AMIDA

Augmented Multi-party Interaction with Distance Access

http://www.amidaproject.org/

Integrated Project IST–033812

Funded under 6th FWP (Sixth Framework Programme)

Action Line: IST-2005-2.5.7 Multimodal interfaces

Deliverable D4.4: WP4 work in year 2

Due date: 30/09/2008	**Submission date:** 30/09/2008
Project start date: 1/10/2006	**Duration:** 36 months
Lead Contractor: USFD	**Revision:** 1

Project co-funded by the European Commission in the 6th Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

D4.4: WP4 work in year 2

Abstract: WP4 is concerned with the development of reliable audio, visual, and audio-visual integration and recognition tools for the automatic extraction of information from raw data streams. This involves multistream fusion, synchronisation, and recognition methods from the different audio-visual information sources. Research topics include automatic speech recognition, keyword and event spotting, visual tracking, speaker diarisation, determining the focus of attention, visual and speaker identification, detecting gestures, actions and social signals.

D4.4 is a report on the implementation and evaluation of the different audio, video, and multimodal algorithms conducted in the second year of AMIDA. The report shows how existing algorithms have been adapted and extended to process data from not just local meeting participants but also remote ones.

Contents

1 Introduction

WP4 is concerned with the development of reliable audio, visual, and audio-visual integration and recognition tools for the automatic extraction of information from raw data streams. This involves multistream fusion, synchronisation, and recognition methods from the different audio-visual information sources. Algorithms have been ported or implemented in the AMIDA domain.The models that have been applied include HMMs, Bayesian networks, neural networks, multistream approaches, and multisource decoding.

1.1 Splitting of work

Instead of dividing the tasks into speech, visual, and audio-visual groups it had been previously decided to split the tasks into problem-based groups. Solutions are not distinguished by their approach (for example visual or audio identification of persons). Work has been conducted in the following seven tasks areas for the analysis of meetings and to support a remote meeting assistant:

1. Automatic Speech Recognition – ASR (Sec. 2)
2. Keyword and Event Spotting (Sec. 3)
3. Speaker Diarisation (Sec. 4)
4. Visual Focus of Attention and Tracking (Sec. 5)
5. Video- and Audio-based Person Identification (Secs. 6 and 7)
6. Gestures and Actions (Sec. 8)
7. Social Signals (Sec. 9)

1.2 Aim in the second year and outline of this deliverable

The expected result of WP4 is a set of multimodal recognisers for the seven different tasks listed in the previous section. This deliverable reports on the implementation and evaluation of the different audio, video, and multimodal algorithms in the second project year. This report shows how new algorithms have been developed or existing algorithms adapted and extended to process data from the AMIDA domain. Furthermore, it shows how such algorithms and been adapted to address realtime requirements.

A key WP4 target for the second year was the development of realtime, online ASR. This has been achieved and is now available; the details are documented in Sec. 2.

Beside the seven main research themes, we also addressed side topics of motion tracking and visualisation in a soccer control room (Sec. 11) and ICT in healthcare team communication (Sec. 10) to show how AMIDA technologies can be transferred to problems outside the meeting domain.

2 Automatic Speech Recognition (ASR)

The primary focus of work during the last period is on achieving a realtime online speech recognition system: a working HUBable on-line ASR system is now available. Other work includes general ASR system improvements, the enhancement of the AMIDA ASR infrastructure so that future ASR R&D work can be performed more efficiently and a continued effort to disseminate AMIDA ASR technology to the general research community (webASR, Bob and CTS tutorial). Furthermore, AMIDA technology has been evaluated through participation in the Dutch LVCSR evaluation N-best 2008 (Karafiát et al., 2008).

2.1 Towards realtime ASR

Significant advances have been made on the realtime/online LVCSR decoder, Juicer, which is now working as a HUBable demonstrator. To allow Juicer to work online, a data flow framework known as Tracter has been developed. Although Tracter has been developed with Juicer in mind, it provides a general modular framework to join together signal processing algorithms.

Tracter and Juicer have been integrated together to produce a real-time capable on-line ASR demo. This was mainly software development, but also involved training suitable acoustic models. A set of scripts has been developed for rapidly training meeting room speech recognition models on new feature sets. A new maximum likelihood set of models can now be trained on 100+ hours meeting room data in half a day. This enables easy investigation and comparison of different front-end processing modules. We have been using these scripts to investigate acoustic front-ends for a real-time, on line meeting recognition system. For instance, techniques like cepstral variance normalization, normally an off-line technique, have been investigated.

Tracter is also able to supply low level feature data via TCP sockets to other producers. Much use has been made of socket connections, allowing the system to be modularised. This also allows the system to be split between Windows and Linux systems easily. Tracter has allowed modules written by distinct groups to be incorporated into a single system: for example, the on-line beamformer is written on a Windows platform and its output is passed through a TCP socket to a linux machine running a HTK front-end module plus other parts of Juicer. Projuicer, a java wrapper for the Tracter and Juicer combination, has been written. This allows the results of recognition to be streamed via the java middleware to the hub. From the hub, metadata is available to all other AMIDA consumers. Thus the demonstrator runs on several computers. Audio is captured on one machine in real time from a single microphone or a microphone array and on-line beamformer. A single audio stream is sent via a TCP port to a second computer which is running Projuicer which sends ASR output to a third computer which serves as the HUB.

The transducers used by juicer to represent the grammar have also been optimised for size and speed (Garner, 2008a,b,c). Medium to large transducers can now be built on 32 bit hardware instead of requiring 64 bit machines, although very large transducers still require 64 bit hardware.

Modern speech decoders are complex with potentially a large number of parameters that allow tuning for speed and accuracy. Methods for automatic optimisation of such parameters have been investigated (El Hannani and Hain, 2008). The objective was to find the optimal configuration of the decoder that yields minimal search errors for a given real-time factor. The approach is based on automatic tracking of that optimal curve. Experiments were conducted using the HTK large vocabulary speech decoder HDecode. Results on NIST 2001 CTS evaluations data show that below real-time speed can be achieved with a degradation of 2.9% word error rate (WER) absolute. This was obtained by joint optimisation of six parameters. No prior knowledge about the interpretation of the parameters is used so the proposed approach should work for any decoder.

2.2 Array based ASR

Array based ASR research has been focussed on the SSC data as was the robust feature extraction. The main focus was Minimum Mutual Information (MMI) beamforming (Kumatani et al., 2007a), but three types have of improvement been done

1. A Nyquist(M) filter bank for adaptive subband processing.
2. Zelinski post-filtering to remove incoherent noise.
3. Binary masking to suppress interference signal further.

Taken together, these achieved a WER 39.6% while MMI beamforming with perfect reconstruction filter banks provided a WER 50.7%.

Following on from these results, we addressed two subjects

1. The filter bank design method for subband beamforming (Kumatani et al., 2007b, 2008a),
2. The subband beamforming algorithm with the Maximum Negentropy (MN) criterion for the scenario where a single speaker is stationary (Kumatani et al., 2008b,c,d).

The properties of the proposed filter banks were thoroughly analyzed and compared with those of other popular filter banks. We also showed that MN beamforming is free from the signal cancellation problem encountered in the conventional adaptive beamforming techniques. Finally, we demonstrated the effectiveness of each technique through a set of automatic speech recognition experiments on the multi-channel data collected by the AMI project.

The work continued to address far-field speech recognition on the Multi-Channel Wall Journal Street Audio-Visual Corpus (MC-WSJ-AV) (Kumatani et al., 2007c). We proposed two beamforming algorithms using high order statistics (HOS), namely Maximum Negentropy (MN) beamforming and Maximum Kurtosis (MK) beamforming. Moreover, as the state-of-art conventional beamformer, the generalized eigen-vector generalized sidelobe canceler (GEV-GSC) has been implemented and compared with our proposed techniques. The GEV-GSC achieved a WER 14.5% which is further reduced to 13.2% by the MN beamforming algorithm. We also demonstrated that the MK beamformer achieved almost the same performance as the MN beamforming with a small computational cost in the case that the sufficient amount of data for the adaptation is available.

2.3 Feature extraction

For feature extraction, bottle-neck features (Grézl et al., 2007) have been improved by replacing phoneme MLP training targets with phoneme states and adding delta features (Grezl and Fousek, 2008). Work was also done in tailoring the bottle-neck feature extraction for the real-time ASR system by reducing the size of the MLP by 75%. This reduced the required real-time factor for the feature extraction to just 0.1 with a loss of accuracy of just 1.2%, which is still outperforming the original, PLP-HLDA system.

Further investigations of pitch adaptive features for ASR were conducted. Experiments to decouple the pitch adaptive and smoothing effect of STRAIGHT features were performed, and the speaker independence of STRAIGHT derived features was investigated (Garau and Renals, 2008a,b).

Other work on robust feature extraction has centered around mapping of multichannel far field speech to single channel near field speech using a neural network (MLP) (Li et al., 2008, 2007a,b). Evaluation has been performed on the Multi-Channel Numbers Corpus (MONC) and subsequently the Speech Separation Challenge (SSC2) task. Initial results showed that the MLP based technique worked well in the log-spectral energy domain — the domain used as the ASR features.

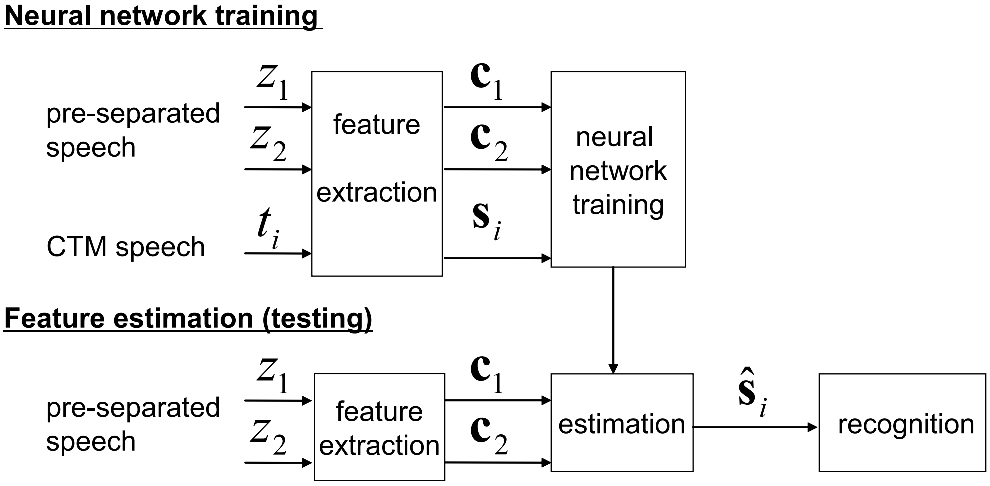


Figure 1: A diagrammatic view of the MLP based feature mapping approach; in this case after beamforming.

To augment the mapping techniques, a postprocessing method has been implemented for single-channel frequency-domain speech enhancement to reduce speech distortion (Li, 2007). This involves use of a standard noise reduction technique followed by a weighted combination with the original distorted signal. The results show that combined method can increase ASR accuracy by 10% relative.

We investigated the use of higher order MFCC features for ASR. Normally, such features contain detailed harmonic information not required for ASR and are ignored. In the context of MLP mapping, however, we find that such features are beneficial, although the very highest order features should still be ignored. Results from a masking post-filter suggest that estimating the parameters of interfering speech also helps in ASR with multiple

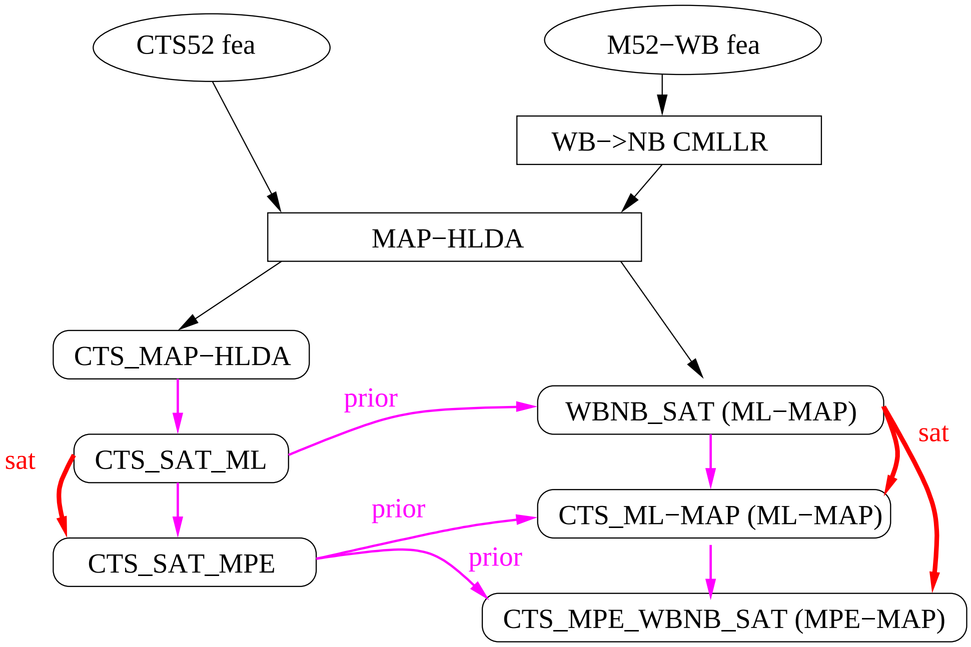


Figure 2: Adaptation scheme of MPE-MAP adaptation into the WB→NB features.

overlapping speakers.

We have also combined the mapping based feature extraction work with the array based work (above) and shown that the approaches are complementary; the mapping further increases ASR accuracy after MMI beamforming.

2.4 Discriminative training with out-of-domain data

Discriminative training coupled with the use of linear transforms allowing for use of conversational telephone speech (CTS) data in the development of meeting recognition system was investigated, focusing on narrow band - wide band adapted systems.

The amount of training data has a crucial effect on the accuracy of Hidden Markov Model (HMM) based meeting recognition systems. One of the largest collections of speech data is conversational telephone speech which was found to match speech in meetings well. However it is naturally recorded with limited bandwidth. In previous work (Karafiat et al., 2007), we presented a scheme that allows to transform wide-band meeting data into the same space for improved model training. This year, we have focused on integration of discriminative adaptation into this scheme. This integration is not straightforward and the process is quite complex (Figure 2).

We successfully implemented an adaptation technique where WB data is transformed to the NB domain by CMLLR feature transform. Here, the well trained CTS models are taken as prior for adaptation. A solution on how to apply this transform for HLDA and SAT systems was given using maximum likelihood where a 4.6% relative improvement against adaptation in the downsampled domain was obtained. Next, ML-MAP was replaced by the discriminative MPE-MAP scheme, where a 2.4% relative improvement over the non-adapted meeting system was shown. In the end, the Fisher corpora were included for improving of the CTS prior model and also some new meeting data resources. In the final MPE-MAP implementation, we obtained a 5.6% relative improvement over non-adapted meeting system. See Karafiat et al. (2008) for details.

2.5 Language modelling

Work has also continued on language modelling. Unsupervised adaptation of language models was studied. The first pass output of the ASR system was used to build a new language model directly and to derive a search model based webdata collection which was then used to build another language model. This yielded a 0.9% absolute reduction in the word error rate on 10 hours of lecture data. The use of Topic and Speaker Role information in hierarchical Bayesian language models has also been investigated (Huang and Renals, 2008a,b,c, 2007).

2.6 BUT system for Dutch 2008 LVCSR evaluations “N-best”

BUT also participated in the Dutch LVCSR evaluations N-best in 2008[1]. Recognition systems for both broadcast news (BN) and conversational telephone speech (CTS) were produced. The whole recognition process operated in 6 passes and was inspired by the AMIDA system (Hain et al., 2007). The system scored excellently in both conditions. See Karafiát et al. (2008) for detailed system description.

2.7 Infrastructure

The research infrastructure for ASR has been improved significantly. A new Resource Optimisation Toolkit was designed and implemented. This toolkit allows a wide range of different ASR systems to be specified, implemented and tested simultaneously while making best use of the available computing resources. Components of an ASR system are written into modules with a set of specified inputs and outputs. These modular components can be connected together flexibly in a graph to produce a complete system. All the components of the 2007 NIST rich transcription evaluation system have been written as modules in the new system. This system forms the basis of the webASR system which is a speech recognition service provided to the general scientific community.

2.8 Dissemination

The webASR system (Hain et al., 2008) is an online interface to the AMIDA ASR system which allows the upload of audio files and, in turn, the download of automatically generated ASR transcripts. It is believed that this – a publicly available end-to-end ASR system – is the first such system of its kind. Depending upon the metadata provided for an audio file (or set of audio files in the case of a microphone array recording), the system will generate the transcription using an appropriate speech recogniser chosen from one of the many available, such as a NIST RT evaluation system. After processing, the transcripts are made available for download in a number of formats which are of use to

[1] http://speech.tm.tno.nl/n-best/

both speech researchers (STM, MLF and LAB) as well as users simply interested in the textual transcription (PDF and HTML). webASR provides a feature-rich interface which allows the user to manage uploads, monitor processing and, according to their individual access rights, reprocess existing audio files with a range of different ASR systems. The system can be accessed from http://www.webasr.com.

AMIDA technology is also being disseminated to the scientific community by publically releasing the tools used to create the lexicons and pronunciation dictionaries. For this a program was specially written in Java and an associated conference paper describing it was submitted (Wan et al., 2008). The paper also highlights, with empirical evidence, how important it is to ensure that words are spelt correctly and that phones are used in consistent manner in pronunciation dictionaries.

2.9 Future work

The real time ASR system will be developed actively to increase speed and robustness. Investigations are ongoing into suitable normalisation techniques for on-line features. Time synchronisation needs to be addressed. In particular, we expect the beamformer to create particular online segmentation difficulties that will need to be addressed. The real time system will also take on a diarisation component.

The webASR system will be tested by a number of individuals (both within AMIDA and external to the project). Once this process has been completed and any necessary alterations have been made, the system will go ‘live’ and be available to the general public. Throughout these two stages (testing and initial release) we will investigate other features which can be incorporated into the system.

3 Keyword Spotting

The work in the area of keyword spotting and spoken term detection in the last period had three important parts: construction of hybrid recognition networks for combined word and sub-word recognition (and hence indexing and search), integration of acoustic keyword spotter into the Hub infrastructure and improvements of the system for detection of out-of-vocabulary words in the output of speech recognizer.

3.1 Hybrid word-subword recognition system for spoken term detection

The first task investigated in the last period was an investigation of hybrid word-subword recognition system for spoken term detection. The decoding is driven by a hybrid recognition network and the decoder directly produces hybrid word-subword lattices. One phone and two multigram models were tested to represent sub-word units. The systems were evaluated in terms of spoken term detection accuracy and the size of index on the NIST STD (Spoken Term Detection Evaluation) data from 2006. We concluded that the best subword model for hybrid word-subword recognition is the multigram model trained on the word recognizer vocabulary.

We achieved an improvement in word recognition accuracy, and in spoken term detection accuracy when in-vocabulary and out-of-vocabulary terms are searched separately. Spoken term detection accuracy with the full (in-vocabulary and out-of-vocabulary) term set was slightly worse but the required index size was significantly reduced. Details of this work were presented at SIGIR/SSCS 2008 – the 2nd workshop on Searching Spontaneous Conversational Speech in Singapore (Szöke et al., 2008).

3.2 Integration of on-line keyword spotting

The second task involved the on-line meeting processing. The on-line acoustic keyword spotter (Szöke et al., 2005) was integrated with the Hub infrastructure to allow for on-line detection of keywords in AMIDA demonstrations. The acoustic keyword-spotter is based on an estimation of phone posterior probabilities by neural networks and on the classical tandem of target word model and background model.

3.3 Detection of out of vocabulary words

Finally, we have extended the work done on the detection of out-of-vocabulary (OOV) words (Burget et al., 2008a). The work in this period concentrated on the approach that was found the most promising in the prior work – the NN-based combination of posteriors from strongly (LVCSR) and weakly constrained phone posteriors into a decision on whether a word at the output of LVCSR is likely to be an OOV (see Figure 3).

Work also focussed on the amount of context required while combining the strong and weak posteriors. While the original work (Burget et al., 2008a) used a fixed context of ±6 frames, this work investigated dynamically assigning the contexts depending on the length of preceding and following phone. Small but stable improvements over the original fixed context were obtained (Kombrink, 2008).

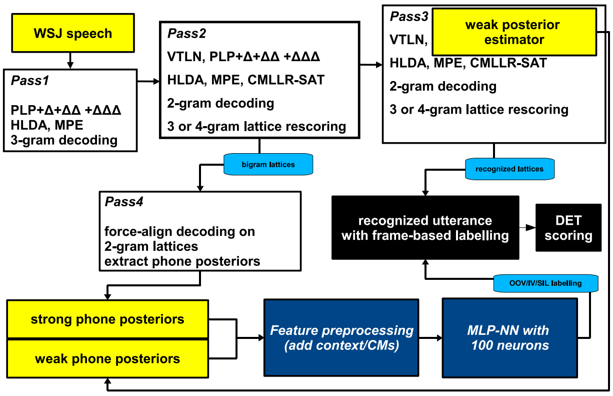


Figure 3: Out of vocabulary detection by the combination of strong and weak posteriors.

4 Speaker Diarisation

Current speaker diarisation approaches are able to detect speaker changes and cluster speaker-homogenous segments. A fundamental goal was to improve both the robustness of the approach as well as to develop the system towards online functionality. Because of its unsupervised nature, however, the results lack a mapping between speaker labels and real participants names. Since there is little control in the choice and placement of microphones in the meeting rooms, a research goal was to investigate approaches that incorporate training of speaker models to make a labeling with real names possible. We intended to perform research on methods and techniques for speaker recognition using both unsupervised approaches (speaker diarisation) and supervised approaches (speaker identification). The following section summarizes the key achievements:

- Overlap Detection. Currently, overlapping speech is not taken into account in neither speaker diarisation nor speaker recognition approaches. This means that if two or more speakers are talking at the same time, current methods only assign one label to it. The detection of overlapping speech is a non-trivial problem (e.g., see Wrigley et al., 2005) and only a very limited number of research projects exist. However, the some AMI meetings contain more than 18% overlapped speech. We found that if we could perfectly assign labels to overlapping speech regions, the Diarisation Error Rate could be decreased by at least 50% (relative). We performed experiments on the detection of overlapped speech using combinations of different short-term and long-term acoustic features. The overlap detection was then used as an independent post-processing step after the diarisation process. This makes the approach also usable as a post processing step after a speaker recognition approach. For the high-quality signal case of a single mixed-headset channel signal, we could demonstrate relative improvement of about 7.4% DER over the baseline ICSI Diarisation system, while for the more challenging case of the single far-field channel signal, the relative improvement is 3.6%.

- Live Speaker Identification. In a first series of experiments we studied how some parts of the ICSI Speaker Diarisation offline system could be simplified or replaced without a significant loss of performance so that online processing would be possible. The main hurdle is the use of a global classification step, namely the Viterbi algorithm on top of a Hidden Markov Model. We found that this step could be replaced by a classification that is only local: tests suggested that one could use a temporal sliding window based on either maximum likelihood or majority voting to segment the audio data into chunks. For window sizes between one and two seconds the performance in terms of Diarisation Error Rate (DER) was the same. The DER on DEV07 and AMI data using a non-overlapping sliding window on frames of 10 ms is compared against the baseline system. For this experiment, a speaker in a given window was detected by majority voting on GMM likelihoods. This also achieves a faster detection since there is a potential saving in computing likelihoods (some likelihood computations can be skipped if someone already has 51% of the votes). A second experiment concerned the use of pre-trained models in order to be able to map speaker clusters to real names. We found that with only 50 seconds of speech per speaker the system is able to perform the diarisation task on a subset of the AMI meetings, a total length of more than 9.5 hours, with a DER better than the offline system. When the pre-trained models are also label with real names, the assignment of speaker clusters to real names becomes trivial. The system was presented at the AMI Knowledge and Know How day at MLMI 2008.

Diarisation work has also continued in the form of an information theoretic *Information Bottleneck* (IB) approach, where the distance between speech segments becomes the Jensen-Shannon divergence. The implications of this approach have been investigated in the context of inferring number of speakers etc. Work continued on multiple features for diarisation, and other system parameters. The system performs around 1% worse than state of the art on the NIST RT06 (Rich Transcription) data set for speaker diarisation of meetings, but *significantly* faster.

5 Visual Focus of Attention and Tracking

This section discusses work conducted towards real time visual focus of attention (VFOA) estimation which requires realtime head pose tracking. Research about view independent head tracking and omni-directional multi-person tracking which are not directly related to VFOA estimation are also presented.

5.1 Focus of Attention

During this year our research activities have been focused mainly in three directions. First we further improved our offline VFOA estimation system using multimodal contextual cues integration. Secondly, the VFOA we estimated with our system have been used by collaborators in WP5 for their study about dominance and status recognition in meetings. Thirdly, research and development have been done towards the development of a realtime head pose estimation module for a single person to be used for VFOA recognition.

5.1.1 VFOA recognition

Visual focus of attention recognition is an important cue to recognize social interactions in meetings. Due to the available camera resolution, however, recognizing the VFOA is a difficult task. The main cue to recognize VFOA is the head pose. However, as the same head pose can be used to gaze at different targets, the VFOA estimation can be improved by exploiting the meeting context represented by people speaking activity and the slide activity. Thus, in 2007 we had started the investigation of the use of multimodal contextual cues to do the VFOA recognition. This year, in order to improve our multi party VFOA recognition method (Ba and Odobez, 2008a,b), we investigated a different way of representing the head pose observation. Rather than just using the estimated head pose from our tracker, represented by a pan and tilt angle, we used the head pose distribution (Ba and Odobez, 2008c). Fig 4 gives illustrations about head pose, head pose pdf and head pose pdf models corresponding to looking at three visual targets in the meeting room. This approach was found to improve performance by 5.4% with respect to the equivalent approach using a single head pose as observation.

5.1.2 VFOA recognition for dominance and status estimation in meetings

As an application of our framework, in collaboration with WP5, we have investigated the use of VFOA cues to the estimation of the dominant people in meetings (Hung et al., 2008; Jayagopi et al., 2008). For the studies in Hung et al. (2008); Jayagopi et al. (2008) VFOA were automatically extracted over 6 hours of recordings from the AMI corpus dataset in which people involved in meetings were sometimes moving to the projection screen and to the white board to make presentations. The presence of moving people makes the VFOA recognition task very challenging.

Hung et al. (2008) conducted a study of the automation of the visual dominance ratio; a classic measure of displayed dominance, which combines both VFOA and speaking

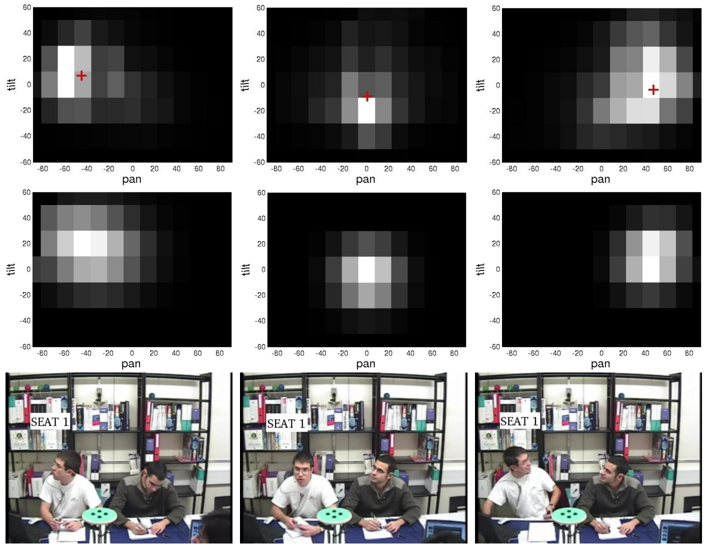


Figure 4: Head pose pdf observation and models. First row: head pose pdf and corresponding estimated head pose (red plus) for a person seated a seat 1 and gazing at the slide screen (first column), at the person at seat 2 (second column), and seat 3 (3rd column). Second row: head pose pdf model for the seat 1. Third row images of a person seated at seat 1 gazing at the slide screen, at the person at seat 2 and at the person at seat 3.

activity cues. Hung et al. (2008) suggest that automated versions of these measures using our recognized VFOA can estimate effectively the most dominant person in a meeting.

In Jayagopi et al. (2008) the automatic estimation of two aspects of social verticality (status and dominance) in small-group meetings using nonverbal cues was addressed. A systematic study about the effectiveness of automatically extracted cues (vocalic, visual activity, and VFOA) to predict both the most-dominant person and the high-status project manager is conducted.

5.1.3 Realtime head tracking and pose estimation

A robust system for VFOA recognition needs the head pose estimation to be accurate. This is a particularly difficult task especially in case of low resolution images and becomes even more challenging if an online VFOA recognition is required since the algorithms used should be designed to be as simple as possible in order to allow low processing time.

The IDIAP head pose estimation system developed in the context of the AMI project (Ba and Odobez, 2005) is reasonably robust since head tracking and pose estimation are considered as two coupled problems in a probabilistic framework. This system has been described in previous AMI/AMIDA deliverables. A drawback of this tracking system is its high computational cost which hinders an online VFOA estimation. The main bottleneck is due to the calculation of the likelihood function which requires to process several image patches (the ”particles”) with Gaussian and Gabor filters. To improve this system with respect to speed a new MSPF has been designed.

Let us denote by X_t the hidden state which represent the object configuration and by Y_t the associated observation extracted by the image at time t. A particle filter aims to estimate

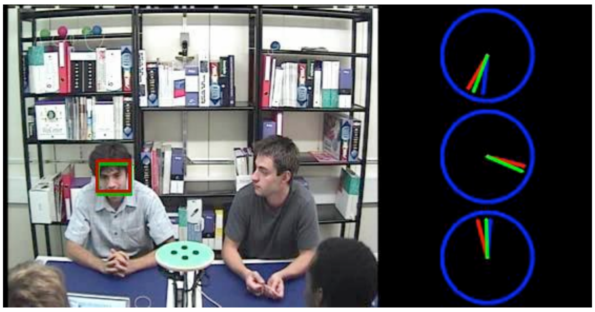


Figure 5: Snapshot of the output of the current head pose tracking system. The three clocks indicates the angle of out-plane (pan and tilt) and in-plane (roll) rotations.

the sequence of hidden parameters $X_{1:t}$ based only on the observed data $Y_{1:t}$. All Bayesian estimates of X_t follow from the posterior distribution $p(X_t|Y_{1:t})$. In the design of a particle filter, once X and Y are defined, a dynamical and an observation model must be specified.

In our specific case the particle filter is said Mixed States because the state model is defined such as X contains both continuous variables (to indicate head location and size) and discrete variables (to indicate the head pose represented by in-plane and out-plane head rotations). A dynamical model similar to the one described in Ba and Odobez (2005) is adopted to represent the evolution of states X_t over time. Instead, w.r.t. the observation model, the likelihood function $p(Y_t|X_t)$, which quantifies the consistency of the reference models with the current observation Y_t, has been modified. An observation $Y = (Y^{tex}, Y^{skin})$ is composed by texture features and skin color features computed on each image. Texture features are based on edge orientation histograms (EOH) (Levi and Weiss, 2004) and they replace the outputs of Gabor and Gaussian filters used in the previous system. This allows to greatly reduce the computational cost since integral histograms (Porikli, 2005) are adopted to compute EOH. Together with texture features a skin color model is learned offline and employed to classify pixels as *skin/not skin*. Then a binary mask of the image, computed by integral images, is used as color features to calculate the likelihood function.

The IDIAP multi-view face detector available with Torch3vision is used to initialize the tracking algorithm and, in case of long videos, to reinitialize it when the value of the likelihood function approaches to zero. The output of the face detector is also used to adapt the skin color model.

The current system, implemented in C++, is able to process both recorded videos and videos acquired by a webcam. An example of the output produced by our system is shown in Fig. 5. From a qualitative analysis of several videos we can say that the system provides quite satisfactory results both in terms of speed and accuracy even in cases of faces with low resolution (30×30 pixels). To quantify the performance of the tracker in terms of head pose estimation accuracy we use the AMI data of the CLEAR'07 task. The same protocol and the same performance measures described in the previous AMIDA deliverable D4.2 are adopted for conducting experiments. The results, shown in Table 1, are slightly worse than the one of the previous system in terms of pose estimation accuracy. However, the tracker runs close to realtime (at about 14fps on a standard laptop).

error	1R	1L	2R	2L	3R	3L	mean	AMI system
pan	13.3	14.1	13.3	11.5	7.4	9.9	11.5	8.8
tilt	8.6	6.1	13.1	7.5	6.4	9.8	8.5	9.9
roll	12.1	10.2	9.2	14.7	12.9	8.6	11.28	9.4

Table 1: Head pose estimation errors in degrees for person left (L) and right (R) in the AMI data. Errors indicate the difference of the angles between the head pose ground truth and the estimation.

5.1.4 Application to addressee detection

We also worked on the definition of possible scenario for user engagement enhancements in collaboration with University of Twente, for addressee detection.

In a meeting room, a fixed camera mounted on the top of the screen displaying the remote participant (RP) captures four co-located participants of the meeting. Two of them are on a side of a table, the other two on the other side. The meeting chairman will be one of the two persons seated near to the screen. The task will be to recognize in realtime when the addressee of the chairman is the RP or not. To achieve this goal the chairman VFOA will be used for addressee detection, together with speech information, as an input for a system developed by University of Twente.

5.1.5 VFOA conclusions and future work

The main results achieved this year are:

VFOA recognition
We introduced a model that uses head pose pdf to infer VFOA. The use of head pose pdf allowed a better head pose information representation.

The VFOA that we automatically extracted were used together with speaking activities to build features that were successfully used to estimate social verticality and the most dominant person in meetings.

In the next period, we will investigate whether the introduction of conversational events (monologue, dialog, group discussions) as context for our model can be profitable for VFOA recognition. Our current model make use of speaking status which might be temporally short and noisy to characterize VFOA dynamics.

Head tracking and pose estimation
A new module for joint head tracking and pose estimation of a single person has been developed. The tracker, implemented in C++, runs close to realtime.

An improved system which uses in the MSPF the output of the face detector to propose some image patches in high likelihood regions is currently under development. Future works will be devoted to achieve realtime performances for medium-high resolution images and to improve head pose estimation accuracy (e.g. through some features selection for texture features or through adapting the observation model to a specific person appearance). Moreover we plan to design a realtime VFOA recognition module to add in cascade to the head pose estimation one. It is worth noting that the current system provides head pose estimation for a single person and for a single camera. An extension to multiple persons and multiple cameras is a very challenging task due to realtime requirements and has not been considered so far.

Scenario for an online VFOA recognition demo
The scenario for a possible demonstration of our system has been defined. In the demo the real time VFOA recognition system will be used for addressee identification in collaboration with University of Twente.

5.2 Other tracking activities

5.2.1 View-independent head tracking

In Schreiber et al. (2008) a view-independent head tracking system is described. To separate background information from the human head, which typically shows a very non-rigid projection in 2D when e.g. turning from profile to frontal view, an active shape model (ASM) is used to parametrize the head. For this reason a head is modeled by 20 landmark points, which have been manually labeled for all training sequences, and an ASM is created by applying principal component analysis to the aligned labeled data. Contrary to the standard ASM approach where the gray values of the pixels are observed along the normal of the contour to detect learned histogram characteristics, in this approach a modified technique is applied directly on the gradient image and thus benefits from not only the fact, that there is an edge at a certain pixel position within the image, but also the direction of this edge. The hypotheses of the particle filter are initialized by a simple skin color detector. The major advantage of this principle is that a decrease in the computation time can be achieved, because of lower precision requirements of the ASM detector. Due to the ability of shape structure improvements during the measurement, where each shape hypothesis is locally adapted to the image data, a significant higher tracking performance than using only a plain ASM structure is achieved.

5.2.2 Omni-directional multiperson tracking

In the meeting scenario a tracking system has to deal with situations of heavy occlusions or reentries of participants. Therefore the system must be robust not only for the determination of human trajectories but also for reliable recovery of all identified persons. In Schreiber and Rigoll (2008) a novel approach has been developed combining a probabilistic particle filter framework with an heuristic simulated annealing technique ported to the tracking domain which is regarding all these needs. While the inter frame correspondence of objects, i.e. the assignment of identities, is handled by the simulated annealing approach, the particle filter architecture will be responsible for both the classification of an object to be a person as well as a stable tracking of the respective trajectory. An active shape model is utilized to create weights for the particles and thus serves as an object classifier. Our system has been evaluated on several video sequences showing meeting scenarios with a different number of participants. Quantitative numbers based on a tracking evaluation scheme show, that our system is capable of not only accurately determining the number of persons visible in each scene but also of precisely tracking each human and correctly assigning a label.

6 Visual Identification

Face recognition actually deals with two tasks: face verification and face identification. Face identification is particularly interesting in the context of AMIDA, as a consequence, we will describe several state-of-the-art face recognition that we will evaluate on a face identification task in the context of meetings, i.e. evaluating on the AMI Meeting Corpus. Several protocols for close-set and open-set identification experiments are proposed, and various experiments are performed with different baseline systems (PCA-LDA, GMM). Performance is then evaluated and compared between the different approaches.

6.1 Face Recognition

6.1.1 General Approaches for Face Recognition

Different approaches of face recognition can be categorized mainly into two main groups such as discriminative approaches and generative approaches. In discriminative approaches, the whole face region is taken into account as input data into face recognition system. Examples of this category are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA). In generative approaches, local features on the face such as nose, and eyes are collected by segmenting the image into several sub-images and then computing the feature vector for each sub-image (Heusch and Marcel, 2007). These features are then used as input data for a generative model such as a Gaussian Mixture Models (GMM), or a Hidden Markov Models (HMM).

6.1.2 Face Identification: open set vs close set

In a face identification system, we differentiate two different modes called close-set and open-set (or watch-list). In close-set mode, all the access images are supposed to belong to the clients, there are no impostor access. The objective of the system is to define the identity of the client, the answer is one of the N clients in the database. In open-set (or watch-list) mode, there are impostor accesses. The system must be able to know whether an access is performed by a client or by an impostor, and in case this is a client access, the system then identifies the client.

The identification system provides a score $\Lambda_I(X)$ corresponding to an opinion on the probe face pattern X to be the identity I. In close-set identification, we can recognize identity I^* corresponding to the probe face pattern X as follows:

$$I^* = \arg\max_I \Lambda_I(X)$$

. In open-set identification, the recognized identity I^* corresponding to the probe face is found as follows:

$$I^* = \begin{cases} \text{unknown} & \text{if } \Lambda_I(X) < \tau \forall I \\ \arg\max_I \Lambda_I(X) & \text{otherwise} \end{cases}$$

, where τ is a rejection threshold.

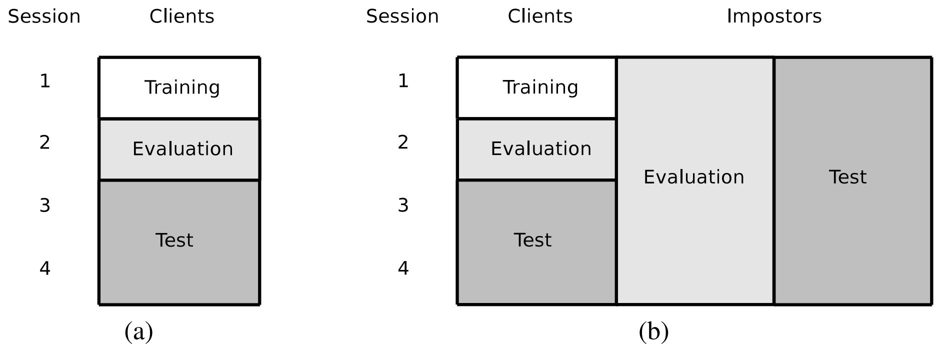


Figure 6: Partitioning of the database according to (a) close-set protocols *IS1*, *ES1* and *TS1* and (b) open-set protocols *IS2*, *ES2* and *TS2*.

6.2 Face Identification in the AMI Corpus

The AMI Meeting Corpus consists of meeting recordings collected in instrumented meeting rooms at the University of Edinburgh (U.K.), IDIAP Research Institute (Switzerland), and the TNO Human Factors Research Institute (The Netherlands). The meeting videos recorded from these three places are identified as ES, IS and TS sites.

Only the videos recorded by the close-up cameras are used in the face identification experiments. Thus we have one video per participant for each meeting session. As a consequence, face identification in the AMI Meeting Corpus is difficult because of the variety of head-pose, expression and occlusion.

6.2.1 Protocols

Protocols are necessary to perform the experiments. Each database could have several protocols and each protocol could have several variations. With different variations of the protocols, we can observe the behavior of the system in different conditions (e.g. the different data distributions of training set, evaluation set and test set). For the AMI Meeting Corpus database, we have created two different protocols, corresponding to close-set mode and open-set mode (see Fig. 6a and Fig. 6b respectively), for each site (IS, ES, TS) and one protocol for the jointed data of three sites (IETS).

6.2.2 Performance Evaluation

The evaluation of an open-set system is different from a close-set system. Open-set identification is also call ”watch list”. We measure the watch list performance using a client set G and two probe sets: client probes P_G and impostor probes P_N. The former is used to state the detection and identification rate equal as the fraction of probes in P_G that are detected at or above threshold t and recognized at rank r or better:

$$P_{DI}(t, r) = \frac{|\{p_j : rank(p_j) \leq r, s_{ij} \geq t, id(p_j) = id(g_j)\}|}{|P_G|} \qquad \forall p_j \in P_G$$

		Identification (close set)				
		I(r=1) (%)	I(r=2) (%)	I(r=3) (%)	I(r=4) (%)	I(r=5) (%)
PCAxLDA + metric	IS1	46	54.75	59.5	62.5	64.5
	ES1	38.31	45.56	51.25	54.81	57.5
	TS1	44.56	53.26	56.09	59.35	61.96
DCT + GMM	IS1	**93.5**	96.62	97.62	98.5	99.5
	ES1	**83.19**	88.69	91.19	92.37	93.25
	TS1	**77.39**	85.33	88.37	90.43	91.96

Table 2: Performance comparison for different rank values for close-set face identification with AMI Meeting Corpus.

where the rank of one probe is defined as the number of identities which have greater than or equal score to the probe than the matching entry:

$$rank(p_j) = |\{g_k : s_{kj} \geq s_{ij}, id(g_j) = id(p_j)\}| \qquad \forall g_k \in G$$

The impostor set is used to compute the false alarm rate as the fraction of probes from P_N whose score to any client model is at or above threshold:

$$P_{FA}(t) = \frac{|\{p_j : max_i s_{ij} \geq t\}|}{|P_N|} \qquad \forall p_j \in P_N \qquad \forall g_i \in G$$

Close-set identification is a special case of the watch list task where the false alarm rate is undefined and a pure identification rate specifies the performance. Formally for each probe p from P_G we sort the scores against all the client models G , and obtain the rank of the match. Identification performance is then computed as follow:

$$P_I(r) = \frac{|C(r)|}{|P_G|}$$

where

$$C(r) = \{p_j : rank(p_j) \leq r\} \qquad \forall p_j \in P_G$$

6.2.3 Experiment Results

This subsection describes the experiments we performed to compare the identification performance of different approaches presented previously: discriminative approach (PCA-LDA) and generative approach (GMM). The algorithms are implemented using Torch3vision [http://torch3vision.idiap.ch/] which is a machine vision library written in C++ and developed at IDIAP.

The face identification experiments are performed both on the close-set and open-set protocols. The idea is to identify who is participating to a meeting. Faces are extracted from the video and then normalized by eyes alignment, scaling and rotating technique. The extracted face images have the size of 64x80 pixels. However, the extracted faces are not always frontal, most of the time they are profiles or rotated faces. Thus, we have manually selected from the videos the ten best frontal faces per video. Because of missing

		Identification (open set / watch list)					
		DI(r=1) (%)	DI(r=2) (%)	DI(r=3) (%)	DI(r=4) (%)	DI(r=5) (%)	FA (%)
PCAxLDA + metric	IS2	31.5	36.5	38.25	39	39	51.56
	ES2	23.5	25.44	27.06	27.87	28.44	46.35
	TS2	37.82	43.04	44.35	45.65	47.17	66.25
	IETS	26.29	28.49	29.64	30.24	30.84	46.74
DCT + GMM	IS2	**82.75**	83	83	83	83	9.69
	ES2	**72.44**	73.06	73.06	73.06	73.06	16.04
	TS2	**71.85**	74.78	75.22	75.22	75.22	25.94
	IETS	**71.29**	72.83	72.89	72.89	72.89	21.61

Table 3: Performance comparison for different rank values for open-set face identification with AMI Meeting Corpus.

data (meeting session missing or some videos missing) or incapacity of finding frontal face from videos, the number of persons whose images are used for the experiments is reduced. The images which belong to participants whose data is not enough will be used for the world model training (named world-model images).

Experiments are performed with protocols *IS1*, *IS2*, *ES1*, *ES2*, *TS1*, *TS2* and *IETS* with PCAxLDA-metric or DCT-GMM. The close-set identification performance comparison is shown in Table 2. The open-set identification performance comparison is shown in Table 3.

From the experiments, we can see that GMM gives a much better result compared to the PCAxLDA in the small sets of data (IS, ES, TS separately) and the larger set (the joint data, IETS). It is clearly observed that PCAxLDA does not work for this kind of data, when the mismatched between training and testing conditions is important. The close-set identification performance of GMM is quite good, the best one for rank 1 is 93.5%. In open-set case, the detection and identification rate is 82.75% for rank 1 with the IS data set, corresponding false alarm rate is 9.68%.

6.2.4 Conclusion and Possible Future Work

We have proposed several protocols for close-set and open-set identification experiments. In all experiments, the performance of the identification system based on GMM is quite good and much better than the one based on PCAxLDA. It shows that the generative approaches (GMM for instance) are more robust to the variation of face pose and background than the discriminative approaches (PCAxLDA).

As future work, we would like to investigate non-frontal face identification. More precisely, it should be possible to design a system with a pose estimator and a generative model per pose. We will then be able to evaluate on the entire videos and not to restrict to a subset of frontal faces manually selected. However, this will require the annotation of the pose of all detected faces in the AMI Corpus.

6.3 Visual identification through AdaBoost video classification

Face detection research focused on improvements in AdaBoost-based methods and oriented in two directions 'training' software, in which we reached good results. Equally more important is the progress in research of novel features which exhibit better performance in face detection.

Local Rank Patterns (LRP; Hradis et al. (2008 (submitted)) – novel features for rapid object detection in images which based on existing features Local Rank Differences (LRD) – were thoroughly tested on frontal face detection task and on the facial part localization task. The performance of the classifiers was compared to the performance of the LRD and the traditionally used Haar-like features. The results show that the LRP surpass the LRD and the Haar-like features in the precision of detection and also in the average number of features needed for classification. Considering recent successful and efficient implementations of LRD on CPU, GPU (e.g., Polok et al. (2008)) and FPGA, the results suggest that LRP are good choice for object detection and that they could replace the Haar-like features in some applications in the future.

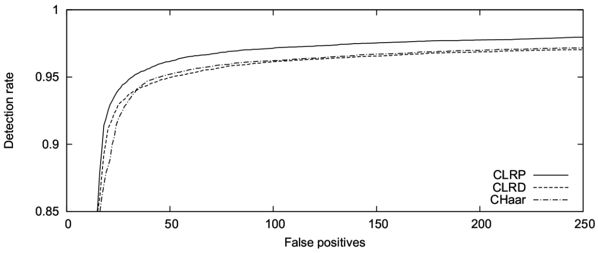


Figure 7: Receiver Operating Characteristics of three face detection classifiers on a dataset of 120 group photos containing 1628 faces. The classifiers were trained using the WaldBoost algorithm. CLRP uses Local Rank Patterns; CLRD uses Local Rank Differences; Chaar uses Haar-like features; Each classifier was trained and tested twelve times and the results were averaged.

6.3.1 NIST TRECVID evaluation participation

AMIDA participated in the 2008 NIST TRECVID evaluation campaign for video summarization task and event detection tasks.

The video summarization task is based on the classification techniques (among others AdaBoost-based classification has been used) and on feature extraction researched in WP4. Above these techniques, a simple mechanism was built that performs unsupervised clustering of video scenes and chooses one representative scene that represents all scenes occurring in the video. Additionally, when this mechanism does not reach the desired shortening ratio, the video shots are speeded up based on the estimated energy contained in the scenes.

The event detection task in TRECVID assumes fixed positioning of the video cameras and pre-defined events to be defined. Due to this fact, the event detection methods can be ‘specialized’. In the approach we used, the event detection is mostly based on tracking and motion analysis studies within AMIDA. Moreover, techniques, such as colour histogram, optical flow, background extraction, etc. are used in the event detection.

The other evaluation tasks in NIST TRECVID campaign, copy detection and video queries, were also covered although the methods themselves were less relevant to AMIDA.

The overall results of NIST TRECVID evaluations achieved through exploitation of AMIDA results are mostly very close to the best ones and the team always placed in the top part of the list.

7 Speaker Identification

In the last period, most of the activities in this area turned around 2008 NIST Speaker Recognition Evaluation http://www.nist.gov/speech/tests/sre/2008/.

7.1 BUT system for NIST SRE 2008

BUT submitted three systems to these evaluations, only to the primary short2-short3 condition. The primary system is a fusion of three sub-systems:

- 2 systems based on MFCC and factor analysis (Kenny et al., 2008)
- one system making use of SVM scoring of CMLLR and MLLR matrices of an ASR system.

The first contrastive systems differs only in calibration and the second contrastive system is a simplified version of the primary one (no ASR use).

The primary system did very well compared to the other submissions – see the black curve in the evaluation of the most important condition *telephone speech in training and test* (Figure 8 – available from 'Official SRE '08 Results' at the above mentioned URL). The presence of BUT group on the AMI/AMIDA ASR team was advantageously used in MLLR/CMLLR system, where we could build up on our knowledge acquired in these two projects. Full system description is available in Burget et al. (2008b).

Several other techniques, such as Heterogeneous Syllable Based Features, Phonotactic speaker identification with SVM modeling and Parametric and derivative kernels for GMM/SVM were tested but did not bring significant improvement over the factor analysis system, so they were not part of the final submission.

7.2 Dealing with different training/test conditions in speaker identification

TNO modified its 2006 speaker recognition system on a number of points. The aim has been to have a single system that can cope with all the different acoustical conditions under evaluation in SRE-2008. In order to deal with non-English, Language Recognition Evaluation 2003 data was used in UBM and background. Additional overall robustness was obtained by using explicit side information, namely non/English, telephone/microphone and gender. This was carried out using the new bilinear fusion tools from Niko Brümmer. Two separate channel compensation projections were trained for telephone and microphone, and combined into a single projection, so that still a single system was obtained. As a final improvement, Wiener filtering of microphone recordings was applied. All these improvements lead to development results dropping from 6% EER to below 4%. The evaluation results showed a very constant performance behavior over all 8 NIST conditions of interest.

In order to deal with the many conditions of interest, a new evaluation methodology is proposed at the NIST SRE-2008 workshop. Here, the trials of different acoustic conditions are weighted according to a pre-defined scheme, such that the actual amount of

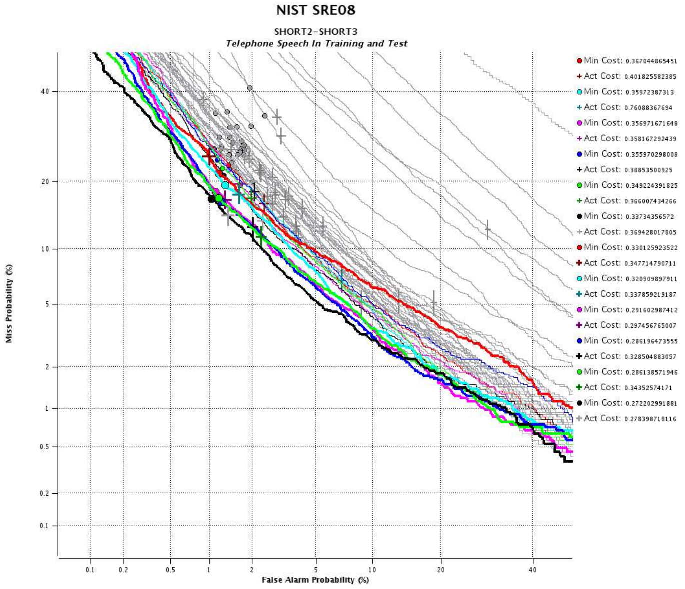


Figure 8: DET curves from NIST SRE 2008 evaluations short2/short3 condition, telephone speech in both training and test. AMIDA's curve is in black.

trials per acoustic condition (targets or non-targets) has no influence on the overall score. This methodology allows for computation of classical Cdet, Cdetmin and EER, as well as application-independent Cllr and Cllrmin. Using this new methodology, the TNO system performed very competitively with a condition-equalized EER of under 5%.

7.3 Speaker identification at the JHU 2008 summer workshop

Other significant work was done at the JHU 2008 summer workshop where L. Burget (BUT Speech@FIT research director) headed work-group Robust Speaker Recognition Over Varying Channels.

The research concentrated on utilizing the large amount of training data currently available to research community to derive the information, that can help discriminate among speakers and discard the information that can not. The world's best researchers in the area (Niko Brümmer, Spescom DataVoice, South Africa; Patrick Kenny, CRIM, Canada; Jason Pelecanos, IBM, USA; Doug Reynolds, MIT, USA; Robbie Vogt, QUT, Australia) participated in the group.

The results in the 4 subgroups (Diarisation using Joint Factor Analysis; Factor Analysis Conditioning; SVM-JFA and fast scoring; Discriminative system optimization) significantly pushed forward the state of the art and are likely to drive the research and development in this area in years to come. The details are summarized in the final workshop talk[2].

[2] http://www.clsp.jhu.edu/workshops/ws08/groups/rsrovc/

8 Gestures and Actions

8.1 Human action recognition

Building on the work described in the AMIDA deliverable D4.2 we extended the approach on pose estimation towards human action recognition. The main ingredients of our approach are HOG-like silhouette descriptors (Section 8.1.1) and Common Spatial Patterns for a discriminative approach to action classification (Section 8.1.2). This approach results in a score of 96% on a standard action data set. Moreover reasonable performance can be obtained by only training the recognizer on only a small set of persons. More details on the results can be found in Section 8.1.3. We conclude and discuss future work in Section 8.1.4. More detailed information can be found in Poppe and Poel (2008)

8.1.1 HOG-like descriptors

The starting point for the HOG-like descriptors are silhouettes, which are assumed to be given, for instance by using the work of Thurau (2007) or Zhu et al. (2006). Based on this silhouette a bounding box is computed in such a way that the height is 2.5 times the width. This bounding box is divided in 4x4 non overlapping cells. For each cell an 8 bin histogram of silhouette gradients is computed, each bin covers a range of 45 degree range, see Figure 9. All these histograms are concatenated and the resulting vector is normalized afterwards. This gives a 128-dimensional descriptor of the silhouette, and each action will result in a temporal sequence of such descriptors.

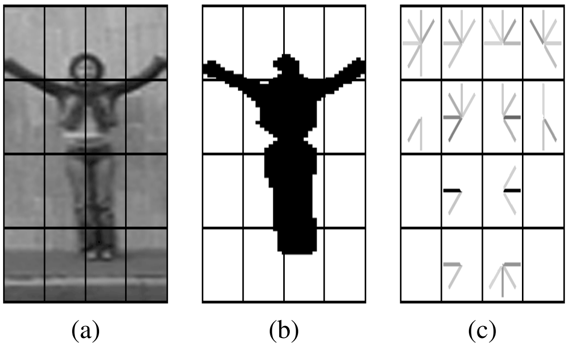


Figure 9: Silhouette descriptor, (a) image, (b) mask and (c) the boundary orientations, spatially binned into cells. Normal vectors are shown for clarity.

8.1.2 Common Spatial Patterns Classifier

Common Spatial Patterns (CSP) is a spatial filter technique often used in classifying brain signals (Müller-Gerking et al., 1999). It transforms temporal feature data by using differences in variance between two classes. After applying the CSP, the first components of the transformed data have high temporal variance for one class, and low temporal variance for the other. For the last components, this effect is opposite. When transforming the feature data of an unseen sequence, the temporal variance in the first and last k components can be used to discriminate between the two classes. It should be remarked that k depends on the classification problem under consideration.

Based on the CSP technique, we design discriminating functions $g_{a,b}$ for every action a and b with $a \neq b$. First we calculate the CSP transformation $W_{a,b}$, then we apply $W_{a,b}$ to each action sequence of class a and b. Afterwards, for each action sequence the normalized temporal variance in the first and last k components is calculated. This results in a single $2k$-dimensional vector, normalized for the length of the sequence. Next, we calculate the mean of these training vectors for action a and b, $\bar{a}$ and $\bar{b}$, respectively. In order to compute $g_{a,b}(x)$ for a new action sequence x, we use the same procedure and first apply $W_{a,b}$ to x. We then calculate then calculate the normalized variance in the first and last k components, which gives a vector x' of length $2k$. Finally, $g_{a,b}(x)$ is defined as follows:

$$g_{a,b}(x) = \frac{||\bar{b} - x'|| - ||\bar{a} - x'||}{||\bar{b} - x'|| + ||\bar{a} - x'||} \tag{1}$$

Evaluation of a discriminant function gives an output in the $[-1, 1]$ interval. Note that $g_{a,b} + g_{b,a} = 0$. Now the action sequence is classified by evaluating all discriminant functions between pairs of a and b over all actions:

$$g_a(x) = \sum_{a \neq b} g_{a,b}(x) \tag{2}$$

and x is classified as action the action a for which $g_a(x)$ is maximal.

8.1.3 Results

The for evaluating the approach described above we used the Weizmann action dataset (Blank et al., 2005). This set consists of 10 different actions, each performed by 9 different persons (see also Figure 10). Each action sequence takes approximately 2.5 seconds. There is considerable intra-class variation due to different performances of the same action by different persons. Most notably, the run, skip and walk actions are performed either from left to right, or in opposite direction. The trials are recorded from a single camera view, against a static background, with minimal lighting differences. Binary silhouette masks are provided with the dataset.

On this dataset we performed a leave-one-out cross validation, where each of the 9 folds corresponds to the all action sequences of one person. This gives 8 training sequences for each of the 10 actions, hence for each discriminating function $g_{a,b}$ there are 16 training sequences. For the value of the hyper-parameter k (c.f. Section 8.1.2) we took the value 5. The performance obtained by this approach is 95.56%. In total 4 action sequences were misclassified.

We also tested our approach on a limited number of training persons and uses the other persons for testing. The results can be found in Table 4.

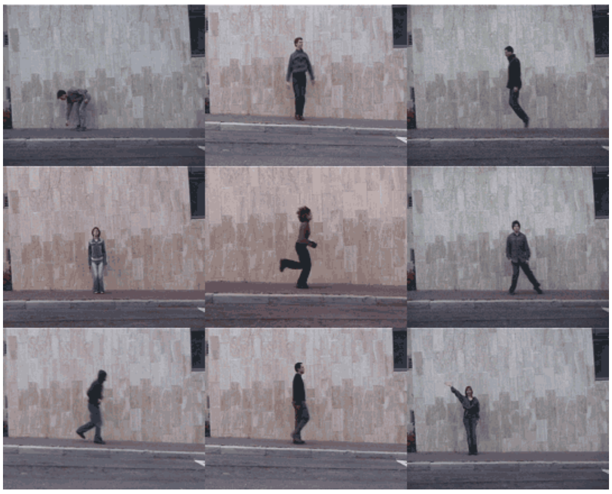


Figure 10: Example frames from the Weizmann dataset. Different subjects perform actions bend, jack, jump, p-jump, run, side, skip, walk and wave1.

Training Subjects	Performance
1	64.72%
2	77.82%
3	81.83%
4	84.60%
5	86.63%
6	89.01%
7	91.39%
8	**95.56**%

Table 4: Classification performance of our CSP classifier on the Weizmann dataset, using different numbers of training subjects.

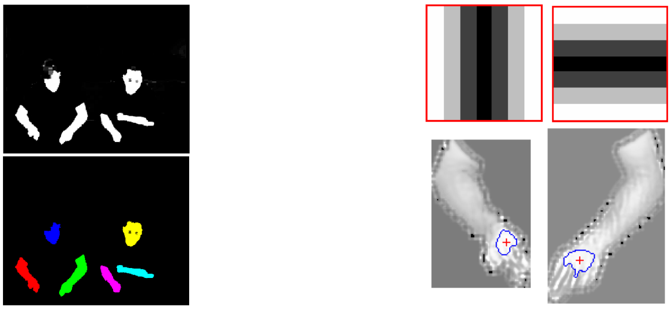
8.1.4 Conclusions and Future Work

The CSP based approach described above shows state of the art performance and generalizes well over unseen persons. Given it's simplicity the training and classification complexity are low, classification can be done in real time.

The next step will be to adapt the approach described towards actions relevant in the AMIDA context.

8.2 Recognition of poses and gestures using motion trajectories

Work has also been conducted on methods for dynamic gesture processing (see Jiřík, 2008). Since the number of all possible classes in AMI data and their inner-class variability are too big we limited our solution to a specific class. This class covers so-called Speech Supporting Gestures (SSG). They were chosen since they provide information about which parts of a speaker's dialog they wish to emphasise.



(a) (b)

Figure 11: (a) Skin probability image and connected components. (b) Convolution kernels (zoomed) and convolved images

As in many other image processing problems, this one can be divided into several stages. The basic framework for dynamic image recognition is as follows: Single Image Processing (region of interest, or ROI, localization) Dynamic Processing (tracking and dynamic characteristics calculation and adaptation) Classification.

8.2.1 Single image processing

The aim here is to find the positions and sizes of image regions that correspond to body parts.

The localization task is carried out using color-based segmentation. A Single Gaussian Method trained from two sets (skin and non-skin pixel colors) outputs the Skin Probability Image (SPI) where connected components are possible occurrences of ROIs (see Fig 11a). Regions whose size falls below a certain threshold are discarded.

The remaining regions need to be assigned to body parts expected to be present in the image. The number of persons in every sequence is assumed to be two and it is also assumed that in the beginning of every sequence both persons remain in regular positions (head above hands and hands not switched while each person occupies their own 'half' of the image). This approach has proven to be sufficient for further steps in the recognition.

The positions of the participants' hands can be refined by localizing palms instead of hands (entire skin-colored regions). Suppose we have a sub-image containing a hand which has been masked by an appropriate SPI. In such a sub-image there will be a flat region of skin color corresponding to the elbow (and potentially arm) and a region with a certain number of protrusions as a consequence of inter-finger distance and shadows. Such finger (and hence hand) regions are detected by convolving the sub-image with a set of kernels as in Fig 11b.

8.2.2 Dynamic processing

Once the hand regions are identified they are consequently tracked so their trajectories are found using the overlapping boxes approach. The essence of this method is that in two subsequent images the two occurrences of the same object lie so close to each other that the bounding boxes overlap.

By examining the trajectories, a certain degree of jitter was uncovered which could cause unsatisfactory results in the recognition stage. To eliminate this, a double exponential filter was employed which ensured that the positions of tracked objects remained stable when in a rest phase, thus preventing unwanted sub-trajectories from being created. In a dynamic phase (hand movement, etc.) it ensured a smoother trajectory.

8.2.3 Classification

Each sub-trajectory may represent some gesture therefore it is formalized by using a feature vector sequence. The two features used are velocities in x- and y-axis directions. Over 30 trajectories were manually segmented and modelled using one Gaussian Mixture Model for each of the two basic SSG classes: one for gestures in a horizontal direction (M_H) and one for gestures in a vertical direction (M_V). For every unknown sequence $O = (o_1, o_2 ... o_n)$ the log-likelihood of being emitted by the horizontal or vertical model can be computed as follows:

$$p(O|M_X) = \sum_{i=1}^{n} \frac{-log(p(o_i|M_X))}{n} \tag{3}$$

where M_X is one of the SSG models.

As an auxiliary metric for validation of O belonging to a particular SSG class, we define the periodicity of O in this manner: assume some vector $w = (w_1, w_2 ... w_n)$ in which (w_i) are indices of winning distributions in model M_X for o_i where M_X is the model with higher log-likelihood $p(O|M_X)$. The periodicity is then the number of sub-sequences $(w_i, w_{i+1} ... w_j)$ for which the following conditions are met:

$i \geq 1,\ j \leq n,\ j - i \geq C$
$w_k = w_{k+1}$ where $k = i ... j - 1$
$w_{i-1} \neq w_i$
$w_{j+1} \neq w_j$

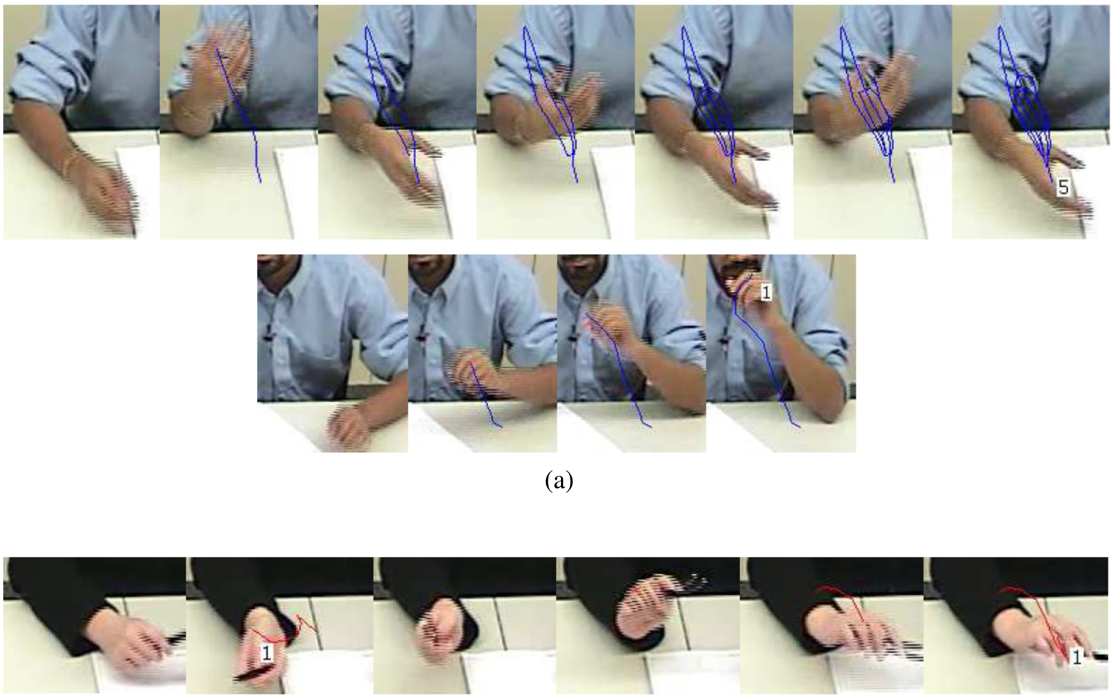
In the first condition we can find a constant C which defines the minimal length of such a sub-sequence (period).

It has been observed that the higher the number of periods, the more likely the unknown sequence will be a SSG class representative.

8.2.4 Results and conclusions

Since the approach outlined in this article was focused only on one specific gesture class, the number of false alarms is low and can be reduced even more by increasing the threshold of periodicity measure for some activity being a Speech Supporting Gesture. Fig. 12a shows the recognition of a vertical gesture.

The performance of our algorithm is also affected by the segmentation method. Fig. 12b shows how a correct gesture trajectory may happen to be split into two separate activities



(a)

(b)

Figure 12: (a) Speech Supporting Gesture and no-gesture activity. Blue trajectory denotes the gesture/activity is led in vertical direction. The number at the end of each trajectory is the periodicity metric. (b) A correct gesture missed due to wrong segmentation. Red trajectory denotes the gesture was led in horizontal direction.

with low periodicity numbers.

Like in many other algorithms and methods on the field of image recognition this one is very sensitive to parameter settings. However, when properly initialized it can give good results. Some important steps remain to be improved: mainly exact effectivity and reliability measures.

9 Social Signals

Not only spoken words are containing information for participants in a meeting, also the tone in the voice, the facial movements or gestures in Pentland (2004). These signals are as important as the spoken words themselves (Pentland, 2007). In this section we describe two research fields in the domain of social signals in AMIDA.

9.1 Multimodal Laughter Detection

We have worked further on laughter detector integrating the information from audio and video. Some new results have been reported, including experiments using alternative machine learning techniques (Reuderink et al., 2008) and proof that using temporal features is highly beneficial (Petridis and Pantic, 2008). The results aimed at determining the level on which the fusion of the two modalities needs to be conducted are still inconclusive as the results attained for decision- and feature-level fusion are highly comparable for the selected portion of AMI data. We intend to conduct a broader experimental study using larger portion of AMI data as well as a portion of SAL data, in order to test the generalisability of the developed method.

9.1.1 Conclusion and Future Work

Continuation of the research on audiovisual laughter detection. The level on which the fusion of audio and visual modality needs to be conducted remains a challenging problem. Modeling temporal correlations between the two modalities is another important and challenging issue to be researched for the rest of AMIDA.

9.2 Dominance/Activity Detection

In every group of people a ranking of the people is established which is derived from the different social signals of the participants after a very short time. Two of the important factors for the ranking are the dominance and activity of the persons. These levels of dominance/activity should be detected by using pattern recognition methods as hidden markov models. Compared to earlier work in AMI (Rienks and Heylen, 2006; Zhang et al., 2005) we are using low level features for the recognition.

9.2.1 Evaluation of Annotation

A small test set for inter-annotator agreement was used for the evaluation. This set was annotated by five different persons and best average kappa value one annotator against the rest is 47.4%. If only two of these five annotators are compared the kappa increases to 60.9% which is a moderate agreement. Therefore the annotation seams to be quiet robust and consistent, so that it can be used for the training of different statistical models.

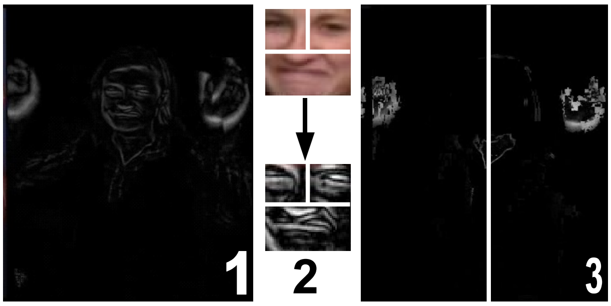


Figure 13: Visualization of applied global motion feature (Arsić et al., 2007)

Table 5: Evaluation of different Hidden Markov Model and feature configurations.

Model and feature configuration	RR (in %)	FER (in %)	AER (in %)
Audio, S=20	54.9	47.7	47.2
Global Motion, S=20	36.6	63.7	64.4
Skinblob, S=20	28.6	71.3	66.6
Audio & GM, S=20	49.2	55.8	63.6
Audio & GM & SK, S=20	42.9	57.0	61.9
GM & SK, S=20	40.5	60.1	60.7

9.2.2 Features

One of the features we use are global motions (Zobl et al., 2003) which are calculated on each full image and on a fixed sub region of each frame. The main drawback of this approach is that the information about the face or hand positions are not applied for adjusting the sub region. The new features which are described in Arsić et al. (2007) solve this problem by using a Rowley approach for finding the face and a condensation algorithm for face tracking in the video stream. The face is split into three parts, shown in stage two of figure 13, and for each of them the global motions are calculated. In a third stage the face is removed from the image and the hand movements are estimated by again using the global motions. And so we got six parts from each video stream where the global motions are calculated from.

An additional semantic feature we are applying to the dominance/activity detection are slide changes. Therefore a region in the centre view is defined, where the slides are shown on the screen, and in this region fast changes are detected. This feature should help us to segment the meeting into smaller parts, where the level of dominance could be estimated.

9.2.3 Results

Table 5 shows the achieved rates for different models and different feature combinations. The table points out that the most relevant information comes from the audio channel. The model with only audio information achieves the best result with 47.7% frame error rate. The models which are using visual stream only are not as good as the audio one. The combination of acoustic and visual features does not lead to further improvements of the results. If both visual features are combined the frame error rate decreases about three percent absolute from 63.7% to 60.1%.

9.2.4 Conclusion and Future Work

Initial results indicate that the models and features are not yet optimized for the problem of dominance/activity detection in meetings. Work will focus on improving these.

Therefore more evaluations about the different feature groups and the influence of them has to be evaluated. Also more complex and more optimized models will be used in the future. An online version of it is currently depending on the development of the algorithm which extract the features from the audio and video streams. The used approach should be capable for a online system but some adaptation will be needed.

9.3 Overview of Methods for Human Spontaneous Behaviour Analysis and Social Signals Processing

An AMIDA team and their colleagues from the MIT and the University of Illinois at Urbana-Champagne published a 'state-of-affairs' paper on how far are we from attaining automatic analysis of human spontaneous behaviour, necessary for realisation of human-centered intelligent interfaces (Pantic et al., 2008). With her colleagues from the IDIAP and MIT, Maja Pantic published a survey about social signals, their function and how they can be analysed automatically (Vinciarelli et al., 2008a,b). To the best of our knowledge, these are the first survey papers ever published on this topic. The team is working on an extended, journal version of this survey.

10 ICT in Healthcare Team Communication

This work presents a different application for AMIDA technologies social science research and intervention in meetings. Traditionally AMIDA has focused on fast meeting review and teleconferencing applications.

Through our manual encoding method presented here we can test the utility of AMIDA technologies for social science research and healthcare team communication interventions.

10.1 Introduction

Healthcare teams are associated with improved outcomes for patients (West et al., 2002). The increasing complexity of healthcare requires teams who have highly developed creative problem solving abilities (McFadzean, 2002). Creative solutions emerge out of the interactions between people (Kinnaman and Bleich, 2004). Consequently observing how health professionals interact is a growing area of research (Ellingson, 2003).

The extraction and processing of observational data from video or audio is resource intensive (Yin, 1984). For example the extraction of basic data on the quantity and sequence of contributions made by meeting participants requires intensive human processing to identify who is speaking and for how long, speaker sequence and turns and who are the dominant group members. Consequently there lies an important role for Information and Communication Technologies (ICT) in this area. This includes the development of algorithms to automatically code interesting events and process the logs of events to extract useful statistics. As well as being more efficient, ICT could facilitate interventional projects where rapid or real-time feedback to participants on their mode of interaction is required.

This work will demonstrate how we can recreate the key events related to quantity of contribution in a team discussion automatically. In Section 10.2 we describe both the manual and automated processes and later we compare the automatic process with the manual process.

10.2 Methods

Meetings of up to eight participants were recorded. They were conducted in a variety of room sizes and shapes. Some participants participated remotely through a speaker phone in conference call mode. Typical meeting length ranged between one and two and a half hours.

10.2.1 Manual Encoding

For manual encoding, two HD video cameras (Sony HDR-SR7, built-in microphone MICREF LEVEL set to NORMAL, built-in NP-FH100 battery pack for power supply, VCL-HG0737C wide conversion lens x0.7) on tripods recorded the meetings from opposite corners of the room.

The video, in highest quality Sony AVCHD format, was converted to 44.1 kHz .wav audio format and SD .wmv video format using ImTOO MPEG Encoder Standard v. 5.1.2.

The two .wav audio files were imported into MathWorks MATLAB R2008a and simple cross-correlation was used to determine the time offset between the two video files. The result was written to a .txt file. This is much faster and more accurate than manual synchronisation.

The two .wmv video files were imported into Noldus Observer XT 7.0, along with the time offset, and speaker times were manually logged from the two video views.

10.2.2 Automated Encoding

Eight (2 x blue, 2 x green, 2 x pink – most popular!, 2 x black) Apple iPod Nano 8GB 3rd Gen. (v. 1.1.3 firmware, voice memo record quality set to HIGH, Sony ECM-C115 tie pin microphone with windshield, Belkin TuneTalk amplifier) were used to record 44.1 kHz .wav audio files for automated encoding.

An iPod was fitted to each speaker, with the microphone placed approximately 10 cm below the mouth pointing upwards. The microphone from another iPod was placed near the telephone speaker.

The .wav files were imported into MATLAB. Simple cross-correlation was used to approximately synchronise each pair of recordings. Each 100 ms of the recording was further cross-correlated to obtain the correct synchronisation time offsets for the dominant speaker/noise during that time. An algorithm was employed that used a threshold at various auto-correlation and cross-correlation metrics for each 100 ms, to obtain the dominant speaker or silence for that 100 ms. The algorithm rejected non-mutual sounds such as microphone movement noise that were negligible in other microphones. Finally silence periods during a single speaker talking such as breath periods were filled in to obtain what humans interpret as continuous speaking.

10.3 Results

Results of manual and automated encoding were in the following form:

Event	Time Start	Time End	Manual Data Extracted	Automated Data Extracted
1	0.1 sec	0.3 sec	A coughs	A speaks
2	4.1 sec	12.1 sec	B speaks	B speaks
3	12.1 sec	13.2 sec	B + C speaks	C speaks
4	13.2 sec	62.7 sec	C speaks	C speaks

Time saved by automated encoding is approximately two to three hours per hour of recorded video.

10.4 Discussion and Conclusion

Efficiency: Automated methods have the potential to significantly reduce the time required to code observational data for team interactions. Use of iPods prohibit real-time processing however they are smaller, simpler and more fashionable than devices that transmit rather than store microphone audio.

Accuracy: Manual coding identifies multiple speakers talking at the same time whereas the automated method presented here only determined the loudest speaker at one time. It also interpreted some mutual non-speech sounds as speech. The automated procedure provides excellent time precision. While the use of iPods is more intrusive than video recorders, they enable close to 100% accuracy in speaker segmentation. Speaker segmentation from a single audio channel is difficult and currently has approximately 20% error rate (AMIDA, 2007).

While at an early stage, we have demonstrated that the automated analysis of audio improves the efficiency of data extraction from video. The next stage for the research will enable faster processing that reliably segments multiple speakers to allow future real time analysis of meetings. In parallel to this will be our identification of proxy signals for team interactions to allow future real time analysis of meetings. The future benefits of real-time ICT in healthcare team communication include interventional studies of team interactions and the rapid review of meetings for late participants. While we are focusing our work on healthcare team communication, the same techniques may be useful for meetings in other contexts.

11 Motion tracking and visualisation

TNO Science & Industry focused on the visualization of information regarding a specific kind of multiparty interaction: a group of professionals working towards a secure environment in a soccer stadium. This topic is introduced in deliverable D4.2, Chapter 11, 'Motion tracking and visualisation'.

11.1 Goals

The first goal of this research is to develop concepts, based on our previous work for AMIDA to assist security staff with handling the data overload around a match. This data overload takes place directly around the game, but also when training, (de)briefing and evaluating the performance of the staff on other moments. Although the events in the game itself might influence the security situation, currently, this research does not concern the game itself. The second goal is to demonstrate innovate ways of visualizing the information in it's context. This is being taken to it's extreme by incorporating state of the art 3D technologies.

11.2 Recording

This year, we focussed on obtaining a dataset from a soccer game. An instrumentation plan for the command room of soccer club ADO The Hague was written. The goal was to obtain data that would enable us to automatically detect and classify events and behaviour inside the command room. Additionally, data from the security camera's is used to present the interaction in the command room in the context of the events outside of the command room.

11.3 Analyzing: motion zones

When collecting these amounts of data of a single event – the soccer game – it is obvious that several types of data analysis can be done. In previous AMIDA work we presented 'motion zones' over time in a 'segmentlist'. This is a very robust approach to analyzing huge amounts of video, while incorporating advanced pattern recognition. The same approach was applied here.

11.4 Visualization

The second use for this data is to show it in the spatial context. To this end we had a 3D model made of the stadium. In this model, we placed the video streams according to their respective camera's in the real stadium. This gives the user the impression that they are looking at a coherent flow of information.

11.5 Dissemination

This model, together with a selection of the video will be put in a visionary short video. To this end we wrote a movie script that clearly conveys the message to interested parties. This video will showcase the broad range of possibilities when a data-overload is turned into an immersive and interactive information management tool. We intend to invite interested parties to discuss the implications of this research and to discuss the value-chain that is in effect here.

11.6 Future work

We intend to build a service that would allow us to analyse a complete soccer game, and provide interested parties with summaries and searchable databases for the purpose of training and evaluation. Additionally, we intend to do research on event detection and signal processing.

12 Summary and future work in WP4

WP4 is concerned with the development of reliable audio, visual, and audio-visual integration and recognition tools for the automatic extraction of information from raw data streams. This involves multistream fusion, synchronisation, and recognition methods from the different audio-visual information sources.

This report shows how new algorithms have been developed or existing algorithms adapted and extended to process data from the AMIDA domain. Furthermore, it shows how such algorithms and been adapted to address realtime requirements.

The progress made in the second year and associated future plans for the seven main research themes is described below.

12.1 Automatic speech recognition

The primary focus of work during the last period is on achieving a real-time on-line speech recognition system: a working HUBable on-line ASR system is now available. Other work includes general ASR system improvements, the enhancement of the AMIDA ASR infrastructure so that future ASR R&D work can be performed more efficiently and a continued effort to disseminate AMIDA ASR technology to the general research community (webASR, Bob and CTS tutorial). Furthermore, AMIDA technology has been evaluated through participation in the Dutch LVCSR evaluation N-best 2008 (Karafiát et al., 2008).

The real time ASR system will be developed actively to increase speed and robustness. Investigations are ongoing into suitable normalisation techniques for on-line features. Time synchronisation needs to be addressed. In particular, we expect the beamformer to create particular online segmentation difficulties that will need to be addressed. The real time system will also take on a diarisation component.

The webASR system will be tested by a number of individuals (both within AMIDA and external to the project). Once this process has been completed and any necessary alterations have been made, the system will go ‘live’ and be available to the general public. Throughout these two stages (testing and initial release) we will investigate other features which can be incorporated into the system.

12.2 Keyword spotting

The work in the area of keyword spotting and spoken term detection in the last period had three important parts: construction of hybrid recognition networks for combined word and sub-word recognition (and hence indexing and search), integration of acoustic keyword spotter into the Hub infrastructure and improvements of the system for detection of out-of-vocabulary words in the output of speech recognizer.

In the next period, the use of multigrams for acoustic on-line keyword spotting will be investigated, which should be significantly better than the current use of phoneme-based models. We will also investigate more deeply the issues of score normalization and calibration.

12.3 Speaker diarisation

The robustness of the offline system was improved further by the use of Modulation-filtered Spectrogram (MSG) features, giving over 20% relative improvement in Diarisation Error Rate (DER).

In a first series of experiments we studied how some parts of the ICSI Speaker Diarisation offline system could be simplified or replaced without a significant loss of performance so that online processing would be possible. An initial form was tested on AMI data using non-overlapping 2 second windows, and gave comparable results to the offline system using only 50 seconds of training for each speaker.

We are planning to make further moves towards online processing, investigating the use of a Universal Background Model (UBM) and its robustness to different meeting rooms. Furthermore we are planning to consider alternatives to conventional beamforming algorithms like feature-based channel combination.

12.4 Focus of Attention and Tracking

VFOA recognition
A new model has been developed that uses head pose posterior distribution over the entire head pose space to infer VFOA instead of single head pose allowing a better head pose information representation. A new module for joint head tracking and pose estimation of a single person has been developed which runs close to realtime and integrates a face detector for automatic (re)initialization. Future work will focus on achieving realtime face detection performance for medium-high resolution images and to improve head pose estimation accuracy. In addition, we plan to design a realtime VFOA recognition module to add in cascade to the head pose estimation one.

In the next period, we will also investigate whether the introduction of conversational events (monologue, dialog, group discussions) as context for our model can be profitable for VFOA recognition. Our current model make use of speaking status which might be temporally short and noisy to characterize VFOA dynamics. Furthermore, recognizing the conversational events will be a first step toward addressee detection.

A demonstration scenario for our system has been defined. In the demo the real time VFOA recognition system will be used for addressee identification in collaboration with University of Twente.

12.5 Identification of persons

12.5.1 Audio-based

Three systems were submitted to the NIST SRE 2008 evaluations. The primary system did very well compared to the other submissions. Future work will investigate the robustness of speaker identification across varying channels and from short speech segments. Work on production version of SpkID is also planned, so that this technology could be integrated into AMIDA demonstrations.

12.5.2 Video-based

Object detection and identification algorithms using the AdaBoost (WaldBoost) classification methods have been improved. The new features' computational aspects have also been investigated on CPU (PC platform), GPU (NVidia platform), and FPGA platforms.

AMIDA participated in the NIST TRECVID evaluations initiative with all four tasks (classification of scenes, video summarization, video queries, and copy detection). From the tasks evaluated so far, our scores were among the top scores and while the evaluation results cannot be disclosed, AMIDA performance is above average. Future work will investigate the automatic identification of more general meeting environments in order to allow automatic recognition of the meeting room topology (and possibly also partially determining its geometry).

Several protocols for close-set and open-set identification experiments have been investigated. A number of experiments were conducted which showed that the generative approaches (GMM for instance) were more robust to the variation of face pose and background than the discriminative approaches (PCAxLDA). As future work, we will investigate non-frontal face identification. More precisely, it should be possible to design a system with a pose estimator and a generative model per pose. We will then be able to evaluate on the entire videos and not to restrict to a subset of frontal faces manually selected.

12.6 Gestures and Actions

Building on the work described in the AMIDA deliverable D4.2 we extended the approach on pose estimation towards human action recognition. The main ingredients of our approach are HOG-like silhouette descriptors (Section 8.1.1) and Common Spatial Patterns for a discriminative approach to action classification (Section 8.1.2). This approach results in a score of 96% on a standard action data set. Moreover reasonable performance can be obtained by only training the recognizer on only a small set of persons. Future work will focus on adapting the approach described towards actions relevant in the AMIDA context.

Further work will also be conducted on the modelling and classification of motion trajectories, including gestures, using HMMs. HMMs of classes of behaviour are created using annotated trajectories. The early results of this research indicate that the information about complex object behaviour of objects – including the motion and gestures of humans – can be discovered (Mlích and Chmelař, 2008; Mlích, 2008). These existing approaches will be applied to the AMIDA domain.

12.7 Social Signals

Work has continued on laughter detection integrating the information from audio and video. Some new results have been reported, including experiments using alternative machine learning techniques (Reuderink et al., 2008) and proof that using temporal features is highly beneficial (Petridis and Pantic, 2008). Continuation of the research on audiovisual laughter detection. The level on which the fusion of audio and visual modality needs to be conducted remains a challenging problem. Modeling temporal correlations between the two modalities is another important and challenging issue to be researched for the rest of AMIDA.

Automatic recognition of participant dominance and activity using pattern recognition methods such as HMMs has been investigated. Initial results indicate that the models and features are not yet optimized for the problem of dominance/activity detection in meetings. Work will focus on resolving this issue. Furthermore, more complex and more optimized models will be used in the future. An online version is being pursued.

12.8 Summary

Important progress has been made in all seven main research themes in the second year of AMIDA, most notably the availability of realtime, online automatic speech recognition.

Beside these seven main research themes, we also addressed side topics of motion tracking and visualisation in a soccer control room (Sec. 11) and ICT in healthcare team communication (Sec. 10) to show how AMIDA research results and algorithms can be transferred to problems outside the meeting domain.

References

M. Karafiát, J. Kopecký, F. Grézl, T. Mikolov, and L. Burget. System description of Brno ASR system for NBEST 2008 Dutch evaluation. In *accepted to Dutch ASR evaluation workshop*, 2008.

P. N. Garner. Silence models in weighted finite-state transducers. In *Interspeech* Garner (2008c). IDIAP-RR 08-19.

P. N. Garner. A weighted finite state transducer tutorial. Idiap-Com Idiap-Com-03-2008, IDIAP, 2008b.

P. N. Garner. Silence models in weighted finite-state transducers. Idiap-RR Idiap-RR-19-2008, IDIAP, 2008c. To appear in Interspeech 2008.

A. El Hannani and T. Hain. Automatic optimisation of speech decoder parameters. *IEEE Signal Processing Letters (submitted)*, 2008.

K. Kumatani, U. Mayer, T. Gehrig, E. Stoimenov, J. McDonough, and M. Wölfel. Adaptive beamforming with a minimum mutual information criterion. Idiap-RR Idiap-RR-74-2007, IDIAP, 2007a.

K. Kumatani, J. McDonough, S. Schacht, D. Klakow, P. N. Garner, and W. Li. Filter bank design based on minimization of individual aliasing terms for minimum mutual information subband adaptive beamforming. Idiap-RR Idiap-RR-77-2007, IDIAP, 2007b.

K. Kumatani, J. McDonough, S. Schacht, D. Klakow, P. N. Garner, and W. Li. Filter bank design for subband adaptive beamforming and application to speech recognition. Idiap-RR Idiap-RR-02-2008, IDIAP, 2008a.

K. Kumatani, J. McDonough, D. Klakow, P. N. Garner, and W. Li. Adaptive beamforming with a maximum negentropy criterion. Idiap-RR Idiap-RR-06-2008, IDIAP, 2008b.

K. Kumatani, J. McDonough, D. Klakow, P. N. Garner, and W. Li. Maximum negentropy beamforming. Idiap-RR Idiap-RR-07-2008, IDIAP, 2008c.

K. Kumatani, J. McDonough, B. Rauch, P. N. Garner, W. Li, and J. Dines. Adaptive beamforming with a maximum negentropy criterion. Idiap-RR Idiap-RR-29-2008, IDIAP, 2008d.

K. Kumatani, U. Mayer, T. Gehrig, E. Stoimenov, J. McDonough, and M. Wölfel. Minimum mutual information beamforming for simultaneous active speakers. Idiap-RR Idiap-RR-73-2007, IDIAP, 2007c.

F. Grézl, M. Karafiát, S. Kontár, and J. Černocký. Probabilistic and bottle-neck features for LVCSR of meetings. In *ICASSP'07*, Hononulu, 2007. ISBN 1-4244-0728-1. URL http://www.fit.vutbr.cz/research/view_pub.php?id=8249.

F. Grezl and P. Fousek. Optimizing bottle-neck features for LVCSR. In *2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4729–4732, 2008. ISBN 1-4244-1484-9. URL http://www.fit.vutbr.cz/research/view_pub.php?id=8601.

G. Garau and S. Renals. Combining spectral representations for large-vocabulary continuous speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(3):508–518, 2008a. doi: http://dx.doi.org/10.1109/TASL.2008.916519. URL http://dx.doi.org/10.1109/TASL.2008.916519.

G. Garau and S. Renals. Pitch Adaptive Features for LVCSR. In *Proc. Interspeech*, 2008b.

W. Li, J. Dines, M. Magimai-Doss, and H. Bourlard. Neural network based regression for robust overlapping speech recognition using microphone arrays. Idiap-RR Idiap-RR-09-2008, IDIAP, 2008. Submitted for publication.

W. Li, J. Dines, and M. Magimai-Doss. Robust overlapping speech recognition based on neural networks. Idiap-RR Idiap-RR-55-2007, IDIAP, 2007a.

W. Li, M. Magimai-Doss, J. Dines, and H. Bourlard. Mlp-based log spectral energy mapping for robust overlapping speech recognition. Idiap-RR Idiap-RR-54-2007, IDIAP, 2007b. Submitted for publication.

W. Li. Effective post-processing for single-channel frequency-domain speech enhancement. Idiap-RR Idiap-RR-71-2007, IDIAP, 2007. Submitted for publication.

M. Karafiat, L. Burget, T. Hain, and J. Cernocky. Application of CMLLR in narrow band wide band adapted systems. In *Proc. INTERSPEECH 2007*, page 4, 2007.

M. Karafiat, L. Burget, T. Hain, and J. Cernocky. Discrimininative training of narrow band - wide band adapted systems for meeting recognition. In *accepted to INTERSPEECH 2008*, page 4, 2008.

S. Huang and S. Renals. Towards the application of hierarchical bayesian models on language models for automatic speech recognition. In *Nonparametric Bayes workshop at ICML'08*, 2008a.

S. Huang and S. Renals. Unsupervised language model adaptation based on topic and role information in multiparty meetings. In *Interspeech*, 2008b.

S. Huang and S. Renals. Modeling topic and role information in meetings using the hierarchical dirichlet process. In *Machine Learning for Multimodal Interaction (MLMI'08)*, 2008c.

S. Huang and S. Renals. Hierarchical pitman-yor language models for asr in meetings. In *In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'07)*, 2007.

T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, V. Wan, and J. Vepa. The AMI system for the transcription of speech meetings. In *In Proceedings of ICASSP 2007*, Honolulu, Hawai USA, April 2007.

T. Hain, A. El Hannani, S. N. Wrigley, and V. Wan. Automatic speech recognition for scientific purposes – webASR. In *Proc. Interspeech 2008*, 2008.

V. Wan, J. Dines, A. El Hannani, and T. Hain. Bob: A lexicon and pronunciation dictionary generator. In *2008 IEEE Workshop on Spoken Language Technology - SLT 2008*, 2008.

I. Szöke, M. Fapšo, L. Burget, and J. Černocký. Hybrid word-subword decoding for spoken term detection. In *SIGIR/SSCS 2008 – 2nd workshop on Searching Spontaneous Conversational Speech*, 2008.

I. Szöke, P. Schwarz, L. Burget, M. Fapšo, M. Karafiát, J. Černocký, and P. Matějka. Comparison of keyword spotting approaches for informal continuous speech. In *Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology*, pages 633–636, 2005. URL http://www.fit.vutbr.cz/research/view_pub.php?id=7886.

L. Burget, P. Schwarz, P. Matějka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, and J. Černocký. Combination of strongly and weakly constrained recognizers for reliable detection of oovs. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008a. URL http://www.fit.vutbr.cz/research/view_pub.php?id=8494.

S. Kombrink. Out of vocabulary detection in large vocabulary continuous speech recognition using neural networks. Technical report, Brno University of Technology, Faculty of Information Technology, 2008.

S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals. Speech and crosstalk detection in multi-channel audio. *IEEE Trans. Speech Audio Processing*, 13(1):84–91, 2005.

S. O. Ba and J.-M. Odobez. Recognizing human visual focus of attention from head pose in meetings: A study. *IEEE Systems, Manand Cybernetics, Special issue on Human Computing (to appear)*, 2008a.

S. O. Ba and J.-M. Odobez. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *International Conference on on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008b.

S. O. Ba and J.-M. Odobez. Visual focus of attention estimation from head pose posterior probability distributions. In *the Inernational Conference on Mutli-media & Expo (ICME)*, 2008c.

H. Hung, D. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *the International Conference on Multimodal Interfaces (to appear)*, 2008.

D. Babu Jayagopi, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez. Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In *the International Conference on Multimodal Interfaces (to appear)*, 2008.

S. O. Ba and J.-M. Odobez. A Rao-Blackwellized mixed state particle filter for head pose tracking. In *ACM ICMI Workshop on Multimodal Multiparty Meeting Processing (MMMP)*, pages 9–16, 2005.

K. Levi and Y. Weiss. Learning object detection from a small number of examples: The importance of good features. In *IEEE International Conference on Computer Vision and Pattern Recognition, 2004 (CVPR 2004)*, volume 2, pages 53–60, 2004.

F. Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. In *IEEE International Conference on Computer Vision and Pattern Recognition, 2005 (CVPR 2005)*, volume 1, pages 829–836, 2005.

S. Schreiber, A. Störmer, and G. Rigoll. Omnidirectional tracking and recognition of persons in planar views. In *Fifteenth International Conference on Image Processing (ICIP)*, 2008. accepted for publication.

S. Schreiber and G. Rigoll. Omni-directional multiperson tracking in meeting scenarios combining simulated annealing and particle filtering. In *Eighth International Conference on Automatic Face and Gesture Recognition*, 2008. accepted for publication.

G. Heusch and S. Marcel. Face Authentication with Salient Local Features and Static Bayesian Network. In *IEEE / IAPR Intl. Conf. On Biometrics (ICB)*, 2007.

M. Hradis, A. Herout, and P. Zemcik. Local rank patterns novel features for rapid object detection. In *ICCVG*, 2008 (submitted).

L. Polok, A. Herout, P. Zemcik, M. Hradis, R. Juranek, and R. Josth. Local rank differences - image feature implemented on gpu. In *Advanced Concepts for Intelligent Vision Systems (ACIVS 2008)*, 2008.

P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):980–988, 2008.

L. Burget, P. Matějka, O. Glembek, P. Schwarz, V. Hubeika, M. Fapšo, M. Karafiát, and J. Černocký. But system description: Nist sre 2008. In *2008 NIST Speaker Recognition Evaluation Workshop*, June 2008b.

R. Poppe and M. Poel. Discriminative human action recognition using pairwise CSP classifiers. In *Proceedings of the 8 th IEEE International Conference on Automatic Face and Gesture Recognition (FGR'08)*, 2008.

C. Thurau. Behavior histograms for action recognition and human detection. In *Human Motion: Understanding, Modeling, Capture and Animation*, number 4814 in Lecture Notes in Computer Science, pages 271–284, Rio de Janeiro, Brazil, October 2007.

Q. Zhu, S. Avidan, M.-C. Yeh, and K.-T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 2*, pages 1491–1498, New York, NY, June 2006.

J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg. Designing optimal spatial filters for single-trial EEG classification in a movement task. *Clinical Neurophysiology*, 110 (5):787–798, May 1999.

M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proceedings of the International Conference On Computer Vision (ICCV'05) - volume 2*, pages 1395–1402, Beijing, China, October 2005.

Leoš Jiřík. Recongnition of poses and gestures. Technical report, Faculty of Information Technology, Brno University of Technology, 2008.

A. Pentland. Social dynamics: Signals and behavior. In *Proc. ICDL*, 2004.

A. Pentland. Social signal processing [exploratory DSP]. *Signal Processing Magazine, IEEE*, 24(4), 2007.

B. Reuderink, M. Poel, K. Truong, R. Poppe, and M. Pantic. Decision-level fusion for audiovisual laughter detection. In *Proceedings of Joint Int'l Workshop on Machine Learning and Multimodal Interaction (MLMI'08)*, 2008. accepted for publication.

S. Petridis and M. Pantic. Audiovisual laughter detection based on temporal features. In *Proceedings of ACM Int'l Conf. Multimodal Interfaces (ICMI'08)*, 2008. accepted for publication.

R. Rienks and D. Heylen. Automatic dominance detection in meetings using easily obtainable features. In *Revised Selected Papers of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI 2005*, pages 76–86, Berlin, Germany, 2006. Springer Verlag.

D. Zhang, D Gatica-Perez, S. Bengio, and D. Roy. Learning influence among interacting Markov chains. In *NIPS*, 2005.

M. Zobl, F. Wallhoff, and G. Rigoll. Action recognition in meeting scenarios using global motion features. In J. Ferryman, editor, *Proceedings of the 4th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-ICVS)*, pages 32–36, 2003.

D. Arsić, B. Schuller, and G. Rigoll. Suspicious behavior detection in public transport by fusion of low-level video descriptors. In *Proceedings of the 8th International Conference on Multimedia and Expo (ICME)*, 2007.

M. Pantic, A. Nijholt, A. Pentland, and T. Huang. Human-centred intelligent human-computer interaction (hci): How far are we from attaining it? *Int'l Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008.

A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signal processing: State of the art and future perspectives of an emerging domain. In *Proceedings of ACM Int'l Conf. Multimedia (MM'08)*, 2008a. accepted for publication.

A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signals, their function and automatic analysis: A survey. In *Proceedings of ACM Int'l Conf. Multimodal Interfaces (ICMI'08)*, 2008b. accepted for publication.

M. A. West, C. Borrill, J. Dawson, J. Scully, M. Carter, A. Anelay, M. Patterson, and Waring J. The link between the management of employees and patient mortality in acute hospitals. *International Journal of Human Resource Management*, 13:1299–1310, 2002.

E. McFadzean. Developing and supporting creative problem-solving teams: part 1 a conceptual model. *Management Decision*, 40:463–475, 2002.

M. L. Kinnaman and M. R. Bleich. Collaboration: aligning resources to create and sustain partnerships. *Journal of Professional Nursing*, 20:310–322, 2004.

L. L. Ellingson. Interdisciplinary health care teamwork in the clinic backstage. *Journal of Applied Communication Research*, 31:93–117, 2003.

R. K. Yin. *Case study research*. Sage Publications Inc., California, USA., 1984.

AMIDA. D4.2: Report on implementation and evaluation of audio, video, and multimodal algorithms. Technical report, AMI Consortium, 2007.

J. Mlích and P. Chmelař. Trajectory classification based on hidden markov models. In *Proceedings of 18th International Conference on Computer Graphics and Vision*, pages 101–105. Lomonosov Moscow State University, 2008. ISBN 595560112-0. URL http://www.fit.vutbr.cz/research/view_pub.php?id=8680.

J. Mlích. Trajectory classification. In *Proceedings of the 14th Conference STUDENT EEICT 2008*, Volume 2, pages 211–213. Faculty of Electrical Engineering and Communication, Brno University of Technology, 2008. ISBN 978-80-214-3615-2. URL http://www.fit.vutbr.cz/research/view_pub.php?id=8681.

W. Li and H. Bourlard. Non-linear spectral contrast stretching for in-car speech recognition. In *Interspeech-Eurospeech* Li and Bourlard (2007b). IDIAP-RR 07-53.

W. Li and H. Bourlard. Non-linear spectral contrast stretching for in-car speech recognition. Idiap-RR Idiap-RR-53-2007, IDIAP, 2007b.

D. Vijayasenan, F. Valente, and H. Bourlard. Integration of tdoa features in information bottleneck framework for fast speaker diarization. In *Interspeech 2008* Vijayasenan et al. (2008d). IDIAP-RR 08-26.

D. Vijayasenan, F. Valente, and H. Bourlard. Agglomerative information bottleneck for speaker diarization of meetings data. In *IEEE Automatic Speech Recognition and Understanding Workshop* Vijayasenan et al. (2007b). IDIAP-RR 07-31.

D. Vijayasenan, F. Valente, and H. Bourlard. Combination of agglomerative and sequential clustering for speaker diarization. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* Vijayasenan et al. (2007c). IDIAP-RR 07-51.

D. Vijayasenan, F. Valente, and H. Bourlard. An information theoretic approach to speaker diarization of meeting data. Idiap-RR Idiap-RR-58-2008, IDIAP, 2008c. Submitted for publication.

D. Vijayasenan, F. Valente, and H. Bourlard. Integration of tdoa features in information bottleneck framework for fast speaker diarization. Idiap-RR Idiap-RR-26-2008, IDIAP, 2008d. Published in Interspeech 2008.

D. Vijayasenan, F. Valente, and H. Bourlard. Agglomerative information bottleneck for speaker diarization of meetings data. Idiap-RR Idiap-RR-31-2007, IDIAP, 2007b.

D. Vijayasenan, F. Valente, and H. Bourlard. Combination of agglomerative and sequential clustering for speaker diarization. Idiap-RR Idiap-RR-51-2007, IDIAP, 2007c.

D. A. van Leeuwen and M. Huijbregts. The ami speaker diarization system for nist rt06s meeting data. *In Machine Learning for Multimodal Interaction*, 4299:371–384, 2006.

D. A. van Leeuwen. A note on performance metrics for speaker recognition using multiple conditions in an evaluation. In *NIST 2008 Speaker Recognition Evaluation Workshop*, 2008.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Secaucus, NJ, USA., 2006.

S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11:305–345, 1999.