light bars represent the $MOTA$ in percent and the dark dots represent the $MOTP$ in mm

using as sole feature a 3D voxelized foreground support map, computed from the various views. The most performant system in terms of precision (91mm) was based on the intelligent tracking and combination of detected faces in the 2D views. According to the runtime information provided in the system descriptions, almost all these systems performed at close to realtime.

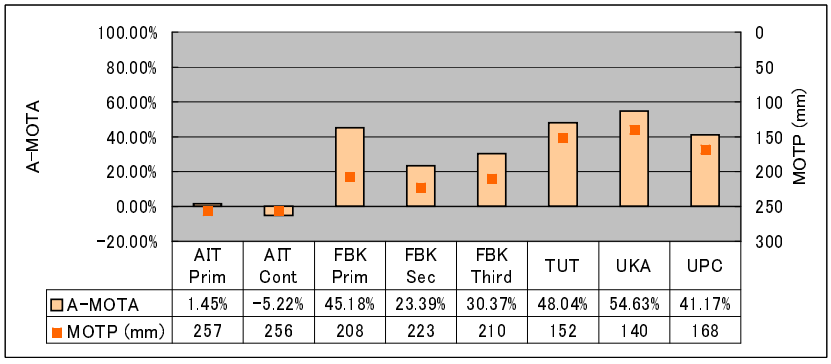


Fig. 7. 3D Person Tracking – Acoustic subtask

Fig. 7 shows the results for the acoustic subtask. A total of 8 systems from 5 sites participated. The approaches were based on the computation of the Generalized Cross-Correlation (GCC-PHAT) between microphone pairs or of a Global Coherence Field (GCF or SRP-PHAT) using the information from all arrays. While some systems still tackled speech segmentation and localization separately, others did use a combined approach. The most performant system overall was a Joint Probabilistic Data Association Filter (JPDAF) - based tracker, which performed speech segmentation by thresholding localization uncertainties.

It reached a precision of 140mm and an accuracy of 54.63%. Most systems proved to be capable of realtime or close to realtime operation.

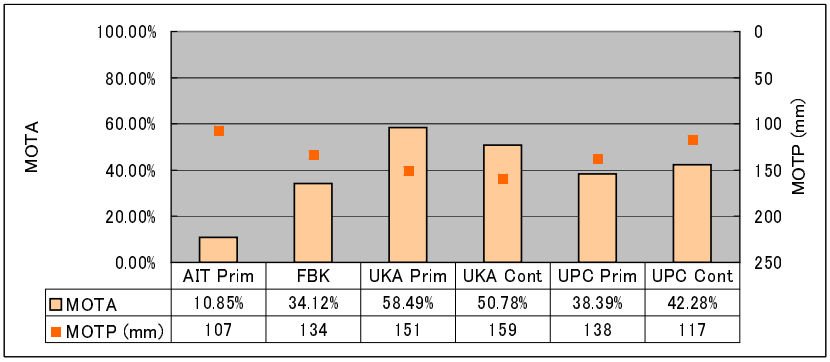


Fig. 8. 3D Person Tracking – Multimodal subtask

Fig. 8 shows the results for the multimodal subtask. A total of 6 systems from 4 sites participated. These systems are a combination of the visual and acoustic trackers presented earlier. Almost all systems perform modality fusion by postprocessing the outputs of the monomodal trackers and combining at the decision level. The only exception is the lead system in terms of accuracy (58.49%), which fused the audio and visual information at the feature level to initiate, update, and terminate person tracks.

In general, the biggest challenge facing visual tracking systems in the CLEAR scenarios is still the reliable detection of persons in various poses, with partial occlusions, from a variety of viewing angles, in natural uncontrolled environments. Compared to systems presented in 2006, the approaches were much more advanced this year, fusing far more features and more types of detectors to achieve higher robustness. The best $MOTA$ score improved from 62.79% to 78.36%, despite the much higher variability caused by the inclusion of more recording sites (for comparison, the best system from 2006, *UKA Cont* [12, 13], achieved only 54.94% $MOTA$ on the 2007 data). Similarly, the challenge on the acoustic side relies on the proper detection and segmentation of speech in the presence of irregular, non-uniform noise sources, reverberation and crosstalk. While the best acoustic performance seems to have dropped from 64% in 2006 to 54.63% in 2007, one must remember that the task this year involved also the automatic segmentation of speech, while last year systems were only evaluated on manually annotated segments of clean speech. On the whole, scores for all systems were much higher, showing that basic difficulties previously encountered could be overcome to some extent. Undoubtedly, though, a great deal of work must still be done to further increase the robustness of acoustic systems. On a last note: The best multimodal $MOTA$ score for 2007, 58.49%, can not be directly compared to the best 2006 multimodal scores (37.58% for condition A, 62.20%

for condition B), as the task definitions, and therefore the goals and difficulties for trackers differ. Also, the 2007 multimodal scores cannot be directly compared to the 2007 monomodal visual or acoustic scores for the same reasons. At the very least, one can observe that an early fusion of audio-visual features seems to bear some advantages, as shown by this year's best performing system. Only the scoring of monomodal acoustic trackers on periods of silence, just as in the multimodal task, could clearly show the advantages gained by the addition of visual features[2]. The problem of objectively measuring the advantages of multimodal fusion, especially in natural, application-near scenarios such as in CLEAR, still poses some difficult questions that must be investigated.

Appendix A graphically shows a more detailed analysis of the results for the CLEAR 2007 3D person tracking task, for the audio, visual and multimodal subtasks.

4.2 2D Face Detection and Tracking

The purpose of this task is to measure the accuracy of face tracking for meeting and lecture room videos. The objective is to automatically detect and keep track of all visible faces in a video sequence, estimating both their position and their extension (see Fig. 9).

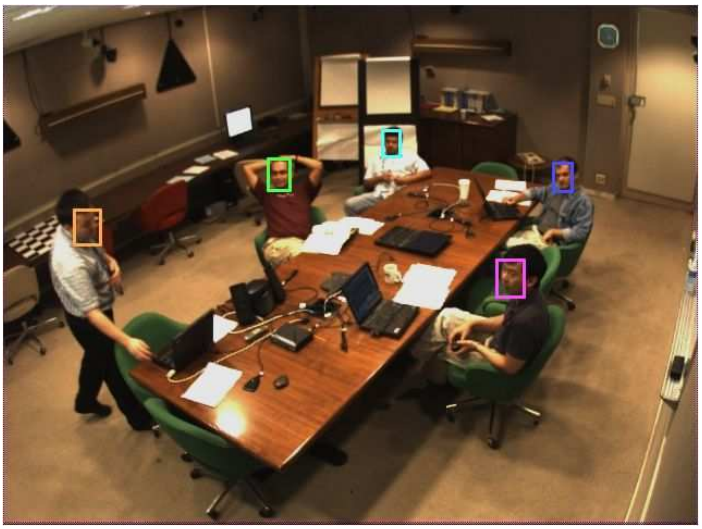


Fig. 9. Example screenshot for the face tracking task on Interactive Seminar data (Image taken from [20]).

[2] While the scoring of such trackers on silence-only periods is useful for diagnostic purposes in determining the contribution from audio tracking to the multimodal task, it is not representative of a real-world task.

The task was evaluated on two databases, the CHIL Interactive Seminars and the VACE Multi-Site Meetings. While for the Multi-Site Meeting database, detection and tracking could only be performed separately in the multiple camera views, the Interactive Seminar database offered exact calibration information between views, allowing to use 3D geometric reasoning about scene locations of faces to increase accuracies (this was not exploited by any of the participating systems, though). In both cases, the overall performance is computed as the average of 2D tracking performances across all views. Face sizes in the CLEAR databases are extremely small (down to 10x10 pixels), faces are rarely oriented directly towards a camera, lighting conditions are difficult and faces are often occluded, making standard skin color segmentation or template matching techniques unusable. Thus, the difficulty of the dataset drives the development of innovative techniques for this research field.

In contrast to CLEAR'06, the task was better harmonized, with respect to the CHIL and VACE datasets, notably concerning the annotation of face extensions, the definition of visible faces, and the metrics used. Faces are considered visible if of the three annotated features, the left eye, the right eye and the nose bridge, at least two are visible. They are regarded as "don't care" objects, which are ignored in scoring, if only one feature is visible. As for all tracking tasks in CLEAR'07, the *MOT* metrics were adopted, using the overlap between annotated and tracked face bounding boxes as distance measure.

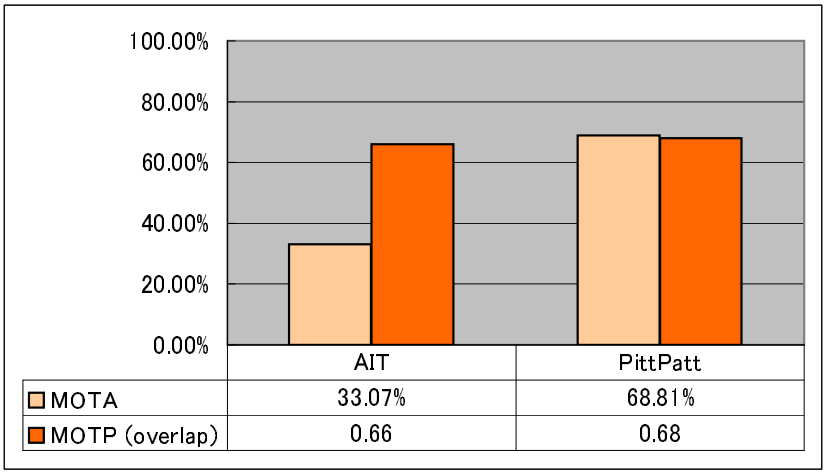


Fig. 10. 2D Face Tracking – CHIL Interactive Seminar database

2D Face Tracking on the CHIL Interactive Seminar Database The results for face tracking on the Interactive Seminar database are shown in Fig. 10.

As in 2006, two sites participated on this dataset. The best system used a 3-stage algorithm consisting of a frame-based face detection step, a motion-based tracking step, and a subsequent track filtering step. It reached a precision of 68% overlap and an accuracy of 68.81%. These results are slightly better than those in 2006 (best $MOTP$: 0.64, best $MOTA$: 68.32%), although in 2006 tracking errors resulting from track ID switches were not counted, and in 2007 the conditions were more challenging due to an increase in the amount of seminar participants involved.

2D Face Tracking on the Multi-Site Meeting Database The results on the Multi-Site Meeting database appear in Fig. 11. 5 systems from 3 different sites participated in the evaluation. The leading system here also used a 3-stage approach consisting of face detection using a hierarchical multi-view face detector, particle filter-based tracking, and filtering of the resulting tracks. It reached scores of 70% $MOTP$ and 85.14% $MOTA$.

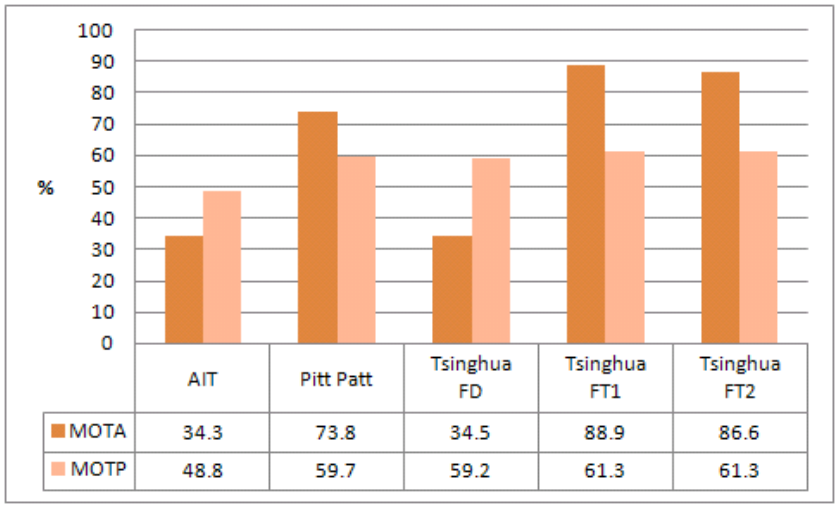
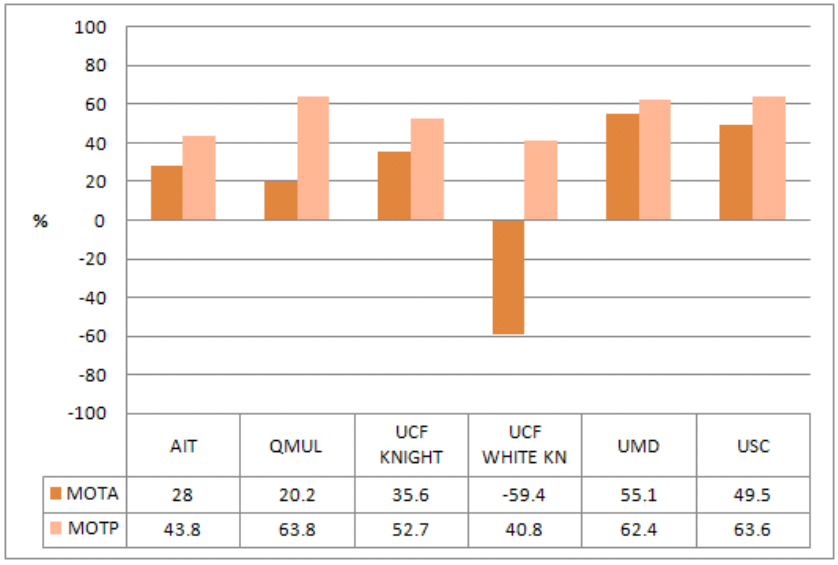


Fig. 11. Face tracking in meeting room

For both datasets, the main difficulties still stemmed from very small or hardly identifiable faces, extreme views of faces, and blurred or highly compressed video. Another important factor is that the quality of annotations is also affected by these same problems. A fair portion of false positives can (for example) be attributed to cases where faces are tracked in extreme poses in which none of the facial features were clearly visible, and were therefore annotated as invisible. The converse also holds for ambiguous cases which were judged visible based on facial feature annotations, but only contain fractions of a face, resulting in a miss by the tracker. In both cases, better guidelines for the

annotation of "don't care" faces, and some form of rating for the difficulty of the underlying video sequence may reveal a much higher performance of presented tracking systems than the actual numbers suggest.

For all VACE-sponsored tasks, a unified evaluation methodology was applied, as task definitions, annotations and metrics were very similar. This methodology should be briefly mentioned here: Each participating site electronically submitted system output for scoring using the USF_DATE software[3]. Submissions were evaluated using a batch process that involved two main stages: a data validation step, followed by application of the metrics.



	AIT	QMUL	UCF KNIGHT	UCF WHITE KN	UMD	USC
MOTA	28	20.2	35.6	-59.4	55.1	49.5
MOTP	43.8	63.8	52.7	40.8	62.4	63.6

Fig. 12. Person tracking in surveillance video

To run the scoring software, it was verified that the submissions were compliant with the Viper Document Type Definition [16, 18] and would successfully be parsed. This required normalization of all submissions to complete validation of their data. In the following cases, sites were notified and asked to resubmit corrected files:

- Wrong object types: this occurred when submissions contained custom keywords for objects.
- No object in submission: either the submission file contained no object at all, or no object relevant to the task being scored was present in the file.
- Using a 2006 index file: cases where submitted files matched CLEAR 2006 index files.

[3] USF_DATE is USF (University of South Florida) DATE (Detection and Tracking Evaluation).

– Incomplete submission: when a submitted system output was not complete, such as a malformed XML file.
– Not a Viper file: some submissions were not in Viper format.

Each submission was evaluated against the ground truth using the metrics described in Section 3. In cases where the submission could not be scored due to limitations in USF_DATE, the clip was marked as being problematic. Finally, the set of all clips that were successfully scored for all submissions was used to obtain the MOTA and MOTP scores, i.e. the same clips were used in these calculations for all submissions, and scores were calculated only with respect to the objects retained in the test set.

4.3 2D Person Tracking

The purpose of this task it is to track persons in a surveillance video clip. The annotation of a person in the Surveillance domain comprises the full extent of the person (completely enclosing the entire body including the arms and legs). Specific annotation details about how a person is marked appear in the guidelines document [21]. The person tracking in surveillance video results appear in Fig. 12.

4.4 2D Vehicle Tracking

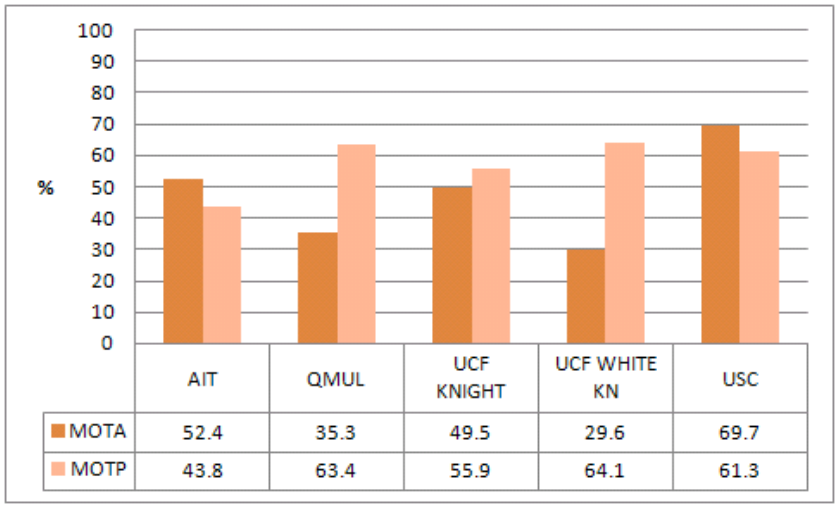


Fig. 13. Vehicle tracking in surveillance video

The goal of the moving vehicle task is to track moving vehicles in a given video clip. For the annotation, only vehicles that have moved at any time during

the clip are marked. Vehicles are annotated at the first frame where they move. For specific details see [21].

For this evaluation task, the vehicle has to be moving and must be clearly visible (i.e., should not be occluded by other objects). In the i-LIDS dataset there are regions where vehicles are not clearly visible due to tree branches or where the sizes of vehicles are very small. These regions are marked accordingly (as "don't care" regions). The vehicle tracking in surveillance video results are summarized in Fig. 13.

4.5 Person Identification

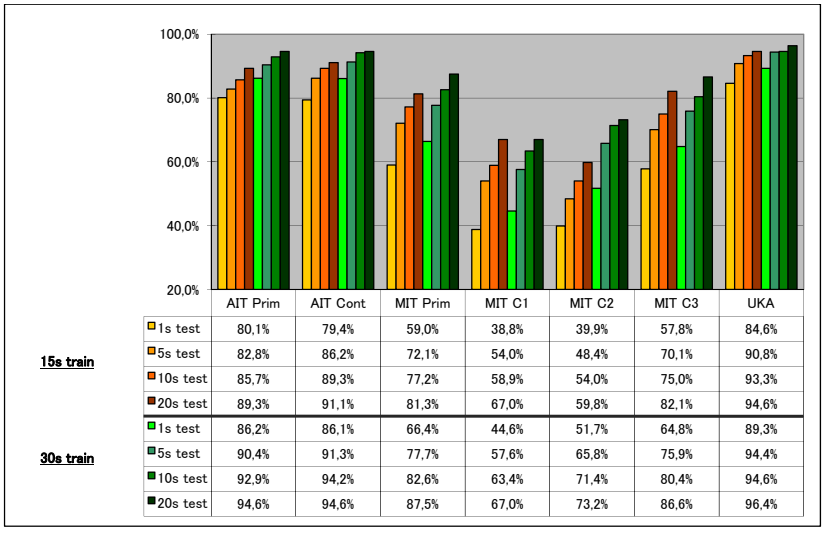
The person identification task in the CLEAR evaluation was designed to measure the performance of visual and acoustic identification systems operating under far-field[4] conditions in realistic meeting and seminar scenarios (see Fig. 14). The task was that of closed set identification and was evaluated on the CHIL

Fig. 14. Example screenshot of a face identification system running on Interactive Seminar data (Image taken from [17]).

Interactive Seminar database. Only corner camera views and the MarkIII microphone array channels were available. For each participant to be identified, training, validation and testing data was provided. The training data consisted

[4] The "far-field" condition implies that only fixed microphones placed on the room table or walls are to be used, as opposed to close talking or lapel microphones, which are worn directly by the users. This causes for a significantly lower signal to noise ratio, making the task much more challenging.

of 15 and 30 second audio-visual data segments extracted from the original sequences. Testing was then made on segments of varying length, from 1 to 20 seconds, to measure the improvements to be achieved by temporal fusion. A major improvement over the CLEAR'06 evaluations is that the evaluation segments were much more carefully chosen to offer a better balance between the audio and visual modalities. Care was taken that, for each segment, at least a certain amount of frontal unoccluded views of the head were available in addition to clean speech, eliminating the artificial bias towards the audio modality observed in 2006. Visual annotations were also of higher frequency and accuracy, with face bounding box, left eye and right eye labels provided every 200ms. The evaluation set comprised 28 individuals in total (up from 26 in 2006). Figs. 15, 16 and 17 show the results for the visual, acoustic and multimodal subtasks respectively.



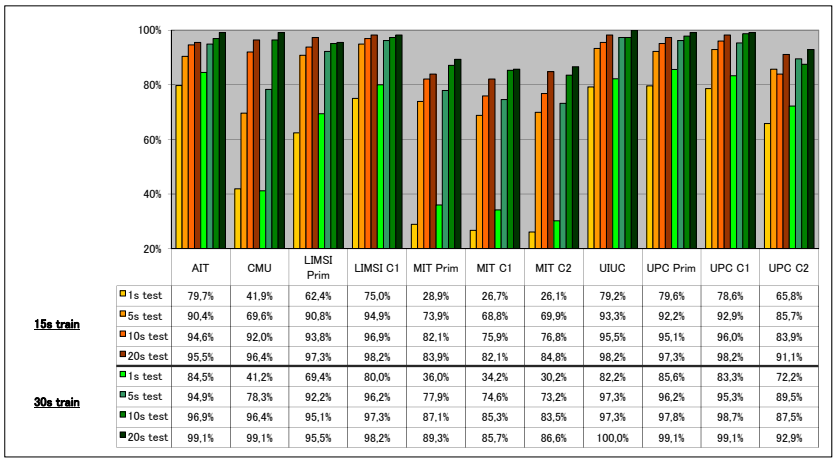
		AIT Prim	AIT Cont	MIT Prim	MIT C1	MIT C2	MIT C3	UKA
15s train	1s test	80,1%	79,4%	59,0%	38,8%	39,9%	57,8%	84,6%
	5s test	82,8%	86,2%	72,1%	54,0%	48,4%	70,1%	90,8%
	10s test	85,7%	89,3%	77,2%	58,9%	54,0%	75,0%	93,3%
	20s test	89,3%	91,1%	81,3%	67,0%	59,8%	82,1%	94,6%
30s train	1s test	86,2%	86,1%	66,4%	44,6%	51,7%	64,8%	89,3%
	5s test	90,4%	91,3%	77,7%	57,6%	65,8%	75,9%	94,4%
	10s test	92,9%	94,2%	82,6%	63,4%	71,4%	80,4%	94,6%
	20s test	94,6%	94,6%	87,5%	67,0%	73,2%	86,6%	96,4%

Fig. 15. Recognition rates for Person Identification – Visual subtask. Results are shown for 15 and 30 second training, and for 1, 5, 10 and 20 second test segment lengths. Shown are the Correct Recognition Rates in percent

For the visual subtask, 7 systems from 3 sites were represented. As manual labels for face bounding boxes and eyes of the concerned participant for a segment were provided, systems did not need to perform tracking, but just to align and crop faces for recognition. Two systems did use some form of preprocessing, though, e.g. interpolating between 200ms label gaps to obtain more facial views. Many types of feature extraction algorithms were used, including PCA, LDA, block-based DCT, and variants or combinations thereof. Classification was mostly done using nearest neighbor classifers. The best results in all train and test conditions were reached by a local appearance based approach

using only labeled faces, DCT features, and nearest neighbor classification. It achieved 84.6% accuracy for the hardest condition in terms of data availability (15s train, 1s test) and 96.4% for the easiest condition (30s train, 20s test). This is a major improvement over 2006, where the best results obtained in the 30s train, 20s test condition were 83.7%.

While some of the improvement stems from algorithm design, some part of it must no doubt also be attributed to better labeling and segmentation of the visual data, as decribed above. Because of the differences in the preprocessing and classification techniques, it is difficult to directly compare the strengths of the feature extraction algorithms. Looking at the results from both CLEAR 2006 and 2007, however, one may find that using local models of appearance does offer some advantages over other techniques. A more thorough experimental investigation is necessary, though, before general conclusions could be made.

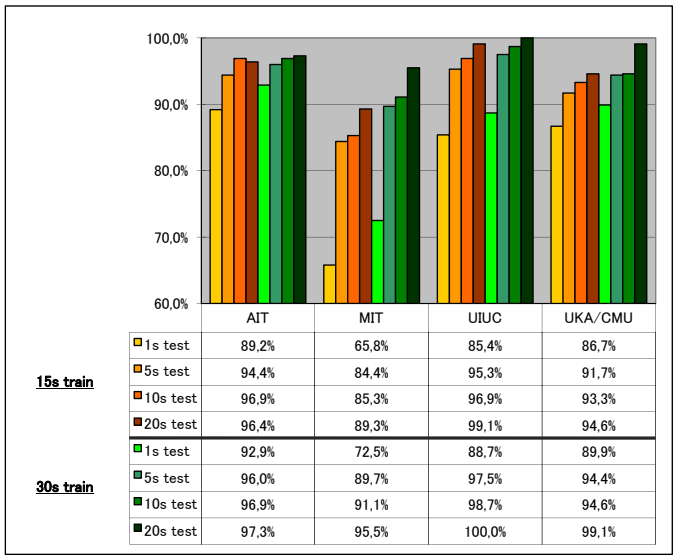


		AIT	CMU	LIMSI Prim	LIMSI C1	MIT Prim	MIT C1	MIT C2	UIUC	UPC Prim	UPC C1	UPC C2
15s train	1s test	79,7%	41,9%	62,4%	75,0%	28,9%	26,7%	26,1%	79,2%	79,6%	78,6%	65,8%
	5s test	90,4%	69,6%	90,8%	94,9%	73,9%	68,8%	69,9%	93,3%	92,2%	92,9%	85,7%
	10s test	94,6%	92,0%	93,8%	96,9%	82,1%	75,9%	76,8%	95,5%	95,1%	96,0%	83,9%
	20s test	95,5%	96,4%	97,3%	98,2%	83,9%	82,1%	84,8%	98,2%	97,3%	98,2%	91,1%
30s train	1s test	84,5%	41,2%	69,4%	80,0%	36,0%	34,2%	30,2%	82,2%	85,6%	83,3%	72,2%
	5s test	94,9%	78,3%	92,2%	96,2%	77,9%	74,6%	73,2%	97,3%	96,2%	95,3%	89,5%
	10s test	96,9%	96,4%	95,1%	97,3%	87,1%	85,3%	83,5%	97,3%	97,8%	98,7%	87,5%
	20s test	99,1%	99,1%	95,5%	98,2%	89,3%	85,7%	86,6%	100,0%	99,1%	99,1%	92,9%

Fig. 16. Recognition rates for Person Identification – Acoustic subtask. Shown are the Correct Recognition Rates in percent

A total of 11 systems from 6 sites participated in the acoustic subtask. The approaches were based on Gaussian Mixture Models (GMMs), adaptated Universal Background Models (UBMs) or Support Vector Machines (SVMs), and used Mel-Frequency Cepstral Coefficient (MFCC) or Perceptually-weighted Linear Predictive (PLP) features, their derivatives, or combinations thereof. Systems also differed in the amount of microphone channels used and the way they were fused in pre- or post-processing. The best overall results in the 15s training condition were achieved by a system using PLP coefficients from one single channel and a UBM model. It reached 98.2% for the 20s test condition. The best overall results in the 30s training condition came from a system using UBM-GMM classifiers separately on 7 channels, and fusing at the decision level. It reached 100% accuracy for the 20s test condition. On the whole, it seems that PLP

features, used stand-alone or in combination with others, outperform other features, and that adapted UBM models outperform speaker specific GMMs. It was also observed that, contrary to expectations, pre-processing multiple channels through beamforming to produce a cleaner signal degrades performance. The more promising path seems to be the combination of classifier outputs at the post-decision level. In comparison with 2006, clear improvements could be noticed only in the 15s training condition.



		AIT	MIT	UIUC	UKA/CMU
15s train	1s test	89,2%	65,8%	85,4%	86,7%
	5s test	94,4%	84,4%	95,3%	91,7%
	10s test	96,9%	85,3%	96,9%	93,3%
	20s test	96,4%	89,3%	99,1%	94,6%
30s train	1s test	92,9%	72,5%	88,7%	89,9%
	5s test	96,0%	89,7%	97,5%	94,4%
	10s test	96,9%	91,1%	98,7%	94,6%
	20s test	97,3%	95,5%	100,0%	99,1%

Fig. 17. Recognition rates for Person Identification – Multimodal subtask. Shown are the Correct Recognition Rates in percent

Four sites participated in the multimodal subtask. All systems used post-decision fusion of the monomodal recognizer outputs, with different strategies for the weighting of audio and visual inputs. The most performant system in the multimodal case was also based on the best overall acoustic system. It used an appearance based technique for face identification, which was not evaluated separately in the visual subtask. For 20s test segments, it reached 99.1% and 100% accuracies for the 15s and 30s train conditions, respectively. Overall, the perfomance of acoustic systems was quite high, such that the advantages of multimodal fusion could only be observed in the 1s test condition, where the best results improved from 79.7% and 85.6% (15s, 30s train) to 89.2% and 92.9%. When the availability of both modalities is guaranteed, the strength of multimodal approaches clearly lies in the smaller amount of observations required, more than in the accuracies to be reached.

Appendix B summarizes the best results for the person identification task in CLEAR 2007, and shows the progress achieved since CLEAR 2006, for the audio, visual and multimodal subtasks, and for all evaluation conditions.

4.6 Head Pose Estimation

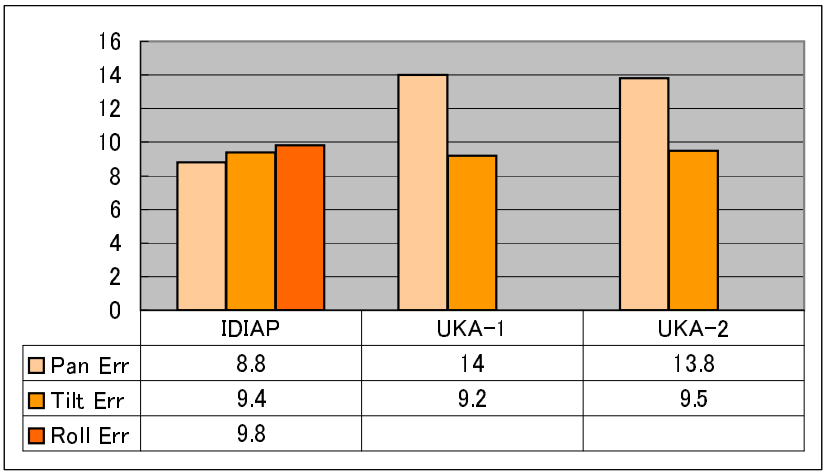
The objective in the head pose estimation task is to continuously estimate the pan, tilt, and roll orientations of a person's head using using visual information from one or more cameras.

Fig. 18. Example screenshot for the head pose estimation task on the AMI Meeting Corpus (Image taken from [10]).

The task was subdivided into two subtasks, determined by the datasets used: The first subtask was built on the AMI Meeting Corpus [10], and offered single views of meeting participants interacting around a table (see Fig. 18). It contained 16 one minute segments, extracted individually for 16 different subjects, of which 10 were to be used for training, and 6 for evaluation. The task required automatically tracking the head of one of the participants, in addition to estimating its orientation. The second subtask involved a data corpus captured in the CHIL-UKA smart room and offered 4 synchronized and calibrated views, which could be combined to derive head orientations in the room coordinate frame. In contrast to the AMI database, head sizes were relatively small and manual annotations for the head bounding box were provided, such that no tracking was necessary. A total of 15 subjects was considered, 10 for training and 5 for evaluation, with a 3 minute segment provided per person. For both subtasks, the

ground truth head orientations were captured with high precision using "Flock of Birds" magnetic sensors. This constitutes a great improvement over the previous evaluation, where only manual annotations into 45° pan classes were available. The metrics used were the mean absolute pan, tilt and roll errors, as well as the mean angular error between annotated and estimated head orientation vectors. Figs. 19 and 20 show the mean absolute pan/tilt/roll errors for the AMI and the CHIL corpus, respectively.

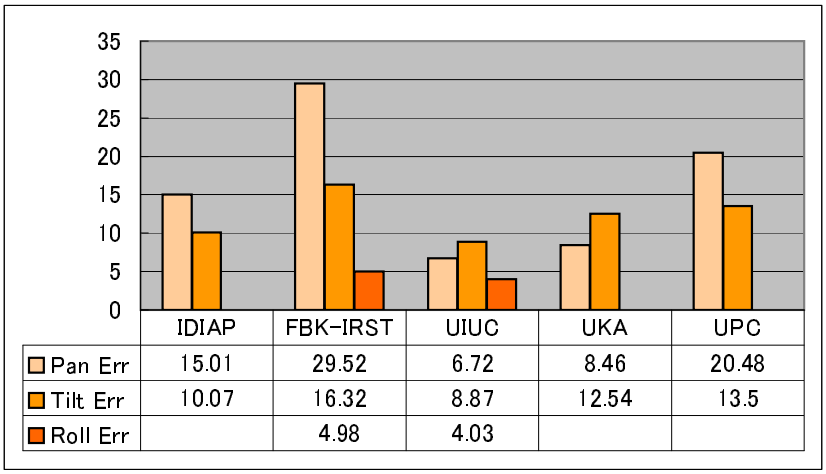
For the first subtask, 3 systems from 2 sites participated. The best systems achieved errors rates of less than 10° in all dimensions. The overall most performant system used a specially designed particle filter approach to jointly track the head location and pose. It reached 8.8° pan, 9.4° tilt and 9.8° roll error.



	IDIAP	UKA-1	UKA-2
Pan Err	8.8	14	13.8
Tilt Err	9.4	9.2	9.5
Roll Err	9.8		

Fig. 19. Head Pose Estimation – AMI Meeting database

A total of 5 sites participated in the second subtask. The best error rates were remarkably low, even compared to the previous subtask, although face sizes in this database were notably smaller. This is due in part to the availability of several camera views for fusion, but undoubtedly also to the availability of head bounding box annotations, which allow for optimal head alignment. Only two systems attempted location and pose tracking jointly, while the best performing systems relied on the manual annotations. The best overall system relied on a special person-independent manifold representation of the feature space, constructed by synchronizing and embedding person-specific submanifolds, and estimated head poses using a k-nearest neighbor classifier. It reached error levels as low as 6.72° pan, 8.87° tilt and 4.03° roll.



	IDIAP	FBK-IRST	UIUC	UKA	UPC
Pan Err	15.01	29.52	6.72	8.46	20.48
Tilt Err	10.07	16.32	8.87	12.54	13.5
Roll Err		4.98	4.03		

Fig. 20. Head Pose Estimation – CHIL database

4.7 Acoustic Event Recognition

As in 2006, the 2007 CLEAR evaluations featured an acoustic event recognition task, in which non-speech noises occuring in a seminar scenario were to be identified. A definite change compared to 2006, is that classification of pre-segmented events was not considered anymore. Instead, evaluation was performed on the CHIL Interactive Seminar database, on the same segments as used in the 3D Person Tracking, Person Identification, and 2D Face Tracking tasks. This is a major extension to the previous evaluation, where only one full-valued seminar was considered, aside from isolated event databases. In addition to *classification*, systems had to automatically *detect* acoustic events, possibly overlapped with speech or other acoustic events. 12 event classes were considered, including "door knock", "steps", "chair moving", "paper work", "phone ring", "applause", "laugh", etc. The recognition of the "speech" and "unknown" classes was not evaluated. For development, one seminar was taken per recording site, as well as the 2 isolated event databases from 2006. The test data was chosen from the remaining seminars and comprised 20 five minute segments from 4 sites, for a total of 6000 seconds, of which 36% were classified as acoustic events of interest, 11% as silence, and 53% as speech or "unknown" events. The Interactive seminars offered a challenging testbed, as 64% of acoustic events in the evaluation data were overlapped with speech and 3% were overlapped with other acoustic events. Two new metrics were defined for this evaluation, the $AED - ACC$, measuring event detection accuracy, and the $AED - ER$, measuring how precisely the temporal boundaries of acoustic events are found. They are defined as follows:

$$AED - ACC = \frac{(1+\beta^2) * Precision * Recall}{\beta^2 * Precision + Recall},$$

where

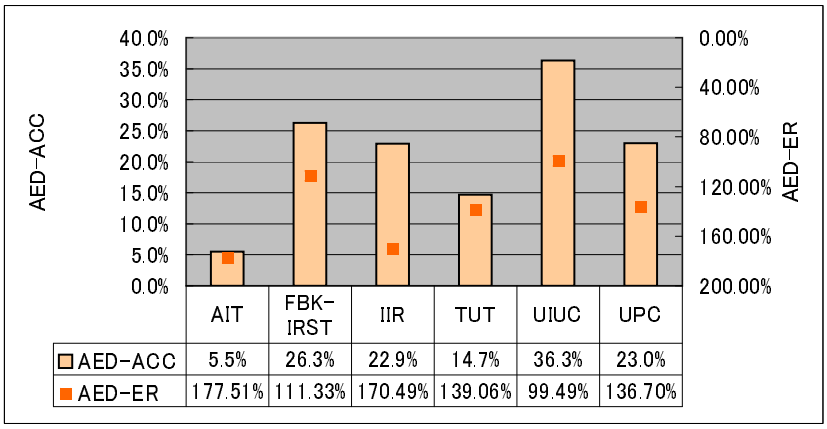
$$Precision = \frac{number\ of\ correct\ system\ output\ AEs}{number\ of\ all\ system\ output\ AEs}$$

$$Recall = \frac{number\ of\ correctly\ detected\ reference\ AEs}{number\ of\ all\ reference\ AEs}$$

and β is a weighting factor that balances precision and recall. In this evaluation, the factor β was set to 1.

$$(AED-ER) = \frac{\sum_{all\ seg} dur(seg) * (max(N_{REF}, N_{SYS} - N_{correct}(seg)))}{\sum_{all\ seg} dur(seg) * N_{REF}(seg)}$$

where, for each segment seg (defined by the boundaries of both reference and hypothesized AEs): $dur(seg)$ is the duration of seg, $N_{REF}(seg)$ is the number of reference AEs in seg, $N_{SYS}(seg)$ is the number of system output AEs in seg and $N_{correct}(seg)$ is the number of reference AEs in seg which correspond to system output AEs in seg. Notice that an overlapping region may contribute to several errors. The results of the Acoustic Event Recognition task are shown in Fig. 21.



	AIT	FBK-IRST	IIR	TUT	UIUC	UPC
AED-ACC	5.5%	26.3%	22.9%	14.7%	36.3%	23.0%
AED-ER	177.51%	111.33%	170.49%	139.06%	99.49%	136.70%

Fig. 21. Acoustic Event Recognition – site-independent systems. The light bars represent the $AED-ACC$ and the dark dots represent the $AED-ER$

Six sites participated in the evaluation. From the presented systems, 5 are Hidden Markov Model (HMM) or Gaussian Mixture Model (GMM) based, and one is based on Support Vector Machines (SVMs). Half of the systems use multiple microphones, and the other half (including the best performing system) use only a single microphone. As can be seen, the overall scores are quite low,

showing that there is still much room for improvement in spontaneous meeting room AED. The best system reached just 36.3% accuracy and almost 100% $AED-ER$ error. An analysis revealed that, on average, more than 71% of errors occur in overlapped segments as, e.g, low-energy acoustic classes, such as "chair moving", "paper work" or "steps", proved difficult to detect in the presence of speech. In occurence, the "step" class accounted for 40% of all acoustic events in the test data. Leaving out segments of overlap, the error rate of most systems would be around 30–40%. No doubt, more research is necessary to overcome the problems caused by overlap. One direction that was not explored could be to build AED systems as a set of isolated recognizers. Other improvements could be expected from the more efficient use of multiple microphones to better isolate events, or from audio-visual analysis.

5 Summary

This paper summarized the CLEAR 2007 evaluation, which started early in 2007 and was concluded with a two day workshop in May 2007. It described the evaluation tasks performed in CLEAR'07, including descriptions of metrics and used databases, and also gave an overview of the individal results achieved by the evaluation participants. Further details on the individual systems can be found in the respective system description papers in the proceedings of the evaluation workshop.

The goal of the CLEAR evaluation is to provide an international framework to evaluate multimodal technologies related to the perception of humans, their activities and interactions. CLEAR has been established through the collaboration and coordination efforts of the European Union (EU) Integrated Project CHIL - Computers in the Human Interactive Loop - and the United States (US) Video Analysis and Content Extraction (VACE) programs. After a succesful first round in 2006, the evaluations were launched again with new challenging tasks and datasets, better harmonized metrics, and with the inclusion of a new head pose estimation task, sponsored by the European Augmented Multiparty Interaction (AMI) project. The CLEAR 2007 workshop took place in May, after more than half a year of preparations, where large amounts of data were collected and annotated, task definitions were redefined, metrics were discussed and harmonized, evaluation tools were developed, and evaluation packages were distributed to participants all over the world. In CLEAR'07, seventeen international research laboratories participated in 13 evaluation subtasks.

An important contribution of the CLEAR evaluations on the whole, is the fact that they provide an international forum for the discussion and harmonization of related evaluation tasks, including the definition of procedures, metrics and guidelines for the collection and annotation of necessary multimodal datasets.

Another important contribution of CLEAR and the supporting programs is also the fact that significant multimedia datasets and evaluation benchmarks have been produced over the course of several years, which are now available

to the research community. Evaluation packages for the various tasks, including datasets, annotations, scoring tools, evaluation protocols and metrics, are available through the Evaluations and Language Distribution Agency (ELDA)[9] and NIST.

While we consider CLEAR'06 and '07 as a great success, we think that the evaluation tasks performed - mainly tracking, identification, head pose estimation and acoustic scene analysis - do yet only scratch the surface of automatic perception and understanding of humans and their activities. As systems addressing such "lower-level" perceptual tasks are becoming more mature, we expect that further tasks, addressing human activity analysis on higher levels, will become part of future CLEAR evaluations.

Acknowledgments

The work presented here was partly funded by the European Union (EU) under the integrated project CHIL, Computers in the Human Interaction Loop (Grant number IST-506909) and partial funding was also provided by the US Government VACE program.

VACE would additionally like to thank the following groups for the use of tools and resources developed under CLEAR 2006:

- The University of Maryland, for their VIPER video annotation tool
- Video Mining, for their video data annotation efforts
- The University of South Florida, for their evaluation scoring tool

The authors also thank Padmanabhan Soundararajan of the University of South Florida for his participation in developing the CLEAR metrics and his kind assistance in the use of the evaluation scoring tool.

Disclaimer: Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose they were presented.

Appendix A: Result Graphs for 3D Person Tracking

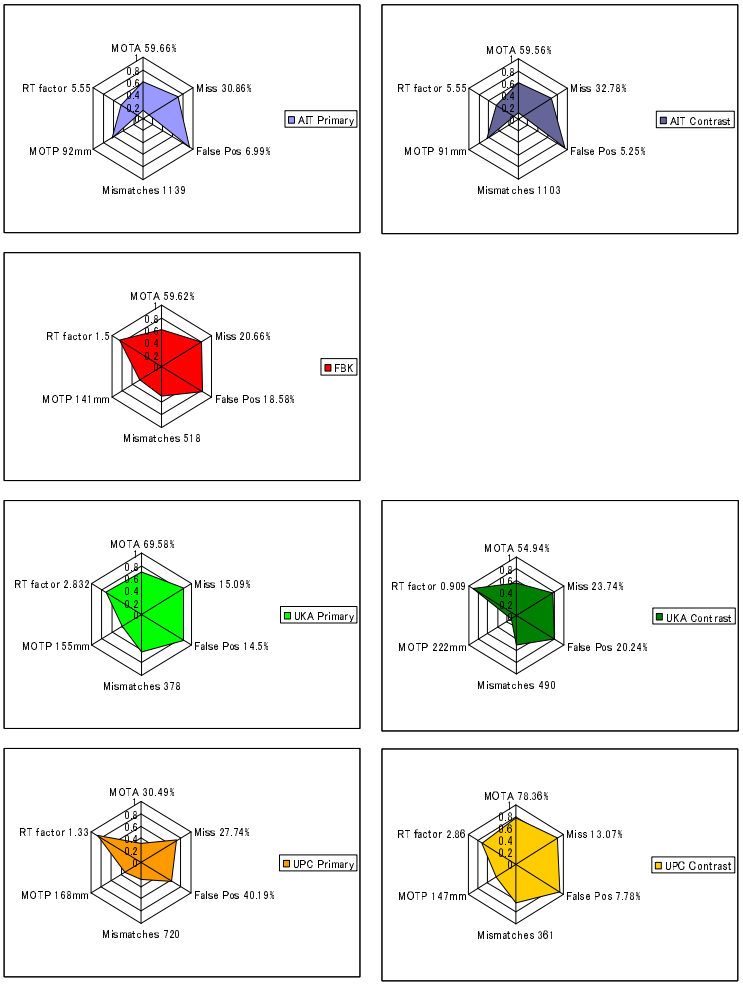


Fig. 22. 3D Person Tracking – Visual subtask: Radar Charts

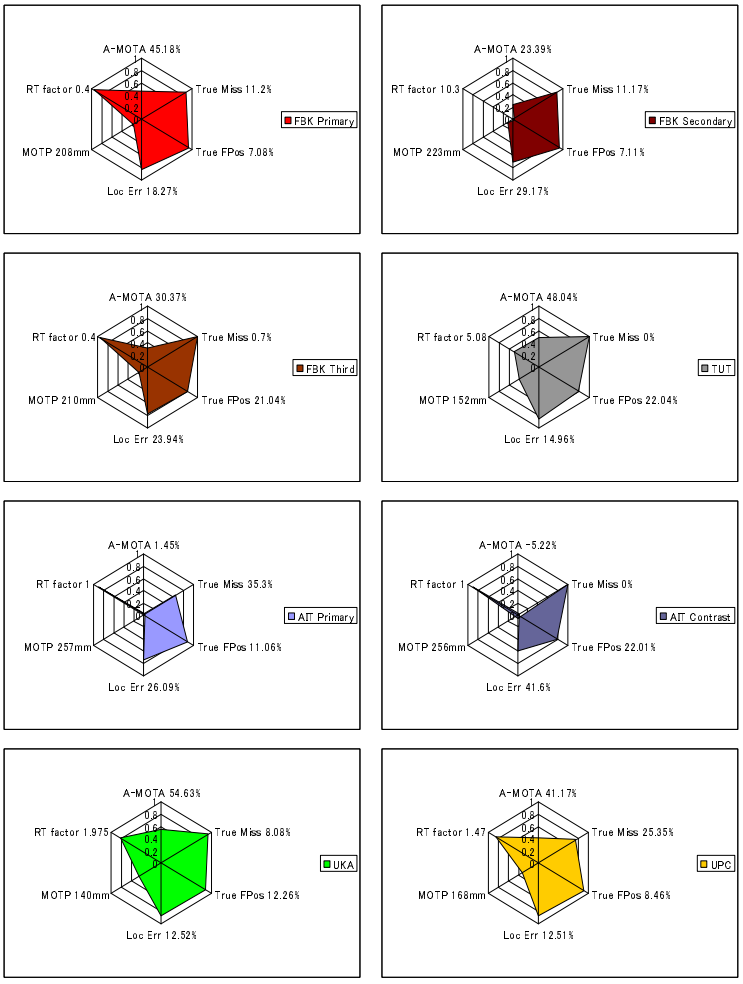


Fig. 23. 3D Person Tracking – Acoustic subtask: Radar Charts

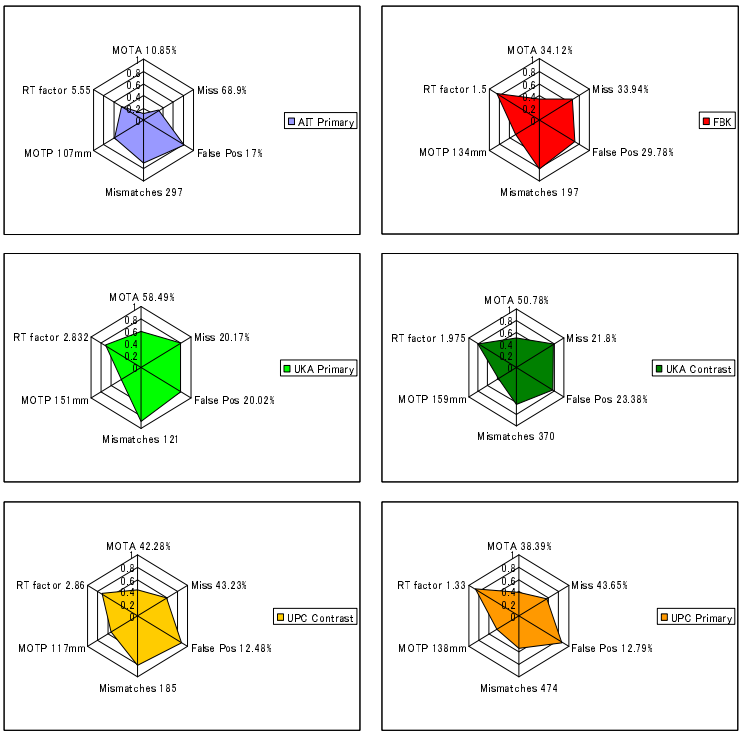


Fig. 24. 3D Person Tracking – Multimodal subtask: Radar Charts

Appendix B: Progress Charts for Person Identification

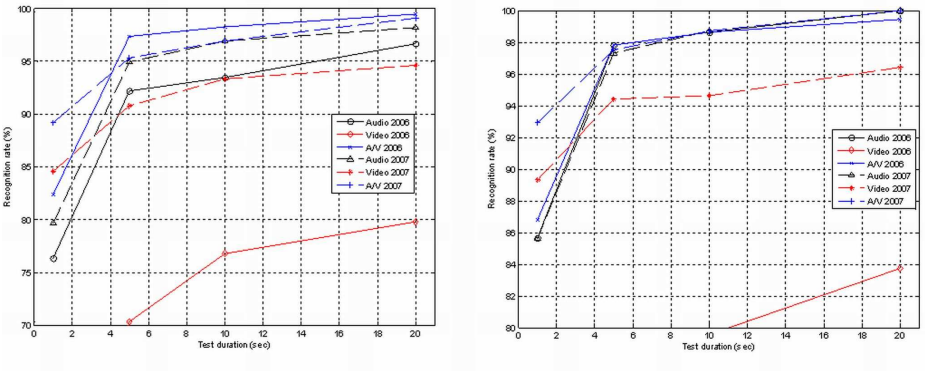


Fig. 25. Person Identification – Visual subtask: Progress Chart

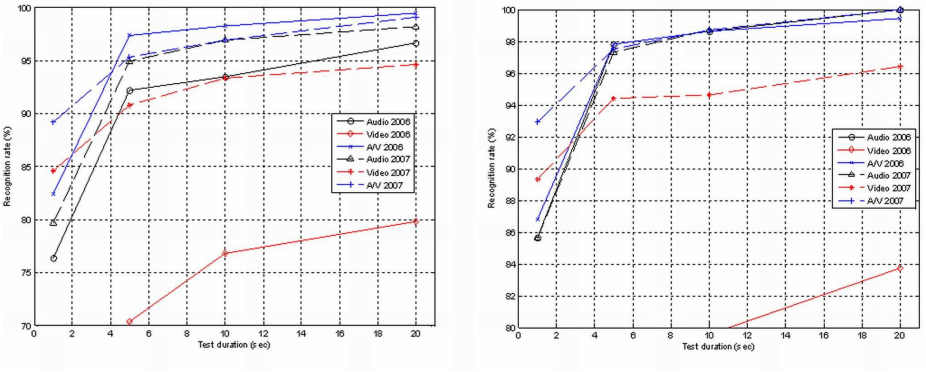


Fig. 26. Person Identification – Acoustic subtask: Progress Chart

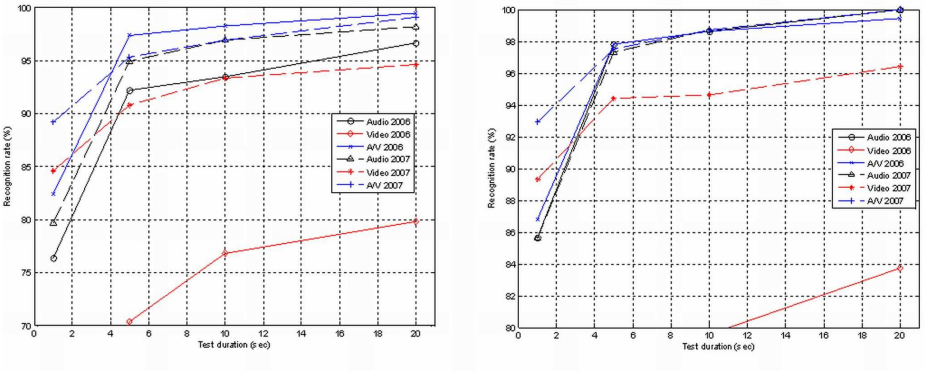


Fig. 27. Person Identification – Multimodal subtask: Progress Chart

References

1. CHIL - Computers In the Human Interaction Loop, http://chil.server.de.
2. AMI - Augmented Multiparty Interaction, http://www.amiproject.org.
3. VACE - Video Analysis and Content Extraction, https://control.nist.gov/dto/twiki/bin/view/Main/WebHome.
4. CALO - Cognitive Agent that Learns and Organizes, http://caloproject.sri.com/.
5. NIST Rich Transcription Meeting Recognition Evaluations, http://www.nist.gov/speech/tests/rt/rt2006/spring/.
6. The i-LIDS dataset, http://scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/video-based-detection-systems/i-lids/.
7. DARPA VIVID, http://www.vividevaluation.ri.cmu.edu/datasets/PETS2005/PkTest02/index.html.
8. CLEAR evaluation webpage, http://www.clear-evaluation.org.
9. ELRA/ELDA's Catalogue of Language Resources: http://catalog.elda.org/.
10. Sileye O. Ba and Jean Marc Odobez. Evaluation of head pose tracking algorithm in indoor environments. In *International Conference on Multimedia & Expo, ICME 2005*, Amsterdam, Netherlands, 2005.
11. Keni Bernardin, Alexander Elbs, and Rainer Stiefelhagen. Multiple object tracking performance metrics and evaluation in a smart room environment. *6th IEEE Int. Workshop on Visual Surveillance, VS 2006*, May 2006.
12. Keni Bernardin, Tobias Gehrig, and Rainer Stiefelhagen. Multi- and single view multiperson tracking for smart room environments. In *Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop*, Southampton, UK, 2006. Springer LNCS 4122.
13. Keni Bernardin, Tobias Gehrig, and Rainer Stiefelhagen. Multi-level particle filter fusion of features and cues for audio-visual person tracking. In *Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops*, Baltimore, MD, USA, 2007. Springer LNCS 4625.
14. Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT Metrics. *Submitted to the EURASIP Journal on Image and Video Processing, Special Issue on Video Tracking in Complex Scenes for Surveillance Applications.*

15. C. Canton-Ferrer, J. Salvador, J.R. Casas, and M.Pardas. Multi-person tracking strategies based on voxel analysis. In *Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops*, Baltimore, MD, USA, 2007. Springer LNCS 4625.
16. D. Doermann and D. Mihalcik. Tools and techniques for video performances evaluation. In *International Conference on Pattern Recognition*, pages 167–170, 2000.
17. Hazim Kemal Ekenel, Qin Jin, Mika Fischer, and Rainer Stiefelhagen. ISL Person Identification Systems in the CLEAR 2007 Evaluations. In *Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops*, Baltimore, MD, USA, 2007. Springer LNCS 4625.
18. V.Y. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, D. Doermann, and T. Drayer. Performance evaluation of object detection algorithms. In *International Conference on Pattern Recognition*, pages 965–969, 2002.
19. D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Christoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet. The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. In *Language Resources and Evaluation*, number 41 in Springer, 2007.
20. M.C. Nechyba, L. Brandy, and H. Schneiderman. Pittpatt face detection and tracking for the clear 2007 evaluation. In *Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops*, Baltimore, MD, USA, 2007. Springer LNCS 4625.
21. Harish Raju and Shubha Prasad. Annotation guidelines for video analysis and content extraction (VACE-II). In *CLEAR Evaluation Workshop*, 2006.
22. Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan. The CLEAR 2006 Evaluation. In *Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop*, Southampton, UK, 2006. Springer LNCS 4122.
23. Rainer Stiefelhagen and John Garofolo, editors. *Multimodal Technologies for Perception of Humans, First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR'06*. Number 4122 in Lecture Notes in Computer Science. Springer, 2007.