The CLEAR 2006 Evaluation

Rainer Stiefelhagen[1], Keni Bernardin[1], Rachel Bowers[2], John Garofolo[2], Djamel Mostefa[3], and Padmanabhan Soundararajan[4]

[1] Interactive Systems Lab, Universität Karlsruhe, 76131 Karlsruhe, Germany
{stiefel, keni}@ira.uka.de
[2] National Institute of Standards and Technology (NIST), Information Technology Lab - Information Access Division, Speech Group
{rachel.bowers, garofolo}@nist.gov
[3] Evaluations and Language Resources Distribution Agency (ELDA), Paris, France
mostefa@elda.org
[4] Computer Science and Engineering, University of South Florida, Tampa, FL, USA
psoundar@cse.usf.edu

Abstract. This paper is a summary of the first CLEAR evaluation on CLassification of Events, Activities and Relationships - which took place in early 2006 and concluded with a two day evaluation workshop in April 2006. CLEAR is an international effort to evaluate systems for the multimodal perception of people, their activities and interactions. It provides a new international evaluation framework for such technologies. It aims to support the definition of common evaluation tasks and metrics, to coordinate and leverage the production of necessary multimodal corpora and to provide a possibility for comparing different algorithms and approaches on common benchmarks, which will result in faster progress in the research community. This paper describes the evaluation tasks, including metrics and databases used, that were conducted in CLEAR 2006, and provides an overview of the results. The evaluation tasks in CLEAR 2006 included person tracking, face detection and tracking, person identification, head pose estimation, vehicle tracking as well as acoustic scene analysis. Overall, more than 20 subtasks were conducted, which included acoustic, visual and audio-visual analysis for many of the main tasks, as well as different data domains and evaluation conditions.

1 Introduction

CLassification of Events, Activities and Relationships (CLEAR) is an international effort to evaluate systems that are designed to analyze people's identities, activities, interactions and relationships in human-human interaction scenarios, as well as related scenarios. The first CLEAR evaluation has been conducted from around December 2005, when the first development data and scoring scripts were disseminated, until April 2006, when a two-day evaluation workshop took place in Southampton, UK, during which the evaluation results and system details of all participants were discussed.

1.1 Motivation

Many researchers, research labs and in particular a number of current major research projects worldwide – including the European projects CHIL, Computers in the Human Interaction Loop [1], and AMI, "Augmented Multi-party Interaction" [2], as well as the US programs VACE, "Video Analysis 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 evaluation standards for such technologies are missing. Until now, most researchers and research projects use their own different data sets, annotations, task definitions, metrics and evaluation procedures. As a consequence, comparability of the 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 forum and framework for such technologies, and to serve as a forum for the discussion and definition of related common benchmarks, including the definition of common metrics, tasks and evaluation procedures.

The outcomes for the research community that we expect from such a common evaluation forum are

- the definition of widely adopted common metrics and tasks
- a greater availability of resources by sharing the data collection and annotation burden
- provision of challenging multimodal data sets for the development of robust perceptual technologies
- comparability of systems and approaches and
- thus faster progress in developing better, more robust technology.

1.2 Background

The CLEAR 2006 evaluation has emerged out of the existing evaluation efforts of the European Integrated project CHIL, which has in previous years conducted a number of evaluations on multimodal perceptual technologies, including tasks such as person tracking and identification, head pose estimation, gesture recognition and acoustic event detection, as well as the technology evaluation efforts in the US VACE program, which conducted several similar evaluations in face, person and vehicle tracking. For CLEAR 2006, the technology evaluations of CHIL and VACE were combined for the first time, and the evaluations were also open to any site interested in participating.

In order to broaden the participation and discussion of evaluation tasks and metrics, representatives from other related projects and evaluation efforts (AMI[2], NIST RT evaluations[5], NIST People-ID evaluations, PETs[6], TrecVid[7], ETISEO[8]) were actively invited to participate in the preparation of the workshop as well as to present an overview about their related activities at the workshop.

1.3 Scope and Evaluation Tasks in 2006

The CLEAR 2006 evaluation and workshop was organized in conjunction with the National Institute of Standards and Technology (NIST) Rich Transcription (RT) 2006 evaluation [5]. 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. CLEAR 2006 and RT06 in particular shared some of their evaluation data sets, so that for example the speaker-localization results generated for CLEAR could be used for the far-field speech-to-text task in RT06. Also the evaluation deadlines of CLEAR and RT 2006 were harmonized so that this would be possible. This is an important first step towards developing a comprehensive multimedia evaluation program.

The evaluation tasks in CLEAR 2006 can be broken down into four categories:

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

These tasks and their various subtasks will be described in Section 3.

Due to the short time frame for preparing the joint technology evaluations in CLEAR, it was decided that the evaluations tasks that had already been defined in VACE and CLEAR, respectively, would be kept as they were, and thus were run independently in parallel, with their slightly differing annotations and on different data sets. As a consequence there were, for example, several 3D person tracking tasks (CHIL) as well as 2D person tracking tasks (VACE) in CLEAR 2006. As a first step of harmonizing evaluation tasks, the participants from CHIL and VACE had, however, agreed on common metrics for multiple object tracking (see section 3.3). The aim for upcoming evaluations is to further harmonize metrics and benchmarks.

1.4 Contributors

CLEAR 2006 would not have been possible without the help and effort of many people and institutions worldwide. CLEAR 2006 was supported by the projects CHIL [1] and VACE [3]. The organizers of CLEAR are the Interactive Systems Labs of the Universität Karlsruhe, Germany (UKA), and the US National Institute of Standards and Technology (NIST), with the support of contractors University of South Florida (USF) and VideoMining Inc. The participants and contributors to the CLEAR 2006 evaluations 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 Project PRIMA of the Institut National de Recherche en Informatique et en Automatique, Grenoble, France (INRIA), the Centro per la ricerca scientifica e tecnologica at the Instituto Trentino di Cultura, Trento, Italy (ITC-IRST), 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 (PPATT), the department of Electronic Engineering of the Queen Mary University of London, UK, (QMUL) the Institute of Signal Processing of the Technical University of Tampere, Finland, (TUT), the Beckman Institute for Advanced Science and Tech. at the University of Illinois Urbana Champaign, USA (UIUC) the Institute for Robotics and Intelligent Systems of the University of Southern California, USA, (USC).

UKA and ITC-IRST provided recordings of seminars (lectures), which were used for the 3D single person tracking tasks the face detection task and for person recognition. AIT, IBM and UPC provided several recordings of "interactive" seminars (basically small interactive meetings), which were used for the multi-person tracking tasks, for face detection, for the person identification tasks and for acoustic event detection. INRIA provided the Pointing'04 database for head pose detection. UKA provided 26 seminar recordings with head pose annotations for the lecturer, which data was used for the second head pose estimation task. UPC, ITC and CMU provided different databases with annotated acoustic events used for acoustic event classification. Visual and acoustic annotations of the CHIL seminar and interactive seminar data were mainly done by ELDA, in collaboration with UKA, CMU, AIT, IBM, ITC-irst and UPC. ELDA also packaged and distributed the data coming from CHIL. The data coming from VACE was derived from a single source for the surveillance data - i-LIDS. The meeting room data was a collection derived from data collected at CMU, University of Edinburgh (EDI), NIST, TNO, and Virginia Tech (VT). The discussion and definition of the invidual tasks and evaluation procedures were moderated by "task-leaders". The task-leaders coordinated all aspects surrounding the running of their given tasks. These were Keni Bernardin (UKA, 3D single- and multi-person tracking), Maurizio Omologo (ITC-irst, 3D acoustic single-person tracking), John Garofolo/Rachel Bowers (NIST, 2D Multi-person tracking tasks, VACE 2D face tracking, vehicle tracking), Hazim Ekenel (UKA, visual person identification), Djamel Mostefa (ELDA, acoustic identification), Aristodemos Pnevmatikakis (AIT, audio-visual identification), Ferran Marques and Ramon Morros (both UPC, CHIL 2D Face detection), Michael Voit (UKA, head pose estimation), Andrey Temko (UPC, acoustic event detection). The tasks leaders were also responsible for scoring the evaluation submissions, which in addition were also centrally scored by ELDA.

This paper aims at giving an overview of the CLEAR 2006 evaluation. In the remainder of this paper we will therefore give a brief overview of the data sets used (Section 2) and the various evaluation tasks (Section 3). In Section 4 we present an overview of the results and discuss some of the outcomes and potential implications for further evaluations.

Further details on the tasks definitions and data sets can be found in the CHIL and VACE evaluation plans [9], [10] and on the CLEAR webpage [11].

2 Datasets Used In CLEAR 2006

2.1 The CHIL Seminar Database

A large mutimodal database has been collected by the CHIL consortium for the CLEAR 2006 evaluation, consisting of audiovisual recordings of regular lecture-like seminars and interactive small working group seminars. For some of the interactive seminars, scripts were used in order to elicit certain activities (e.g., opening doors, taking a coffee break), which were to be automatically detected in one or more evaluation tasks (e.g., acoustic event detection).

The database contains audio and video recordings segments from 47 seminars recorded at the following sites:

- AIT, Athens, Greece,
- IBM, New-York, USA,
- ITC-IRST, Trento, Italy,
- UKA, Karlsruhe, Germany,
- UPC, Barcelona, Spain.

These seminars were given by students and lecturers of each institution or by invited speakers on topics concerning technologies involved in the CHIL project, such as speech recognition, audio source localization, audio scene analysis, video scene analysis, person identification and tracking, etc. The language is English spoken by mostly non native speakers. A detailled description of the CLEAR database can be found in [9].

Non-Interactive Seminars versus Interactive Seminars

- **Non-interactive seminars** were provided by ITC-IRST and UKA. These seminars consist of lecture-like presentations in a small seminar room. One presenter is talking in front of an audience of 10 to 20 people. In these recordings, the focus is mainly on the presenter (he's the only one wearing a close talking microphone, moving, ...). As a consequence, only the presenter has been annotated for the different tasks using this database (tracking, identification, ...). An example of non-interactive seminars is given by the first two pictures in Fig. 1.
- **Interactive seminars** were recorded by AIT, IBM and UPC. The recording room is a meeting room and the audience is made up of only 3 to 5 people. The attendees are sitting around a table and are wearing close-talking microphones, just as the presenter. There is a higher degree of interaction between the presenter and the audience. During and after the presentation, there are questions from the attendees with answers from the presenter. Moreover there is also activity in terms of people entering or leaving the room, opening and closing the door. AIT and UPC seminars have been scripted in order

to elicit certain activities (e.g., opening doors, taking a coffee break). These activities were to be automatically detected in one or more evaluation tasks (e.g., acoustic event detection). The last 3 pictures of Fig. 1 show examples of interactive seminars.

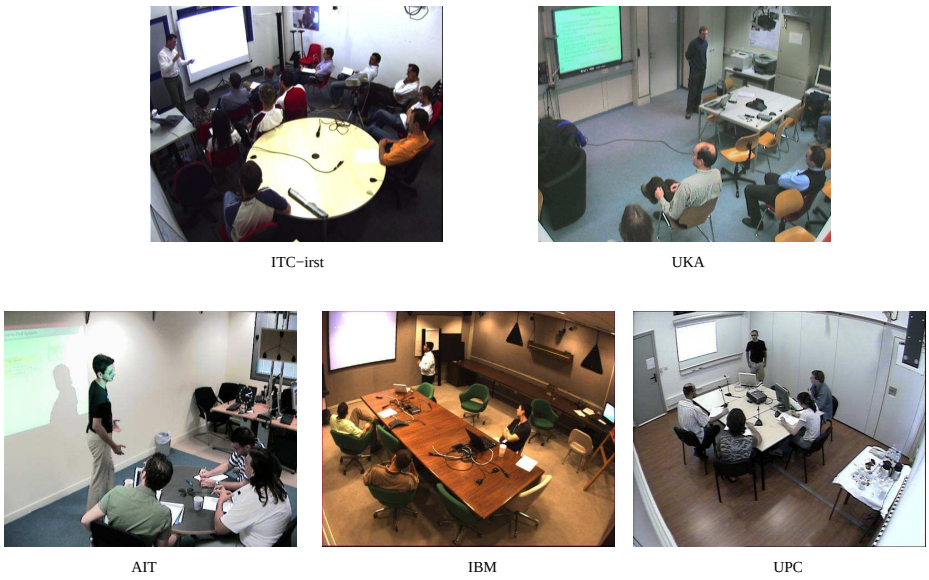


Fig. 1. Scenes from non-interactive and interactive seminars

Data Description

- **Raw data:** Each seminar is composed of synchronized audio and video streams. The video streams consist of 4 to 5 JPEG sequences recorded at 15 to 30 frames per second by 4 fixed corner and a ceiling camera. Acoustic sounds are recorded using a great variety of sensors. High quality close talking microphones are used by every participant in interactive seminars and by the presenter only in non-interactive seminars. In addition, omnidirectional table top microphones and several T-shaped arrays are used for far-field recordings. All these microphones are synchronised at the sample level by a dedicated sound card. Moreover, far field recordings are also captured by a NIST markIII 64 channel microphone array. Fig. 2 shows an example of a recording room setup.
- **Audio transcription:** For a single audiovisual data element (a seminar), two transcriptions were produced. The first one is the speaker transcription which contains the speech utterances of all intervening speakers, including human-generated noises accompanying speech. This is done by transcribing

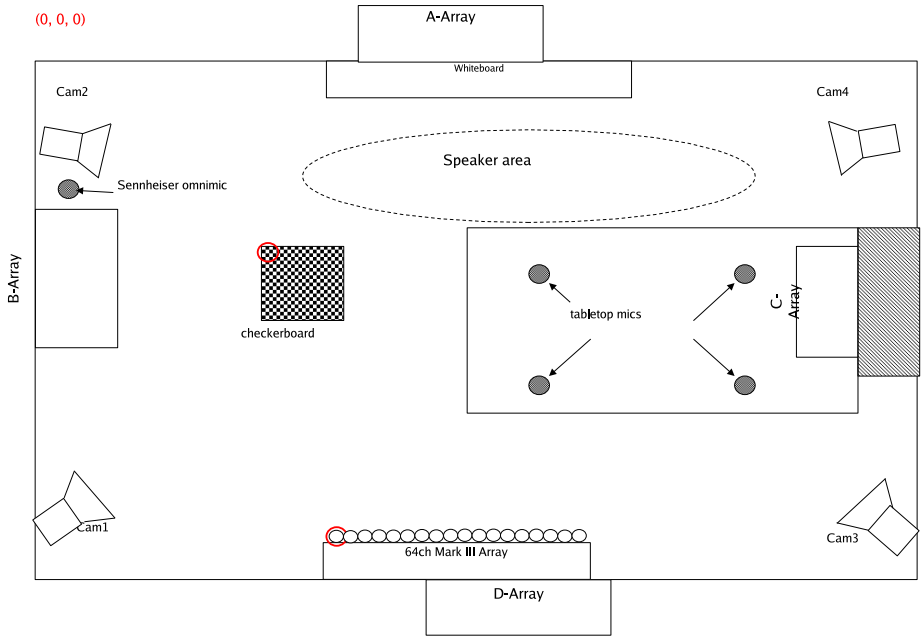


Fig. 2. Example recording room setup (source: UKA)

the close-talking microphone recording of the main speaker. The second one is the environment transcription which contains all noises not produced by the speaker(s). Environment transcriptions are realized on far-field recordings. All environmental noises (human and non-human) and all speaker utterances are transcribed. Both transcriptions were produced with Transcriber [12] and are in native XML format.

Acoustic event noises annotations are made on the far field recordings with AGTK annotation tool [13]. This tool enables the annotations of overlapping noises in a simple XML format.

– **Video labels:** Video annotations were realized using an *in house* developed tool. This tool allows to sequentially display video frames to be annotated, for the 4 corner cameras. On each displayed picture, the annotator was to click on the head centroid (the estimated centre of the head), the left eye, right eye, and nose bridge of the annotated person. In addition to these four points, a face rectangle bounding box was used to delimit the person's face. These annotations were done on the lecturer for non-interactive seminars and on each participant for interactive seminars. The 2D coordinates within the camera planes were interpolated among all cameras in order to compute the real "ground truth" location of the speaker within the room. Fig. 3 shows an example of video labeling. Displayed are the head centroid, the left eye, the nose bridge, the right eye and the face bounding box.

Fig. 3. Example of video annotations

Development Data The development data is made of segments used in previous CHIL evaluations and of new seminars provided by new recording sites. 17 seminars from UKA used in the first CHIL evaluation and the NIST Rich Transcription 2005 were used as development data for CLEAR 2006. For each UKA seminar, two segments of 5min each were used. The first one is taken from the talk of the presenter and the other one is selected from the question-answering session at the end of the talk. The second segment usually contains more spontaneous speech and involves more speakers than the first one. In addition to the UKA seminars, around 1 h of data coming from AIT, IBM, ITC-IRST and UPC was added to the development set. The first 15min of the first seminar recorded by each site was used. In total, the development set duration is 204min with 80 % non-interactive seminars and 20 % interactive seminars. This imbalance is mainly due to the fact that only 3 interactive seminars were recorded and labeled at the time the development set was released. Table 1 gives an overview of the composition of the development set.

Evaluation Data As for the development set, the evaluation set is composed of segments from interactive and non-interactive seminars. Due to the availability of more data recorded at each site, the evaluation data is much more balanced between interactive and non-interactive seminars. The total duration of the CLEAR'06 evaluation set is 190min, of which 14 seminars, representing 68 %, are non-interactive and 12 seminars, representing 32 %, are interactive. Table 2 gives an overview of the composition of the evaluation set.

Site	Type	Number	Total length (in minutes)
ITC-irst	non interactive	1	15
UKA	non interactive	17	148
AIT	interactive	1	13
IBM	interactive	1	15
UPC	interactive	1	13
TOTAL		21	204

Table 1. The CLEAR'06 development set

Site	Type	Number	Total length (in minutes)
ITC-irst	non interactive	2	10
UKA	non interactive	12	120
AIT	interactive	4	20
IBM	interactive	4	20
UPC	interactive	4	20
TOTAL		26	190

Table 2. The CLEAR'06 evaluation set

2.2 VACE Related Databases

For tasks coordinated and led by the VACE community, the evaluations were conducted using two main databases, the Multi-Site Meetings and the i-LIDS Surveillance data (see Table 3).

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

Table 3. The VACE related databases

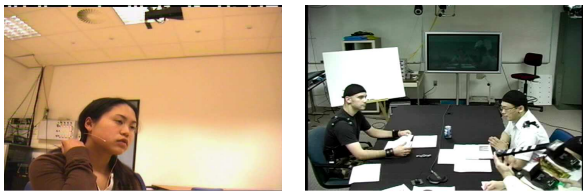
All the raw data is in MPEG-2 format with either 12 or 15 I-frame rate encoding. The annotations are specifically done using the ViPER tool developed by UMD by VideoMining.

The Multi-Site Meetings are composed of datasets from different sites, samples of which are shown in Fig 4.

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. 4. Scenes from Multi–Site Meetings

Each site has their own independent camera setup, different illuminations, viewpoints, people and topics in the meetings. Most of these datasets also figured High-Definition (HD) recordings but were subsequently formatted to MPEG-2 standard for evaluation purposes. Fig. 2.2 shows an example of the recording room setup for the NIST meeting data collection laboratory. The room has seven HD cameras, the table has one quad microphone and three omni-directional microphones. Each meeting room participant is equipped with one wireless lapel mic and one headmounted mic. The room is equipped with both traditional and electronic whiteboards as well as a projector for presentations. All cameras are synchronized using the *NIST Smart Data Flow* synchronization software. For more details on the individual room setup for all the sites, please refer to [14]. Specific annotation or labeling details can be found in Section 3.4.

i-LIDS is a video surveillance dataset that has been developed by the United Kingdom Government as a "benchmark for video-based detection systems"[15]. VACE has obtained permission to use this data for their person and vehicle detection and tracking evaluations. The dataset for the CLEAR evaluation includes outdoor views of roadways with walking paths. Though night scenes were available for the data training test set the actual evaluation was limited to day scenes. The dataset was composed of two different scenes with various shapes

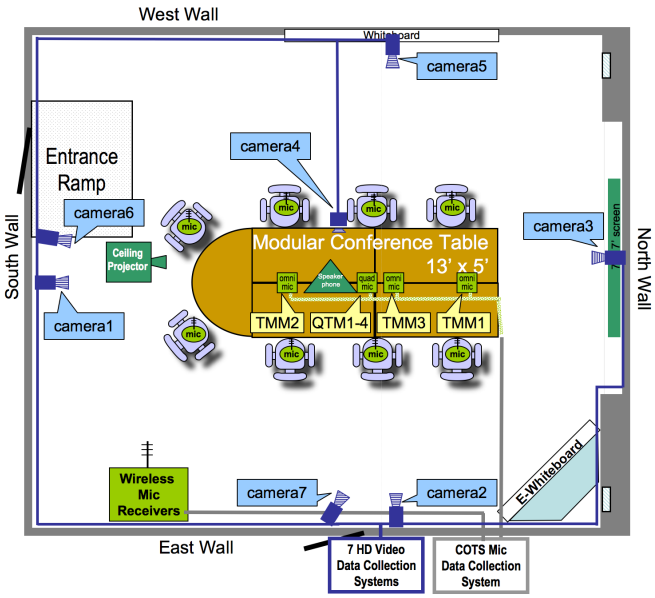


Fig. 5. Example recording room setup (source: NIST)

and sizes of vehicles and people, making for a challenging evaluation task. Specific annotation/labeling details for a person or vehicle in the video can be found in Section 3.5 and 3.6.

2.3 Other Databases

In addition to the two main databases mentioned above, specific datasets attuned to the head pose estimation and the acoustic scene analysis tasks were also used in the CLEAR'06 evaluation. These databases will be explained in more detail together with the corresponding task descriptions in section 3.

3 CLEAR Tasks and Metrics

This section gives an overview of the different tasks evaluated in the CLEAR'06 evaluation. Three main databases were evaluated on: The first is a series of recordings made in CHIL smartrooms, using a wide range of synchronized sensors, and useful for multimodal analysis in indoor environments. The second, originally used for the VACE tasks, comprises a set of single camera surveillance videos used for visual outdoor detection and tracking scenarios. The third is a set of multi-camera meeting room recordings used mainly for face detection tasks (see Section 2 for details on the used data sets).

The CLEAR tasks can be broken down into four main categories: tracking tasks, identification tasks, head pose estimation tasks and acoustic scene analysis tasks. Table 4 shows the different CLEAR tasks.

Table 4. Overview of CLEAR'06 tasks

Task name	Organizer	Database
Tracking		
3D Single Person Tracking (A,V,AV)	CHIL	Non-interactive Seminars
3D Multi-Person Tracking (A,V,AV)	CHIL	Interactive Seminars
2D Face Detection & Tracking (V)	CHIL/VACE	All Seminars/Multi-Site Meetings
2D Person Tracking (V)	VACE	Surveillance Data
Vehicle Tracking (V)	VACE	Surveillance Data
Person Identification (A,V,AV)	CHIL	All Seminars
Head Pose Estimation (V)	CHIL	Seminars[1], Pointing04 DB
Acoustic Scene Analysis		
Acoustic Event Detection	CHIL	Isolated Events, UPC Seminars
Acoustic Environment Classification	CHIL	AATEPS corpus

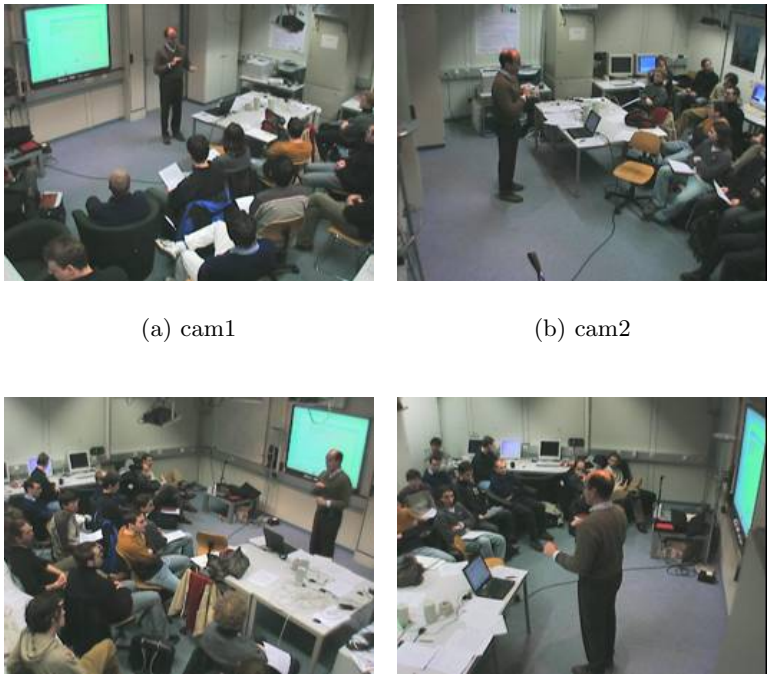
3.1 3D Single Person Tracking

One of the main tasks in the 2006 CLEAR evaluation, in terms of participation, was the 3D single person tracking task. The task definition was inherited from previous evaluations made in the CHIL project. The objective was to track a presenter giving a talk in front of an audience in a small seminar room (see Fig. 6). The database to be evaluated on consisted of recordings made at two CHIL sites, UKA and ITC-IRST, with different room sizes and layouts, but with a common sensor setup. The video streams from the four corner cameras of the room and the audio streams from the four T-shaped arrays and the MarkIII microphone array were available to do the tracking. In addition to the raw data, only the calibration information for the cameras and the locations of the microphones could be used. No explicit knowledge about the initial position of the presenter, the location of the whiteboard, of the room doors, of the audience, etc. was provided. However, participants were able to tune their systems on data from a separate development set, showing different seminars recorded in the same rooms.

Whereas in earlier CHIL evaluations the visual and acoustic tracking tasks were evaluated separately, here, for the first time, it was possible to compare the performance of trackers from both modalities, through the use of common datasets and metrics. A multimodal tracking task was also newly introduced, where the combined audio-visual streams could be used.

As opposed to the CLEAR 2D person tracking task, or similar tasks from other evaluations, such as e.g. PETS [6], where the objective is typically to track the positon or bounding box of moving objects in 2D images, the objective here was to track the actual location of a person in a room coordinate frame (typically with the origin at one of the bottom corners of the room and the

[1] For this task, a number of non-interactive seminars, which were recorded in 2004, were annotated and used. These seminars, however, were not part of the dataset used for the tracking and identification tasks.



(a) cam1 (b) cam2

(c) cam3 (d) cam4

Fig. 6. Example scene from a UKA seminar recording

axes parallel to the walls). This is possible because the CHIL seminar recordings offer 4 overlapping, synchronized and calibrated camera views, allowing for video triangulation, and at least 4 sets of microphone arrays, allowing for precise sound source localization. As it was not intended to track specific body regions, such as the head or the feet, a person's position was defined as his or her x,y-coordinates on the ground plane. This proved a reasonable approximation usable for both standing and sitting persons and allowing to evaluate all types of trackers across modalities.

The ground truth person locations for error calculations were obtained from manual annotation of the video streams. In each of the four corner camera streams, the presenter's head centroid was marked. Using calibration information, these 2D positions were triangulated to obtain the 3D head position, which was then projected to the ground to yield the person's reference position. If the presenter's head was not visible in at least 2 camera views, the frame was left unmarked. Note that due to this annotation scheme, slight errors could be introduced in the labeled positions, for ex. when the presenter bends forward to change his presentation slides. Nevertheless, the annotation of the head centroid

was found to be the easiest, most precise, and least error prone for this kind of task. To further reduce the cost of annotations, it was chosen to label video frames only in intervals of 1s (i.e. every 15, 25, or 30 frames, depending on the actual framerate of the recording). Tracking systems could be run using all video frames and audio samples, but were to be evaluated only on labeled frames. This helped reduce the cost of evaluation dramatically with only little impact on the accuracy of results.

For the acoustic tracking task, an additional restriction was made. The evaluation of tracking performance was to be decoupled from that of speech detection and segmentation. That is why acoustic tracking systems, although run continuously on all data, were evaluated only on segments of non-overlapping speech where the presenter is speaking and no greater source of noise (e.g. clapping) is audible. These segments were defined by manual annotation.

For the multimodal tracking task, two separate conditions were defined, to offer better comparability to the visual and acoustic tracking tasks. In condition A, multimodal tracking systems were evaluated on segments of non-ovelapping speech only, just as in the acoustic task. This could serve to measure what increase in precision the addition of the visual modality would bring to acoustic tracking, given an already accurate speech segmentation. In condition B, they were evaluated on all labeled time frames, as in the visual task, regardless if the speaker was active or not. This served to measure the enhancement brought by the fusion of modalities in the general case.

The metrics used to evaluate single person tracking performance are explained in section 3.3 and the results for all subtasks and conditions summed up in section 4.1.

3.2 3D Multi-Person Tracking

As opposed to the 3D single person tracking task, where only the main speaker had to be accounted for, ignoring the audience, the objective in the 3D multi-person tracking task is to simultaneously track all the participants in a small interactive meeting. To this effect, a set of recordings was made at three CHIL sites, IBM, UPC, and AIT, with a slightly modified scenario involving 4 to 6 people (see Fig. 7). While there is still a main speaker presenting a topic to the other participants, there is much more interaction as participants take turns asking questions or move around while entering the room or during coffee breaks. These recordings proved quite challenging compared to the single person tracking task due to the number of persons to track, the relatively small size of the meeting rooms and the high variability of the scenario.

The same sensor setup as for single person tracking was used. Additionally, video streams from a ceiling mounted panoramic camera were available. The annotations were also made in the same manner, with the exception that for each time frame, the head centroids of all participants were labeled.

In contrast to single person tracking, the definition of the multi-person tracking task is quite dependent on the chosen modality.

(a) cam1 (b) cam2 (c) cam3

(d) cam4 (e) cam5

Fig. 7. Example scene from a UPC interactive seminar recording

For visual tracking, the objective is to track every participant of the interactive seminar for all labeled frames in the sequence.

For the acoustic tracking task, on the other hand, the objective was to track only one person at a time, namely the active speaker, because tracking during overlapping speech was considered to be too difficult at this time. While in single person tracking, this was limited to the presenter, here it could also be anyone in the audience. Systems are evaluated only on manually defined segments of non-overlapping speech with no considerable noise sources.

For multimodal tracking, again, two conditions were introduced: In condition A, the objective is to audio-visually track only one person at each point in time, namely the active speaker. This is best comparable to the acoustic tracking task, and is evaluated only on manually defined active speech segments. In condition B, the goal is to track all persons in all labeled time frames using streams from both audio and visual modalities.

Evaluating the performance of systems for tracking multiple persons, and allowing for comparative results across modalities and tasks required the definition of a specialized set of metrics. These same metrics are also used in single person tracking, and in modified form in most other tracking tasks. They are explained in detail in section 3.3. The results for the 3D multi-person tracking task are summarized in section 4.2.

3.3 Multiple Object Tracking Metrics

Defining measures to express all of the important characteristics of a system for continuous tracking of multiple objects is not a straightforward task. Various measures, all with strengths and weaknesses, currently exist and there is no consensus in the tracking community on the best set to use. For the CLEAR workshop, a small expressive set of metrics was proposed. In the following, these metrics are briefly introduced and a systematic procedure for their calculation is shown. A more detailed discussion of the metrics can be found in [16].

The MOT Precision and Accuracy Metrics For the design of the CLEAR multiple object (person) tracking metrics, the following criteria were followed:

- They should allow to judge a tracker's precision in determining exact object locations.
- They should reflect its ability to consistently track object configurations through time, i.e. to correctly trace object trajectories, producing exactly one trajectory per object (see Fig. 8).

Additionally, we expect useful metrics

- to have as few free parameters, adjustable thresholds, etc, as possible to help make evaluations straightforward and keep results comparable.
- to be clear, easily understandable and behave according to human intuition, especially in the occurence of multiple errors of different types or of uneven repartition of errors throughout the sequence.
- to be general enough to allow comparison of most types of trackers (2D, 3D trackers, acoustic or visual trackers, etc).
- to be few in number and yet expressive, so they may be used e.g. in large evaluations where many systems are being compared.

Based on the above criteria, we define a procedure for systematic and objective evaluation of a tracker's characteristics. Assuming that for every time frame t a multiple object tracker outputs a set of hypotheses $\{h_1 \dots h_m\}$ for a set of visible objects $\{o_1 \dots o_n\}$, we define the procedure to evaluate its performance as follows:

Let the correspondence between an object o_i and a hypothesis h_j be valid only if their distance $dist_{i,j}$ does not exceed a certain threshold T (for CLEAR'06, T was set to 500mm), and let $M_t = \{(o_i, h_j)\}$ be a dynamic mapping of object-hypothesis pairs.

Let $M_0 = \{\}$. For every time frame t,

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

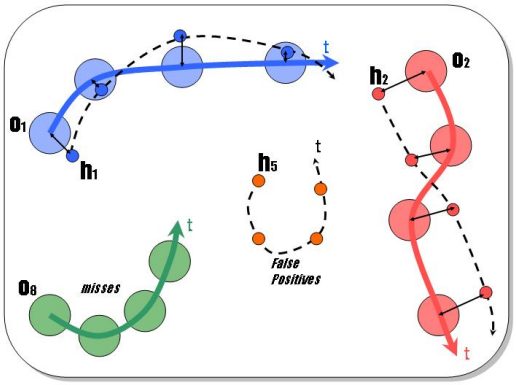


Fig. 8. Matching multiple object tracks to reference annotations

2. For all objects 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 object o_i and its corresponding hypothesis.
4. All remaining hypotheses are considered false positives. Similarly, all remaining objects 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 objects 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.

Based on the matching strategy described above, two very intuitive metrics can be defined: The Multiple Object Tracking Precision ($MOTP$), which shows the tracker's ability to estimate precise object positions, and the Multiple Object Tracking Accuracy ($MOTA$), which expresses its performance at estimating the number of objects, and at keeping consistent trajectories:

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

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

The $MOTA$ can be seen as composed of 3 error ratios:

$$\overline{m} = \frac{\sum_t m_t}{\sum_t g_t}, \quad \overline{fp} = \frac{\sum_t fp_t}{\sum_t g_t}, \quad \overline{mme} = \frac{\sum_t mme_t}{\sum_t g_t},$$

the ratio of misses, false positives and mismatches in the sequence, computed over the total number of objects present in all frames.

For the current run of CLEAR evaluations, it was decided that for acoustic tracking, it was not required to detect speaker change or to track speaker identities through time. Therefore, the measurement of identity mismatches is not meaningful for these systems, and an separate measure, the $A - MOTA$ is computed, by ignoring mismatch errors in the global error computation:

$$A - MOTA = 1 - \frac{\sum_t (m_t + fp_t)}{\sum_t g_t} \tag{3}$$

The above described $MOTP$ and $MOTA$ metrics were used in slightly modified form throughout the CLEAR tracking tasks and proved very useful for large scale comparisons of tracker performance across tasks and modalities.

Multiple Object Detection Precision and Accuracy In contrast to the point-wise distance metric described above, for the **Multiple Object Detection Precision (MODP)** the spatial overlap information between the ground truth and the system output is used to compute an Overlap Ratio as defined in Eq 4.

Here, the notation $G_i^{(t)}$ denotes the i^{th} ground truth object in the t^{th} frame and D_i^t denotes the detected object for G_i^t.

$$\text{Overlap Ratio} = \sum_{i=1}^{N_{mapped}^t} \frac{|G_i^{(t)} \bigcap D_i^{(t)}|}{|G_i^{(t)} \bigcup D_i^{(t)}|} \tag{4}$$

A threshold of 0.2 for the spatial overlap is used, primarily to compute the misses and false alarms (required while computing the MODA measure).

Using the assignment sets, the Multiple Object Detection Precision (MODP) is computed for each frame t as:

$$MODP(t) = \frac{(Overlap\ Ratio)}{N_{mapped}^t} \tag{5}$$

where, N_{mapped}^t is the number of mapped object sets in frame t. This gives us the localization precision of objects in any given frame and the measure can also be normalized by taking into account the total number of relevant evaluation frames. If $N_{mapped}^t = 0$, then the MODP is forced to a zero value.

$$N - MODP = \frac{\sum_{t=1}^{N_{frames}} MODP(t)}{N_{frames}} \tag{6}$$

The thresholded approach for the Overlap Ratio is meant to minimize the importance of the spatial accuracy. The N-MODP hence gives the localization precision for the entire sequence.

The **Multiple Object Detection Accuracy (MODA)** serves to assess the *accuracy* aspect of system performance. Here, only the missed counts and false alarm counts are used. Assuming that in each frame t, the number of misses are indicated by m_t and the number of false positives indicated by fp_t, the Multiple Object Detection Precision (MODA) can be computed as:

$$MODA(t) = 1 - \frac{c_m(m_t) + c_f(fp_t)}{N_G^t} \tag{7}$$

where, c_m and c_f are the cost functions for the missed detects and false alarm penalties. These cost functions are used as weights and can be varied based on the application at hand. If misses are more critical than false alarms, c_m can be increased and c_f reduced. N_G^t is the number of ground truth objects in the t^{th} frame.

The computation of the N-MODA, the normalized MODA for the entire sequence, is made as:

$$N - MODA = 1 - \frac{\sum_{i=1}^{N_{frames}} (c_m(m_i) + c_f(fp_i))}{\sum_{i=1}^{N_{frames}} N_G^i} \tag{8}$$

Differences in the VACE Detection and Tracking Metrics In November 2005, the evaluation teams from the CHIL and VACE projects both had their own sets of individual metrics. It was decided that in order to harmonize the CLEAR evaluation tasks, the metrics also have to be harmonized. In the CHIL Project, the tracking metrics used were:

- MOTP (Multiple Object Tracking Precision)
- MOTA (Multiple Object Tracking Accuracy)

On the other hand, the VACE side used the following detection and tracking metrics:

- SFDA (Sequence Frame Detection Accuracy) for Detection
- ATA (Average Tracking Accuracy) for Tracking

and a whole set of diagnostic metrics to measure individual components of the performance.

The key differences between the MODP/A and MOTP/A metrics, used in VACE-related tasks, and the standard MOTP/A used in CHIL-related tasks are:

- The metrics use the *spatial* component instead of the *distance*. We believe that for this evaluation we can keep this additional dimensionality.

– The *mapping* differs as in we use an Hungarian matching to map ground truth and system output boxes and this again uses the *spatial* component (as in spatial overlap between two objects). The idea is to maximize the metric score based on these individual components.

3.4 2D Face Detection and Tracking

The goal of this evaluation task was to measure the quality and accuracy of face detection techniques, both for meeting and for lecture scenarios. As opposed to the person tracking tasks, the objective here was not to estimate the trajectories of faces in real world coordinates, but rather to correctly detect as many faces as possible within the separate camera views. To this effect, no triangulation or 3D computation between views and no continuous tracking were required.

The main difficulty - and at the same time the scientific contribution - of this task stems from the nature of the database itself. In the CLEAR seminar and meeting databases, faces sizes are extremely small, in some cases down to (10x10) pixels, faces are rarely oriented towards cameras, lighting conditions are extremely difficult and faces are often partly occluded, making standard skin color segmentation or template matching techniques inapplicable. This drives the development of new techniques, which can handle very difficult data recorded under realistic wide camera view conditions. As in person tracking tasks, for the lecture scenario, only the presenter's face was to be found, whereas for interactive seminar and meeting scenarios, all faces had to be detected (see Fig. 9).

(a) UKA seminar (b) AIT interactive seminar

Fig. 9. Scenes from the Face Detection & Tracking database

A correct face detection should deliver not only the position of the face in the image, but also its extension, as this information can be valuable for subsequent identification or pose estimation processes. The ouput of face detection systems are therefore the bounding boxes of detected faces, which are compared to manual annotations. The guidelines for annotating the face bounding boxes differed

very slightly for the CHIL and VACE databases, resulting in somewhat larger face boxes in the CHIL data. Also, the criteria for considering a face as visible differed. Whereas in the VACE data it depended on the visibility of at least one eye, the nose, and part of the mouth, in the CHIL data, only visibility of at least one eye or the nose bridge was necessary. For future evaluations, it is planned to harmonize the annotation guidelines, to produce more uniform databases. As for the person tracking task, a face label was created only for every second of video.

To evaluate the performance of face detection and tracking algorithms, five measures were used: The percentage of correctly detected faces, wrong detections, and non-detected (missing) faces, the mean weighted error (in pixels) of the estimated face center, and the mean (face) extension accuracy.

For a correctly detected face in a frame i, the mean weighted error is defined as:

$$we_i = \frac{\left\| C_i^d - C_i^l \right\|_2}{R_i}$$

with C_i^d and C_i^l, the centers of the detected and labeled faces respectively, and R_i the face size, calculated as the average of the vertical and horizontal face bounding box lengths.

The mean extension accuracy is defined as:

$$\frac{A((BB^l \cup BB^d) - (BB^l \cap BB^d))}{A(BB^l)}$$

the ratio of the area $A(.)$ of the symmetric difference of the detected and labeled bounding boxes BB^d and BB^l with respect to the labeled bounding box BB^l.

The resulting errors are averaged over all faces in all frames. The results of the face detection and tracking task, evaluated on the CHIL recording database, are presented in section 4.3.

In the VACE Multi-Site Meeting database, the face is marked horizontally bound to the extent of the eyes and vertically bound from just above the eyes to the chin. The face must have *at–least* one eye, part of the nose and lips seen to be annotated. For specific annotation guidelines, please refer to [17]. The MODA/MODP metrics for detection and MOTA/MOTP metrics for tracking are used.

3.5 2D Person Detection and Tracking

The goal of the person detection task is to detect persons in a particular frame, while for the tracking task it is to track persons in the entire 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

Fig. 10. Sample annotation for a person in the Surveillance domain.

annotation details about how a person is marked are given in the annotation guidelines document [17].

Fig 10 shows a sample person annotation. When at least 25 % of a person is visible, the person is annotated. Each person is marked with a bounding box and each box has a rich set of attributes to enable sub–scoring if needed.

For formal evaluations though, the simplest setting is used: the person must be clearly visible (should not be occluded by any other object, e.g. being occluded by the branches of the tree. If a person walks behind a bigger object the annotations are stopped temporarily until the person is visible again. Depending on how long it takes for this person to re-appear the objectID is maintained accordingly. The specific guidelines can be found in [17].

The metrics used are the MODA/MODP and the MOTA/MOTP.

3.6 Vehicle Tracking

The goal of the moving vehicle task is to track any moving vehicle in a given clip. During annotations, only vehicles that have *moved* at any time during the clip are marked. Vehicles which are completely stationary are not marked. Vehicles are annotated at the first frame where they move. For specific details about the annotations please refer to the annotation guidelines document [17].

For a vehicle to be annotated, at least 25 % of the vehicle must be visible, and it is marked with a bounding box. Each box has a rich set of attributes, essentially recording if the vehicle is currently moving and whether it is occluded (a vehicle is marked as occluded if it is more than 50 % occluded). Fig 11 shows a sample vehicle annotation.

For this evaluation, the simplest setting was used: the vehicle has to be moving and must be clearly visible (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 particular regions are marked accordingly and are not evaluated. Also, since this is purely a tracking task, the metrics used here are the MOTA and MOTP.

Fig. 11. Sample from the moving vehicle tracking in Surveillance domain

3.7 Person Identification

In a smart meeting or lecture room environment, where many sensors and perceptual components cooperate to provide rich information about room activities, the tracking algorithms presented in the previous sections can serve as building blocks, providing necessary person locations, aligned faces, or localized speech segments for subsequent identification processes. The goal of the CLEAR person identification task is to measure what identification accuracies can be reached, and how fast they can be reached, using only far-field microphones and cameras, assuming person locations are already well known (see Fig. 12).

For this purpose, in addition to the head centers and the face bounding boxes, three additional marks have been annotated in the video images: The positions of the left and right eye and that of the nose bridge. These labels serve to achieve an exact alignment and cropping of face images necessary for face identification routines, clearly decoupling the identification task from the detection and tracking task. While all other features were marked for every second of video, the eye labels were produced every 200 ms, for better precision.

As for the face detection task, one of the big challenges - and the novelty - of the CLEAR visual identification task comes from the database itself. The seminar videos contain extremely low resolution faces, down to (10x10) pixels with eye distances ranging from 4 to 16 pixels, which are very difficult to detect with

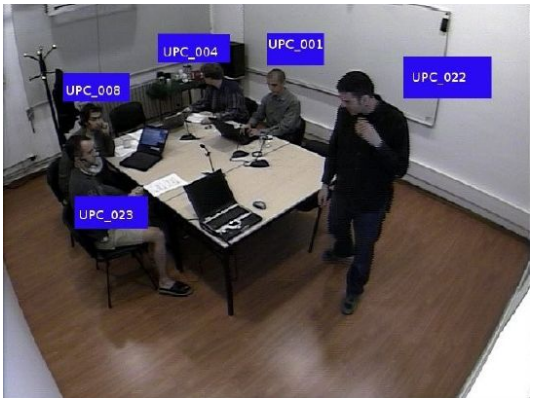


Fig. 12. Sample from the CLEAR person identification database

conventional techniques, let alone to identify. This is also why a decoupling from the tracking task becomes necessary, if the performance of identification techniques alone is to be accurately measured. Similarly, the acoustic identification is to be made solely on far-field microphones, arrays and tabletops, which can be very distant from the speaker and include all kinds of room noises, murmurs, cross-talk, etc.

The above mentioned difficulties in the data led to a task definition requiring the identification over time windows of varying length, as opposed to identification on single frames, allowing for enough evidence for correct recognition to be accumulated. For CLEAR 2006, a closed set identification task was proposed. The data consisted of synchronized audio-visual segments cut out from the CHIL seminar recordings and containing in total 26 different subjects. In the seminar scenario, only the presenter was to be identified, whereas in the interactive seminar scenarios, recognition was to be done for all participants. For the visual task, images from the four corner cameras, for the acoustic task, all the signals from the far-field microphones could be used for identification. In the multimodal task, all information from the audio-visual streams was available.

The data for the person identification task was partitioned into training (enrollment) and test segments of varying lengths, to assess the effect of temporal information fusion: For training, two conditions, A and B, with segment lengths of (15 and 30)s respectively, were evaluated. The test conditions comprised segments of (1, 5, 10 and 20)s, allowing to measure the increase in recognition accuracy as more information becomes available.

Identification systems are required to output one recognized ID per test segment, which is compared to the labeled identity. The error measure used is the percentage of wrongfully recognized persons for all training and testing conditions. The results of the person identification task are presented and discussed in detail in section 4.6.

3.8 Head Pose Estimation

As for the person identification tasks, the main condition in the CLEAR head pose estimation task builds on the results of person and head detection techniques and aims at determining the head orientations of seminar or meeting attendees using only the information provided by room corner cameras.

The head pose estimation task in CLEAR'06 was split into two conditions, based on two very different databases. The first is the INRIA 2004 Pointing Database figuring studio quality close-up recordings of 15 persons providing 93 images each (see Fig. 13). The objective for this database is to determine the pan and tilt of the user's head in still images. The reference annotations are made in 15 degree intervals in the range from $-90°$ to $+90°$, and the error measures used are the mean absolute error in pan and tilt, and the rate of correct classification to one of the discrete pan and tilt classes.

Fig. 13. Samples from the INRIA Pointing'04 Database

A more natural and challenging problem is addressed in the second condition. Here, the goal is to estimate the pan orientation of the presenter's head in a CHIL seminar room using the room corner cameras (see Fig. 14). Again, the low resolution of heads in the camera views and the difficult lighting conditions, as well as the availability of multiple synchronized video streams are what make this task novel and challenging. The recordings consist of 12 training and 14 test seminars recorded in the Karlsruhe seminar room, with a length of 18min to 68min each. The manual annotations are made for every tenth frame of video, and mark the presenter's head orientation as belonging to one of 8 pan classes (north, north-west, west, south-west, ...), of $45°$ width each.

The goal in this subtask is to continuously track the presenter's horizontal viewing direction in the global room coordinate frame. As for the visual person identification task, the problem of estimating the head pose is decoupled from the head tracking problem by the availability of manually annotated head bounding boxes in the camera images. The error measures used are the mean absolute pan error and the correct classification rate into one of the eight pan classes. In addition, the classification rate into either the correct pan class or one of its neighboring classes (representing at most $90°$ absolute estimation error) is also measured.

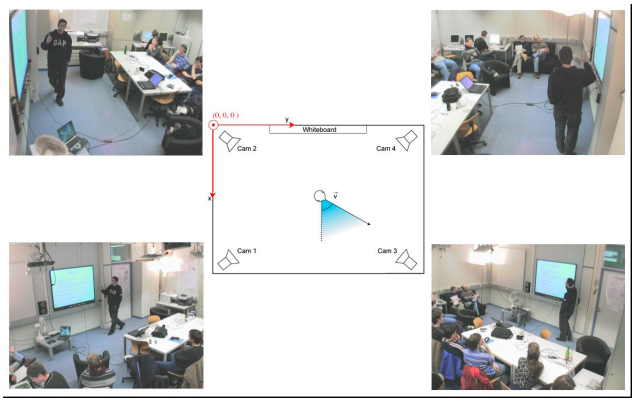


Fig. 14. Sample from the CHIL seminar recordings for head pose estimation

The results for the head pose estimation task can be found in section 4.7.

3.9 Acoustic Event Detection and Classification

To gain a better understanding of the situations occuring in a room and of the activities of its occupants, the recognition of certain events can be very helpful. In particular, the detection of acoustic events, such as keyboard clicks, door slams, speech, applause, etc, in a meeting or seminar can be used to focus the attention of other systems on particular persons or regions, to filter the output of speech recognizers, to detect phases of user interaction, and so forth. The CLEAR acoustic event detection (AED) task aims at measuring the accuracy of acoustic detection systems for this type of scenario, using the input from wall-mounted or table top microphones.

A total of 12 semantic classes are to be recognized: Knock (door, table), door slam, steps, moving chair, spoon (cup jingle), paper wrapping, key jingle, keyboard typing, phone ringing/music, applause, coughing, and laughing. Two additional classes, namely speech and an "unknown event" class are also considered.

Two types of databases are used in this task: One consisting of isolated events, where the goal is solely to achieve a high classification accuracy, and another consisting of scripted seminars recorded in UPC's smart meeting room, where the goal is to detect the time of occurence of an event, in addition to making a correct classification. For the subtask of isolated AED, only the isolated event database is used in training and testing. For the subtask of AED in real environments, both databases are used in training, and testing is made on dedicated segments of scripted seminars.

The error metric used is the Acoustic Event Error Rate (AEER):

$$AEER = \frac{D + I + S}{N} * 100$$

with D, I, S, the number of deletions, insertions, and substitutions respectively, and N the number of events to detect. Here, an event is considered correctly detected when its hypothesized temporal center is situated in the appropriate time interval of one or more reference events and the hypothesized and reference labels match. If none of the labels match, it is counted as a substitution error. An insertion error occurs when the hypothesised temporal center of the detected event does not coincide with any reference event's time interval. A deletion error is counted when a referenced event was not detected at all.

Section 4.8 sums up the results for the acoustic event detection task and briefly describes the challenges and difficulties encountered.

3.10 Acoustic Environment Classification

In contrast to the acoustic event detection task, where the recognition of small, temporally constricted acoustic events is aimed at, the goal in this task is to gain a high level understanding of the type of recording environment itself using audio information. This high level knowledge can be used to provide context awareness in *mobile* settings where large suites of sensors are not available. One example of an application where such knowledge is useful is the CHIL Connector [18] service, in which the environment is used as an information source to help a smart mobile telephone decide whether the user is available for communication. Knowledge of the environmental type can also be useful to boost the performance of basic perceptual algorithms, e.g., by providing appropriate preprocessing or context dependent grammars for speech recognition modules.

In the CLEAR'06 evaluation, classification was tested on a fairly specific set of environments. These environments included airport, bus, gallery, park, restaurant, street, plaza, train, and train platform. Many of these environmental types are self-explanatory. "Gallery" refers to a large indoor space in which people gather, e.g., malls, museums, etc. "Street" is any urban outdoor space with streets dominated by vehicular traffic, while "plaza" refers to an urban outdoor space with streets dominated by pedestrian traffic, e.g., a city square or outdoor marketplace. "Park" is an outdoor space not dominated by urban accoutrements. Finally, "train platform" refers specifically to that part of a train or subway station where passengers board and exit train cars.

The environmental recording database used for this evaluation, the Ambient Acoustic Textures in Enclosed Public Spaces (AATEPS) corpus, consisted of a set of 10min audio recordings made with identical recording equipment in these environments; recordings were made in 2004 and 2005 in North America, Europe, Asia, and Africa. A total of 10.5 h of data, divided into 5s segments, was used in this evaluation, with 5400 segments used for training and 2160 for testing, with half of the test segments taken from recordings *not* part of the training set. Classification results attained in this evaluation are reported in section 4.9.

4 CLEAR Results and Lessons Learned

This section gives an overview of the CLEAR evaluation results and offers a brief discussion based on the attributes of the evaluated systems, and the underlying problems in the tasks and databases. It also hints at future directions to be followed in the next evaluation run, based on the experiences made. For each of the CLEAR tasks and conditions, participants were asked to submit hypothesis files, which were then centrally scored against the reference ground truths. Sites could submit several sets of results for each task, coming from different systems, with the condition that there were basic differences in the concerned systems' algorithms themselves, as opposed to simple differences coming from parameter tweaking. Because of the great number of evaluated systems, no deep insight into the individual approaches could be given here. The interested reader is referred to the individual system publications for details.

4.1 3D Single Person Tracking

The 3D single person tracking task solicited the greatest number of interest and participation. A total of 21 systems were evaluated for the different audio and visual conditions. This was due in part to the traditional nature of the task - person tracking -, allowing for a great variety of approaches, from well known techniques to cutting edge algorithms, to be applied even though the difficulty of the data and the availability of multiple sensors posed new challenges which demanded their share of innovation. The evaluation was made for 4 conditions, the acoustic, the visual, as well as two audio-visual conditions, and the systems were scored using the MOT metrics described in section 3.3. The common database and metrics allowed for an easier comparison of the advantages of different modalities for tracking on the realistic CLEAR data.

Fig. 15 shows the results for acoustic tracking. As the systems are only scored on segments of active speech without noticeable noise, and there is only one target to track, the acoustic subtask very closely resembles a source localization problem, with the difference that the actual detection of speech is not being evaluated. For this reason, and for easier analysis of the results, two additional error measures to the MOT metrics are shown in Fig. 15: The rate of misses caused by localization errors exceeding the 500mm threshold, and the rate of misses attributed to missing speaker hypotheses. Many techniques were presented, mostly based on the calculation of a generalized cross correlation (GCC) or global coherence field (GCF) function, accompanied by Kalman, particle, or data association filtering. The best overall result was achieved by a joint probabilistic data association filtering technique using as features the TDOA between microphone pairs. Overall, the MOTP measure shows that, given correct speech segmentation, very high localization accuracies of up to 14cm can be achieved. For comparison, the expected error in manual annotation of the speaker's head is also of the order of 8-10cm. The MOTA measure, on the other hand, shows us that even for the best systems, in roughly 20 % of all cases the presenter is still to be considered missed. While for most systems, this stems from gross

localization errors in problematic segments, for others it comes from the failure to produce a location hypothesis, hinting at where considerable improvements could still be achieved.

Site/ System	MOTP	Miss Rate (dist > T)	Miss Rate (no Hypo)	A-MOTA
AIT (Inf. Theory)	226 mm	51.16 %	0.0 %	-2.32 %
ITC (GCF)	144 mm	5.17 %	41.44 %	48.22 %
TUT (Part+Int)	245 mm	27.84 %	0.03 %	44.29 %
TUT (Particle)	245 mm	27.86 %	0.07 %	44.21 %
UKA (JPDAF)	137 mm	10.28 %	0.0 %	79.43 %
UKA (IEKF)	138 mm	16.88 %	0.0 %	66.23 %
UKA (Particle)	186 mm	22.58 %	0.0 %	54.84 %
UPC (SRP-PHAT)	145 mm	9.43 %	5.10 %	76.04 %

Fig. 15. Results for the acoustic single person tracking task

The results for the visual condition can be found in Fig. 16. Overall, they are quite high, showing a good match of the task definition to the current state of the art, and the appropriateness of the available sensors.

The highest accuracy (91 %) was reached by a particle filter based approach using color and shape features acquired prior to tracking by a fully automatic procedure. The advantage of particle filters for this type of task is that they are robust to noise and allow to easily integrate a variety of features from several sensor streams. Indeed, they have enjoyed a steady growth in popularity over the past years due to their flexibility. The appearance model adopted here allows efficient particle scoring, resulting in a fast system appropriate for online applications.

The best localization precision (a noteworthy 88mm), on the other hand, was reached by a joint 2D-face and 3D-head tracking system, using adaboost-trained classifier cascades for finding faces in the 2D images. Using faces and heads as the base for tracking, as opposed to full-body tracking, ensures that the system hypothesis is always very close to the annotated ground truth, which explains the high score. It also explains the somewhat higher miss rate, as faces can not always be found in the corner camera images.

This system illustrates another popular trend, the use of boosted classifier cascades using simple features (as presented in [19]), trained for specific detection tasks, and that serve as high confidence initialization steps in combination with

other fast but less reliable tracking techniques. It may be useful to remind here that no background images of the empty room were supplied for this task, and no training was allowed on the test set itself, which made it hard to use foreground segmentation based techniques.

The evaluation also revealed a problem in the visual tracking task definition itself, namely the loose definition of the tracking object. In some cases, it can not be unambiguously decided which of the room occupants is the presenter without using prior scene knowledge or accumulating enough tracking statistics. While this is a minor problem, it will most likely lead to changes in the tracking task definition or annotations in future evaluations.

Site / System	MOTP	Miss Rate (dist > T)	Miss Rate (no Hypo)	MOTA
AIT (Kalman)	246 mm	88.75 %	2.28 %	-79.78 %
IBM (Face det)	88 mm	5.73 %	2.57 %	85.96 %
INRIA (2)	168 mm	15.29 %	3.65 %	65.44 %
ITC (Particle)	132 mm	4.34 %	0.09 %	91.23 %
UKA (Particle)	127 mm	14.32 %	0.0 %	71.36 %
USC (Body det + head)	161 mm	9.64 %	0.04 %	80.67 %
USC (Body det + feet)	207 mm	12.21 %	0.06 %	75.52 %

Fig. 16. Results for the visual single person tracking task

Figs. 17 and 18 show the results for the multimodal tracking task, conditions B and A. As a reminder, for this task the two multimodal conditions differ only in the data segments to be evaluated. In condition B, all time frames, whether they contain speech or not, were scored. For condition A, only the time frames in which the presenter is speaking, without loud noise or crosstalk were scored. This is to better decouple the task from the speaker segmentation problem, accounting for the fact that single modality acoustic trackers are not usable in longer periods of silence.

Compared to the visual tracking results, the numbers for multimodal condition B show no significant improvement. This should by no means imply that audio-visual fusion bears no advantages, but rather that for this type of scenario, with the current visual sensor coverage, the addition of acoustic features could

[2] Results submitted one month after the official deadline and printed here for completeness

not help maintain tracks in the seldom events where visual tracking fails. In contrast, condition A shows that, considering only cases where both modalities are present, the addition of visual features helps improve performance, compared to acoustic tracking alone. For comparison, the best system for this task, a realtime-capable system using a particle filter framework, reached 90 % accuracy using both streams, and just 55 % and 71 % respectively using only acoustic and visual streams. These examples also show us that a modified task description, e.g. limiting the numbers of available cameras or making automatic speech segmentation a requirement, or a slightly more compex scenario might be advantageous in order to better measure the improvement audio-visual fusion can bring when single modalities more frequently fail.

Site / System	MOTP	Miss Rate (dist > T)	Miss Rate (no Hypo)	MOTA
AIT (Kalman)	377 mm	93.90 %	0.02 %	-87.82 %
ITC (Particle)	134 mm	4.27 %	0.32 %	91.13 %
UKA (Particle)	143 mm	14.63 %	0.0 %	70.75 %

Fig. 17. Results for the multimodal single person tracking task, condition B

Site / System	MOTP	Miss Rate (dist > T)	Miss Rate (no Hypo)	(A-)MOTA
AIT (Kalman)	379 mm	94.41 %	0.0 %	-88.83 %
ITC (Particle)	132 mm	3.43 %	6.35 %	86.80 %
UKA (Particle)	140 mm	5.11 %	0.0 %	89.77 %

Fig. 18. Results for the multimodal single person tracking task, condition A

In conclusion, the results for the single person tracking task overall were quite satisfying, although there is still room for improvement. Accounting for the lessons learned in this evaluation run, a move towards a more complex task definition and a shift away from scenarios involving the tracking of just one person becomes very likely in the future.

4.2 3D Multi-Person Tracking

Compared to the single person case, the multi-person tracking task offers a variety of new challenges requiring different systems and strategies. As the number

of tracking objects is no longer fixed, new techniques for determining person configurations, for deciding when to create or destroy a track, for avoiding track mismatches, merges, etc, have to be designed. Compared to seminar recordings, which were used for the single person case, the scenarios in the interactive seminar database used here are also more challenging, including e.g. coffee breaks where all tracked persons move and interact in very close proximity. A total of 5 sites participated in the various subtasks for a total of 11 acoustic and visual tracking systems.

For the acoustic tracking subtask, the objective was quite similar to the single person case, in the sense that only one speaking person needs to be tracked at every point in time. As a consequence, the presented approaches did not differ significantly from the algorithmic point of view. The results are shown in Fig. 19.

Site / System	MOTP	Miss Rate (dist > T)	Miss Rate (no Hypo)	A-MOTA
AIT (Inf. Theory)	230 mm	56.19 %	0.0 %	-12.38 %
ITC (GCF)	218 mm	19.32 %	45.71 %	15.65 %
TUT (Particle)	334 mm	83.22 %	0.10 %	-66.53 %
UKA (JPDAF)	240 mm	52.45 %	0.0 %	-4.90 %
UKA (IEKF)	247 mm	54.39 %	0.0 %	-8.78 %
UPC (SRP-PHAT)	157 mm	15.05 %	5.90 %	64.00 %

Fig. 19. Results for the acoustic multi-person tracking task

On the whole, the scores were quite low, compared to the single person case. Except for the leading system, which reached 64 % accuracy and 16cm precision, all other results were well below expectations. While for the second ranking system, this again comes from a large number of missing hypotheses, for all other systems, the error lies in large inaccuracies in localization itself. The comparatively poor performance of systems can be attributed to several factors: In part it comes from the difficult data itself, including very small rooms with severe reverberations, and in part from the interactive seminar scenario, including frequent speaker switches, coffee breaks, etc.

The visual subtask, requiring the simultaneous tracking of all room occupants, posed a problem of much higher complexity. Three sites participated in the evaluation, which was split in two conditions: The main condition involved data from three sites, for which no previously recorded background images of the empty room were available. This made it much harder for trackers based

on conventional foreground segmentation to acquire clean tracks. The second condition involved data from just two sites, for which such background images were supplied. In addition to the four room corner cameras, a ceiling-mounted panoramic camera, delivering a wide angle view of the room was available. The results can be found in Figs. 20 and 21.

Site / System	MOTP	Miss Rate	False Pos. Rate	Mismatch Rate	MOTA
AIT (Kalman)	233 mm	59.87 %	31.74 %	4.06 %	4.33 %
UKA (Top View)	217 mm	27.62 %	20.29 %	0.97 %	51.12 %
UKA (Multi-View)	203 mm	45.97 %	24.89 %	2.79 %	26.35 %

Fig. 20. Results for the visual multi-person tracking task (3-site dataset)

Site / System	MOTP	Miss Rate	False Pos. Rate	Mismatch Rate	MOTA
AIT (Kalman)	225 mm	50.40 %	29.10 %	5.13 %	15.37 %
UKA (Top View)	201 mm	13.79 %	21.98 %	1.45 %	62.79 %
UKA (Multi-View)	210 mm	35.94 %	34.49 %	4.23 %	25.35 %
UPC (Voxel)	195 mm	21.24 %	46.19 %	4.22 %	28.35 %

Fig. 21. Results for the visual multi-person tracking task (2-site dataset)

Despite the associated problems, all submitted systems were based on foreground segmentation features at the lower level, with the main differences in the higher level data fusion and tracking schemes. The leading system was a realtime-capable foreground blob tracking algorithm using just the single input stream from the top view camera. It reached 51 % and 63 % MOT accuracies for the two conditions respectively, with precisions of about 20cm. The other approaches were based on the fusion of multiple camera streams and the results revealed the still not satisfactorily solved problem of data association for such highly cluttered scenes. Perhaps the extension of one of the probabilistic tracking schemes, which proved very effective in the single person tracking task, to the multi-person case will allow to achieve a jump in performance for the next evaluation runs.

Another important observation is that for all systems the relative amount of track identity mismatches made over a complete recording sequence is very

low, compared to other error types. Although this is explained in part by the nature of the data itself, with only few crossing person tracks, it does considerably diminish the influence of the mismatch rate on the general MOTA score. This observation is likely to lead to a redefinition or modification of the metric for future evaluations, e.g. by the addition of separate weighting factors for the different error ratios.

Fig. 22 shows the results for the audio-visual condition B, which is very similar to the visual tracking subtask, with the exception that acoustic information could be opportunistically used whenever available to increase the confidence in the currently active speaker's track. All presented systems used decision level fusion on the outputs of single modality trackers. The figures show no significant increase compared to visual tracking alone, which can in part be explained by the low accuracies of the acoustic systems, and by the fact that usually only one of the multiple persons to track is speaking at any point in time, considerably decreasing the importance of audio features for the global tracking task.

Site / System	MOTP	Miss Rate	False Pos. Rate	Mismatch Rate	MOTA
UKA (Top View+JPDAF)	204 mm	13.22 %	23.00 %	1.58 %	62.20 %
UKA (Multi-View+JPDAF)	227 mm	34.68 %	35.57 %	4.76 %	24.99 %
UPC (Voxel)	195 mm	21.71 %	46.71 %	4.31 %	27.28 %

Fig. 22. Results for the multimodal multi-person tracking task, condition B (2-site dataset)

The results for condition A, in contrast, are better suited for analyzing the effectiveness of data fusion techniques, as the importance of the single modalities for tracking is better balanced. Here, the objective is to track just the active speakers and to keep a correct record of their identities through time. The results, on the whole, stay relatively poor, due to the low performance of the acoustic component in most systems, which did not allow to filter out the correct speaker track, and of the visual component for the leading system. More work is no doubt required on the single modalities before a synergetic effect can be obtained for the combined systems. It would also be interesting to see if a robust feature level fusion scheme, such as the ones presented in the single person tracking scenario, could lead to heightened performance.

In conclusion, it may be said that the CLEAR multi-person scenario still poses a number of unmet challenges, which will keep driving cutting edge research on new and versatile techniques. Although the CLEAR 3D multi-person tracking task featured a novel and unconventional problem definition, the sub-

Site / System	MOTP	Miss Rate	False Pos. Rate	Mismatch Rate	MOTA	A-MOTA
AIT (Kalman+Inf. Theory)	256 mm	88.64 %	11.43 %	1.97 %	-2.03 %	-0.07 %
UKA (Top View+JPDAF)	237 mm	86.37 %	9.19 %	1.16 %	3.27 %	4.44 %
UKA (Multi-View+JPDAF)	213 mm	86.40 %	9.23 %	1.95 %	2.42 %	4.36 %
UPC (Voxel)	118 mm	41.18 %	16.13 %	5.12 %	37.58 %	42.70 %

Fig. 23. Results for the multimodal multi-person tracking task, condition A (2-site dataset)

mitted results for this first evaluation run were in part very encouraging and the experiences made should prove valuable for future runs.

4.3 2D Face Detection and Tracking

Three sites participated in the face detection and tracking task, where the evaluation was performed separately for the single person seminar scenario and the multi-person interactive seminar scenario. The results can be seen in Fig. 24. For both conditions, the leading systems built on the use of boosted classifier cascades, specially trained for use on CHIL recordings, delivering initial detection hints which where then used by more elaborate multiple pass tracking and filtering techniques.

Site / System	%Correct	%Wrong	%Missing	Dist. error (pixels)	Ext. error (pixels)
Non-interactive seminars					
AIT	12.08 %	137.38 %	1.39 %	0.33 pix	76.17 pix
IBM	54.44 %	37.18 %	18.95 %	0.20 pix	95.76 pix
Interactive seminars					
AIT	11.08 %	94.50 %	17.62 %	0.34 pix	132.97 pix
PITTPATT	71.55 %	10.60 %	23.06 %	0.14 pix	140.99 pix

Fig. 24. Results for the 2D face detection and tracking task

For the seminar scenario, the same system as already presented in the 3D visual single person tracking task achieved best scores, with a correct detection rate

of 54 %, and moderate miss and false positive ratios. For the interactive seminar scenario, a three-stage system involving high acceptance detection, motion-based tracking, and track filtering achieved a remarkable 72 % correct detection, with relatively low miss and false positive ratios. In both cases, the average localization error was in the sub-pixel domain at under 0.2 pixels and face extension errors reached from 96 pixels to 141 pixels.

When judging these numbers, one must bear in mind that these results are averages computed over several seminars featuring multiple faces of different sizes. Detection accuracy was in fact nearly perfect for larger faces, which were located close to the recording camera, while small, far away faces were very often missed. This also explains why systems run on the seminar database, involving only the presenter's face, tended to produce somewhat lower scores: The presenter's face in this database was rarely visible (meaning an eye or the nose bridge is visible) from the closer cameras and face sizes were typically very small. To better assess the effectiveness of face detection and tracking techniques in future evaluations, perhaps a categorization of the visual data into classes of increasing difficulty, with annotated face sizes as the selection criterion, and the separate scoring of results for each class could be a worthwile extension to the task definition.

Similar conclusions were obtained in the VACE run evaluations, the results of which are shown in Fig 25. Smaller faces are harder to detect and track. The best score is about 71 %. Further analysis on how the sites performed on different datasets from the Multi–Site Meetings revealed that the data from VT was the hardest, possibly because faces were smaller in that set.

Site / System	MODA	MODP	MOTA	MOTP
PPATT	71.09 %	0.5385	71.09 %	0.5375
QMUL	-33.90 %	0.0496	-33.95 %	-0.0493

Fig. 25. Results for the 2D face detection task (Multi-Site Meetings[3])

4.4 2D Person Tracking

The results for the 2D person detection and tracking task are shown in Fig 26. Four sites participated in this challenging evaluation and the best performance for both detection and tracking in terms of accuracy is about 42 %. The dataset is challenging, figuring person of different sizes and different viewpoints.

A sub-analysis using the person size as parameter revealed that eliminating small objects gave a boost to the scores compared to including all sizes. In

[3] Scoring differs slightly to the method presented in Section 3.3. Please see [9, 10]

Site / System	MODA	MODP	MOTA	MOTP
AIT	3.56 %	0.5287	2.49 %	0.5273
QMUL	-18.73 %	0.4328	-19.22 %	0.4352
UMD-L	-33.98 %	0.2979	-34.10 %	0.2925
USC	41.98 %	0.5864	41.65 %	0.5862

Fig. 26. Results for the 2D Person Detection and Tracking task (Surveillance)

conclusion, it can be said that smaller persons are harder to detect. Also, performance on one particular viewpoint was much better compared to the other, possibly because of lighting condition differences.

4.5 Vehicle Tracking

The evaluation results for Vehicle Tracking in the Surveillance domain are shown in Fig 27. The best performance for tracking in terms of accuracy is about 64 %.

Site / System	MOTA	MOTP
AIT	29.52 %	0.3592
QMUL	-23.98 %	0.5805
UCF [4]	-8.27 %	0.1980
UMD	10.49 %	0.2888
USC	63.88 %	0.6183

Fig. 27. Results for the Moving Vehicle Tracking Task (Surveillance)

The dataset is challenging figuring different viewpoints and vehicle sizes. A sub-analysis using the vehicle size as parameter revealed that eliminating small objects gave a boost to the scores compared to including all object sizes. In conclusion, it can be said that smaller vehicles, with respect to the frame, are harder to detect and track. Performance on both viewpoints was about equal in contrast to the 2D person detection and tracking evaluation (where performance on one was better than on the other). This could possibly be due to the fact that vehicles are in general bigger, with respect to the frame, most of the time.

[4] Problems with extracting video frames

4.6 Person Identification

Among the 2006 CLEAR tasks, the person identification task was no doubt one of the most complex to organize and carry out, from the point of view of database preparation and annotation, task definition, harmonization of acoustic and visual metrics, and weighting and fusion of multiple audio-visual information streams. 6 different sites participated in the evaluation and a total of 12 audio and visual systems were presented.
For the acoustic identification subtask, most systems built on Mel-frequency cepstral analysis of a single microphone stream, combined with filtering, warping or reverberation cancellation, to reduce environmental effects. Fig. 28 shows the results for the 15s and 30s training conditions.

	15s training				30s training			
Site / System	1s test	5s test	10s test	20s test	1s test	5s test	10s test	20s test
AIT	26.92	9.73	7.96	4.49	15.17	2.68	1.73	0.56
CMU	23.65	7.79	7.27	3.93	14.36	2.19	1.38	0.00
LIMSI	51.71	10.95	6.57	3.37	38.83	5.84	2.08	0.00
UPC	24.96	10.71	10.73	11.80	15.99	2.92	3.81	2.81

Fig. 28. Error rates for the acoustic person identification task

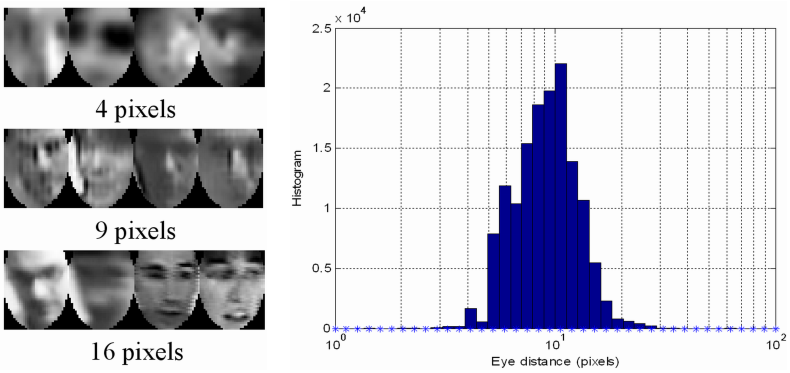
In both cases, identification systems show a big drop in error rates from the 1s to the 5s testing conditions, followed by a steady decrease as more data becomes available. For the 30s train and 20s test condition, the best systems already reach 0 % error. This shows us that for a closed set identification task, with the current indoor seminar scenario and even using just one microphone, acoustic speaker identification can be a very powerful and robust tool. The next worthwhile challenge would be an open set task involving also the automatic detection and segmentation of speech from multiple persons, and the evaluation of identification hypotheses e.g. on a speaker turn basis.

The visual identification task proved much harder for the participating sites, in spite of manual face annotations to alleviate the alignment problem. There were three main difficulties:

- The dataset contained many tiny faces; the median eye distance was just 9 pixels (see Fig. 29).
- There was no regularity in the number or visibility of faces in the (1, 5, 10, and 20)s test sets. This is because the visual data was segmented synchronously to the acoustic data, in view of the multimodal task, and a higher priority was put on producing segments containing speech. Due to this fact,

some small segments contained few or no usable frontal faces in any of the four available camera views. This problem is especially severe for the 1s tests: more than 10 % of them contained no usable frontal faces.
– The frequency of the provided labels (every second for face bounding boxes and nose bridges, and every 200 ms for eyes) proved insufficient for problem-free face alignment.

Three systems were submitted for this subtask and the results are shown in Fig. 30.



(a) Face samples (b) Eye distance histogram

Fig. 29. Examples of frontal faces at various eye distances and histogram of the eye distances in the training and testing faces of the CLEAR database.

	15s training				30s training			
Site / System	1s test	5s test	10s test	20s test	1s test	5s test	10s test	20s test
AIT	50.57	29.68	23.18	20.22	47.31	31.14	26.64	24.72
UKA	46.82	33.58	28.03	23.03	40.13	23.11	20.42	16.29
UPC	79.77	78.59	77.51	76.40	80.42	77.13	74.39	73.03

Fig. 30. Error rates for the visual person identification task

The best system for the 15s training case used two classifiers (PCA and LDA) fused together with temporal confidence accumulation and reached 20 % error rate for the 20s test condition. The leading system for the 30s training case used

a local appearance technique based on DCT features. It reached a minimum 16 % error rate. Both systems showed the expected steady decrease in error rates as the test segment lengths increase, although minimum rates still stayed well above those reached using the acoustic modality.

Fig. 31 shows the results for the combined audio-visual identification systems. 4 different approaches were presented, all using decision-level fusion of single modality system outputs. As could be expected from the single modality results, the weighting of the two modalities played an important role, with systems favoring the acoustic side clearly outperforming those which assigned equal weights. The best system, which was not represented in the acoustic subtask, used a fusion scheme incorporating streams from multiple microphones in addition to temporal information. It reached remarkably low error rates of 0.56 % for the 20s test condition, in both 15s and 30s test cases.

	15s training				30s training			
Site / System	1s test	5s test	10s test	20s test	1s test	5s test	10s test	20s test
AIT	23.65	6.81	6.57	2.81	13.70	2.19	1.73	0.56
UIUC primary	17.61	2.68	1.73	0.56	13.21	2.43	1.38	0.56
UIUC contrast	20.55	5.60	3.81	2.25	15.99	3.41	2.42	1.12
UKA / CMU	43.07	29.20	23.88	20.22	35.73	19.71	16.61	12.36
UPC	23.16	8.03	5.88	3.93	13.38	2.92	2.08	1.12

Fig. 31. Error rates for the multimodal person identification task

In conclusion, it may be said that although acoustic identification techniques seem to outperform visual ones, this is largely due to the nature of the data at hand and the definition of the task. Perhaps the only fair way of comparing modalities would imply completely automatic detection and segmentation of speech in an open set for the acoustic side, and fully automatic tracking, alignment, and identification for the visual side. This would however also greatly increase the complexity of the tasks and required metrics. In a lesser case, a careful selection of the database, with equal audio and visual complexities, and a redefinition of the multimodal task to better reflect the advantage of fusion during single modality failure, could also help reduce the apparent imbalance and drive the development of novel fusion techniques.

4.7 Head Pose Estimation

The two subtasks of the CLEAR head pose estimation task offered two very distinct levels of challenge to evaluation participants. While the frame based estimation on the studio database, featuring close-up images, was the more conventional task which has been adressed before, the pose tracking task on the seminar database with multi-view low resolution head captures opened a new field with new challenges for head pose estimation.

For the former condition, three systems were presented, based on a variety of approaches, from PCA classification to feed-forward or auto-associative neural nets. The results can be seen in Fig. 32.

Site / System	Pan Mean Error (deg)	Tilt Mean Error (deg)	Pan Correct Class	Tilt Correct Class
INRIA	10.1	16.7	50.2 %	44.37 %
UIUC	14.1	14.9	55.2 %	84.3 %
UKA	12.3	12.8	41.8 %	51.11 %

Fig. 32. Results for the head pose estimation task (Pointing'04 data)

The best systems reach an error rate of 10.1° and 12.8° for pan and tilt respectively, which is well in range of a human's estimation error on such images. The correct classification rate into 15° orientation classes is also shown in Fig. 32, with the leading system achieving 55 % pan and 84 % tilt classification accuracy.

For the more difficult subtask involving real seminar data, a series of new techniques were explored, including 3D head texture modeling, fusion of neural net classifiers, and combination of boosted classifier cascades. The results are shown in Fig. 33. For the leading system, based on sequential multi-view face detection and HMM filtering, 45 % correct classification (into 45° classes) was reached. When allowing also classification into the neighboring class, the score reaches 87 %. To better view these numbers in context, it must be said that even human annotation of 45° head pose classes in the room corner camera images proved very difficult, since it was often ambiguous to the annotators, which orientation class to choose. Here, an analysis of inter-annotator agreement is needed in the future.

In conclusion, one can say that although the head pose estimation task on CHIL seminar data presented a novelty to the field, the results achieved in this first evaluation run proved very encouraging. The availability of several camera views alleviates the problem of small head sizes with respect to the frame and drives the development of more sophisticated fusion schemes. One must also note that a big part of the difficuly in the current recordings came from the

Site / System	Pan Correct Class	Pan Correct Class + Neighbor
UIUC	44.8 %	86.6 %
UKA	34.9 %	72.9 %
UPC	19.7 %	48.8 %

Fig. 33. Results for the head pose estimation task (Seminar data)

difficult lighting conditions in the seminar room, affecting the performance of all algorithms.

4.8 Acoustic Event Detection and Classification

For the two conditions of the acoustic event detection task, the classification of isolated events, and the detection and classification of events in seminars, a total of 11 systems were presented by 3 sites. The systems are based on the HMM or SVM classification of spectral features gained from a single audio channel. The results are shown in Figs. 34 and 35.

Systems / Databases	CMU-C1	CMU-C2	ITC-C1	ITC-C2	UPC-C
ITC isolated DB	7.5 %	----	12.3 %	----	4.1 %
UPC isolated DB	----	5.8 %	----	6.2 %	5.8 %

Fig. 34. AEER error rates for the acoustic event detection task (classification only)

Systems / Databases	CMU-D1	CMU-D2	ITC-D1	ITC-D2	ITC-D3	UPC-D
ITC isolated DB	45.2 %	----	23.6 %	----	----	64.6 %
UPC isolated DB	----	52.5 %	----	33.7 %	----	58.9 %
UPC seminar DB	----	177.3 %	----	----	99.3 %	97.1 %

Fig. 35. AEER error rates for the acoustic event detection task (detection and classification)

The error rates show that, while for the recognition of isolated events, current techniques are already appropriate, reaching about 4 % error in the best case, the detection of low-energy events in a complex seminar scenario, on the background of speech, is still an unsolved problem. The best system, using a two step SVM approach for detection of silence/non-silence and subsequent recognition of the 12 event classes, delivered 97 % error rate on unsegmented seminar data, and about 60 % error on presegmented event databases. One of the main difficulties no doubt came from the presence of speech in the recordings, showing that a better coupling with SAD systems could yield some improvement. Additionally, the use of multiple microphones to better handle noise and room acoustics yet has to be explored, and may constitute one of the main research directions for the future.

4.9 Acoustic Environment Classification

For the acoustic environment classification task, only one site participated. The results for the seen test condition, the unseen test condition, and the average of these two conditions are shown in Fig. 36. The system performed much better in identifying environments from locales specifically seen in the training data; however, the error rate for unseen locales is still much better than chance. These results indicate that while practical systems might be fielded to identify a user's frequently-visited locales, work still needs to be done on improving generality and adapting to new locales.

Site / System	Seen Error	Unseen Error	Total Error
CMU	15.4 %	25.4 %	5.4 %

Fig. 36. Results for the Acoustic Environment Classification Task

5 Summary

This paper summarized the CLEAR 2006 evaluation, which started early 2006 and was concluded with a two day workshop in April 2006. It described the evaluation tasks performed in CLEAR'06, 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 used 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. In CLEAR'06, sixteen international research laboratories participated in more than 20 evaluation subtasks.

An important contribution of the CLEAR evaluation is the fact that it provides 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.

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. From a decision made in mid November 2005 by CHIL and VACE to establish CLEAR, to the actual CLEAR workshop in April 2006, over 20 evaluation subtasks were performed. In that period of four months, evaluation tracking metrics between CHIL and VACE were harmonized, several hours of multimedia data were annotated for the various evaluation tasks, large amounts of data were distributed to 16 participants worldwide, and dozens of teleconferences were held to help coordinate the entire evaluation effort.

An additional important contribution of CLEAR 2006 and the supporting programs is that significant multimedia datasets and evaluation benchmarks have been produced and made available to the research and community. Evaluation packages for the various tasks, including data sets, annotations, scoring tools, evaluation protocols and metrics, are available through the Evaluations and Language Distribution Agency (ELDA)[20] and NIST.

While we consider CLEAR 2006 as a remarkable success, we think that the evaluation tasks performed in CLEAR 2006 - mainly tracking, identification, head pose estimation and acoustic scene analysis - 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 more challenging tasks, addressing human activity analysis on higher levels will become part of future CLEAR evaluations.

In order to keep CLEAR focused, the coordinators are committed to working together to synergize more aspects of the CLEAR evaluations. This synergy will allow evaluation assets developed to be greater than if they were developed independently by each participating evaluation program. For instance synergy in areas of data annotations and format will positively impact future evaluations by providing a lasting data resource whose development is cost-shared across evaluation programs and projects, while useful for numerous tasks due to the commonalities.

Acknowledgments

The authors would like to thank the following people for all their help and support in organizing the CLEAR evaluation and for their help in revising this paper:

Matthew Boonstra, Susanne Burger, Josep Casas, Hazim Ekenel, Dmitry Goldgof, Rangachar Kasturi, Valentina Korzhova, Oswald Lanz, Uwe Mayer,

Rob Malkin, Vasant Manohar, Ferran Marques, John McDonough, Dennis Moellmann, Ramon Morros, Maurizio Omologo, Aristodemos Pnevmatikakis, Gerasimos Potamianos, Cedrick Rochet, Margit Rödder, Andrey Temko, Michael Voit, Alex Waibel.

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.

Disclaimer

The here presented tests are designed for local implementation by each participant. The reported results are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U. S. Government.

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. PETS - Performance Evaluation of Tracking and Surveillance, http://www.cbsr.ia.ac.cn/conferences/VS-PETS-2005/.
7. TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid/.
8. ETISEO Video Understanding Evaluation, http://www.silogic.fr/etiseo/.
9. Mostefa, D., Garcia, M.N., Bernardin, K., Stiefelhagen, R., McDonough, J., Voit, M., Omologo, M., Marques, F., Ekenel, H.K., Pnevmatikakis, A.: Clear evaluation plan. Technical report, http://www.clear-evaluation.org/downloads/chil-clear-v1.1-2006-02-21.pdf (2006)
10. The VACE evaluation plan, http://www.clear-evaluation.org/downloads/ClearEval-Protocol-v5.pdf.
11. CLEAR evaluation webpage, http://www.clear-evaluation.org.
12. Transcriber Labeling Tool, http://trans.sourceforge.net/.
13. AGTK: Annotation Graph Toolkit, http://agtk.sourceforge.net/.
14. Fiscus, J.G., Ajot, J., Michel, M., Garofolo, J.S.: The rich transcription 2006 spring meeting recognition evaluation. In: 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI-06), (Springer)
15. The i-LIDS dataset, http://scienceandresearch.homeoffice.gov.uk/hosdb/physical-security/detection-systems/i-lids/ilids-scenario-pricing/?view=Standard.
16. Bernardin, K., Elbs, A., Stiefelhagen, R.: Multiple object tracking performance metrics and evaluation in a smart room environment. In: Sixth IEEE International Workshop on Visual Surveillance, in conjunction with ECCV2006, Graz, Austria (2006)

17. ViPER: The Video Performance Evaluation Resource, http://viper-toolkit.sourceforge.net/.
18. Danninger, M., Flaherty, G., Bernadin, K., Ekenel, H., Kohler, T., Malkin, R., Stiefelhagen, R., Waibel, A.: The Connector — facilitating context-aware communication. In: Proceedings of the International Conference on Multimodal Interfaces. (2005)
19. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - CVPR 2001. Volume 1. (2001) 511–518
20. ELRA/ELDA's Catalogue of Language Resources: http://catalog.elda.org/.