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 D4.5: WP4 work in year 3

Due date: 30/09/2009 **Submission date:** 30/09/2009
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)	

D4.5: WP4 work in year 3

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.

D4.5 is the final report on the implementation and evaluation of the different audio, video, and multimodal algorithms conducted in the final year of AMIDA. The report shows how new algorithms have been developed and existing algorithms have been adapted and extended to process data from the AMIDA domain.

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 to, or implemented for, the AMIDA domain with particular emphasis on realtime requirements.

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 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 (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 final year and outline of this deliverable

The outcome of WP4 is a set of multimodal recognisers for the 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 final 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 result is the provision of realtime ASR. Since the last deliverable, the realtime ASR system has been significantly overhauled with many substantial improvements throughout the whole of the system. The details are documented in Sec. 2.

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

2 Automatic Speech Recognition

2.1 Introduction

The description of the work conducted this year is split into several parts. Sec. 2.2 describes the work focussing on real-time and on-line aspects of the AMIDA requirement. Sec. 2.3 describes continued work in improving speech recognition in general, which includes our participation in the NIST Rich Transcription 2009 meeting transcription evaluation. Sec. 2.4 describes the work done in exploiting AMIDA speech recognition technology.

2.2 Real-time ASR

This year the real-time ASR system has been massively overhauled with many substantial improvements throughout the whole of the system. Sec. 2.2.1 describes improvements made to the real-time ASR framework, Sec. 2.2.2 describes an implementation of VTLN and CMLLR for real-time ASR, Sec. 2.2.3 describes the improvements made to the Juicer speech decoder and Sec. 2.2.4 describes the automatic decoder optimisation work.

2.2.1 Tracter – the real-time framework

One year ago, Tracter contained processing components for a simple single (lapel or close-talking) mic ASR system. During the last year, we have progressed this to support the full array microphone system. One major aspect of this has been the provision of wrappers for external components to be seamlessly inserted into the data flow chain, and in turn enables more advanced features such as bottleneck MLP, fast VTLN (Sec. 2.2.2) and HLDA. Wrappers have also been written for the Torch3 machine learning library enabling support for MLP based voice activity detection (VAD). Wrappers for HTK have also been written which import many of the features of HTK into Tracter including (C)MLLR speaker adaptation, full-covariance decoding and run-time feature expansion, normalisation and transformation.

Tracter's internal cepstral normalisation components have been continually improved, allowing on-line cepstral normalisation. The result of all the above is that several stages of feature processing and re-combination, including VAD, can now run in a single process.

Tracter is also now able to collate speaker information from the ICSI online-speaker ID system (see AMIDA D4.4). This is obtained over a TCP socket and propagated to the hub via the decoder and java wrapper.

The concept of time-stamping has been overhauled throughout the Tracter chain and into the decoder. Time stamps are received from the source (a beamformer implemented as a VST plug-in) and propagated through Tracter to the decoder and the hub. This in turn enables hub consumers to receive words with accurate word timing information and a tag indicating who is speaking.

2.2.2 Fast GMM based VTLN and Fast CMLLR

Most off-line ASR systems run in more than one pass. The first pass generates "preliminary" word hypotheses which are used for vocal tract length normalization (VTLN) and (constrained) maximum likelihood linear regression (C)MLLR. Later passes use VTLN and (C)MLLR to achieve better accuracy.

Work was done to find different implementations of VTLN and CMLLR which do not require the first pass decode.

Fast VTLN: We choose an implementation published in Welling et al. (1999). It is based on training warping factor specific HMMs.

First, warping factors for training data were estimated using standard iterative search of best maximum likelihood giving warping factor (WF). The data was split into the set of warp factors (WF) based bins in range from 0.85 to 1.15 with step 0.02. First part was to train WF specific HMMs. We used MAP adaptation from UBM (single GMM with 32 diagonal Gaussians trained on all data) to derive specific models for each WF. Acoustic models are retrained using MMI (Maximum Mutual Information) criterion. Features are the same as for the system (PLP) but without CMN/CVN.

During the offline recognition, we accumulate likelihoods for each WF specific model. The winning model gives the resulting WF.

For our experiments we chose a simplified version of the AMI LVCSR system built for NIST Rich Transcription benchmark in 2007. The decoding process could be split into these parts:

- P1 - 1 pass decoding
- P2 - VTLN estimation
- P3
    1. PLP+LCRC features generation
    2. Phoneme Alignment using P1 output
    3. CMLLR estimation
    4. Lattice generation HDecode - 2gram LM
    5. Lattice expansion 3g LM
    6. Lattice expansion 4g LM
    7. Generation one best output from 4g lattices

We investigated the replacement of VTLN estimation by Fast VTLN implementation. Table 1 shows comparison with CMLLR Adaptation and without. It seems that degradation of performance could be partly corrected by CMLLR.

CMLLR: Typically, input for CMLLR adaptation algorithm is a phoneme string generated from output of first decoding pass. Our approach is to replace LVCSR decoding by phoneme recognizer based on a neural network (NN). We used simply 4 Layer NN and played with dependency of WER and size of NN.

The test system was the same as above using fast VTLN based features.

System	AMI segmentation WER [%]	reference segmentation WER [%]
LVCSR VTLN - no CMLLR,2gram LM	32.3	29.2
Fast VTLN - no CMLLR,2gram LM	33.0	30.2
LVCSR VTLN - CMLLR,2gram LM	30.8	27.7
Fast VTLN - CMLLR,2gram LM	31.0	28.0
LVCSR VTLN - CMLLR, 4gram LM	27.8	24.6
Fast VTLN - CMLLR, 4gram LM	27.9	24.8

Table 1: Comparison of VTLN based on GMM and standard search algorithm

Phoneme string generation	RT factor	reference segmentation - WER [%]
From 1 pass decode	4.9	24.8
250k NN	0.03	25.4
500k NN	0.04	25.3
1M NN	0.06	25.1

Table 2: Phoneme recognizer as an input for adaptation.

2.2.3 Juicer

Whilst our speech decoder, Juicer, has always been very configurable and capable, it was known to be a little slow. This year a major effort was made to speed up Juicer. This has resulted both in a faster version of the standard Juicer decoder, and a completely new "pull" based decoder capable of even faster performance.

Furthermore the tight integration of Tracter with HTK means that Juicer is able to use all of the features and functionalities what come with HTK. This makes Juicer a fast alternative to HDecode. Fig. 1 shows the performance of the latest version of Juicer against HDecode using trigram language models. At low RTFs the degradation in WER is much smaller in Juicer than that of HDecode. Lower word error rates (WERs) at all real time factors (RTFs) can be achieved simply by using 4-gram language models instead.

The composition of WFSTs has been enhanced by removing a WFST's dependence upon a particular set of acoustic models. WFSTs built for juicer can now be used with any set of acoustic models without having to recompose the complete WFST.

These performance boosts contributed directly to the RT09 evaluation in the Spring of 2009, enabling more experiments to be performed faster on fewer computers (Sec. 2.3).

The functionality of Juicer has also been improved through the addition of partial trace back which, in a real-time system, means that transcriptions can be sent to the output as soon as a winning hypothesis is determined rather than waiting for the end of the segment. This reduces the overall latency of the ASR enormously.

The performance increases allow the system to be run in real time in demonstration environments.

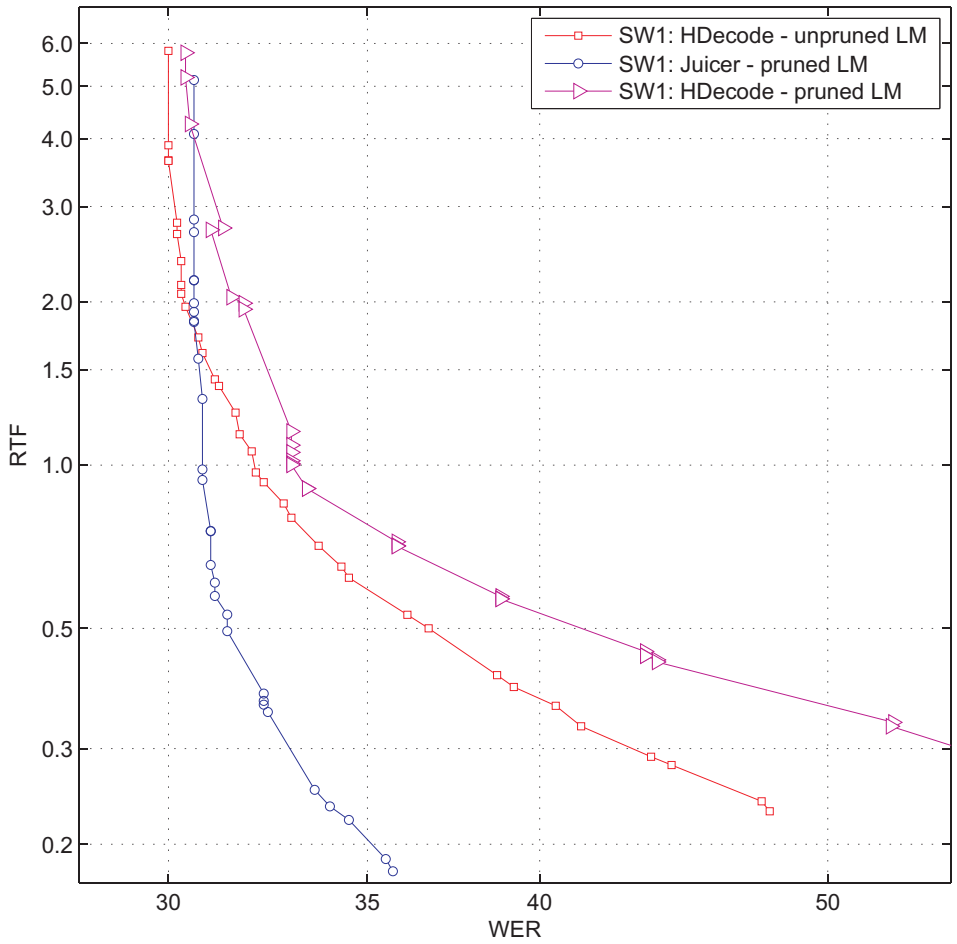


Figure 1: A comparison of Juicer and HDecode at various decoding speeds – using automatic decoder optimisation. The language model is a trigram.

2.2.4 Decoder optimisation

Speech decoders are complex with a large number of parameters to tune for performance and speed. Since different parameters affect the system differently, while possibly interacting with each other, then it is a complicated process to balance them all such that the system runs quickly and with the best possible accuracy. We investigated methods for automatic optimisation of such parameters. The objective is to find the optimal configuration of the decoder that yields minimal search errors for an given real-time factor. The approach investigates a method based on automatic tracking of that optimal curve (El Hannani and Hain, 2009). Experiments were conducted using HDecode and Juicer decoders on test set of conversation telephone speech and meeting data. Results such as in Fig. 1 have shown that high quality performance curves can be found allowing us to tune decoders exactly for a certain RTF value. In practice the search has found better configurations than those obtained through local optimisation by hand.

Task	RT07s WER [%] using sys07	RT07s WER using sys09	RT09s WER using sys09
IHM	24.9	23.4	27.4
MDM	33.7	29.3	33.2

Table 3: Performance of AMIDA system on RT09s eval.

2.2.5 Projuicer

The java wrapper that links the real-time ASR system to the Hub has been continuously maintained to incorporate both changes to the ASR system and to the hub infrastructure. The main tangible difference is the provision of time stamps and speaker metadata. Much of the work is summarised in Garner et al. (2009).

2.3 NIST RT09s evaluation

Once again AMIDA participated in the NIST RT09s evaluation. Advances this year have been made in beamforming, neural network based features and mappings, use of discriminatively trained features, speech activity detection, language modelling and improvements to decoding already described above. The improvement in performance between the 2007 and 2009 systems on RT07s eval is shown in table 3 along with the final result on RT09. A significant improvement this year is the speed of the overall system. The improvements made to the Juicer decoder meant that it was possible this year to eliminate the use of the much slower HDecode decoder from the system. The complete system runs in less than 13 times real-time with earlier outputs available in about 8xRT and 6xRT. Other improvements include the use of full meeting adaptation and a framework that runs the system completely automatically with one button push (in contrast, previous year's evaluations had to be run manually).

2.3.1 Digital microphone array

An eight element microphone array has been constructed using Knowles Acoustics SPM0205HD4 MEMS Digital microphones - to our knowledge the first of its kind. The MEMS digital microphones integrate the microphone, amplifier and analogue to digital convert on a single chip, producing a digital PDM output of the incident acoustic signal. This removes the need for the costly and bulky audio interface required with traditional analogue microphone arrays and means the device can be smaller, cheaper, and more robust to noise (since the signal is always in the digital domain) than its analogue equivalent. In order to evaluate the effect of the digital microphones on ASR performance, an analogue array, and the new digital array are used to simultaneously record test data for a speech recognition experiment. Initial results employing no adaptation show that performance using the digital array is significantly worse (14% absolute WER) than the analogue device. Subsequent experiments using MLLR and CMLLR channel adaptation reduce this gap, and employing MLLR for both channel and speaker adaptation reduces the difference between the arrays to 4.5% absolute WER.

2.3.2 Beamforming

We have experimented with various beamforming tools for MDM meeting channels. Mdmtrain07 training setup were beamformed and standard VTLN PLP features were extracted. New HMM sets were made by single pass retraining approach from pretrained models based on AMI beamformed data.

NIST Rich Transcription 2007 data, which we used for testing, were processed in same way as in training. The AMI bigram LM trained during RT07 evals was taken for decoding. Results are shown in table 4.

train beamforming	test beamforming	WER
ami	ami	42.6
ami	icsibeam3.3	42.3
ami	icsibeam2.0	41.9
icsibeam3.3	icsibeam3.3	41.5
icsibeam2.0	icsibeam2.0	40.4

Table 4: Dependency of WER on beamforming

2.3.3 Meeting room selection

A number of the RT09 recordings include participants joining the meeting by means of a remove video conference system. High quality synchronised recordings were available for both rooms and therefore a system was developed to associate each speaker (as defined by the automatic segmentation) with a room. The system relies on there being a delay in the audio transmission over the video conference system and proceeds as follows :

1. Perform beamforming on the audio from each room to produce a single enhanced audio track for each room.
2. Perform speaker segmentation on room 1 audio.
3. For each speakers audio, on a frame by frame basis, calculate the maximum of the cross correlation between the audio from room 1 and room 2 (ie the delay between the two rooms). if delay > 0 (ie the audio occurs in room 1 before room 2), increment the room 1 count if delay < 0 (ie the audio occurs in room 2 before room 1), increment the room 2 count
4. Assign the speaker to the room with the highest count.
5. Discard segments from the speakers assigned to room 2
6. Repeat from 2) using segmentation from room 2 audio, and discard segments assigned to room 1 in step 5)

Because of the relatively long delays in the video conference transmission, a large frame size (2.5 seconds) was used.

NN training data	WER
30h	33.5
180h	31.9

Table 5: Effect of training data

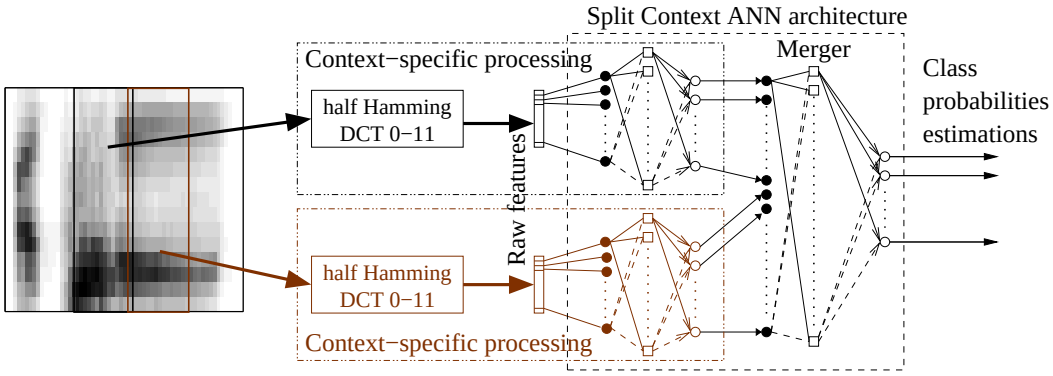


Figure 2: Block diagram of Split Context ANN architecture.

2.3.4 Neural Net based features

Using output of neural networks (NN) as features in Automatic Speech recognition system is becoming a widely used technique to improve a system accuracy (Hermansky et al., 2000). Typically, a NN produces a stream of phoneme or phoneme state posteriors which are further decorrelated by log followed by PCA or (H)LDA matrix. This feature stream is concatenated with standard features (PLP + derivatives) and used as a final features for speech recognizer.

In further research we have developed the Bottle-Neck (BN) architecture (Grézl et al., 2007). This is 5 Layer NN where the output is taken from middle, bottle-neck, layer. Inner product of NN is more suitable for full-covariance modeling and this structure allows better data compression.

To study the effect of varying the amount of NN training data we used VTLN PLP_0_D_A_T features with a HLDA transform concatenated with NN based features. HMM acoustic models were trained on the whole of ihmtrain07 training set (180 hours of speech). The NN was trained on a 30 hour subset and the full 80 hour set.

Table 5 shows 1.6% improvement if NN training data were increased to the same data amount used to train the acoustic models.

A split context architecture (SC) is shown in Fig. 2. We considered two different configurations of neural network:

- Context NNs which are standard probability estimators with a merger that had bottle-neck structure (SC-M BN).
- Bottle-neck in all stages.

Recognition results are shown in table 6.

HLDA-PLP	36.0
HLDA-PLP + baseline BN	31.7
HLDA-PLP + SC-M BN	30.6

Table 6: RT07 (ihmref) WER [%] of PLP features, baseline BN features and Split Context architecture with bottle-neck in the merger.

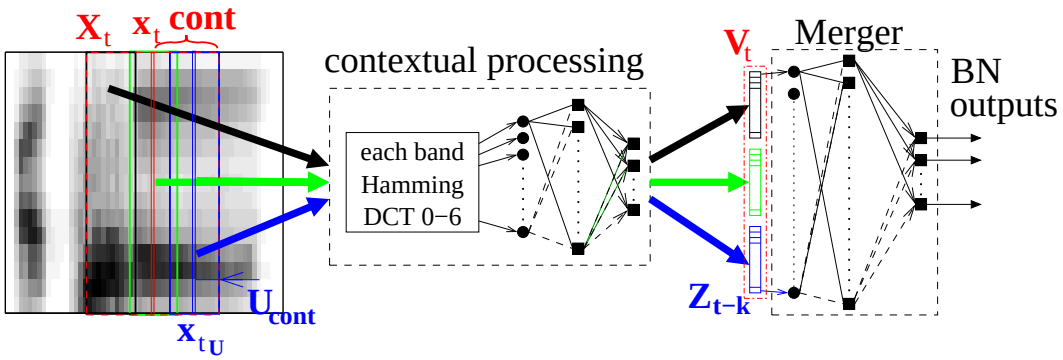


Figure 3: Block diagram of Universal Context approach.

Fig. 3 shows the universal context (UC) architecture. By replacing context-dependent ANNs by a general one a significant simplification was achieved: it is obvious, that the ANN does not have to be in the system several times. Instead, processing of the smaller – contextual – block is done frame by frame and stacked, and only desired frames are taken to form merger input. The number of trainable parameters in the system is therefore reduced, allowing for training of larger ANNs to reach the same number of trainable parameters in whole architecture. The stacking of context-independent ANN outputs is also convenient when experimenting with different numbers of temporal splits. Recognition results are shown in 7.

	Contextual ANN bottle-neck size				
features	50	60	70	80	90
HLDA-PLP + UC BN	29.5	29.4	29.5	29.4	29.4

Table 7: WER [%] of BN features generated by universal context architecture.

2.3.5 Discriminative training of NN features

Bottle neck is a feature extraction scheme based on discriminative training. Therefore, it is interesting to compare and combine our feature extraction technique with other discriminative training techniques used in speech recognition. Specifically, we have examined Minimum Phone Error (MPE) training of model parameters (Povey, 2003) and feature-level MPE (fMPE) (Povey, 2005). The comparison and combination with fMPE is particularly interesting as fMPE is an alternative discriminative feature extraction technique. However, while the neural net is trained to estimated phoneme state posterior probabilities for each frame in the case of BN features, in case of fMPE, an ensemble of linear

feature transformations (Zhang et al., 2006a) is discriminatively trained to optimize the MPE criterion, which is believed to be better related to our task of speech recognition.

Table 8 presents the results for three different feature sets:

- HLDA-PLP – baseline
- UC BN70_D – UC with 70 neurons in contextual ANN bottle-neck augmented with delta coefficients – one of our best performing feature sets based purely on BN processing
- HLDA-PLP+UC BN70 – feature set concatenating both the HLDA-PLP and the UC BN70 (no deltas) stream

features	Training			
	ML	MPE	fMPE	fMPE+MPE
HLDA-PLP	35.6	32.6	31.4	29.7
HLDA-SC BN50BN30	30.4	28.1	26.7	26.3
UC BN70_D	29.9	27.9	27.8	27.6
HLDA-PLP + UC BN70	29.4	27.5	26.9	26.1

Table 8: WER [%] of BN and HLDA-PLP features using different techniques.

Comparing the two discriminative feature extraction schemes, we see that the ML results obtained with UC BN70_D features (29.9% WER) compare favorably to fMPE HLDA-PLP (31.4% WER). Applying fMPE on top of BN feature extraction and MPE training of the models brings further significant gains. Highest gains are, however, obtained with fMPE and MPE applied on HLDA-PLP+UC BN70 features consisting of both BN and HLDA-PLP feature streams. This suggests that fMPE is able to extract additional complementary discriminative information contained in the "raw" features that was already lost during the BN processing.

2.3.6 Neural network based feature mapping

The neural network based feature mapping approach for overlap speech recognition was further investigated along two directions (Li et al., 2009):

1. Mapping across feature domains, i.e., unlike our previous work the input feature domain and output feature domain are not same. For instance, training the neural network which learns the mapping from log mel filter bank energy domain (extracted from delay-sum beamformed speech signals for target speaker and interfering speaker) to MFCC domain (extracted from closed talking microphone signals). Overlap speech recognition studies on MONC corpus showed that mapping log mel filter bank energies to MFCC yields the best system performance.

Segmenter	Number of segments	RT07s IHM WER
Reference segmentation	4527	29.3
RT07 IHM segmenter	2717	32.6
RT09 IHM segmenter	4541	31.7

Table 9: Improvement in speech activity detection

2. Regression based speech separation, where the neural network is trained to map jointly log magnitude spectrum estimated from target speaker speech signal and interfering speaker speech signal to the log magnitude spectrum estimated from closed talking microphone speech. During testing, given the target speaker speech and interfering speaker speech the neural network output is used to reconstruct speech signal (an estimate of clean speech signal). The reconstructed speech signal was evaluated in terms of "source-to-distortion ratio", and the features extracted from the reconstructed speech were used for overlap speech recognition. This approach was compared against standard delay-sum beamforming approach and binary masking approach, and was found to be consistently better than the two standard approaches.

2.3.7 Speech activity detection

The AMIDA IHM speech/silence detector is based on an MLP which is post processed by an HMM. Improvements were made in the speech/silence segmenter by training the MLP on the extra AMIDA data that was available this year and by tuning it better for the evaluations (specifically RT07s). Experiments were performed on a first pass adapted system which used PLP features and a CMLLR transform for speaker adaptation. Results in table 9 show an improvement in 0.9% WER from using the new segmenter.

2.3.8 Speech non-speech detection for privacy features

Modeling real-life social interactions using multi-modal sensor data is the central goal of this work. Capturing spontaneous, multiparty conversations, also referred to as personal audio logs, is a step towards this. However, recording and storing raw audio could breach the privacy of people whose consent has not been explicitly obtained. One way to address this is to store features instead of raw audio, such that neither intelligible speech nor lexical content can be reconstructed (Parthasarathi et al., 2009).

One of the key preprocessing in conversation analysis is speech/nonspeech detection. State-of-the-art speech/nonspeech detection use spectral-based feature estimated from short-term signal. From these features it is possible to reconstruct fairly intelligible speech or recognize lexical content. We investigated a set of four different privacy-sensitive (or privacy preserving) features estimated solely by temporal domain processing of speech signal, namely energy, zero crossing rate, spectral flatness, and kurtosis, for speech detection in multiparty conversations. Due to lack of an appropriate personal audio log data, we created a personal audio log scenario from the meeting room scenario, and the meeting

datasets and annotations were defined accordingly. The features were studied by modeling them individually and modeling them in combination with each other, with or without temporal context. We compared these features against state-of-the-art speech/nonspeech MLP-based detector (developed earlier in AMI) which uses standard spectral-based features. Our studies show that the privacy-sensitive features when modeled jointly with time context (around 500ms) can achieve a performance close to that of the spectral-based feature.

2.3.9 Full covariance modelling

The Gaussian mixture models used in acoustic modelling are usually assumed to have diagonal covariance matrices; however, it is known that increasing the number of covariance parameters can improve recognition performance. However, standard approaches to parameter estimation for full covariance matrices can lead to models which are poorly conditioned and over-fitted to training data. To remedy this we investigated a smoothing technique using diagonal covariance prior models, with a varying smoothing weight, τ.

We developed the technique using the AMIDA system for Conversational Telephone Speech recognition and carried out experiments on the 2001 Hub5 evaluation task. Full covariance models were trained with maximum likelihood (ML) training, and we compared the effects of using ML-trained diagonal priors, and also discriminatively trained diagonal priors using maximum mutual information (MMI). Results are shown in Table 10 and Fig. 4. It can be seen that the smoothed full covariance models achieve substantially improved performance over both the diagonal models and the full covariance models with no smoothing, with the best performance achieved when a discriminatively trained prior is used.

The full covariance technique was integrated into the AMIDA decoder for RT09, although it is not possible to identify the affect this had on the performance of the system as a whole.

Table 10: Selected WER results with full covariance models

System	ML prior	MMI prior
Diagonal	-	31.2%
Naive full covariance	31.5%	31.1%
$\tau = 10$	30.7%	30.7%
$\tau = 20$	30.5%	30.5%
$\tau = 40$	30.3%	30.2%
$\tau = 100$	30.3%	30.1%
$\tau = 200$	30.6%	30.1%
$\tau = 400$	30.9%	30.2%

2.3.10 Language modelling

Language models (LM) were trained using the same corpora as per RT07: no additional data of any kind was used. The only change was that the cut-offs were lowered slightly:

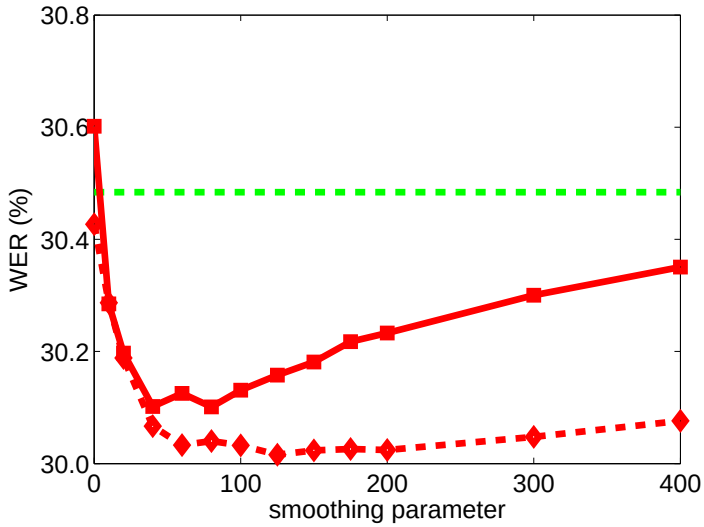


Figure 4: WER of ML-trained full covariance models initialised with varying smoothing parameter τ: using an ML prior (solid red) and an MMI prior (dashed red); compared with diagonal MMI-trained models (dashed green)

	2007 LM WER	2009 LM WER
RT07s IHM	37.9	36.7
RT07s MDM	36.2	35.9

Table 11: Improvement between 2007 and 2009 4-gram language models, evaluated on RT07s

i.e. the minimum count of 4-grams was lowered from 4 in previous evaluations to 3 in RT09. This led to larger but better LMs. Table 11 shows the 4-gram language model performance on RT07s and RT09s evals, while table 12 shows the perplexity and out of vocabulary rates of both LMs on both evals

We also investigated the application of a Bayesian language model based on nonparametric priors called the Pitman-Yor processes — the hierarchical Pitman-Yor process language model (HPYLM) (Teh, 2006) — for multiparty conversational meetings. In the HPYLM, Pitman-Yor processes are recursively placed as priors over the predictive probabilities in a n-gram LM, and the posterior distribution is inferred from the ob-

	2007 LM PPL	2007 LM OOV	2009 LM PPL	2009 LM OOV
RT07s IHM	87.8	0.74%	86.4	0.62%
RT09s IHM	73.1	0.30%	71.0	0.29%

Table 12: Out of vocabulary (OOV) rates and perplexity (PPL) of 2007 and 2009 LMs on RT07 and RT09 evals

served training data. By integrating out latent variables and hyperparameters, we are able to estimate smoother predictive probabilities in the HPYLM. We demonstrated the practical application of the HPYLM to large vocabulary automatic speech recognition (ASR) of conversational speech in multiparty meetings, indicating that this model can offer consistent and significant reductions in perplexity and word error rate (WER), compared to both an interpolated Kneser-Ney LM (IKNLM) and a modified Kneser-Ney LM. (MKNLM) (Huang and Renals, 2007). Moreover, we proposed a parallel training algorithm for the HPYLM (Huang and Renals, 2009), which enables us to efficiently work on large corpora using a large vocabulary for ASR.

However, it is still expensive to estimate an HPYLM, even with a distributed training algorithm. More recently, we presented a power law discounting language model (PLDLM) (Huang and Renals, 2010), which approximates to the HPYLM while not requiring a computationally expensive approximate inference process. The PLDLM maintains the power law distribution over word tokens, one important property of natural language, which enable the PLDLM to produce statistically significant reductions in perplexity and WER compared to the IKNLM and the MKNLM.

Table 13 briefly shows the experimental results on NIST RT06s and the AMI Corpus. For RT06s, we trained trigram LMs on a corpus of 211.4M words and evaluated on NIST RT06s test data *rt06seval* (31,810 words). For the AMI Corpus, we trained trigram LMs on a corpus of 157.3M words and tested on 32 AMI scenario meetings (175,302 words). A vocabulary of 50,000 words was used when training LMs. We found that, on both RT06s and the AMI Corpus, the PLDLM and the HPYLM reduced the perplexity and the WER significantly, in comparison to the IKNLM and MKNLM.

Table 13: Perplexity and WER results of the PLDLM and the HPYLM, on NIST RT06s and the AMI Corpus.

DATA	Metric	IKN	MKN	PLD	HPY
RT06s	perplexity	107.0	105.2	100.7	**98.9**
	WER	27.0	26.8	26.6	**26.5**
AMI	perplexity	168.6	163.9	**157.9**	158.8
	WER	38.6	38.5	38.3	**38.2**

2.4 Dissemination and exploitation

2.4.1 Recognition Applications

The realtime recognition system has been used as an integral part of the Content Linking Device evaluation. We continue to routinely use the recognition system as part of our mini projects, transcribing data from HP and RBS. The availability of an easy to use, accurate, conversational speech recognition system has also allowed us to provide transcriptions on a one off basis for a number of projects and organisations: for the EU funded C-Cast project we have transcribed some sample broadcasts from Deutsche Welle, Germany's international broadcaster; we have transcribed some videos for VidioWiki, a small startup

company from the University of Edinburgh; as part of a dissemination event CISCO requested a transcription of their 'techwise tv' podcast; a number of school level meetings are now regularly recorded, transcribed and made available in browsable form internally at UEDIN.

2.4.2 webASR

In the last year webASR[1] has been improved and extended in many ways. The output of the Real-time and RT09 evaluations work have fed directly into webASR. For example, Juicer, optimised for decoding at speeds faster than 0.5xRT, coupled with the new discriminatively trained NN feature extraction technique has led to faster and more accurate systems in webASR.

webASR has been extended to provide transparent access to its services via a web-based API. The basic API interactions use XML/HTTP communication between the client application and the webASR servers. In order to facilitate the integration of webASR services into commercial applications (i.e., minimal programming overhead to developers, etc.), a generic software plug-in was provided. This plug-in provides full, programmatic access to the functionality of the API whilst at the same time customising the communication and services available on a per client basis. The plug-in was developed using C# targeting the Microsoft .NET 2.0 framework. The plug-in provides access to the following features:

- File management
  - Manages upload of audio files from the client machine to the webASR servers.
  - Supports uploading of a wide range of audio formats including WMA and MP3 instead of just WAV.
  - Supports the upload of additional information in XML format such as speaker ID or speech/silence segmentation and other metadata. Such additional uploaded information is associated with a particular audio file upload.
- Audio transcription
  - Allows access to transcriptions created at various stages of a multipass speech recognition process.
  - Allows transcription of earlier speech recognition passes to be retrieved when the later passes are not yet finished.
- Service management
  - Client authentication
  - Enumeration of supported audio file types
  - Gives feedback on how busy the webASR service is.
  - Gives feedback on the ASR processing status of a particular audio file.

[1] http://www.webasr.com

- Ability to assign different processing priorities for particular clients when running jobs on webASR servers.
- Fully integrated with webASR: audio uploaded via the plug-in/API also appears in the users uploads section of the webASR web browser interface.

Work done in collaboration with community of interest partners Dev/Audio and Outside Echo have produced two working product demonstrations that incorporate webASR.

3 Keyword Spotting

The work on keyword spotting and spoken term detection in the last period encompassed the following activities:

3.1 Comparison of keyword spotting systems

While research on automatic speech recognition (ASR) has several standard databases the results are reported on, for keyword spotting (KWS) and spoken term detection (STD), such standard corpora are scarce. The last international effort – NIST Spoken Term detection evaluation – was organized in 2006[2], and the reference transcripts for the evaluation corpus were still not published. BUT team working on KWS and STD therefore welcomed the activity of Czech Ministry of Interior to organize such comparison on Czech CTS data in 2008 and 2009.

The official evaluation took place at the end of 2008, but the work continues also in 2009 in tight cooperation with our security and defense partners. Systems are compared using standard metrics such as FOM (figure of merit) and EER (equal error rate).

BUT tested 4 systems in this evaluation: FastLVCSR was based on LVCSR with insertion of keywords into language model; HybridLVCSR used full-fledged word and subword recognition and indexing; and two acoustic systems were based on GMM/HMM and NN/HMM. While LVCSR systems are more precise, the advantage of acoustic ones is in their speed. HybridLVCSR is worth mentioning as it allows for pre-processing large quantities of data off-line with subsequent very fast searches, including OOVs. The effect of length of keywords, keywords being sub-parts of another ones, and effects of phonetic content were studied. A paper on this work is being prepared.

3.2 KWS viewer

In order to facilitate research and demonstrations of keyword spotting, we have developed an Interactive viewer for Keyword spotting output. This tool can load an output of a keyword-spotting system (KWS) and reference file in HTK-MLF format and show detections in a tabular view. It can be also used to replay detections, tune and visualize scores, hits, misses and false-alarms using sliders on the right-side panel. In case reference MLF file is available, the tool can plot histograms of scores for true detections and false alarms. The tool is available on BUT's web[3], the installation is extremely simple, as it requires only Qt version 4 (4.5 is recommended) and FMOD Ex Programmers API, that are both freely available. The screen-shot of the application is shown in Fig. 5.

3.3 Detection of out of vocabulary words

In cooperation with EU-FP6 project DIRAC, work continued also on the detection of out-of-vocabulary words (OOV) in the output of speech recognizer. In this reporting period,

[2] http://www.itl.nist.gov/iad/mig/tests/std/

[3] http://speech.fit.vutbr.cz/en/software/kwsviewer-interactive-viewer-keyword-spotting-output

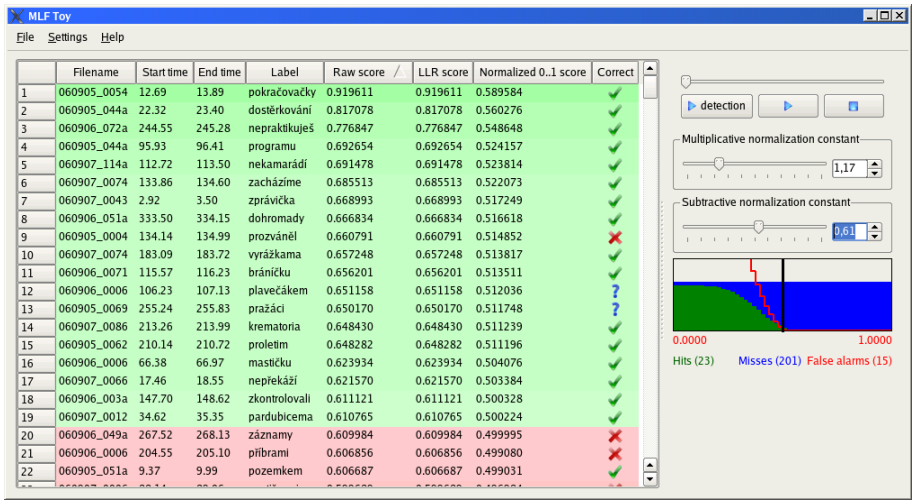


Figure 5: Interactive viewer for Keyword spotting output.

we have followed the approach based on phone posteriors created by a Large Vocabulary Continuous Speech Recognition system and an additional phone recognizer (combination of strongly and weekly constrained posterior estimators). We have compared the performances on challenging continuous telephone speech (CTS) task from CallHome English corpus to previously reported Wall Street Journal data. We have also conducted a study on finer output classes.

The conclusion is that the posterior-based OOV word detection approach generalizes across data (clean speech, 16kHz vs. noisy speech, 8kHz) and across varied language models (read speech, 5k words vs. spontaneous speech 2k8 words) with some performance degradation. The 4-class neural net improves classification performance and allows scoring with a single class or any compound class probability, see Kombrink et al. (2009).

Future work in this direction will concentrate on comparison of the NN-based OOV detector with our hybrid word/subword recognizers, taking into account very good results obtained by Tejedor et al. (2009).

3.4 Sub-space acoustic modeling

Finally, we have participated at the John's Hopkins University workshop group "Low Development Cost, High Quality Speech Recognition for New Languages and Domains"[4]. The group investigated mainly into an acoustic modeling approach in which all phonetic states share a common Gaussian Mixture Model structure, and the means and mixture weights vary in a subspace of the total parameter space. It is called Subspace Gaussian Mixture Model (SGMM) Povey et al. (2010). Globally shared parameters define the subspace. This style of acoustic model allows for a much more compact representation and gives better results than a conventional modeling approach, particularly with smaller amounts of training data.

[4] http://www.clsp.jhu.edu/workshops/ws09/groups/ldchqsrnld/

Although this approach was developed primarily for acoustic modeling in large vocabulary continuous speech recognition (LVCSR), it has a great potential in keyword spotting and in rapid prototyping and development of systems for new languages, see Burget et al. (2010).

4 Speaker Diarisation

The AMIDA speaker diarisation work has been carried out mainly at ICSI and UTwente this year. New approaches have been a low-latency system and a multi-modal diarisation system by ICSI, and speech overlap detection and fusion of systems by UTwente. The ICSI and UTwente systems are separate system developments, but they are based on a common architecture, which to a large extent is due to the participation of UTwente developer Marijn Huijbregts to the AMIDA trainee programme which was carried out at ICSI in 2007.

The task of speaker diarisation can be summarized as to automatically determine *who* spoke *when* from a recording in which several people take part in a conversation.

4.1 Common system design

The global architecture of the systems can be summarized as follows. First, for every available microphone channel a Wiener filter is applied to the signal in order to reduce the noise. For this purpose, the noise reduction algorithm developed by ICSI, OGI and Qualcomm is used (Adami et al., 2002). Then, when multiple synchronous microphone recordings are available, e.g., as can be obtained from microphone arrays, the various signals are combined to a single, enhanced, signal by automatic beamforming. This process estimates the relative delays of the wavefront as originating from the speaker arriving at the various microphone positions, and sums the signals using these delays to further enhance the signal to noise ratio. The delays are computed over limited duration intervals and therefore are capable of following the dynamics of the change in the speech source location due to the turn taking of the speakers. Hence, the delays themselves can be used as features at later stages (Pardo et al., 2006). For beamforming, version 2.0 of the `BeamformIt` toolkit is used.

From the enhanced signal, speech features are extracted. These are typically 19 Mel Frequency Cepstral Coefficients (MFCC). This feature stream is used for speech activity detection (SAD). For this, an algorithm is used that automatically finds a silence and speech class, and an optional "audible non-speech" class. Details of this algorithm can be found in (Huijbregts et al., 2007). Non-speech frames are removed, and the remaining feature stream is used for the final speaker segmentation and agglomerative clustering steps.

Speakers are represented by a Hidden Markov Models (HMMs) that consist of a string of states sharing the same output probability density function, a Gaussian Mixture Model (GMM). The string guarantees a minimum duration of a HMM being assigned to speech. The HMMs are linearly initialized on the speech feature stream, in the hope that most HMMs will contain speech from predominantly one speaker. Initially the number of HMMs is chosen to be much larger than the maximum expected number of speakers in the conversation, so that hopefully all speakers will be represented by at least one HMM. The HMMs are trained on the initial linear segmentation, and then the models are used in a Viterbi decoding to assign the speech frames to one of the HMMs. With the new segmentation, new models can be trained, and the process repeats several times.

In case of multiple microphones, a single Gaussian model is used to model the delay

features, and the likelihoods of these models are interpolated with the MFCC likelihoods.

After a segmentation step, a merging step is carried out by finding clusters that are most similar according to some criterion. At this stage it is necessary to decide whether the best merging candidate are actually the same or different speakers, which also needs some criterion. In practice we use the BIC criterion (Schwartz, 1978) for both, making sure that the number of model parameter in both the separate and merged models are the same in order to avoid the BIC penalty weight parameter.

In case of a merge, the re-segmentation and possible merging steps are carried out. This process is repeated, effectively clustering the HMMs in an agglomerative fashion, until the number of HMMs presumably is the same as the number of speakers in the conversation. Finally, the silences are inserted in the segmentation output and smoothed.

The development of diarisation systems depend heavily on the NIST Rich Transcription (RT) evaluation series, which contain speaker-annotated recordings using multiple distant microphones (MDM) in a variety of recording rooms. For development, a set of 27 meetings is used. Several versions of both ICSI and UTwente (under the name of AMIDA) have been evaluated in NIST RT 2009.

4.2 Multi-modal Diarisation

In this research effort, we performed an experiment to demonstrate the idea of using a simple activity detector to improve speaker clustering. The approach uses a subset of twelve meetings from the AMI corpus.

Using a skin-color detector (Chai and Ngan, 1999), we find the faces and hands of participants in the meetings. Then we extract the motion vectors of the MPEG-4 compressed version of the video. The magnitudes of the motion vectors within the detected skin regions are then summed. The frames are averaged over 400 ms. The experiments where performed both on the 4 close-up views and on the single camera view. We obtain one number for each of the four close-up videos per frame. In the single-camera case we partition the video image into 8 equally-sized regions. Fig. 6 illustrates the creation of the features. The approach we used for combining the compressed-domain video features and MFCC is similar to the one in (Pardo et al., 2006). Using agglomerative clustering each cluster is modeled by two GMMs, one for the audio MFCC features and one for the video activity features, where the number of mixture components varies for each feature stream (we use 5 for audio and 2 for video). We assume that the two sets of features are conditionally independent given a speaker. The combined log-likelihood of the two streams is defined as:

$$\log p(x_{MFCC}, x_{VID}|\theta_i) \doteq (1-\alpha)\log p(x_{MFCC}|\theta_{i1}) + \alpha \log p(x_{VID}|\theta_{i2})$$

where x_{MFCC} is the 19-dimensional MFCC vector, x_{VID} is the 4- or 8-dimensional video feature vector, θ_{i1} denotes the parameters of a GMM trained on MFCC features of cluster i, and θ_{i2} denotes the parameters of a GMM trained on video features of cluster i.

The audio-only baseline Diarisation Error Rate (DER) is 32.09 % for these meetings. Adding the video features from the single-camera view resulted in a DER of 27.52 % (which is a 14 % relative improvement) and adding the video features from the 4-camera

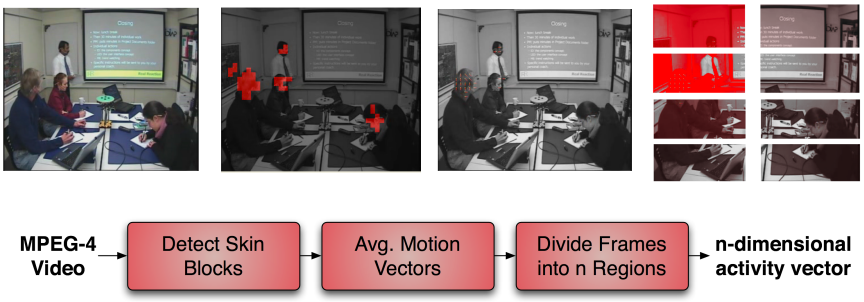


Figure 6: This figure illustrates how the video features where extracted for the single-camera view diarisation experiments. The frames are averaged over 400 ms.

view resulted in a DER of 24.00 % (25 % relative improvement). In both cases the combination with video performed better than the audio-only baseline. Not surprisingly, the higher resolution of the four camera-view provided better results. A more detailed discussion on the techniques can be found in Friedland et al. (2009a,b).

The experiment (Friedland et al., 2009a) shows that it is possible to treat audio and video data as part of the same optimization problem to help improve the diarisation task. However, the combined training of audio and video models allows for more than just improved accuracy. In a second pass over the video, we obtain the location of the current speaker's skin patches by "inverting" the visual models. Given the current speaker (as output in the previous step), we assume activity for each individual skin patch in a frame to calculate the likelihoods of belonging to a speaker using the visual Gaussian Mixture Models for each of the patches. Those skin patches that belong to the camera (four-camera case) or the part of the frame (single-camera) that is most likely to be active given the speaker determined by the diarisation in the first step are tagged (for visualization or further processing). This enables a completely unsupervised diarisation and tracking of the speakers in meetings. This "visual tracking using acoustic models" example shows that the proper integration of acoustic and visual data can lead to new synergistic effects: not only was the accuracy of the diarisation improved but another feature was added to the system at very little engineering cost.

4.3 Meeting duration

This research effort concerns novel initialization methods for the ICSI speaker diarisation engine. We know that the current ICSI Speaker Diarisation engine performs suboptimal on shorter segment durations than the usual 600 seconds occurring in NIST RT meeting data, e.g. 300 seconds, 200 seconds, or even 100 seconds. Being able to diarise shorter meeting chunks is a very important preliminary step for online diarisation. Therefore the goal was to improve the performance on short segment durations. The reason for the degradation in performance was that some of the main parameters of the engine, namely the number of initial Gaussians and the number of initial clusters, are highly dependent on the length of the overall meeting recording. Therefore we performed experiments to reduce the number of manually tunable parameters and get a more accurate Speaker Diarisation result at the same time.

We started the investigation by analyzing how many seconds of speech would be optimally represented by one Gaussian so that the Diarisation Error Rate is minimized. If the value of this parameter is too small, there is not enough speech available to train the system and if the value is too high there may be not enough Gaussians per speaker model to be able to separate different speakers accurately. By running a set of experiments, we found that we could use a linear regression to calculate the proper trade-off between having enough speech per Gaussian and having enough Gaussians. The resulting engine outperforms the current ICSI GMM/HMM-based approach using agglomerative clustering significantly.

We also experimented with a prosodic clustering approach to estimates the number of initial clusters. Rather than estimating the number of speakers in the meeting, we maximize the negative log-likelihood of the clustering of all the 12-dimensional prosodic feature vectors with a GMM with diagonal covariance. Thus, this approach groups speech regions that are similar in terms of prosodic features and deliberately overestimates the number of speakers in the meeting. This number and the model is used to perform a non-uniform initialization of the diarisation engine which leads to both a significant speed improvement of the engine (factor 3) and a much higher robustness against different recording length. The Diarisation Error Rate on meetings of 100s duration was improved by more than 50% relative, measured on the development data.

The details of this work are presented in a Master Thesis by David Imseng and in Imseng and Friedland (2009b,a); Imseng (2009). The work was further used to participate in the NIST Rich Transcription 2009 evaluation in several different versions of the system, varying the inclusion of prosodic features, the inclusion of a video channels, and the number of microphones used (single or multiple distant microphones, the Mark-III array or all microphones).

For the prosodic features we are currently using a 10-dimensional feature vector. The prosodic features are extracted every 10 ms using an analysis window of 500 ms. This system has three parameters, which are the weight of the different feature streams (of which only 2 are free, because of the normalization of the weights).

4.4 Overlap detection

Overlapping speech in meetings hurts speaker diarisation for two reasons: it spoils the purity of the speakers models that are trained during the clustering process, and since the Viterbi decoding only outputs one speaker at every time instant, the overlapping speech is missed by definition.

At UTwente we developed a two-pass system for detecting overlapping speech and attributing it to a second speaker (Huijbregts et al., 2009a) Overlapping speech was modeled by a single GMM. We initialized this overlapping speech model by assuming overlapping speech is most likely to occur at speaker changes. After a first full diarisation run, 1 s of speech around every speaker change is aggregated and a single "overlapping speech" model is trained from this material. This model is added to the earlier found models, and another three segmentation/retraining iterations followed. This model is then used in a second diarisation pass, where it is used for Viterbi segmentation, but is not retrained. Finally, the detected regions of overlap need to be assigned to speakers. For this we used the heuristic that if both speakers before and after the overlapping speech segment were

different, the segment was assigned to these two speaker. If the speakers were the same (as could occur in, e.g., backchanneling), the segment was assigned to only this speaker and overlap was effectively ignored. On the development test set, improvements from 19.07 % to 18.27 % and from 16.02 % to 15.29 % were observed for SDM and MDM without delay features, respectively. However, for the full diarisation system, MDM including delay features, no significant improvement was observed on the development test set.

4.5 Fusing multiple diarisation systems

One of the problems of building speaker diarisation systems is that the performance on individual meetings tends to vary rapidly with the change of some hyperparameters of the system (SAD usage, number of segmentation iterations or initial clusters, etc). With a small change of one of the parameters, the performance for one meeting may go up, while for another it may go down.

The idea in this research effort is to actually run various slightly different version of the diarisation system on the same meeting, and employ a "voting" strategy for deciding which segmentation is the best for this meeting (Huijbregts et al., 2009b). This idea is similar to the ROVER approach utilized in automatic speech recognition (Fiscus, 1997). Since there is no reference segmentation, we have to use some heuristic to come up with the best candidate. The idea is to used a symmetricized Diarisation Error Rate between two segmentations as a basic distance measure of the similarity of the two segmentations. By using these distances, we can agglomeratively cluster these segmentations until two clusters are left. Then we choose the biggest cluster, and from this cluster we choose the segmentation with the lowest average distance to the other segmentations in the cluster.

By using various slightly different preparations of the overall speaker diarisation system we could lower the DER of the development set for various microphone conditions. For SDM, using three versions of SAD, the DER could be lowered from 18.74 % for the best SAD condition to 17.27 % for the voted combination. In the MDM condition, adding 3 levels of variation in delay window length in beamforming led to 9 different systems. Where the best of those nine performed at 12.21 %, the voted combination resulted in 11.77,% DER. Even though this drop is not spectacular, the idea is that the voting system will be more robust to unseen data, for which the hyperparameters might be slightly off the optimum settings.

4.6 Other RT contributions

ICSI participated in four of the speaker-attributed speech-to-text tasks by combining the respective ICSI speaker diarisation system with a speech recognizer from SRI. UTwente contributed its segmentation and diarisation system to the AMI speech-to-text system with the purpose of speaker adaptation of the acoustic models.

The results were presented and discussed together with the other participants at the NIST RT workshop at Florida Institute of Technology. NIST plans a special issue in the Transactions on Audio, Speech, and Language Processing about the workshop.

4.7 IDIAP Diarisation

During the last year, work at IDIAP has focused on two themes:

1. Mutual information channel selection
2. Multiple stream diarisation

1) In the meeting case scenario, recording is performed using Multiple Distant Microphones (MDM). Beamforming is performed in order to obtain a single enhanced signal out of the multiple channels. We investigated the use of mutual information for selecting the channel subset that produces the lowest error in a diarisation system. Typically the selection is done on the basis of signal properties such as SNR, cross correlation. We propose the use of a mutual information measure that is directly related to the objective function of the diarisation system. The proposed algorithms are evaluated on the NIST RT 06 evaluation dataset. Channel selection improves the speaker error by 1.1% absolute (6.5% relative) w.r.t. the use of all channels. Results were published in Vijayasenan et al. (2009a).

2) This work investigated the use of Kullback-Leibler (KL) divergence based realignment with application to speaker diarisation. The use of KL divergence based realignment operates directly on the speaker posterior distribution estimates and is compared with traditional realignment performed using HMM/GMM system. We hypothesize that using posterior estimates to re-align speaker boundaries is more robust than gaussian mixture models in case of multiple feature streams with different statistical properties. Experiments are run on the NIST RT06 data. These experiments reveal that in case of conventional MFCC features the two approaches yields the same performance while the KL based system outperforms the HMM/GMM re-alignment in case of combination of multiple feature streams (MFCC and TDOA). This work was published in Vijayasenan et al. (2009b).

Furthermore experiments on extending multistream diarisation behind the combination of two acoustic models were performed. The use of Information Bottleneck principle combined with the KL divergence based re-alignment has produced reduction in terms of Diarisation Error when combining up to four different feature streams (MFCC,TDOA, MOdulation and FDLP). The use of four reduces by half the Speaker Error on the RT06 evaluation data.

4.8 Audio Visual Speaker Diarisation

We investigated audio-visual speaker diarisation (the task of estimating "who spoke when" using audio and visual cues) combining audio features with two psychology inspired visual features such as: Visual Focus of Attention (VFoA) and motion features. VFoA features are motivated by language and social psychology studies on the role of gaze in a conversation (Novick et al., 1996; Vertegaal et al., 2001): listeners are more likely to look at the person who is talking and they request turn shifts using gaze; speakers are likely to look at the person they are addressing and to shift their attention towards the next speaker before a speaker turn occurs. Thus VFoA features were defined as a measure of

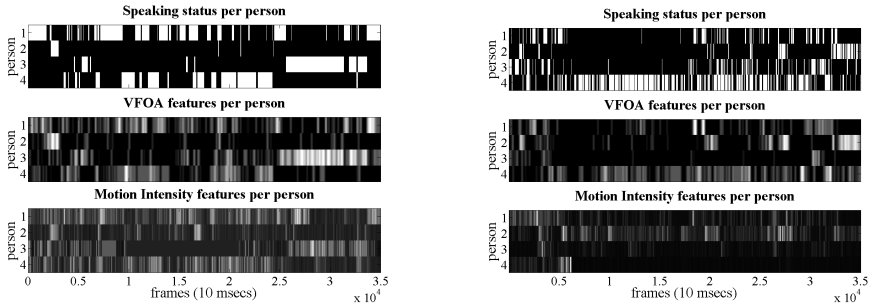


Figure 7: Comparison of the Speech/Non-Speech status (top), reference VFoA features $f_{vfoa}(i,t)$ (middle) and Motion Intensity features $f_{mot}(i,t)$ (bottom).

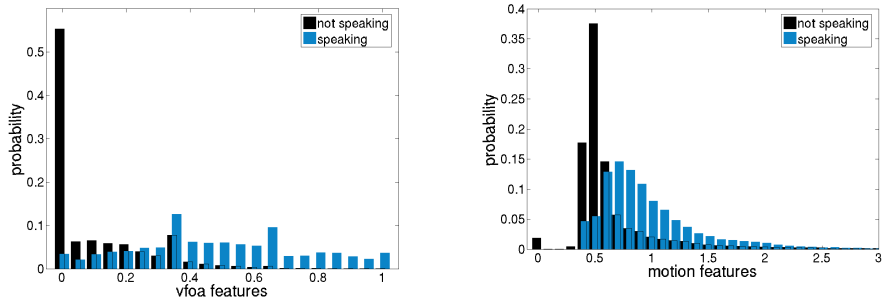


Figure 8: Distributions of $f_{vfoa}(i,t)$ and $f_{mot}(i,t)$.

the number of persons who are looking at each meeting participant. We experimented both with the VFoA obtained from manual annotation and with the VFoA automatically estimated by 3 different systems (Ba et al., 2009a) (see Sec. 5 on VFoA estimation for more details), relying on several graphical model structures and integrating different cues playing the role of context. In the first system (VFoA(1)) only the participant head pose is used to estimate his focus. The second system (VFoA(2)) exploits a slide activity cue to detect when looking at slides is more likely. The third system (VFoA(3)) exploits in addition visual activities at each seat and the whiteboard to detect who are more visually active, and hence more likely to speak, and thus the visual focus of others. We also investigated motion intensity features, which take into account both speaker's movement for speech production and the speaker's use of gestures to maintain the conversation floor. Our motion features measure global motion activities in each close-up and are computed as the average of the pixel by pixel difference of subsequent gray images. The VFoA features $f_{vfoa}(i,t)$ and the motion features $f_{mot}(i,t)$ are compared to the speaking status for each participant i in Fig. 7 for an excerpt of meeting IS1008a and IS1006d while their distributions for speaking and not speaking frames are shown in Fig. 8.

Experiments on audio visual speaker diarisation have been performed based on the ICSI speaker diarisation system where multiple feature streams were integrated by training separate models for each audio and video stream. Combining MFCCs and reference VFoA features we obtained a relative improvement of around 13% compared to the baseline au-

dio only system. Interestingly, from the 3 automatic VFoA systems, the one using only head pose resulted to be the best, providing an overall 10% relative improvement w.r.t. the MFCC only system. Indeed, although this system performs worse for VFoA estimation, it treats all the targets equally since it is not biased by any other information. This might be the reason why VFoA(1) performs better than the other two automatic systems, which are biased by priors on the slide change or on the participant visual activities (see Section 5). We observed different performances for static meetings (where people are seating all the time) and dynamic ones (where people move around for example to go to the whiteboard). For the automatic VFoA features larger improvements are observed on static meetings (26% reduction for VFoA(1)), for which the accuracy of the estimated VFoA is also higher. For dynamic meetings the best performances are achieved by the reference VFoA features (17% relative improvement). The combination of motion intensity features with MFCCs provided an overall 7% relative DER reduction. Similar DER reductions are achieved both on static and dynamic meetings.

5 Focus of Attention

During this final year research activities have been focused in two main directions: first, improving visual focus of attention (VFOA) recognition and investigating its use for speaker diarisation. Secondly, working on different head pose tracking systems. More concretely, the following achievements have been made:

- the work on multi-party visual focus of attention (VFOA) recognition from multimodal cues (Ba and Odobez, 2009) has been further extended by investigating the use of visual activity (how much people are gesturing with their hand, head and body) as context for VFOA recognition, and by improving the VFOA recognition model to handle the case of looking at moving people (Ba et al., 2009b).
- as an application of our work on gaze estimation, we have investigated the use of VFOA cues in a speaker diarisation system, by implicitly associating speech segments to people which are more looked at in a multi-stream approach (G.Garau et al., 2009).
- investigation of large margin likelihoods in a mixed-state particle filter for real-time head-pose processing of low to mid-resolution head videos (Ricci and Odobez, 2009).
- in addition, research on robust 3D head pose and facial expression estimation using structure and appearance features has been conducted to handle videos with higher head resolution (Lefèvre and Odobez, 2009). This work received the best student paper award at the IEEE-ICME conference.
- combination of the real-time head-pose tracker with a VFOA recognition module, and integration into the Hub demonstrator. This real-time VFOA estimation module has further been used in the User Engagement and Floor Control demonstrator for addressee detection.

These achievements are described in the rest of this Section with more details.

5.1 VFOA recognition: moving targets and visual activity context

In previous years, we had investigated the joint recognition of the VFOA of all meeting participants from multimodal cues. This was achieved by modeling the interactions existing between the gaze (VFOA) and speaking status of people in meetings as well as group activity contextual cues (slide presentation).

This work has been extended in two ways. First, we have addressed the problem of recognizing the VFOA of people in dynamics meetings in which people do not remain seated all the time. To account for the presence of moving visual targets (mainly participants standing at the whiteboard or the slide-screen during presentation), the location of people is tracked in the meeting room and used as context in the VFOA dynamics and our cognitive model that maps people's estimated head pose to VFOA semantic targets. Secondly, we have investigate the use of the visual activity of participants as a way of modeling the

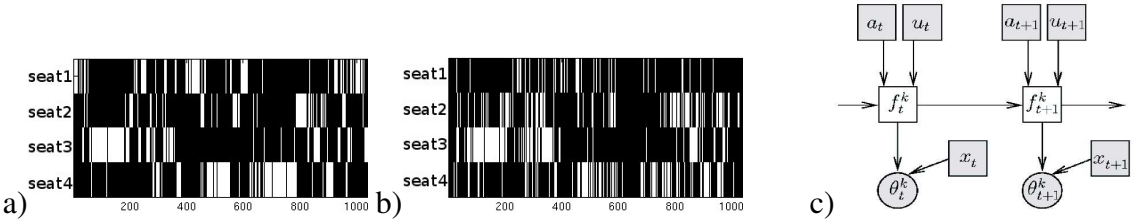


Figure 9: a) Speaking statuses and b) motion statuses of the 4 meeting participants extracted from speaking and motion energy. Note the similarities between the speaking and motion patterns.c) VFOA graphical model.

Figure 10: a) Head location and pose tracking from one side camera view. b: Areas A, B,C in the central camera view used for standing people localization.

conversation context in the case where audio information would not be accessible. This is motivated by the fact that although speech and visual activity are not equivalent, the production of speech is often associated with visual activity (lip, head, hand and body gesturing). This fact is illustrated in Fig. 9a,b), which shows the high correlation existing between the two cues. On the overall dataset, the percentage of time a person speaking is visually active is 66%. This shows that visual activity captures a significant proportion of speaking activity. Reversely, the percentage of time a person visually active is speaking is 47%. Compared to 25%, the approximate chance that a person is speaking in a meeting, it shows that visual activity is a good indicator of speech. More details are given below.

The VFOA modeling.
Our VFOA model is shown in Fig. 9. As can be seen, in the current study, the VFOA f_t^k of participant k was recognized independently of the VFOA of the other participants (but taking into account activities of all participants). In this model, the main difference with our previous work are the following terms.
The observation model: When person k looks at person j, we need to consider that person j may occupy different positions in the room. Hence, looking at the *semantic* target j corresponds to looking in the direction of the location x_t^j occupied by this person[5]. The observation model was thus defined as the Gaussian distribution:

$$p(\theta_t^k | f_t^k = j, x_t) = \mathcal{N}(\theta_t^k; \mu_{k,x_t^j}, \Sigma_k^j) \quad (1)$$

where μ_{k,x_t^j} is the Gaussian mean which models the mean head pose when the person at seat k looks at person j located at position x_t^j, and Σ_k^j is the covariance of the Gaussian. For other visual target j, the observation model is defined as a Gaussian distribution around

[5]For simplicity in our framework, people candidate position where discretized into a predefined set of 4 positions: their seat, and three positions near the slide screen, see Fig. 10. The location of people was tracked using simple features, assuming that a maximum of one person was standing at a given time.

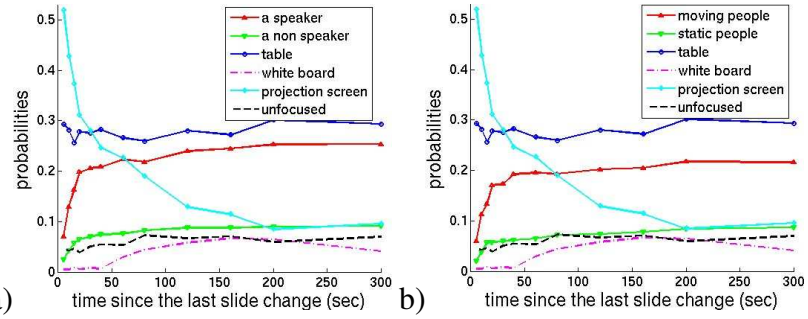


Figure 11: Probabilities of focusing at visual targets given the time elapsed since the last slide change. In a) the probability of looking at a person is spread into looking at a speaker (red) or a listener (green). In b), the probability of looking at a person is spread between looking at a person visually active (red) or static (green).

experimental setup	seat1	seat2	seat3	seat4	mean
slide-speaking	56.2	58.2	49	47.5	52.7
slide-motion	**56.1**	**56.9**	**49.3**	**50.3**	**53.2**
slide	53.7	55.2	43.5	48.6	50.2

Table 14: VFOA recognition performance over 12 meetings involving moving people.

the mean pose associated with looking at the target j.
The state dynamics is defined as follows

$$p(f_t^k|f_{t-1}^k, a_t, u_t) \propto p(f_t^k|f_{t-1})p(f_t^k|a_t, u_t) \quad (2)$$

where $p(f_t^k|f_{t-1}^k)$ models the temporal transitions between focus states (high probability to remain in the same state) and $p(f_t^k = l|a_t, u_t)$ models the contextual probability to observe a VFOA target l given the slide activity a_t context, and u_t denoting either the speaking or the motion activities of all people, according to the experimental conditions. The later term is our main interest here, and is illustrated in Fig. 11 for the two types of activity context. These figures show that right after a slide change, the probability of looking at the projection screen is high and gradually decreases. Inversely the probability of looking at the people increases. The probability of looking at an active person, either speaking or visually active, is higher than the probability of looking at non-active people. Also, the probability of looking at a person speaking is higher than the probability of looking at a visually active person, indicating that there is a higher correlation between gaze and speaking behaviours than between gaze and visual activity behaviours.

Results. Performance was evaluated on the 12 meetings of the AMIDA database for which FOA annotation is available, using frame recognition rate (FRR) as performance measure (percentage of video frames for which the VFOA is correctly classified.
The results are in Tab. 14. They show that the use of contextual activity is always beneficial. The best FRR is achieved when slide and visual activities are used as contextual cues (FRR of 53.2%). A significance T-tests showed that, at a p-values of 1% the performances of the methods based on speaking and visual activity context are not significantly different, but are significantly better than the method using only the slide context. Thus, for VFOA estimation in meetings, visual activity is as effective as speaking activity as modelisation context.

Comparing the VFOA recognition performances on meetings involving only seated persons (4 recordings) and on meeting involving moving persons (8 recordings) shows that for all experimental conditions the recognition rates were always higher on the former than on the latter. The main reason is that people standing for presenting slides are often the focus of attention, and by moving to the slide screen area, they increase the level of confusion between these two important targets (the slide screen and the presenter).

5.2 Speaker diarisation using Visual Focus of Attention

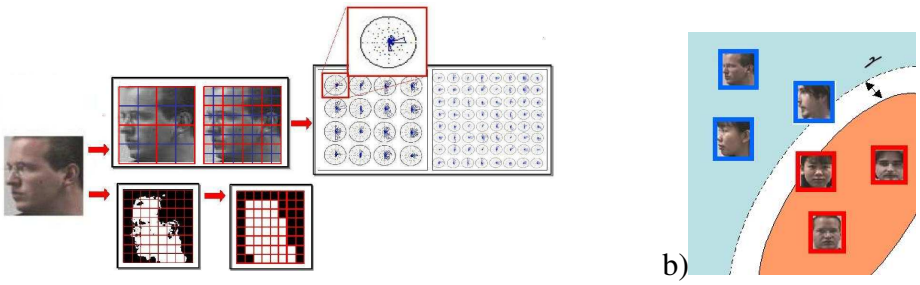
Conversation is multimodal in nature. Exploiting this basic fact, we have explored in G.Garau et al. (2009) the use of psychologically inspired visual features as additional cues to audio in a speaker diarisation task (determine who speaks when). More precisely, the role of gaze in conversation (and hence of Visual Focus of Attention) and more specifically the fact that listeners usually look more at speakers than listeners during conversation, was exploited. This was done by defining VFOA cues as the amount of visual attention received by each person in the meeting over a temporal window, and exploiting these cues in a multi-stream approach along with traditional audio cues. Experiments were performed both with the reference and 3 automatic VFoA estimation systems, based on head pose and visual activity cues, of increasing complexity. VFoA cues in combination with audio features yielded consistent speaker diarisation improvements. More details can be found in Sec. 4.8.

5.3 Large margin likelihoods for realtime head pose tracking

We have pursued our efforts towards the development of a real-time VFOA recognition system. The main computational bottleneck is the estimation of the head pose of participants. To develop a real-time head pose tracker, we have followed the same Bayesian filtering with sampling approximation framework that we used in the past and that was presented in last year report D44. One of the main strength of this approach (w.r.t. pose estimation accuracy) is to perform the joint head tracking and pose estimation, rather than doing the head tracking first and then estimating the head pose. The main issues that were dealt with are the following:

- real-time: to achieve this, we have selected features that can be computed in a fast way using integral images. These are the Histogram of Oriented Gradients (HOG), which are the main cue for pose estimation, and skin features which are important to avoid tracking failures.
- likelihood modeling: modeling appropriately the likelihood of the data for different head pose is crucial for the accurate the head pose estimation. To this end, we have designed an examplar-based and large margin approach to design and learn the parameters of the likelihood function. It optimizes a criterion that jointly enforces that head images (with different appearances) of the same pose get a high likelihood for the correct pose, and a smaller one for other poses.

The proposed modeling yielded the best results on the CLEAR06 benchmark data (static head images), and close to the best results on the CLEAR07 benchmark data (but with



a) b)

Figure 12: a) features: original image, multilevel HOGs (cells in blue, blocks in red) and skin mask. b) distance learning principle: distances between images of the same pose should be smaller than the smallest distance between images of different poses.

a real-time tracker). The approach is summarized below. More details can be found in Ricci and Odobez (2009).

Head pose representation: We consider the head orientation θ as described by 3 angles, pan and tilt (to represent out-of-plane rotations) and roll (for in-plane rotations). We discretize the space of all possible orientations into $\Theta = 273$ poses. For the purpose of tracking and likelihood modeling, we use an examplar-based approach. More precisely, we consider multiple reference models for each pose and we denote by $\mathcal{R}_\theta^k$ the k-th reference of pose θ. We use the PRIMA-POINTING head pose database (Gourier et al., 2004) to build this reference set. It contains 15 images of different individuals for 91 poses. This approach allows to alleviate the problems due to the large variations of head appearance models corresponding to the same pose. For each of these reference image $\mathcal{R}_\theta^k$, we compute the feature vectors $\mathbf{r}_\theta^k$. We adopt two types of features in order to discriminate between different head orientations (Fig. 12).
Texture features. We use multi level HOG descriptors as texture features: we partition the image into 2×2 (first level) and 4×4 (second level) non overlapping blocks of 2×2 cells and compute the histograms of gradient orientation on each cell. As suggested by Dalal and Triggs (2005) we employ unsigned orientation of the image gradient, and histograms are normalized locally i.e. considering all the cells in the same block. The final HOG descriptor is obtained by the concatenation of the small histograms. Fig. 12 give a example of an image and its HOG representation. Color features. This is done by modeling skin color in the normalized RG space as a gaussian model, learned online, and used to obtain a binary image of *skin/not skin* pixels. The resulting binary image is divided into 8×8 cells in the region of interest and used to build the color features.

Tracking and likelihood model: We follow a particle filtering framework, where the goal is to estimate the object state $\mathbf{s}_t$ at time t given the sequence of observation $\mathbf{o}_{1:t}$. As our state space, we choose a rectangular box as head tracking region described by the vector $\mathbf{s} = (t_x, t_y, s_x, e_y, \theta, k)$, which contains both continuous variables ($x = \left(t_x, t_y, s_x, e_y\right)$ to indicate head location and size) and discrete variables (θ, k representing respectively the pose and the k-th reference model of pose θ).
The innovative part of our method lies in the modeling of the likelihood term. An observation $\mathbf{o} = (\mathbf{o}^{tex}, \mathbf{o}^{col})$ is composed by texture and skin color features. Under conditional independence assumption, we have

$$p(\mathbf{o}_t|\mathbf{s}_t) = p^{col}(\mathbf{o}_t^{col}|\mathbf{s}_t)p^{tex}(\mathbf{o}_t^{tex}|\mathbf{s}_t) \text{ with } p^{tex}(\mathbf{o}_t^{tex}|\mathbf{s}_t) = e^{-\lambda_T D_W(\mathbf{o}_t^{tex}, \mathbf{r}_{\theta_t}^{k_t tex})}$$

Table 15: Average error in degrees with CLEAR06 setup. Numbers in parenthesis correspond to all weights set to 1.

	This paper	Ba	Voit	Tu	Gourier
pan	9.1 (13.7)	11	12.3	14.1	10.3
tilt	10.5 (14.2)	11.5	12.7	14.9	15.9

where λ_T is a user define constant and the distance $D_{\boldsymbol{W}}(\mathbf{o}, \mathbf{r}_\theta^k)$ which allows to compare the observation to the examplar $\mathbf{r}_{\theta_t}^{k_t tex}$ is a linear function of some parameter vector $\boldsymbol{w}_\theta^k \in R^M$ e.g. $D_{\boldsymbol{W}}(\mathbf{o}, \mathbf{r}_\theta^k) = \boldsymbol{w}_\theta^{kT} \boldsymbol{d}_\theta^k(\mathbf{o})$. The vector $\boldsymbol{d}_\theta^k(\mathbf{o})$ contains the concatenation of elementary distances between features of an observation $\mathbf{o}$ and the corresponding features in the reference model $\mathbf{r}_\theta^k$. Arbitrary distances can be used as elementary distances: in our experiments we used χ_2 distances between histograms of corresponding HOG cells. A similar (but simplified) model is used for the skin likelihood modeling.

Learning the likelihood, or equivalently in our case, the distance function, is a crucial step. Intuitively the distance between the features of the observed and reference head images should be low when their pose are similar, and high otherwise. A difficulty in practice is that often, the distance between the features extracted from head images of the same person but at different pose are smaller than features of heads from the same pose but of different persons. Our learning algorithm explicitly deal with this issue.
Given a training set $\mathcal{T} = \{(\mathbf{o}_1, y_1), (\mathbf{o}_2, y_2), \ldots, (\mathbf{o}_\ell, y_\ell)\}$ of pairs of texture feature $\mathbf{o}_i$ with their associated poses y_i, and the set of reference models $\mathbf{r}_\theta^k$ (sampled from the training set), we learn the distance functions $D_{\boldsymbol{W}}$ in order to impose that exemplars $\mathbf{o}_i$ associated to pose y_i should be closer to all reference models of the same pose $\mathbf{r}_{y_i}^k$ and separated at least by a margin of 1 from reference models $\mathbf{r}_{\theta'}^{k'}$ of different pose ($\theta' \neq y_i$). In formulas:

$$\min_{y_i \neq \theta', k'} \boldsymbol{w}_{\theta'}^{k'T} \boldsymbol{d}_{\theta'}^{k'}(\mathbf{o}_i) - \max_k \boldsymbol{w}_{y_i}^{k\,T} \boldsymbol{d}_{y_i}^k(\mathbf{o}_i) \geq 1 \quad \forall \mathbf{o}_i$$

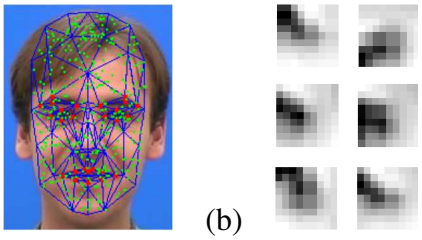
In other words we impose that for each image $\mathbf{o}_i$ the difference between the minimal distance from references of different poses and the maximal distance from references of the same pose should be larger than one. This principle is illustrated in 12. To learn the weights, the above constraints are reformulated as a constrained optimization problem, and solved with an efficient iterative algorithm based on stochastic gradient descent which is a variation of an optimization strategy recently proposed in the literature.

Results. The efficiency on the distance learning algorithm has been evaluated on two benchmark datasets: static images from the tracking results the international CLEAR evaluation workshop 2006 (*http://isl.ira.uka.de/clear06/*), and video sequences from the CLEAR evaluation workshop 2007.

In the first cases, images of 93 poses are used and split in two sets: the first series is used as training set, the second as test set. Faces in the images are cropped automatically by a skin color model and rescaled into 64×64 pixels. In the distance learning algorithm we use $K = 15$ reference models per pose. Classification is performed with 1-NN classifier. As shown in Table 15 our method achieves better accuracy than state-of-the approaches This demonstrates that multilevel HOGs are effective descriptors: data are clustered with

Table 16: Pose estimation errors (degrees) for person left (L) and right (R), on the CLEAR07 benchmark data (number in parentheses: no distance learning).

	1L	1R	2L	2R	3L	3R	mean
pan	16.9	11.2	16.6	12.1	11.3	7.2	12.5 (15.8)
tilt	8.4	5.7	7.1	13.1	11.5	5.1	8.5 (11.3)
roll	6.9	9.6	11.7	8.4	9.8	5.1	8.5 (9.6)



(a) (b)

Figure 13: (a) *Structure* (red dots) and *appearance* (green dots) features. (b) Samples of the training set for the *structure* feature located on the right corner of the right eye (before removing the patch mean).

respect to pose and classification performance are already good. Moreover with distance learning the accuracy is significantly improved.

Secondly, we show the validity of our approach for joint tracking and pose estimation using the CLEAR07 benchmark datasets and protocol, which uses videos from the IDIAP Head pose database. The pose estimation errors corresponding to $K = 5$ reference models per pose are shown in Table 16. It is evident that using our large margin learning approach the estimation accuracy significantly improves with respect to the baseline (no distance learning). Comparing our results with the best method (Ba and Odobez, 2007) on this dataset, we see that we achieve higher performance in term of tilt and roll estimation while the pan recognition is less accurate. However, our tracker runs close to realtime (at about 20fps) while the system in Ba and Odobez (2007) is very slow (about 2fps) due to the likelihood computation (which heavily relies on particles resizing, histogram equalization and Gabor filters) and to Rao-Blackwellization. From the analysis of the output videos we observe that the major cause of pose estimation errors is probably the fact that we do not model large in-plane rotations since in these cases it is difficult to compute features with integral images.

5.4 Head pose and facial expression estimation using 3D deformable models

Our previous study showed that VFOA estimation from head pose is heavily dependent on the accuracy of the estimated head pose. View-based approaches that we had used in the past are well adapted to handle mid-resolution images, but are intrinsically limited to improve accuracy when higher resolution images are available (e.g. with webcams). This year, we have investigated the use of 3D deformable models to perform robust head pose and facial action estimation. In particular, we have addressed the challenging case of 3D head pose tracking under large head pan rotations (up to profile and beyond) which constitute a common limitation using such models.

To this end, we have designed several tracking algorithms relying on an hybrid set of features to represent the face: structure features, which are illumination-invariant local

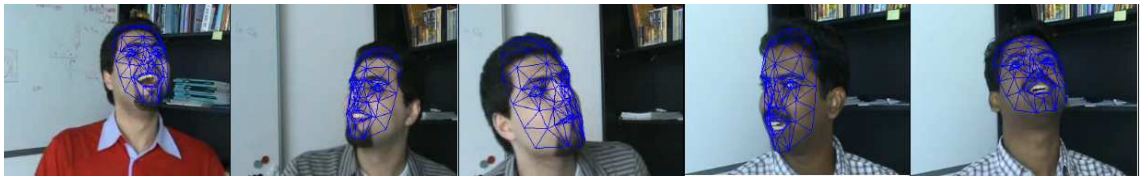


Figure 14: Sample images from various sequences obtained with our tracker.

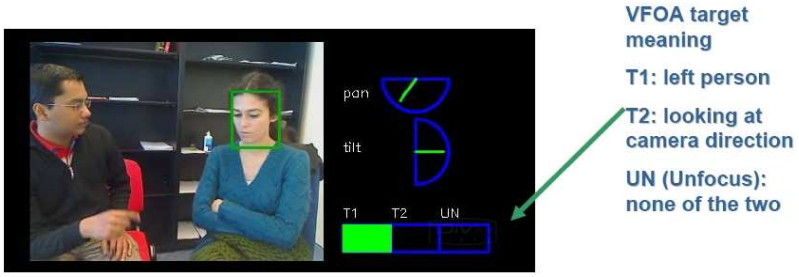


Figure 15: Output of the real-time VFOA demonstration module. Both the head pose and the recognized VFOA are sent to the HUB, and can further be used by the User Engagement and Floor Control monitoring module (for online addressee detection).

features located around salient regions of the face, with a fixed pose-invariant appearance model learned by generating a set of virtual samples from a single training image; and appearance features, that is, sparse facial texture points located throughout the face and compared to a reference or online generated template. These features are illustrated in Fig. 13.

Innovations of our approach come from the combination of these features, the extension of the 3D mesh to appropriately cover face sides, and the online learning of a set of pose-based template and the way of interpolating them to generate the template used for comparison with the extracted appearance features.

Evaluation on the Boston University Face Tracking benchmark showed that our approach provides results as good as state-of-the-art of the algorithms. We have also shown the stability of the system as well as its ability to track facial actions in long real video sequences of natural and animated conversation (cf Fig. 14). Computationally, the algorithm, implemented in C++ with non-optimized code runs at 3 frames per second. Initialisation has not been made automatic yet. More details can be found in Lefèvre and Odobez (2009).

5.5 Real-time VFOA recognition module for addressee detection

The real-time head pose estimation algorithm described above, implemented in C++, runs at approximately 10 to 15 frames on a standard laptop. It was further modified to incorporate a VFOA recognition algorithm, and communicate to the HUB. A standalone version of this system was made, and an example of output is displayed in Fig. 15.

This standalone software has been successfully transferred to the University of Twente, who used it for the multi-modal addressee detection module of the user engagement and floor control demonstrator. It was demonstrated at the last review meeting.

6 Visual Identification

In this section, we consider that visual identification could be achieved by the visual recognition of a particular object: the face. We will thus address the problems of face detection and recognition.

Face detection is the very first step in many visual processing systems like face recognition, emotion recognition or lip reading. It consists in locating faces in an arbitrary image, irrespective of variations in illumination conditions, background, pose, scale, expression and the identity of the person.

In Sec. 6.1, we present an original work based on a novel feature called Haar Local Binary Pattern (HLBP) for fast and reliable face detection, particularly in adverse imaging conditions. This binary feature compares bin values of Local Binary Pattern histograms calculated over two adjacent image subregions. These subregions are similar to those in the Haar masks, hence the name of the feature. They capture the region-specific variations of local texture patterns and are boosted using AdaBoost in a framework similar to that proposed by Viola and Jones. Results obtained on several standard databases show that it competes well with other face detection systems, especially in adverse illumination conditions.

Face recognition refers to the automatic recognition of individuals based on their face image. Research in this area has been conducted for more than 30 years; as a result, the current status of face recognition technology is well advanced. Face recognition actually deals with two tasks: face authentication (also called verification) and face identification. Although, face identification is directly relevant in the context of AMIDA, we will mainly focus on face authentication. Indeed, face identification shares the same components (feature extraction, classification, ...) with face authentication. As an example, it has already be shown in the previous deliverable D44 that face authentication algorithms can be applied to a face identification task on AMIDA meeting videos. Furthermore, face authentication is much easier (1) to evaluate because most well known benchmark face databases have been designed for this task and have been provided with unbiased evaluation protocols, and (2) to compare because baseline results exist on these benchmark face databases using these precise protocols. As a consequence, we will evaluate the proposed algorithms on a face authentication task.

Sec. 6.2 thus presents an extension of the work on face recognition using Bayesian Networks initially proposed in the deliverable D42. The reader is asked to refer to this deliverable as an introduction. Hence we present generative models dedicated to face recognition considering data extracted from color face images and using Bayesian Networks to model relationships between different observations derived from a single face. Specifically, the use of color as a complementary observation to local, grayscale-based features is investigated. This is done by means of new generative models, combining color and grayscale information in a principled way. Color is either incorporated at the global face level, at the local facial feature level, or at both levels. Obtained results show that integrating color in an intelligent manner improves the performance over a similar baseline system acting on grayscale only, but also over an Eigenfaces-based system were information from different color channels are treated independently.

6.1 Fast Illumination Invariant Face Detection using Haar Local Binary Pattern Features

6.1.1 Introduction

The main challenge for a face detection system is to successfully detect faces in an arbitrary image, irrespective of variations in illumination conditions, background, pose, scale, expression and the identity of the person. Numerous approaches have been proposed to counter these issues. Most of these approaches can be organized in three categories: feature-based approaches (Heisele et al., 2001), appearance-based approaches (Yang et al., 2000) and boosting-based approaches (Viola and Jones, 2001). The third approach, which involves the boosting of simple local features called Haar features in a cascade architecture, was introduced in 2001 by Viola and Jones (2001). It has become very popular since then because it shows very good results both in terms of accuracy and speed (with the use of Integral Image concept), and is quite suitable for real-time applications. Since the initial work of Viola and Jones, most of the research in face detection has focused on the improvement of their cascade architecture. Related works can be classified in mainly two possible directions: alternative boosting algorithms (Lyu, 2005; Sun et al., 2004) or alternative architecture designs (Luo, 2005; Sochman and Matas, 2005).

However, most of these boosting-based methods which are derived from the Haar feature set have a common limitation. This is the *vulnerability of the Haar feature set to variations in illumination conditions*, for example, where there is a strong side illumination either from left or right, or the dynamic range of the image intensity varies from region to region over the face (see Sec. 6.1.2, Fig.21). Thus, there is a need to improve the robustness of the system to take into account these illumination variations, but retaining the richness of the feature set, and the advantages of efficient feature selection by boosting and fast evaluation of the features using the Integral Image concept.

The Local Binary Pattern (LBP) introduced by Ojala et al. (1996) is one such operator which is robust to monotonic illumination variations (see Fig. 16). Thus, various face detection systems have been proposed using LBP or its variants, such as Improved Local Binary Patterns (ILBP) (Jin et al., 2004), Multi-Block Local Binary Patterns (Zhang et al., 2007), the Modified Census Transform (MCT) (Froba and Ernst, 2004; Rodriguez, 2006) and the Locally Assembled Binary (LAB) features (Yan et al., 2008).

In this work, we propose a new type of feature called the Haar Local Binary Pattern (HLBP) feature which combines the advantages of both Haar and LBP. This feature compares the LBP label counts in two adjacent image subregions, i.e. it indicates whether the number of times a particular LBP label occurs in one region is greater or lesser than the number of times it occurs in another region, offset by a certain threshold. These two subregions are represented by a set of masks similar to Haar masks (Viola and Jones, 2001). Thus, our features are able to capture the region-specific variation of local texture patterns. This makes our features more robust to illumination variations, which may be quite complex and concentrated over certain subregions of the image only (strong side illumination), compared to Haar and LBP individually. Since each LBP label count is actually a particular bin value of the spatial histogram (Zhang et al., 2005), our features are also robust to slight variations in location and pose.

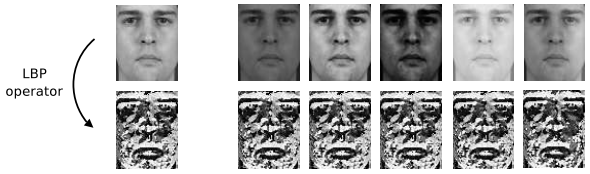


Figure 16: LBP robustness to monotonic gray-scale transformations. On the top row, the original image (left) as well as several images (right) obtained by varying the brightness, contrast and illumination. The bottom row shows the corresponding LBP images which are almost identical. Please see Fig. 21 for more complex illumination changes considered in our experiments.

To our knowledge, this is the first time individual LBP label counts have been combined with Haar features for face detection. Since each HLBP feature is linked with exactly one LBP label, there is no need to consider the entire LBP histogram in training and test, as in (Froba and Ernst, 2004). Thus our system is more efficient in terms of storage requirements as well as speed (see Sec.6.1.3). This makes it more suitable for use on mobile devices for instance. We use a variation of the Integral Histogram (Wang et al., 2006) to calculate our features, which further increases the speed.

We tested our proposed approach using several standard databases against two standard face detection systems. The first is the baseline system based on Haar features (Viola and Jones, 2001). The second is the system based on MCT (Froba and Ernst, 2004) which is one of the best performing systems representing the state of the art today. [6]

The rest of these sections is organized as follows: we first introduce the proposed HLBP features in Sec. 6.1.2. We report the experiments and discuss the results in Sec. 6.1.3. Finally, conclusions are given in Sec. 6.1.4.

6.1.2 The Proposed Framework : Face Detection using HLBP features

In the current work, we unite the two popular concepts of Boosted Haar features (Viola and Jones, 2001) and Local Binary Patterns (Ojala et al., 1996), so as to use the advantages of both in the task of face detection.

General Boosting Framework

The central concept of our framework (as in the Viola and Jones' face detector) is to use boosting, that linearly combines simple weak classifiers $f_j(I)$ to build a strong ensemble, $F(I)$ as follows :

$$F(I) = \sum_{j=1}^{n} \alpha_j f_j(I). \tag{3}$$

The selection of weak classifiers $f_j(I)$ as well as the estimation of the weights α_j are learned by the boosting procedure. An input image I is detected as a face if $F(I)$ is higher

[6] A public demonstration of the MCT-based face detection system can be found at *http://www.idiap.ch/onlinefacedetector* .

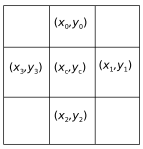


Figure 17: The $LBP_{4,1}$ label for a particular pixel (x_c, y_c) is calculated by comparing its intensity with each one of its four neighbors (vertical and horizontal only), $\{x_i, y_i\}_{i=0}^{3}$, and forming an 4-bit word. Unlike the $LBP_{8,1}$ case, the 4 diagonal neighbors are not considered.

than a certain threshold Θ which is also given by the boosting procedure (Viola and Jones, 2001) and is rejected otherwise. Each weak classifier f_j is associated with a weak feature, called the Haar feature in Viola and Jones' system. Here, instead of the Haar feature, we use a different set of weak features which we call Haar Local Binary Pattern (HLBP) features.

The proposed HLBP features

We assume that our input is an $N \times M$ 8-bit gray-level image, which can be represented as an $N \times M$ matrix I, each of whose elements satisfy, $0 \leq I(x, y) \leq 2^8$. In the first stage, we calculate the LBP image I_{LBP} (Ojala et al., 1996) from the original input image I. The LBP operator can be applied at different scales. However, after extensive preliminary testing, we have found the $LBP_{4,1}$ operator as the optimal LBP operator in our case. At a given pixel position (x_c, y_c), the $LBP_{4,1}$ operator is defined as an ordered set of binary comparisons of pixel intensities between the center pixel (x_c, y_c) and its four surrounding pixels, $\{(x_i, y_i)\}_{i=0}^{3}$ (see Fig. 17). The decimal form of the resulting 4-bit word is called the LBP code or LBP label of the center pixel and can be expressed as,

$$I_{LBP}(x_c, y_c) = \sum_{n=0}^{3} s(I(x_n, y_n) - I(x_c, y_c))2^n. \tag{4}$$

where $I(x_c, y_c)$ is the gray-level value of the center pixel (x_c, y_c) and $\{I(x_n, y_n)\}_{n=0}^{3}$ are the gray-level values of the 4 surrounding pixels. The function $s(x)$ is defined as,

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases} \tag{5}$$

In the second stage, we calculate the Integral Histogram set $\{I_k^H\}_{k=1}^{N_{labels}}$ (Wang et al., 2006) of the LBP image I_{LBP}. Here, N_{labels} indicates the number of LBP labels depending on the LBP operator used, and here it has a value of 16 (2^4). Thus the Integral Histogram set consists of $N_{label} = 16$ Integral Histograms. The individual pixels $I_k^H(x, y)$ of the k-th Integral Histogram I_k^H is calculated as the number of pixels above and to the left of the

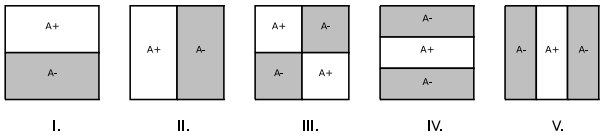


Figure 18: The five types of masks used for the calculation of both Haar and HLBP features, I. Bihorizontal, II. Bivertical, III. Diagonal, IV. Trihorizontal, V. Trivertical.

pixel (x, y) in the LBP image I_{LBP} which have a label k, as follows,

$$I_k^H(x, y) = \sum_{u \le x, v \le y} \delta_k(u, v) \tag{6}$$

where $\delta_k(u, v) = 1$ if the label of the pixel at location (u, v) in the LBP image I_{LBP} is k, and is zero otherwise. Using the following pair of references, for all $k \in \{1, N_{label}\}$:

$$i_k^H(x, y) = \quad i_k^H(x, y-1) + \delta_k(x, y) \tag{7}$$
$$I_k^H(x, y) = \quad I_k^H(x-1, y) + i_k^H(x, y) \tag{8}$$

where $i_k^H(x, 0) = 0$ for any x and k, the Integral Histogram set can be calculated by one pass over the LBP image. In the third and final stage, the Integral Histogram set will enable us to calculate the proposed HLBP features directly in an efficient and fast way as with Integral Image for the original Haar features. A particular HLBP feature is defined by the following parameters : mask type T (one out of five, see Fig. 18), LBP label k (one out of sixteen for $LBP_{4,1}$), position (x, y) of the mask inside the image plane, size (w, h) of the mask, a threshold θ and a direction p (either +1 or -1). It can be observed that a HLBP feature has exactly the same definition as a Haar feature except the addition of the parameter k. To calculate the value of a particular feature $f_{T,k,x,y,w,h,\theta,p}(I)$, its corresponding mask of size (w, h) is placed on the LBP image I_{LBP} at the location (x, y). Like in Viola and Jones' system, each mask type divides the mask region into two areas (see Fig. 18), a positive (A_+) and a negative (A_-) region. If we define,

$$S_{A+} = \quad \textstyle\sum_{(u,v)\in A_+} \delta_k(u, v) \tag{9}$$
$$S_{A-} = \quad \textstyle\sum_{(u,v)\in A_-} \delta_k(u, v) \tag{10}$$

with $\delta_k(u, v)$ as defined[7], then the HLBP feature value is given simply by,

$$f_{T,k,x,y,w,h,\theta,p}(I) = \begin{cases} 1 & \text{if } p.(S_{A_+} - S_{A_-}) > p.\theta, \\ -1 & \text{if } p.(S_{A_+} - S_{A_-}) \le p.\theta \end{cases} \tag{11}$$

Thus, the HLBP feature is a binary feature, as the normal Haar feature. In other words, the HLBP feature indicates whether region A_+ (region A_-) has θ pixels more with the LBP label k compared to region A_- (region A_+), given $p = 1$ ($p = -1$), i.e. the spatial count differences of the LBP label k (see Sec. 6.1.1). However, to calculate S_{A+} and S_{A-} we do not need to use the above equations 9 and 10. They can each be calculated directly by only a few references to the corresponding Integral Histogram I_k^H as in usual Haar

[7]For the Viola and Jones' system, $\delta_k(u, v)$ is replaced by $I(u, v)$, the pixel intensity at location (u, v).

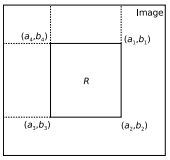


Figure 19: Calculation of the sum of LBP label counts within region R using Integral Histogram (see Eqn. 12).

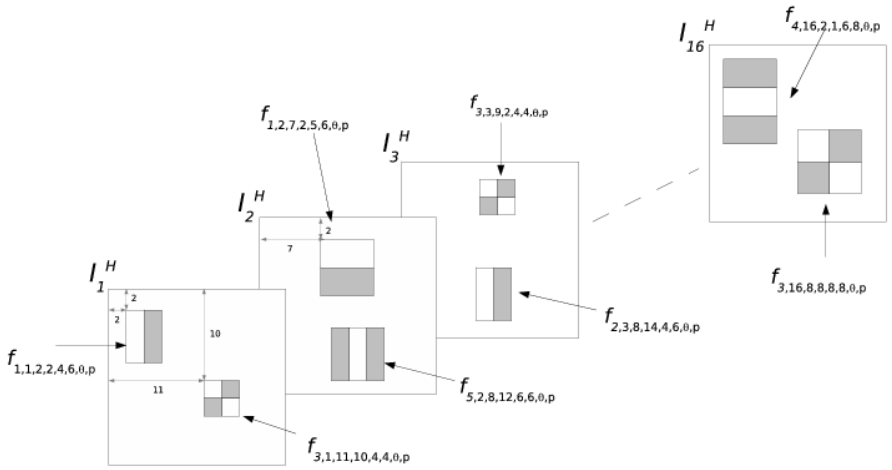


Figure 20: The HLBP features $f_{T,k,x,y,w,h,\theta,p}$ are calculated by placing the corresponding mask at the specified location (x, y) inside the Integral Histogram I_k^H and with the specified size (w, h). Examples of eight different masks corresponding to eight different features have been shown in the figure.

features, as follows. Let us denote by $(a_1, b_1), (a_2, b_2), (a_3, b_3), (a_4, b_4)$ the four corners of a generic rectangular region R, like A_+ or A_- (see Fig. 18). Then the sum S_R (as in Eqns.9 and 10) can be calculated directly as (see Fig. 19),

$$S_R = I_k^H(a_2, b_2) - I_k^H(a_3, b_3) - I_k^H(a_1, b_1) + I_k^H(a_4, b_4) \tag{12}$$

Thus finally, each such HLBP feature can also be calculated with just a few references to the pertinent Integral Histogram I_k^H, allowing our algorithm for real time implementation just as with normal Haar features.

Advantage of HLBP features over Haar features

The HLBP features involve counting the number of pixels in a region having a certain LBP label k, instead of summing over pixel intensities as with Haar features. Now, due to adverse illumination conditions, the pixel intensities in an image I may change. However, the LBP label of a pixel is much more robust to illumination changes as shown in Fig. 16. Thus, the number of pixels within a region having a particular LBP label will also remain more or less constant with varying illumination. More precisely, if we observe footnote[7], the term $I(u, v)$, the pixel intensity at location (u, v), changes with varying illumination.

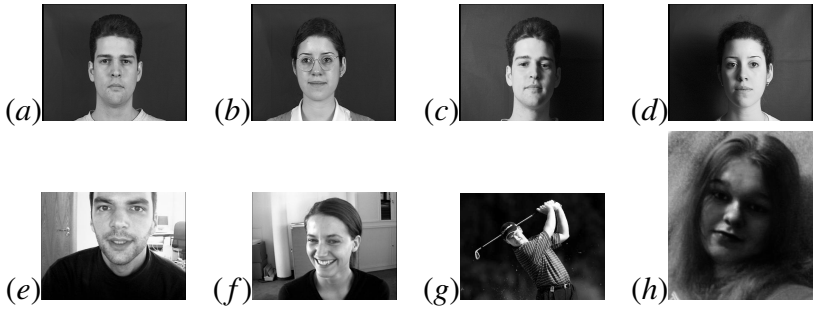


Figure 21: Example images from the databases used in our experiments: (a)-(b) XM2VTS Normal set, (c)-(d) XM2VTS Darkened set, (e)-(f) BioID database, (g)-(h) Fleuret database.

Hence the final Haar feature value will also change. In contrast, if we observe the defining Eqns. 9 and 10, in Sec. 6.1.2 for the calculation of HLBP features, we see that $I(u, v)$ has been replaced by $\delta_k(u, v)$, which is 1 if the LBP label of pixel (u, v) is k, the feature parameter, and 0 otherwise. According to definition of LBP, since LBP code is robust to illumination changes, $\delta_k(u, v)$ is also robust to illumination changes. Thus the final HLBP feature value, as defined in Eqn. 11, remains robust too. This observation has motivated us to combine the LBP concept with the Haar feature framework to obtain the advantages of both.

6.1.3 Experiments

We implemented a face detection system using our proposed HLBP features, and compared its performance against two other reference face detection systems.

Reference systems and databases used

The first reference system is the one by Viola and Jones (2001) using normal Haar features. It provides the baseline for Haar feature-based systems. The second reference system is the one by Froba et al. (Froba and Ernst, 2004; Rodriguez, 2006) using Modified Census Transform (MCT). It is one of the LBP variants representing the current state of the art. To calculate the MCT, Froba et al. compare each pixel in a 3×3 grid against the average of the intensity values within that grid, instead of the center pixel as in LBP (see Sec.6.1.2). This leads to a 9-bit code and a $511(2^9 - 1)$-bin Lookup table (LUT), each entry of which stores the log-likelihood ratio of a particular code. This LUT has to be stored for each feature. The face detector is implemented as a cascade of classifier stages, where each stage calculates the sum of LUT bins corresponding to the MCT-codes at particular locations in the test image.

We implemented our system and both the reference systems as cascades of 5 stages. Each stage had a strong classifier boosted from the set of weak classifiers (see Sec.6.1.2). The stages had 5, 10, 20, 50 and 200 weak classifiers respectively. Thus, the number of features is the same for all the 3 systems.

Database	Number of images	Illumination conditions	Other challenging aspects
XM2VTS Normal set (Messer et al., 1999)	2360	Uniform illumination	-
XM2VTS Darkened set (Messer et al., 1999)	1180	Strong side-illumination	-
BioID (Jesorsky et al., 2001)	1521	Non-uniform illumination	Images were obtained in real world conditions featuring a large variety of illumination, background and face size.
Fleuret (Fleuret, 2004)	580	Non-uniform illumination	Images from real life situations were collected from the web, showing large variations in illumination, background and face size and slight variations in pose.

Table 17: Description of the databases used in our experiments

For training, we used two internally created databases consisting of face and non-face images extracted from BANCA(Spanish Corpus) (Bailly-Bailliere et al., 2003), Essex, Feret (Phillips et al., 2000), ORL (Samaria and Young, 1994), Stirling and Yale (Belhumeur et al., 1997) databases. For testing, we used 1) the standard XM2VTS database (Messer et al., 1999; Luettin and Maitre, 2000), taking into account two cases, the Normal set with normal lighting conditions and the Darkened set with adverse or side illumination, 3) the BioID database (Jesorsky et al., 2001) and 4) an additional database from Fleuret (2004). Examples from each database used for testing are shown in Fig.21. A brief description of each database is given in Table 17.

Results and discussions

The face detection performance of the three systems are given in Fig.22 in terms of ROC curves on each of the four databases. We discuss these results and various other aspects of the system below.

Performance From Fig.22, we observe that our system (HLBP) performs reasonably well on all the four databases. However, its performance is noteworthy especially for the three cases with adverse imaging conditions, i.e., XM2VTS Darkened set, BioID database and the Fleuret database (please refer to Table 17 for more details). For the XM2VTS Darkened set, it outperforms Haar by a wide margin. Although MCT is able to achieve an initial higher True Positive Rate (TPR), HLBP is able to outperform MCT as soon as the number of false positives are allowed to reach 50. From this point onwards, MCT is not able to improve its TPR further, while HLBP is able to improve it by a significant amount. For the BioID database, HLBP performs as well as Haar and soon outperforms MCT after an initial higher TPR by MCT. MCT is not able to handle the variation in face size and pose as well as HLBP and keeps rejecting some of the faces. For the Fleuret database also, HLBP outperforms Haar by a wide margin, and outperforms MCT also, after an initial higher performance by the latter. It is true that HLBP is not able to outperform the two systems for the XM2VTS Controlled set, however this is not so significant since most real world situations would correspond to the other three cases.

Storage requirements and number of parameters In Table 18, we enlist all the parameters required to define a Haar, HLBP and MCT feature respectively. We observe that the number of parameters required is within 10 for Haar and HLBP, while it is 513 for MCT. The major difference for MCT comes from the 511-bin LUT (see Sec.6.1.3) which

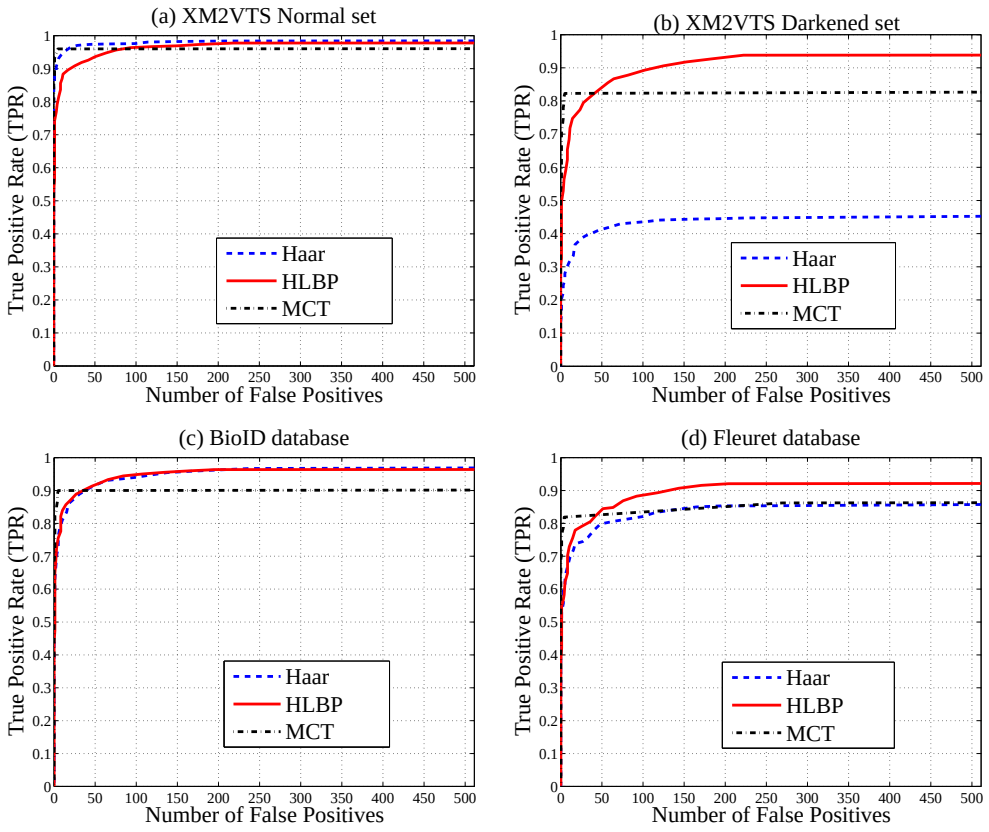


Figure 22: Comparison of face detection performance on different datasets by the three systems using Haar, HLBP and MCT features: (a) XM2VTS Normal set, (b) XM2VTS Darkened set, (c) BioID database and (d) Fleuret database.

is not required for Haar and HLBP. Thus a single MCT feature is much more complex to represent than a Haar or HLBP feature. We also give an estimate of the minimum number of bits required to store these parameters based on their ranges and types. For Haar and HLBP, it is around $26 + 2 \times N_f$ bits, where N_f is the number of bits required to store a floating point number. For MCT, it is $10 + 511 \times N_f$. With $N_f = 32$ bits or 4 bytes, the value used in our system, Haar requires 86 bits, HLBP 90 bits and MCT requires 16362 bits. Thus, MCT has a much higher storage complexity than HLBP and Haar in terms of bits per feature and also in terms of total number of bits to represent the model, since exactly the same number of features were used for all the three systems (see Sec. 6.1.3). Thus HLBP is able to achieve comparable results with MCT using a model as simple as Haar but much simpler than MCT. This justifies the use of HLBP in low memory applications involving embedded devices and mobile phones rather than MCT. Further, a model with higher number of parameters (MCT) entails a higher classification risk at test time due to overfitting on the training set (Vapnik, 1989).

Training and test time At first glance, the total number of possible features should be 16 times more for HLBP than for Haar since every Haar feature can be associated with one out of 16 possible LBP labels to give one HLBP feature. However, since HLBP is derived from histograms or counts of the $LBP_{4,1}$ labels and not the pixel intensity themselves, we

Parameter Type	Number of parameters	Range/Type of each parameter	Minimum number of bits per parameter	Total number of bits required	Haar	HLBP	MCT
Location	2 (x, y)	1-19	5	10	✓	✓	✓
Size	2 (w, h)	6-19	4	8	✓	✓	-
Mask Type, T	1	1-5	3	3	✓	✓	-
Direction, p	1	$\{-1, 1\}$	1	1	✓	✓	-
LBP Label, k	1	1-16	4	4	-	✓	-
Feature weight, α	1	`float`	N_f	N_f	✓	✓	-
Threshold, θ	1	`float`	N_f	N_f	✓	✓	-
Lookup Table (LUT)	511	`float`	N_f	$511 \times N_f$	-	-	✓
			Total number of parameters per feature		8	9	513
			Total number of bits per feature		22+ $2 \times N_f$	26+ $2 \times N_f$	10+ $511 \times N_f$

Table 18: Comparison of storage requirements (in bits) and the number of free parameters per feature of the 3 systems, Haar, HLBP and MCT (Froba and Ernst, 2004). Each row lists a parameter and a checkmark (✓) in a particular column indicates that this parameter is required for the definition of the corresponding feature. Please refer to Sec.6.1.2 (Eqn.11) and Sec.6.1.3 for more details about each parameter. Here N_f denotes the number of bits required to store one floating point number. It is compiler-dependent. In our setup it is 32 bits or 4 bytes, a typical value.

do not use all possible windows at all locations and scales, but only use windows which have a minimum size of 6 pixels. This is because smaller sized windows would not be useful in filling up the histogram. This reduced the number of features to around 100,000 which compares favorably with the Haar feature set which number around 64,000, for a window size of 19×19. This leads to comparable training times for the two algorithms. In fact, HLBP is able to reject about 81.2% of the non-faces in the first stage compared to 75.5% for Haar, leading to a further reduction in its training time. For MCT, a 511-bin LUT needs to be calculated for each individual feature (see Sec.6.1.3) which is avoided by our system, thus making it faster. For testing, we use exactly the same setup (number of stages and number of classifiers at each stage) for the three systems, the only difference from Haar being the calculation of the $LBP_{4,1}$ image as a preprocessing in HLBP. However, the calculation of the $LBP_{4,1}$ image can be done in one pass over the image using only two relational operations per pixel. Also, this operation is only needed once per scale. Hence, the relative increase in computation time is negligible. MCT also requires a similar preprocessing step as for HLBP (see Sec.6.1.3).

Originality of proposed method Certain other systems also involve either Local Binary Patterns and / or boosted Haar-like features, similar to Viola and Jones. However, they are different from our proposed system. The Multi-Block Local Binary Pattern (Zhang et al., 2007) and Locally Assembled Binary Feature (Yan et al., 2008) extend the idea of LBP by comparing sums of intensities over image patches to calculate the LBP label itself. The object detection framework by Zhang et al. (2005) uses the concept of spatial histograms of Local Binary Patterns. Their features measure the similarity between model and test histograms using histogram intersection (Schiele, 1997). However, none of these methods compare counts of individual LBP labels in two regions as we do. Our method tries to capture the region-specific variation of certain local texture patterns, which is not done

in Zhang et al. (2007); Yan et al. (2008); Zhang et al. (2005). Wang et al. (2006) have used Fisher Linear Discriminant on Histogram features for Face Detection. However, there is no use of LBP concept which is the major contribution of our work. Furthermore, the inclusion of Fisher Linear Discriminant increases the computational complexity at test time.

6.1.4 Conclusion

In this work, we have introduced a new type of feature called the HLBP feature which combines the concepts of Haar feature introduced by Viola and Jones, with Local Binary Patterns, harnessing the advantages of both for the problem of face detection. Our features are able to model the region-specific variations of local texture and are relatively robust to wide variations in illumination, pose and background, and also slight variations in pose. Experiments have shown that our system performs significantly better in such adverse imaging conditions than normal Haar features and performs reasonably better than MCT features with much less storage and computation requirements.

6.2 Face Recognition using Bayesian Networks to combine intensity and color information

6.2.1 Introduction

Face recognition is an active research area, probably because of its numerous applications, ranging from video surveillance to human-computer interaction for instance. Hence, there exists numerous systems allowing to recognize people based on their face image. The vast majority of such existing approaches typically act on grayscale images only, since color is usually considered to introduce high variability. Nevertheless, it was shown that color plays an important role in human face recognition (Russell et al., 2006; Sinha et al., 2006). It is thus likely that it may also carry useful information for computer-based systems. Surprisingly, only a few studies are using color in automatic face recognition systems. Torres et al. (1999) developed a color Eigenfaces system, where Principal Component Analysis (PCA) is independently applied on each color channels and results are then combined for final classification. They showed that an improvement is obtained over traditional Eigenfaces acting on grayscale images (Turk and Pentland, 1991). This result was later confirmed in a study by Gutta et al. on the larger FERET database (Gutta et al., 2001). Another interesting study is due to Sadeghi et al. (2007): different channels from numerous colorspaces are first classified independently thanks to Linear Discriminant Analysis (LDA). An optimal subset of such classifiers is then found, and selected classification scores are combined using Support Vector Machines. Another approach proposed in Jones and Abott (2006) consists in extracting color features to use them as input to an Elastic Graph Matching algorithm. Again, color features were shown to perform better than grayscale-based ones.

Bayesian Networks provide an elegant framework to describe relationships (and hence correlations) between different pieces of information. In this work, our aim is to derive models describing the process that generates observations of different nature extracted from face images. Such generative models will then be used for recognition purposes.

Specific generative models for face recognition were recently proposed by Heusch and Marcel (2007). The authors proposed a tree-structured Bayesian Network to describe data extracted from grayscale face images. More precisely, they assumed that observations derived from salient facial features are related to each other, and hence tried to model correlations between such observations. Going one step further, we believe that such models are suitable to model correlations between local, grayscale-based features and other information, such as color. Hence, in this contribution, new models integrating color at the global face level, the local facial feature level but also at both levels are derived. Experimental evaluation on face authentication is carried out on the XM2VTS (Messer et al., 1999) and the BANCA (Bailly-Baillière et al., 2003) databases. Results show that integrating color intelligently into dedicated generative models may help at reducing the authentication error rate, at least when the training and testing acquisition conditions are quite similar. The best proposed model, correlating grayscale and color at both the local and the global level show a significant improvement in performance as compared to a similar baseline model acting on the luminance channel only (Heusch and Marcel, 2007), but also performs better than the Eigenfaces-based system acting on color channels independently (Torres et al., 1999). Obtained results thus suggest that color is of valuable information when combined to grayscale in a coherent manner.

In the next section, Bayesian Networks are briefly introduced, before describing the proposed models and the features in Sec. 6.2.3. Sec. 6.2.4 details the experimental framework and Sec. 6.2.5 describes the databases and discusses the obtained results. Finally, a conclusion is drawn in Sec. 6.2.6.

6.2.2 Bayesian Networks

A Bayesian Network is a probabilistic graphical model representing the joint probability distribution over a set of random variables, and having the ability to encode dependencies among these variables (Pearl, 1988). It is specified as a directed acyclic graph, where nodes represent random variables and directed links represent causal relationships between these variables. Defining a set of random variables $\mathbf{U} = (x_1, ..., x_n)$, the joint probability defined by a Bayesian Network is given by the following chain rule:

$$P(\mathbf{U}) = \prod_{i=1}^{n} P(x_i|Parents(x_i)) \tag{13}$$

where $Parents(x_i)$ denotes the set of parents of the node representing the variable x_i. Hence, a Bayesian Network is fully defined by the structure of the graph, and by its parameters: the conditional distribution of each variable given its parents. Computing probabilities in the network is referred to as *inference*. It is typically done to update the state knowledge of a set of hidden variables when other variables, referred to as *evidence*, have been observed. In our case, inference is carried out thanks to the Junction Tree Algorithm (Cowell et al., 1999). Another important issue is how to learn such models from data. Learning in Bayesian Networks may either refer to structure learning, parameters learning or both. Since in this work, the structure is derived according to prior domain knowledge, the focus is made on learning the parameters. As the proposed

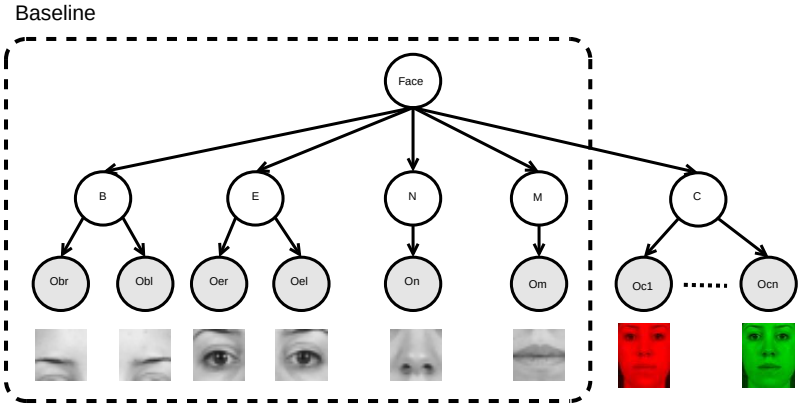


Figure 23: Bayesian Network model for the face incorporating color information at the global face level. Gray nodes represent the observations extracted from the face image. White nodes are the hidden variables describing 'types' of observations.

models contain hidden variables, a natural choice is the Expectation-Maximization (EM) algorithm (Dempster et al., 1977).

6.2.3 Proposed Models

Reference Model

The model presented in Heusch and Marcel (2007) relies on the assumption that facial features are related to each other. To model such relationships, a tree-structured Bayesian Network is proposed. This generative model assumes that there exists pairwise relationships between observations derived from grayscale images. Namely, relationships between eyebrows and eyes, eyes and nose and nose and mouth are considered. This model performs better than a simpler generative model (based on Gaussian Mixtures Model) were independence between facial features is assumed. However, it only acts on local features derived from intensity images and does not take advantage of the Bayesian Networks framework to integrate other source of information in a smart manner.

Color at the Global Level The first proposed model is depicted in Fig. 23 and should be understood as follow: the root node is used to relate various information describing the face. A face thus consists in a relationship between different 'types' of facial features (nodes **B**: eyebrows, **E**: eyes, **N**: nose and **M**: mouth). In addition, a 'type' of color is also modelled through node **C**. This hidden node *causes* observations derived in each color channel: it is hence assumed that information coming from different color channels are explicitly correlated. Finally, the different types of facial feature, as well as the type of color, generates the corresponding observations extracted from the face image. Note also that, unlike the reference model (Heusch and Marcel, 2007), a single hidden node is used to model the relationship between the different observations. Actually, global color information has to be related to the whole face rather than to pairwise relationships between facial features.

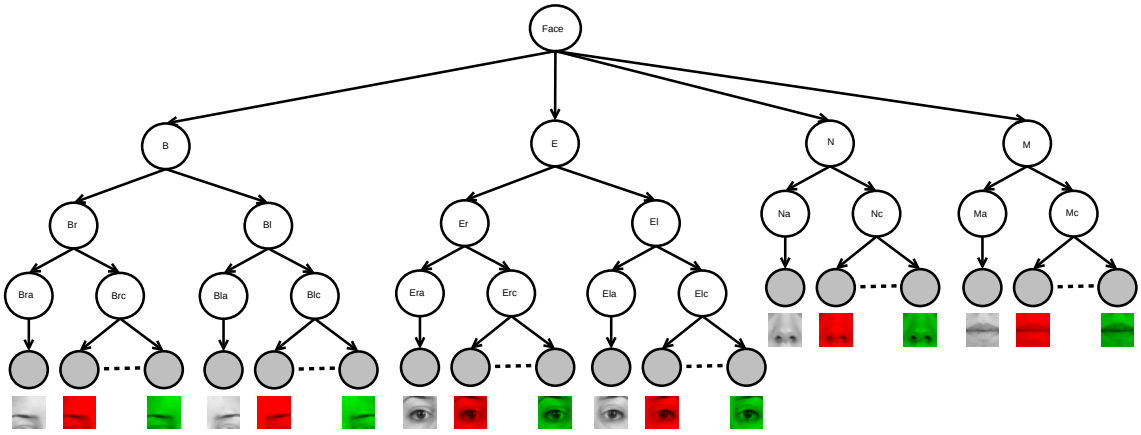


Figure 24: Bayesian Network model for the face incorporating color information at the local facial feature level.

Color at the Local Level To model the face more accurately, we also propose to incorporate color information at the local facial feature level. For this purpose, every type of facial feature is not only *explained* using grayscale appearance, but also with color information, as depicted in Fig. 24. Note that two additional layers of hidden nodes were introduced in this model. The first one aims at separating the left and right part of 'symmetric' observations (such as eyebrows and eyes, the separation is made through nodes **Br**, **Bl**, **Er** and **El**). This is done to incorporate color information directly at the facial feature level: we choose to correlate the grayscale appearance of a particular facial feature with its color information instead of correlating color of symmetric features together. The purpose of the second additional layer is to separate the appearance of the facial feature from its color.

Color at Both Levels Color information at the global level is used to represent the global skin color of the face. On the other hand, color information at the local level aims at modelling the color of the associated facial feature. Hence, these two different observations may provide complementary information. As a consequence, we also propose a model combining both global and local information. This model consists in the local model depicted in Fig. 24, where the branch of the global model (Fig. 23) corresponding to color information is added.

Feature Extraction

As observations are derived around facial features, they are first located in the face image using an Active Shape Model (ASM) (Cootes et al., 1995). For local grayscale observations, the same feature extraction scheme as in Heusch and Marcel (2007) is applied here. The original color image is converted to grayscale. Multiple squared windows are then cropped around each facial feature, by adding shifts of a variable amount of pixels. Each extracted window is preprocessed using histogram equalization in order to enhance its contrast. Finally, a feature vector is obtained by applying a two-dimensional Discrete Cosine Transform (2D-DCT) on each preprocessed window.

Global Color Observations In this framework, our aim is to extract skin color information. Hence, a bounding box containing only the inner part of the face is cropped from the color image, based on eyes position. After being preprocessed by histogram equalization, it is subsampled to yield a low-resolution representation of the face, which discards details and thus mainly contains skin-colored pixels. Finally, feature vectors representing color are obtained by decomposing each color channel in terms of 2D-DCT.

Local Color Observations Regarding the color observations at the local feature level, the same windows as for the grayscale observations are cropped (using shifts as well), but from the original color image. Then, each extracted window is preprocessed by histogram equalization. Finally, feature vectors are obtained by decomposing each color channel in each window in terms of 2D-DCT.

6.2.4 Face Authentication and Performance Measures

Face authentication consists in confirming (or denying) a client's claim supported by its face image. In such a framework, either the claimant provides its real identity, either it is trying to fool the system (it is then referred to as an *impostor*). The system has thus to make a decision on whether the claimant is a true client or an impostor. Since modelling all possible impostors is not feasible, a so-called *world model* is trained thanks to the EM algorithm with the Maximum Likelihood (ML) criterion (Dempster et al., 1977) using data coming from different identities. In face authentication, there are usually few training examples available for each client, and hence Maximum Likelihood estimates of the parameters for the client specific models may be inaccurate. To tackle this problem, a form of Maximum A Posteriori (MAP) adaptation (Gauvain and Lee, 1994) is used to adapt client models from a nearby distribution, given by the world model. This approach was already successfully applied to this task Cardinaux et al. (2005); Heusch and Marcel (2007).

When using generative models, authentication decision is typically performed by taking the likelihood ratio between the model corresponding to the claimed identity and the world model, which is used to represent arbitrary impostors. Given a client's claim supported by its face representation $\mathbf{X}$ (i.e. the set of observations derived from the face image, as depicted on Figures 23 and 24), the decision is made according to:

$$\Lambda(\mathbf{X}) = \log p(\mathbf{X}|\theta_C) - \log p(\mathbf{X}|\theta_{world}) \tag{14}$$

where $p(\mathbf{X}|\theta_C)$ is the probability that the client's model θ_C has generated the data $\mathbf{X}$ and $p(\mathbf{X}|\theta_{world})$ is the probability that the data were generated by an impostor. Based on a threshold τ, the claim is accepted if $\Lambda(\mathbf{X}) \geq \tau$.

In a face authentication framework, two kinds of error can occur: either the true claimant is rejected (false rejection), or an impostor is accepted (false acceptance). Hence, authentication results are typically presented using the Half Total Error Rate, which combines

the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) and is defined as:

$$HTER(\tau) = (FAR(\tau) + FRR(\tau))/2 \quad [\%] \tag{15}$$

6.2.5 Experiments & Results

In this work, we are interested in seeing if color is a valuable source of information for the face authentication task and, more importantly, we also would like to know if combining color and grayscale intelligently is better than treating such information independently. Color is here encoded in the HSV colorspace, since it was previously shown to be suitable for various computer vision tasks. Results are reported using the three proposed models (referred to as *Global-HSV*, *Local-HSV* and *Both-HSV*), but also with three baseline algorithms. The first one is similar to Heusch and Marcel (2007) and consists in the first proposed model, but where the color part has been discarded (see Fig. 23) and is referred to as *Local-gray*. The second and third one are our own implementations of the traditional Eigenfaces algorithm (Turk and Pentland, 1991) (*PCA-gray*) and of the color-based Eigenfaces (*PCA-HSV*) proposed in Torres et al. (1999).

Results presented throughout this work are obtained using the following settings for feature extraction: windows of size 24x24 pixels were extracted from the original images based on the results given by the ASM, with shifts of 2, 4 and 6 pixels in each directions. Regarding global color information, faces windows of size 64x80 pixels were first cropped and then subsampled to 24x24 pixels. We kept the first 64 DCT coefficients to build the final feature vectors. For the Eigenfaces-based system, 95% of the variance was kept, and the standard correlation was used as the metric. All these hyper-parameters, as well as the threshold τ, were selected by minimizing the Equal Error Rate (EER: when the FAR equals the FRR) on a separate validation set.

Experiments on the XM2VTS Database

The XM2VTS database (Messer et al., 1999) contains 295 identities, among which 200 are used as true clients and the remainder are used to simulate impostors. Recordings were acquired during four sessions under controlled conditions and covering a time period of five months. Along with the database, two experimental protocols, specifying which images have to be used for training, validation and testing have been defined. In Table 19, HTER performance with its 95% confidence interval is reported for the various systems using both XM2VTS protocols.

One can first remark that color is indeed a valuable information. Actually, results obtained with the color-based Eigenfaces algorithm significantly outperforms the classical Eigenfaces, as expected. However, it reaches almost the same performance as the reference generative model acting on grayscale only. This suggest that treating color channels as independent sources of information is not really successful. Note that, on the other hand, the proposed generative models integrating color in a principled way yields better performance than the similar model based on grayscale features only (*Local-gray*), again emphasizing the usefulness of color information in face processing.

System		HTER [%]	
		LP1	LP2
Baselines	Local-gray	2.74 (0.96)	2.43 (0.84)
	PCA-gray	5.32 (1.07)	4.28 (1.07)
	PCA-HSV	3.09 (0.87)	1.94 (0.49)
Proposed	Global-HSV	2.33 (0.90)	1.25 (0.69)
	Local-HSV	1.98 (0.87)	0.98 (0.60)
	Both-HSV	**1.22** (0.59)	**0.59** (0.42)

Table 19: HTER Performance on the test set of XM2VTS.

Correlating color and grayscale information (*Global-HSV*) seems to be better than treating such information independently (*PCA-HSV*). This is again evidenced by the results obtained with the proposed generative models: they all perform better than the color-based Eigenfaces system. Regarding the different proposed generative models, we can see that modelling color at the facial feature level consistently yields better results than using information derived from the whole face image. This result is not surprising since the local model is inherently more accurate than the global one. Note finally that the best performance is achieved with the model taking both global and local color information into account (*Both-HSV*), suggesting that both clues are valuable and complementary to describe an identity.

Experiments on the BANCA Database

To assess the validity of our approach, experiments were also carried out with the baseline systems and the proposed generative models on the more challenging BANCA database (Bailly-Baillière et al., 2003). This database contains 52 clients (English corpus), equally divided into two groups $g1$ and $g2$ used for validation and test respectively. An additional set with 10 images of 30 other subjects is also provided as the world model. In this database, image acquisition was performed according to three different scenarios: Controlled (high-quality camera, uniform background, controlled lighting), Degraded (web-cam, non-uniform background) and Adverse (high-quality camera, arbitrary conditions). Examples of acquisition conditions can be seen on Fig. 25. There exists several experimental protocols defining which scenarios and which images have to be used for enrollment and testing. In this study, the protocols Mc, Ua, Ud, P and G have been considered.

In Table 20, HTER performance with its 95% confidence interval for the different protocols is reported on the test set $g2$. Obtained results show that when the acquisition conditions are not well-controlled, global skin color is not a reliable clue anymore. This is evidenced by the performance obtained with the Eigenfaces-based system, but also with generative models taking this information into account (i.e. *Global* and *Both*). Indeed, the baseline generative model based on grayscale features (*Local-gray*) performs better when there is a strong mismatch between training and testing conditions. However, modelling the color at the local facial feature level achieves good results: when the training and testing conditions are the same (protocols Mc and G), local color information improves the

(*controlled*) (*degraded*) (*adverse*)

Figure 25: Example of the different scenarios in the BANCA database.

System		HTER [%]				
		Mc	Ua	Ud	P	G
Baselines	Local-gray	2.24 (0.93)	20.51 (2.64)	**19.90** (2.59)	**16.52** (1.38)	6.85 (0.95)
	PCA-gray	20.38 (2.62)	41.60 (3.17)	37.02 (3.17)	34.51 (1.80)	21.84 (1.56)
	PCA-HSV	14.71 (2.30)	34.07 (3.11)	32.34 (3.08)	29.78 (1.72)	18.48 (1.47)
Proposed	Global-HSV	5.19 (1.41)	28.11 (2.94)	32.08 (3.04)	24.22 (1.63)	11.50 (1.22)
	Local-HSV	**1.89** (0.87)	**17.24** (2.43)	20.77 (2.65)	18.80 (1.41)	**5.58** (0.86)
	Both-HSV	3.21 (1.09)	21.79 (2.72)	22.02 (2.66)	19.94 (1.52)	6.31 (0.93)

Table 20: HTER on the test set $g2$ of the BANCA database.

performance, and clearly outperforms the Eigenfaces-based systems. Another interesting result is obtained with protocol Ua: in this case, even if the training/testing conditions are different, the model integrating color performs better than the baseline system. This can be explained by the fact that the same acquisition device was used in both scenarios, and thus color seems to remain consistent across controlled and adverse conditions (see Fig. 25).

6.2.6 Conclusion

In this contribution, new generative models based on Bayesian Networks were proposed to tackle the face authentication task. The purpose of these models was to integrate color in a principled way into a local-feature based model acting on grayscale observations only. To do so, new models were derived: they combine color and grayscale information either at the global face level, at the local facial feature level or at both levels. Face authentication experiments were conducted on two different benchmark databases. Obtained results showed that improvement can be gained when color is combined to grayscale as additional information. Namely, we showed that the proposed models are suitable for the face authentication task, at least when the acquisition conditions between enrollment and testing are quite similar. In particular, the model taking color into account at both the global and the local level significantly outperforms a similar baseline system acting on grayscale-based features, as well as a color-based Eigenfaces algorithm (Torres et al., 1999). Correlating different sources of information thus seems to be more effective than treating them independently. However, when there is a strong mismatch between training and testing conditions, color information may become confusing. An obvious possible

future direction is hence the investigation of other colorspaces, and particularly the one taking the illuminant into account (i.e. CIE-XYZ and its derivatives). Besides, it would also be interesting to combine more than one colorspace representation, since this approach was shown to yield good results (Sadeghi et al., 2007).

7 Speaker Identification

Most of the work conducted on speaker identification in the last reporting period by TNO and BUT concentrated on Joint Factor analysis — a framework, that is nowadays in the center of interest of the whole speaker identification community. The work followed closely the 2008 NIST speaker recognition evaluation[8] and 2008 JHU workshop workgroup Robust Speaker Recognition Over Varying Channels that was headed by Lukas Burget[9].

7.1 JFA in speaker identification

In Burget et al. (2009a), we have brought a consolidated report of BUT system from 2008 NIST SRE evaluations. JFA systems built according to Kenny's recipe (Kenny et al., 2008) perform excellently – it was hard to find another complementary system that would contribute to fusion of our two JFA systems. Especially for the matched condition, a single JFA system is as good as system combination. Although our system was primarily trained on and tuned for telephone data, JFA subsystems can be simply augmented with eigenchannels trained on microphone data, which makes the system performing well also on microphone conditions. Another significant improvement was obtained by training additional eigen-channels on data with matching channel condition, even thought there was very limited amount of such data.

In Burget et al. (2009b), we have investigated into variants of Joint Factor Analysis for speaker recognition. We performed systematic comparison of full JFA with its simplified variants and confirmed superior performance of the full JFA with both eigenchannels and eigenvoices. We investigated into sensitivity of JFA on the number of eigenvoices both for the full one and simplified variants. We studied the importance of normalization and found that gender-dependent zt-norm was crucial. The results were also reported on NIST 2006 and 2008 SRE evaluation data.

Finally, in Glembek et al. (2009), we have compared scoring methods used in speaker recognition with Joint Factor Analysis. Different log-likelihood scoring methods, that different sites used in the latest state-of-the-art Joint Factor Analysis (JFA) Speaker Recognition systems were studied. The algorithms use various assumptions and have been derived from various approximations of the objective functions of JFA. We have compared the techniques in terms of speed and performance. We show, that approximations of the true log-likelihood ratio (LLR) may lead to significant speedup without any loss in performance.

7.2 JFA Matlab Tutorial Demo

To facilitate the research into JFA (with its non-trivial mathematics), BUT prepared a Matlab toolkit for JFA including training and test data (permission was obtained from LDC and NIST) and made it publicly available[10]. The tutorial already received very

[8] http://www.nist.gov/speech/tests/sre/2008/
[9] http://www.clsp.jhu.edu/workshops/ws08/groups/rsrovc/
[10] http://speech.fit.vutbr.cz/en/software/joint-factor-analysis-matlab-demo

positive reactions at major speech conferences (ICASSP and Interspeech in 2009).

7.3 Speaker verification as a target-nontarget trial task

In Hubeika (2009), we presented a preliminary study on the formulation of speaker verification as a target-nontarget trial task. In the standard approach, each speaker is modeled by their own model and the task is to decide whether the test speech segment was generated by the given model or not. In this work, only two models are used: one represents the target trials and the other represents nontarget trials, where the trial is represented by two speech segments, both from the same speaker, and two from different speakers, respectively. As the input features, fixed-length low-dimensional vectors derived from speaker factors generated by Joint Factor Analysis are used. Gaussian Mixture Models framework is used to model the feature distribution. The achieved results are compared to the state of the art systems.

7.4 Study of feature extraction and implementation issues

7.4.1 Final independence of Abbot PLP-features

On long-awaited effort at TNO was to make speaker and language recognition systems independent of a particular feature extraction implementation, SoftSound's (now Autonomy's) tool `plp`. This implementation uses a bi-linear transform for approximation of the Bark frequency scale, and has always led to best results in any experiments carried out at TNO. By running TNO's complete SRE-2008 implementation it was possible to investigate the various options of Dan Ellis's implementation of PLP feature extraction. Finally we could reproduce the baseline performance measure of EER = 7.17 % with Abbot's `plp` to 7.15 % with ICSI's `feacalc`, by using Mel-scale frequency warping and turning off RASTA processing. Results are for the SRE-2008 telephone-telephone condition including all trials (AKA "NIST condition 6"), using only a single TNO sub-system.

7.4.2 Improvements in efficiency

TNO's system was redesigned to collect all relevant statistics of an utterance in a single file, the dependence on Matlab's `eigs()` for computing the NAP matrix was removed by re-linking this ARPACK function to GPL'ed Octave, and system efficiency was improved by computing UBM Gaussian occupation indexes in C and linking to AMD's ACML libraries.

7.4.3 Dot-scoring

On the algorithmic level, the new organization of speech segment statistics allowed for an easy implementation of dot-scoring, a UBM-GMM approach where scoring is approximated by a Taylor expansion of the Gaussian likelihood function. Experiments comparing the baseline GMM-SVM with the dot-scoring system were carried out, where the SVM's negative-example cohort were used as Z-norm segments for dot-scoring, and the NAP

speaker segments were used for estimating channel factors for dot-scoring. The best dot-scoring system scored 7.57 % on the test mentioned above, compared to the GMM-SVM system performing at 7.15 %. Interestingly, the dot-scoring did not improve by using gender-dependent models and scoring, contrary to what is reported in literature. This may be a result of still using a gender-independent UBM from which all statistics are computed. In the TNO system, different UBMs can be used in parallel, conditioned on speaker male or female speakers or gender-independent, which can be effectively fused.

7.5 Forthcoming events – Odyssey 2010 and NIST SRE 2010

From 28 June – 1 July 2010, BUT will organize ISCA Odyssey 2010: The Speaker and Language Recognition Workshop[11]. The aim of this workshop is to continue to foster interactions among researchers in speaker and language recognition as the successor of previous successful events held in Martigny (1994), Avignon (1998), Crete (2001), Toledo (2004), San Juan (2006) and Stellenbosch (2008).

The NIST 2010 SRE evaluation[12] workshop will precede Odyssey and will take place 24 - 25 June 2010.

[11] http://speakerodyssey.com/
[12] http://www.itl.nist.gov/iad/mig/tests/sre/2010

8 Gestures and Actions

Research on vision-based gesture and action recognition aims at real-time recognition of human activity in smart environments. This requires invariancy to viewpoint, lightning, background and person appearance in a video. As taking into account all these sources of variation would severely hinder efficient training of human action models, we propose to use recovered poses as an intermediate representation. There is significant variation in the performance of motion of a certain class, especially in natural settings. A discriminative classification approach can effectively distinguish between classes, rather than modeling each. We discuss here an approach that addresses full-body motion.

8.1 Approach

We describe an approach that is a combination of previous work (Poppe, 2007; Poppe and Poel, 2008) on example-based pose recovery and action classification based on Common Spatial Patterns (CSP). Given a video sequence containing human motion, the pose is estimated for each frame. A pose consists of the 3D locations of 20 key joints, relative to the root. The temporal variance in the pose is used to classify the action. This classification is based on a pair-wise comparison between action prototypes.

8.1.1 Pose recovery

The image region within a located window is described as a histogram of oriented gradients (HOG), where only those pixels that belong to the foreground are taken into account. The total 270D descriptor is normalized to unit length. This makes the representation somewhat invariant to person appearance whereas the use of foreground masks ignores the influence of the environment. To recover the pose, the descriptor is matched against image descriptors in the training sets. The estimated pose is the weighted interpolation of the poses that correspond to the best matches. Details are in Poppe (2007, 2009).

To make the pose representation invariant to viewpoint and person dimensions, we normalized the pose for rotation around a vertical axis, and scaled the distances between all joints uniformly so that the height of the person was constant.

8.1.2 CSP based classification

CSP is a spatial filter technique often used in classifying brain signals (Müller-Gerking et al., 1999). It is a two-class discriminative approach based on differences in variance of temporal features. After applying 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. The value of 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 ($a \neq b$). First, we calculate the CSP transform $W_{a,b}$, then we apply $W_{a,b}$ to

each training sequence of 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 an observed 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:

$$g_{a,b}(x) = \frac{\|\bar{b} - x'\| - \|\bar{a} - x'\|}{\|\bar{b} - x'\| + \|\bar{a} - x'\|} \tag{16}$$

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

$$g_a(x) = \sum_{a \neq b} g_{a,b}(x) \tag{17}$$

and x is classified as the action a for which $g_a(x)$ is maximal (Poppe and Poel, 2008).

8.2 Experimental results

This combined approach was evaluated on the HumanEva-I dataset (Sigal and Black, 2006). This dataset contains sequences with different actions performed in an uncontrolled manner by three persons and from different viewpoints. Moreover, human pose information is present to validate the accuracy of the pose recovery step. The different actions within each sequence are manually labeled with action labels. Table 21 summarizes the number of action segments and the total number of frames for each action class.

After recovering and normalizing the poses, we applied the CSP transform. To avoid singularity problems, the first 30 principal components were selected. The $2k = 6$ components are used to describe the action prototype. The sequences are recorded with 60 frames per second. We evaluated our approach on several different sublengths. The number of training and test sequences for these lengths are summarized in Table 22. Some actions, such as Punch and Throw, are relatively short and only sub-sequences of length 30 and 60 are available. These correspond to 0.5 and 1 second respectively.

The results of the combination of pose estimation and CSP-based classification are summarized in Table 23. For comparison, we calculated also the performance without CSP, with action sequences reduced to 30 or 6 PCA components. The baseline based on *a priori* class probabilities are 36.18%, 43.48%, 48.25% and 53.59% for 30, 60, 90 and 120 frames respectively.

An important factor in the classification performance is the high number of walking and jog sub-sequences, c.f. Table 22. In general, these were classified correctly more often than other classes. For sub-sequence lengths of 30 frames, the total number of walking and jog sub-sequences is 64.35% of all sub-sequences. For a sub-sequence length of 120, this share is 95.42%. The fact that these walking motions are performed in a circle shows the ability of our method to cope with different viewpoints. A factor that contributed to the higher performance for longer sub-sequences

	Training			Test		
Action	S1	S2	S3	S1	S2	S3
Rest	5 (453)	7 (624)	8 (373)	5 (370)	13 (780)	10 (461)
Walking	1 (1203)	2 (974)	1 (939)	3 (2162)	4 (2078)	3 (1515)
Jog	1 (740)	1 (795)	1 (842)	2 (1603)	2 (1328)	2 (1390)
Punch r.	5 (225)	7 (352)	9 (448)	3 (163)	7 (348)	5 (210)
Punch l.	4 (200)	6 (258)	9 (328)	3 (120)	6 (335)	5 (187)
Uppercut r.	2 (89)	1 (39)	3 (195)	3 (146)	2 (134)	2 (133)
Uppercut l.	2 (136)	1 (45)	1 (36)	3 (177)	0 (0)	0 (0)
Wave r.	4 (449)	2 (100)	5 (511)	5 (516)	4 (252)	2 (224)
Wave l.	0 (0)	1 (80)	0 (0)	0 (0)	2 (117)	0 (0)
Beckon r.	3 (352)	4 (301)	4 (516)	5 (554)	4 (346)	3 (329)
Beckon l.	0 (0)	2 (128)	0 (0)	0 (0)	1 (93)	0 (0)
Throw low r.	2 (139)	1 (108)	3 (240)	1 (104)	3 (322)	3 (240)
Throw side r.	0 (0)	1 (104)	0 (0)	1 (83)	0 (0)	0 (0)
Throw high r.	2 (180)	1 (120)	2 (172)	2 (175)	2 (193)	2 (169)
Catch	4 (243)	3 (246)	5 (317)	4 (219)	5 (358)	6 (353)

Table 21: Number of segments for different actions and subjects in the training and test set of the HumanEva-I dataset. Numbers between brackets are the total numbers of available frames. Abbreviation r. stands for right, l. stands for left.

contain more information. The 30 frames, or half a second, often contain only part of the whole action. For example, a throw action can take two seconds. When an action is only observed for half a second, characteristic movement might not be taken into account. A more detailed analysis can be found in Poppe (2009).

8.3 Conclusion and Future Work

We have combined example-based pose recovery with a CSP classifier to recognize human actions using recovered poses. Combining these two algorithms solves several challenges that are difficult to address in a single step, such as dealing with variations in viewpoint, different background and changes in lightning conditions. Moreover, the combined approach has the advantage that action models can be trained using motion capture data, if the motion capture data and pose recovery adhere to a similar pose representation. We have evaluated our approach on the HumanEva-I dataset. Using CSP proved to be advantageous, both in recognition performance and in the number of dimensions that was used to describe the action prototype vectors.

These experiments show the efficacy of our method. The next step is to apply this work in the area of smart meeting rooms. This will require the assembly of a dataset that is both annotated in poses and action labels. Given such a set, we can construct the pose recovery mapping and calculate the CSP transforms.

Sub-sequence length	30 frames		60 frames		90 frames		120 frames	
Action	Train	Test	Train	Test	Train	Test	Train	Test
Rest	68	65	21	11	10	5	5	2
Walking	202	368	99	178	63	111	46	82
Jog	155	280	75	136	48	88	36	64
Punch r.	39	27	5	3	2	0	0	0
Punch l.	25	23	3	1	0	0	0	0
Uppercut r.	13	18	2	4	0	0	0	0
Uppercut l.	10	7	2	1	0	0	0	0
Wave r.	56	51	19	18	8	8	2	2
Wave l.	4	5	1	1	0	0	0	0
Beckon r.	60	65	21	22	10	11	3	3
Beckon l.	6	5	1	2	0	0	0	0
Throw low r.	24	34	7	11	2	4	0	0
Throw side r.	5	4	2	1	1	0	0	0
Throw high r.	25	27	9	9	3	3	1	0
Catch	36	38	9	8	0	0	0	0

Table 22: Number of available training and test sub-sequences for different sub-sequence lengths.

Sub-sequence length	30 frames	60 frames	90 frames	120 frames
Number of classes	15	15	7	5
CSP classifier	76.89%	87.20%	92.17%	93.46%
Without CSP (30D)	68.04%	71.50%	69.13%	81.70%
Without CSP (6D)	63.72%	65.46%	69.57%	71.24%

Table 23: Classification performance on the HumanEva-I human action dataset, for different sub-sequence lengths and corresponding action subsets.

9 Social Signals

9.1 Multimodal Laughter Detection

Efforts have concentrated on an audiovisual approach to distinguishing two types of laughter from speech and we show that integrating the information from audio and video channels leads to improved performance over single-modal approaches.

Two types of experiments were performed: 1) discrimination between laughter and speech, and 2) discrimination between two types of laughter (voiced and unvoiced) and speech. We experimented with different combinations of cues with the most informative being the facial expressions and the cepstral features for the first experiment and prosody and cepstral features for the second experiment. We used decision- and feature-level fusion to integrate information from the audio and video channel with feature fusion achieving slightly better performance than decision fusion.

When tested on 206 audiovisual sequences (8590 video frames), depicting spontaneously

displayed (as opposed to posed) laughter and speech episodes (AMI corpus), in a person independent way, the proposed audiovisual approach achieves for frame / window-based classification over 90% classification rate for speech vs laughter and over 80% classification rate for the discrimination of voiced laughter, unvoiced laughter and speech.

A full paper containing more details of this work has been submitted to *IEEE Transactions on Multimedia* (Petridis and Pantic, submitted).

9.2 Dominance/Activity Detection

In every face-to-face meeting – even if the participants do not know each other – an order of dominance is established after a short period of time. However, not only a dominance level will be found in the meeting, also the activity of the different participants is observable. These social signals are correlated to each other.

In Zhang et al. (2006b) and Rienks and Heylen (2006) mostly high level features as speech transcriptions are used for the dominance detection in meetings. In year two of AMIDA we presented a Hidden-Markov-Model (Rabiner (1989)) using low level features for the task of automatic activity detection during meetings. In year three the Hidden-Markov-Model was extended to a two layer graphical model which is using additional semantic features.

9.2.1 Additional semantic information

Not only acoustic and visual low level features are applied to the detection task, but also features that contain more semantic information are used. These features are interesting because of the close relation between what a person or the group is doing and the level of activity. The features, which have been applied are group action, person action and person movement.

The group action has been deeply investigated in the research community over the last couple of years, for example in et al. (2004, 2006); Reiter et al. (2007). The systems are working directly on audio and video streams and achieve reliably results, but they are currently not real time capable. The meeting is segmented into a sequence of labels like monologue participant one to four, discussion, presentation, whiteboard and note taking.

Moreover, a person action detection system has been applied in Zobl et al. (2003); Wallhoff et al. (2004) to meeting data to extracted additional semantic features. These systems create a sequence of actions for each of the participants, thus four features for each time frame are available. The labels used, are similar to the group actions but contain some more classes: sitting down, standing up, nodding, shaking the head, writing, pointing, using a computer, giving a presentation, writing on the white-board, manipulation of an important item and idle. Idle for example is used if the person is speaking or listening to the meeting. The classes nodding or shaking should help to find points in the meeting where a decision is made or a person is highly active.

The last semantic feature which is currently used is the person movement. It describes what each person is currently doing in the meeting as the person action does, but only the labels off camera, sit, other, move, stand whiteboard, stand screen and take notes are

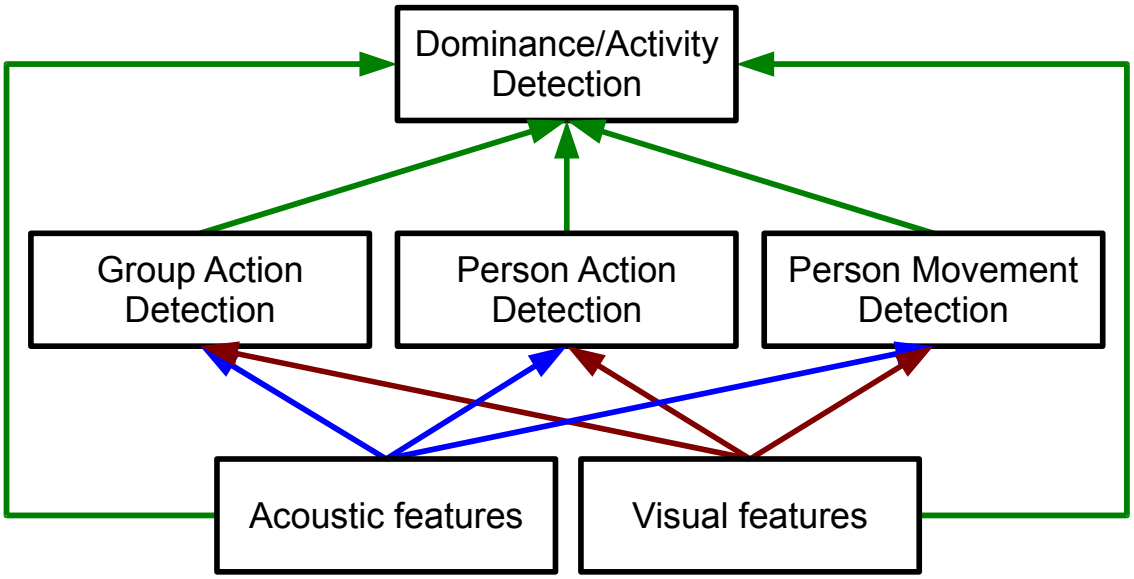


Figure 26: Symbolic structure of the two layer graphical model for activity/dominance detection.

available. Thus, it should improve the results for the activity detection, as it contains information about what the participants are currently doing.

9.2.2 Two layer graphical model

The approach adds semantic information to the single layer model, see Fig. 26. In this the first layer semantic features as group action, person action or person movement are classified. This is done by using the similar HMMs as for the single-layer model which have been trained with the annotated semantic information. The second layer of the model is also similar to the single-layer model, but additional semantic features are added to the input feature stream for the training as well as for the decoding. For the training the annotated semantic information is used again, but the decoding is performed differently. For the decoding the first layer is decoded and the output is added to the observation of the second layer and then this layer is decoded.

9.2.3 Results

The results shown in Table 24 show that the low level fusion of the different modalities does not lead to an improvement of the results. The best result achieves a simple Hidden Markov Model by only using the audio features. The integration of more semantic features at the feature level increases the frame error rate by more than 15%. It seems to be that the early fusion is not the right approach for a feature fusion for the activity detection in meetings.

10 Motion tracking and visualisation

TNO Science & Industry focused on the visualization of information regarding a specific kind of multiparty interaction: an operator or security staff working towards a secure environment in a security setting. This topic is introduced in deliverable D4.2, Chapter 11, "Motion tracking and visualization".

10.1 Goals

Last year, we focused on a group of professionals maintaining a secure environment in a soccer stadium. A situation where a security staff is handling large amounts of (video-) data is very common. For instance, in city centers, in airports or (train-) stations, and at events like soccer matches or festivals.

The first goal of our research is therefore to develop concepts, based on previous work for AMIDA, to assist security staff efficiently handling (video-) data overload. In other parts of the AMIDA project, effort is put into interpreting and classifying data, e.g. in speech recognition, language processing and communication modeling. In our work, as we reported last year, we approached the video data of the soccer match in a different way. What if, without interpreting what we see, we know what normal behavior in and around the stadium is, can we then identify abnormal behavior solely by doing anomaly detection? This would mean, according to the NAIHS model (Kester, 2008), that the object level is skipped, and that a situation assessment is done based on signals. By doing so, as we demonstrate here, the data overload can then be reduced to a fraction which contains interesting data.

A second goal is demonstrating innovative ways of visualizing the interesting data in

Table 24: Evaluation of different modality combinations. The model has 20 states for all single modalities and the fusion models has 15 states per class. The two layer models always have 20 states per class. As additional semantic features in combination with the acoustic feature have been used: movement (M), person actions (P) and group actions (G). AER stands for action error rate, FER means frame error rate and RR is the recognition rate.

Model	AER	FER	RR
Audio (A)	47.2	47.7	54.9
Global Motion (GM)	64.4	63.7	36.6
Skinblob (SK)	66.6	71.3	28.6
Single stream (A&GM)	48.4	48.6	51.8
Multi stream (A&GM)	63.6	55.8	49.2
Multi stream (A,GM&SK)	60.7	57.6	44.1
Two layer (M)	87.5	64.2	34.2
Two layer (P)	87.7	65.4	34.0
Two layer (G&M)	87.5	64.3	34.3
Two layer (G&P)	87.7	65.3	34.0

real-time to the security staff.

10.2 Results this year

10.2.1 Clustering

Accurate classification of events requires knowledge of the location, the calibration and the viewing angle of the cameras. In large public places or stadiums with many cameras present, this requires a huge effort in preparing and maintaining these camera parameters. Using cameras as a motion sensor, this effort is not necessary.

In previous AMIDA work we presented ”motion zones” over time in a ”segment list”. This is a very robust approach to analyzing huge amounts of video, while incorporating advanced pattern recognition. Last year, we collected motion data of a single event, a soccer game. From that event the presence of motion is plotted for 28 cameras over time in the following figure.

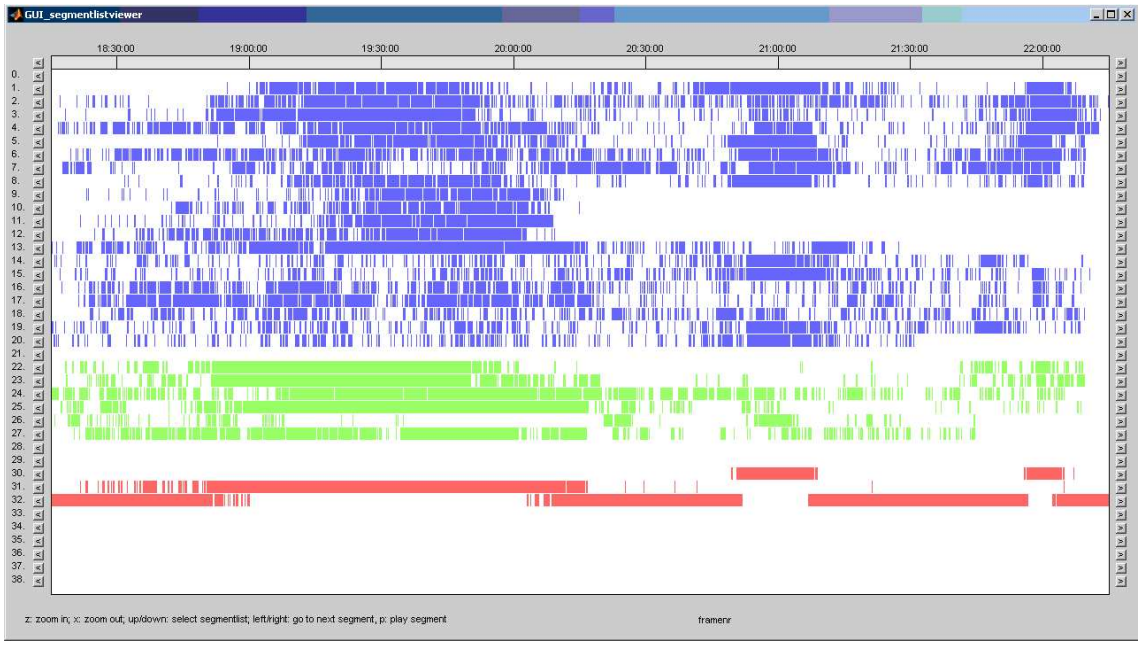


Figure 27: A segment list indicating the activity in 28 cameras of a soccer match.

At the top of the figure the real time can be seen. The game started at 20h. The blue plots mark the passages between the entrances and the standings of the stadium. The green plots mark the entrances of the stadium. Last year we presented a similar figure, and drew some conclusions from this plot. We continued by representing the motion of 28 cameras in a 28-dimensional space, clustering this data in e.g. three states. This led to a surprising result. The bottom (red) segment lists in the following figure represent three different states. We can see a clear visualization of the course of the match in the states: between 19h and 20h people are arriving at the stadium. Between 20h and 20h45 was the first half of the match. The halftime break was from 20h45 until 21h10. The second half of the match was from 21h10 until around 22h, after which people left the stadium again.

10.2.2 Correlation

The analysis in the previous section determines states based on instantaneous motion at different camera positions. We also expanded the data analysis by the use spatio-temporal correlations. The goal was to obtain data that would enable us to automatically detect abnormal events and behavior inside the stadium. However, for reasons of better availability we used the cameras in our TNO field lab at the Stieltjesweg in Delft, consisting of 41 security cameras in our building, for the development and demonstration of the system. The security staff is in this case the operator of all security cameras in the building.

We developed a system which can record motion in (part of) cameras and generate a sort of fingerprint of the normal motion patterns. This happens during a 'training period'. With this knowledge of the normal situation, the system can then real-time monitor the current situation and label anomalous motion in the camera images. This allows the security operator to take extra care in looking at the labeled video data. Alternatively also in real-time an alarm could be given, focusing the attention immediately on anomalous movement. This is illustrated in the following figure.

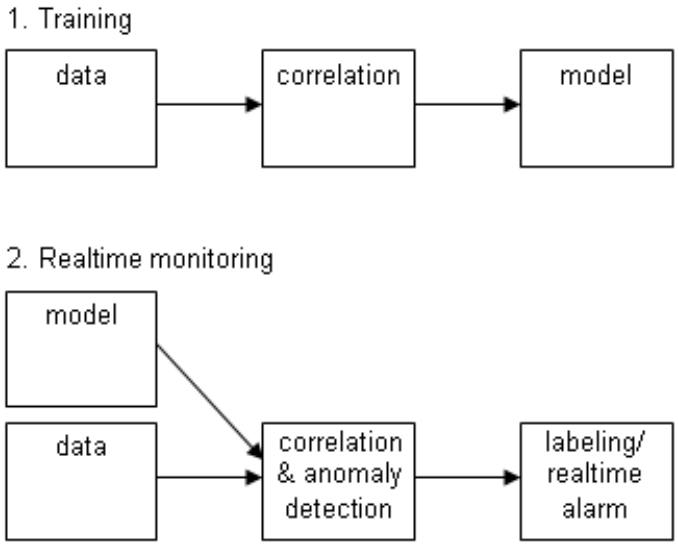


Figure 28: Schematic illustration of the training and the realtime use of the developed application.

Since the system should learn the 'normal' situation, it will need to record a number of normal days first. This led us to choose our TNO field lab as the best place for the recordings. We stress that the cameras in the ADO stadium might have been used as well, as we showed in deliverable D4.4.

In total we recorded (and still record) many weeks continuously on 41 security cameras in our building. Two times per second the motion was measured in 24 motion zones (six columns times four rows) in each camera image.

The collection of the data allows for many different analyses. We chose to gather all data of all motion zones in to segment lists, as we introduced in previous AMIDA work. The segment lists are binary, and represent motion in the camera images: motion or no motion. The next step is the comparison of all segment lists in all motion zones with

each other: ”Is there in general simultaneous motion in different motion zones?” For instance, cameras which are partially looking at the same room will see simultaneous motion. We also calculated the cross-correlation of all segment lists: ”Is there in general subsequent motion observed in different motion zones?” People which enter a staircase will in general be seen leaving a staircase at a different floor a few seconds later. People walk at an average speed and, with the cameras at fixed positions in corridors, will arrive at average times at different cameras images. The normal behavior patterns of people in the field lab can thus be found and learned.

An example of a correlation matrix (representing correlations between motion zones) is shown in the following graph. The graph shows four cameras at one staircase on Friday, August 7, 2009, where camera 1 is on the ground floor, 2 on the first floor, 3 on the second floor, and camera 4 is on the third floor.

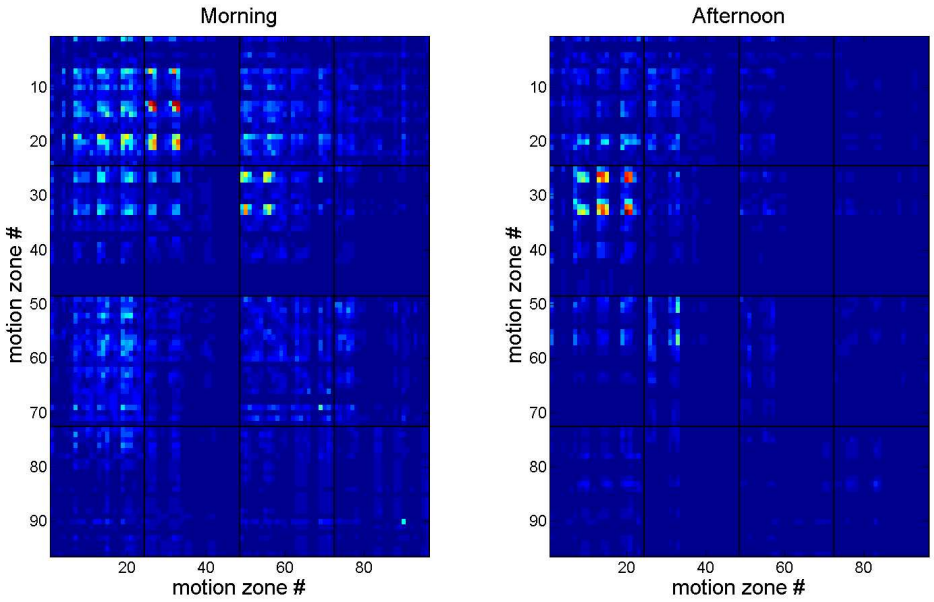


Figure 29: Graph showing the correlation intensity of all motion zones at four different floor of one staircase in the TNO building. Blue means low intensity, and red means high intensity.

Each camera has 24 motion zones which are correlated with themselves. Only the correlation between 10 and 15 seconds is summed. The left graph shows the correlation in the morning, between 8 and 9 am. The right graph shows the correlation in the afternoon, between 5 and 6 pm. The color indicates the correlation number, blue being very low correlation, and red indicating a high correlation. One can see that in the morning people tend to move up the stairs (since in general between 10-15s movement is seen on the first floor, when there was movement on the ground floor: people arrive at work), and in the afternoon people tend to move down the stair (in general there is between 10-15 seconds movement on the ground floor when there was movement on the first floor: people go home).

10.2.3 Visualization: real-time alarm or labeling

The figure in the previous section shows that without knowing what we see in the camera images, we can see differences in the correlations at different hours during the day. With this fingerprint of what is normal at each hour, we have trained the system to recognize odd motion patterns. An operator can be given a real-time alert when unknown or unexpected motion is happening. There can however there can be a delay, due to the fact that every day the motion is slightly different, and these differences should be significant before an alarm can be given. Examples of detectable different motion are:

- staircases or corridors being obstructed
- large meetings in certain areas
- faster moving of people (maybe due to an emergency)

Next to alarms, certain parts of the video data can also be labeled as 'potentially interesting'. Security staff can be provided with a segment list as shown in Fig. 27, and be able to quickly browse through large amounts of video data, identifying what has happened.

With the above described system present, the visualization can also be implemented in the 3D model that was created last year. When in the correlations significant differences arise compared to normal situations, these camera images can be shown to the security staff incorporated in the 3D model.

Figure 30: Figure showing rendering screenshots of the 3D stadium which was devised in the AMIDA project.

In March of 2009 a workshop was organized for delegates from the police, the fire department, the GHOR (a government organization responsible for the Medical Assistance for Accidents and Disasters), ADO (a soccer club from The Hague), and E-Semble (an SME which develops simulation software for the education and training of safety and security professionals). During this workshop the 3D model was introduced and the potential of the crowd fingerprinting methods was demonstrated and this was followed by great interest.

10.2.4 Conclusion and outlook

In the AMI and AMIDA project tools were developed that allow users to find the information that they want quickly from a recorded meeting, by automatic annotation and recognition of events. The tools we developed this year in the AMIDA project are capable of automatic annotation of unknown and unexpected events in a large security setting with many video streams. The meeting participants are replaced by visitors to a certain place, and the people who need to be informed and brought up to speed are in this case the operator or the security staff. The tools use fingerprinting and anomaly detection, and are developed in such a way that they are also able to handle live video streams and label interesting parts in real-time.

A small step forward could even be content linking, where a fingerprint of unexpected or unknown behavior can be searched for and matched to past events, allowing a better estimate of the importance of the anomalous situation.

11 Summary

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 final year is described below.

11.1 Automatic speech recognition

The real-time ASR system has seen a major overhaul with many substantial improvements achieved throughout the whole of the system. The real-time framework (Tracter; see Sec. 2.2.1) has been enhanced in order to support a full microphone array microphone system. One major aspect of this has been the provision of wrappers for external components to be seamlessly inserted into the data flow chain, and in turn enables more advanced features such as bottleneck MLP, fast VTLN and HLDA to be used. The result of this is that several stages of feature processing and re-combination, including VAD, can now run in a single process. Tracter is also now able to collate speaker information from the ICSI online-speaker ID system.

Another major effort during this final year involved increasing the speed of the speech decoder, Juicer (Sec. 2.2.3). This has resulted both in a faster version of the standard Juicer decoder, and a completely new 'pull' based decoder capable of even faster performance. Furthermore, the tight integration of Tracter with HTK means that Juicer is able to use all of the features and functionalities what come with HTK. This makes Juicer a fast alternative to HDecode. The addition of partial trace back within Juicer has significantly reduced the overall latency of the ASR allowing the system to be run in real time in demonstration environments.

AMIDA participated in the NIST RT09s evaluation (Sec. 2.3). Significant improvements were made this year in terms of the speed and accuracy of the overall system. The complete system runs in less than 13 times real-time with earlier outputs available in about 8xRT and 6xRT.

The realtime recognition system has been used as an integral part of the Content Linking Device evaluation. We also continue to routinely use the recognition system as part of AMIDA mini projects (Sec. 2.4).

AMIDA ASR technology has also been made available to the community and commercial partners via webASR (Sec. 2.4.2). During the final year, webASR has been improved and extended in many ways and now provides transparent access to its services via a web-based API. In addition, a software plug-in was developed to reduce the overhead of integrating webASR into their products.

11.2 Keyword spotting

AMIDA participated in a keyword spotting (KWS) and spoken term detection (STD) evaluation organised by the Czech Ministry of Interior using Czech CTS data in 2008 and 2009 (Sec. 3). Four systems developed by AMIDA partners were evaluated and the results are due to be published in a forthcoming paper.

In order to facilitate research and demonstrations of keyword spotting, an Interactive viewer for Keyword spotting output has been developed (Sec. 3.2). This tool can load an output of a keyword-spotting system (KWS) and reference file in HTK-MLF format and show detections in a tabular view. It can be also used to replay detections, tune and visualize scores, hits, misses and false-alarms. The tool is available to the wider community from BUT's website[13].

In cooperation with EU-FP6 project DIRAC, work continued on the detection of out-of-vocabulary words (OOV) in the output of the speech recognition system. We followed an approach based on phone posteriors created by a Large Vocabulary Continuous Speech Recognition system and an additional phone recognizer. Following a number of studies, it was found that the posterior-based OOV word detection approach generalizes across both data and varied language models.

11.3 Speaker diarisation

AMIDA speaker diarisation work has concentrated on a number of distinct themes during the final year and a number of systems were submitted in the NIST RT 2009 evaluation.

For multi-modal diarisation, a combination of features extracted from the video stream as well as the audio stream was investigated with the aim of improving speaker clustering (Sec. 4.2). Studies established that the combination of video and audio performed better than the audio-only baseline.

An alternative approach to multi-modal diarisation was also investigated (Sec. 4.8) in which audio features were combined with two psychology inspired visual features: Visual Focus of Attention (VFoA; see Sec. 5) and motion features. A number of experiments were conducted and showed a performance improvement when combining audio features with the reference VFoA features compared to the baseline audio-only system.

Work was also conducted on novel initialization methods for the ICSI speaker diarisation engine (Sec. 4.3). The current ICSI Speaker Diarisation engine performs suboptimally on short segment durations (less than 600 seconds). Significant speed improvements and a much higher robustness against different recording lengths were achieved.

A system for detecting overlapping speech was developed (Sec. 4.4). Overlapping speech in meetings is detrimental to speaker diarisation for two reasons: it spoils the purity of the speakers models that are trained during the clustering process, and since the Viterbi decoding only outputs one speaker at every time instant, the overlapping speech is missed by definition. On the development test set, improvements from 19.07 % to 18.27 % and from 16.02 % to 15.29 % were observed for SDM and MDM without delay features, respectively.

[13] http://speech.fit.vutbr.cz/en/software/kwsviewer-interactive-viewer-keyword-spotting-output

Finally, work was also conducted on limiting the influence of hyperparameter selection on performance (see Sec. 4.5). By using various slightly different preparations of the overall speaker diarisation system we could lower the DER of the development set for various microphone conditions.

11.4 Focus of Attention

The work during the final project period (see Sec. 5) focused on two main themes. Firstly, improving visual focus of attention (VFOA) recognition and investigating its use for speaker diarisation. Secondly, working on different head pose tracking systems.

Work on multi-party visual focus of attention (VFOA) recognition from multimodal cues (Ba and Odobez, 2009) was further extended by investigating the use of visual activity (how much people are gesturing with their hand, head and body) as context for VFOA recognition, and by improving the VFOA recognition model to handle the case of looking at moving people (Ba et al., 2009b).

The use of VFOA cues in a speaker diarisation system has been investigated, by implicitly associating speech segments to people which are more looked at in a multi-stream approach (G.Garau et al., 2009).

Research on robust 3D head pose and facial expression estimation using structure and appearance features was conducted to handle videos with higher head resolution (Lefèvre and Odobez, 2009).

The real-time VFOA estimation module has been used in the User Engagement and Floor Control demonstrator for addressee detection.

11.5 Speaker Identification

Much of the work conducted on audio-based speaker identification in the last reporting period concentrated on Joint Factor analysis (Sec. 7). Variants of Joint Factor Analysis for speaker recognition were investigated. A systematic comparison of full JFA with its simplified variants was performed and confirmed superior performance of the full JFA with both eigenchannels and eigenvoices.

A number of different log-likelihood scoring methods used by different research groups in state-of-the-art Joint Factor Analysis (JFA) Speaker Recognition systems were also studied with specific reference to speed and performance. It was shown that approximations of the true log-likelihood ratio (LLR) may lead to significant speedup without any loss in performance.

A Matlab toolkit for JFA, including training and test data, has been made publicly available[14].

[14] http://speech.fit.vutbr.cz/en/software/joint-factor-analysis-matlab-demo

11.6 Visual Identification

Work was conducted on a novel feature called Haar Local Binary Pattern (HLBP) for fast and reliable face detection, particularly in adverse imaging conditions (Sec. 6.1). Results obtained on several standard databases show that it competes well with other face detection systems, especially in adverse illumination conditions.

Effort was also directed towards an extension of the work on face recognition using Bayesian Networks initially proposed in the AMIDA deliverable D4.2 (Sec. 6.2). This focussed on generative models dedicated to face recognition considering data extracted from color face images and using Bayesian Networks to model relationships between different observations derived from a single face. Results show that integrating color in an intelligent manner improves the performance over a similar baseline system acting on grayscale only, but also over an Eigenfaces-based system were information from different color channels are treated independently.

11.7 Social Signals

The HMM-based system described in the previous deliverable was extended to a two layer graphical model using additional semantic features (Sec. 9). These features were group action, person action and person movement (see Sec. 9.2.1 for more details). However, it was found that the low level fusion of the different modalities did not lead to a performance improvement.

11.8 Gestures and Actions

Work concentrated on combining previous work (Poppe, 2007; Poppe and Poel, 2008) on example-based pose recovery and action classification based on Common Spatial Patterns (CSP). Combining these two algorithms solved several challenges that are difficult to address in a single step, such as dealing with variations in viewpoint, different background and changes in lightning conditions. Moreover, the combined approach had the advantage that action models could be trained using motion capture data, if the motion capture data and pose recovery adhere to a similar pose representation. The approach was evaluated on the HumanEva-I dataset. Using CSP proved to be advantageous, both in recognition performance and in the number of dimensions that was used to describe the action prototype vectors (Sec. 8).

11.9 Summary

Important progress has been made in all research themes in the final year of AMIDA, most notably the significant enhancement of realtime, online automatic speech recognition.

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

References

A. Adami, L. Burget, S. Dupont, G. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas. Qualcomm-ICSI-OGI features for ASR. In *Proc. ICSLP*, 2002.

S. Ba and J.-M. Odobez. A probabilistic head pose tracking evaluation in single and multiple camera setups. In *CLEAR Evaluation and Workshop*, 2007.

S. Ba and J.-M. Odobez. Recognizing human visual focus of attention from head pose in meetings. *IEEE Trans. on System, Man and Cybernetics: part B, Man*, 39(1):16–34, Feb. 2009.

S. Ba, H. Hung, and J.-M. Odobez. Visual Activity Context for Focus of Attention Estimation in Dynamic Meetings. In *Proc. of ICME*, 2009a.

S. Ba, H. Hung, and J.-M. Odobez. Visual activity context for focus of attention estimation in dynamic meetings. In *IEEE Proc. Int. Conf. on Multimedia and Expo (ICME)*, New-York, june 2009b.

E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, B. Ruiz, and J. Thiran. The BANCA database and evaluation protocol. In *Proc. of the 4th Intl. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, pages 625–638, 2003.

E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The Banca Database and Evaluation Protocol. In *4th Intl. Conf. Audio- and Video-based Biometric Person Authentication, AVBPA'03*. Springer, 2003.

P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

L. Burget, M. Fapso, V. Hubeika, O. Glembek, M. Karafiat, M. Kockmann, P. Matejka, P. Schwarz, and J. Cernocky. But system for nist 2008 speaker recognition evaluation. In *Proc. Interspeech 2009*, pages 2335–2338, 2009a. URL http://www.fit.vutbr.cz/research/view_pub.php?id=9041.

L. Burget, P. Matejka, V. Hubeika, and J. Cernocky. Investigation into variants of joint factor analysis for speaker recognition. In *Proc. Interspeech 2009*, pages 1263–1266, 2009b. URL http://www.fit.vutbr.cz/research/view_pub.php?id=9040.

L. Burget, P. Schwarz, D. Povey, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiat, D. Povey, A. Rastrow, R. Rose, , and S. Thomas. Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models. In *submitted to ICASSP 2010*, 2010.

F. Cardinaux, C. Sanderson, and S. Bengio. User Authentication via Adapted Statistical Models of Face Images. *IEEE Trans. on Signal Processing*, 54(1):361–373, 2005.

D. Chai and K. N. Ngan. Face segmentation using skin color map in videophone applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(4):551–564, 1999.

T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active Shape Models: Their Training and Applications. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

G. Cowell, P. Dawid, L. Lauritzen, and J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer Verlag, 1999.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood From Incomplete Data via the EM Algorithm. *The Journal of Royal Statistical Society*, 39:1–37, 1977.

A. El Hannani and T. Hain. Automatic optimization of speech decoder parameters. *Accepted for publication in IEEE Signal Processing Letters*, 2009.

D. Z. et al. Modeling individual and group actions in meetings: a two-layer HMM framework. In *Proceedings of the Second IEEE Workshop on Event Mining: Detection and Recognition of Events in Video, in Association with CVPR*, 2004.

M. A.-H. et al. Multimodal integration for meeting group action segmentation and recognition. In *Proceedings of the 2nd Joint Workshop on MLMI)*, 2006.

J. Fiscus. A post-processing system to yield reduced word-error rates: Recognizer output voting error reduction(ROVER). In *Proc. Workshop Automatic Speech Recognition and Understanding*, pages 347–352, Santa Barbara, 1997. IEEE.

F. Fleuret. Fast Binary Feature Selection with Conditional Mutual Information. *Journal of Machine Learning Research*, 5:1531–1555, 2004.

G. Friedland, H. Hung, and C. Yeo. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. In *Proc. ICASSP*, pages 4069–4072. IEEE, April 2009a.

G. Friedland, C. Yeo, and H. Hung. Visual speaker localization aided by acoustic models. In *Proc. ACM Multimedia*, Bejing, China, October 2009b. full paper.

B. Froba and A. Ernst. Face detection with the modified census transform. In *Sixth IEEE International Conference on Face and Gesture Recognition*, pages 91–96, 2004.

P. N. Garner, J. Dines, T. Hain, A. El Hannani, M. Karafiát, D. Korchagin, M. Lincoln, V. Wan, and L. Zhang. Real-time ASR from meetings. In *Proceedings of Interspeech*, Brighton, UK, September 2009.

J.-L. Gauvain and C.-H. Lee. Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298, 1994.

G.Garau, S. Ba, H.Bourlard, and J.-M. Odobez. Investigating the use of Visual Focus of Attention for Audio-Visual Speaker Diarisation. In *Proc. ACM Multimedia*, 2009.

O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny. Comparison of scoring methods used in speaker recognition with joint factor analysis. In *Proc. ICASSP 2009*, 2009. ISBN 978-1-4244-2354-5. URL http://www.fit.vutbr.cz/research/view_pub.php?id=9035.

N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. In *Int. Work. on Visual Observation of Deictic Gestures*, 2004.

F. Grézl, M. Karafiát, S. Kontár, and J. Černocký. Probabilistic and bottle-neck features for LVCSR of meetings. In *Proc. ICASSP 2007*, pages 757–760, Honolulu, Hawaii, USA, Apr 2007. IEEE Signal Processing Society. ISBN 1-4244-0728-1. URL http://www.fit.vutbr.cz/research/view_pub.php?id=8249.

S. Gutta, J. Huang, L. Chengjun, and H. Wechsler. Comparative Performance Evaluation of Gray-Scale and Color Information for Face Recognition Tasks. In *Intl Conf. on Audio- and Video-based Biometric Person Authentication (AVBPA)*, volume 2091 of *Lecture Notes in Computer Science*, pages 38 – 43. Springer, 2001.

B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based Face Detection. In *Computer Vision and Pattern Recognition*, pages 657–662, 2001.

H. Hermansky, D. P. W. Ellis, and S. Sharma. Tandem connectionist feature extraction for conventional HMM systems. In *Proc. ICASSP 2000*, Turkey, 2000.

G. Heusch and S. Marcel. Face Authentication with Salient Facial Features and Static Bayesian Network. In *Intl Conf. on Biometrics (ICB)*, volume 4642 of *Lecture Notes in Computer Science*, pages 878–887. Springer, 2007.

S. Huang and S. Renals. Hierarchical Pitman-Yor language models for ASR in meetings. In *Proc. IEEE ASRU'07*, pages 124–129, Dec. 2007.

S. Huang and S. Renals. A parallel training algorithm for hierarchical Pitman-Yor process language models. In *Proc. Interspeech'09*, Sep. 2009.

S. Huang and S. Renals. Power law discounting for n-gram language models. In *Submitted to ICASSP'10*, March 2010.

V. Hubeika. Speaker verification as a target-nontarget trial task. In *Proceedings of the 15th Conference and Competition STUDENT EEICT 2009*, 2009. ISBN 978-80-214-3870-5. URL http://www.fit.vutbr.cz/research/view_pub.php?id=9036.

M. Huijbregts, C. Wooters, and R. Ordelman. Filtering the unknown: Speech activity detection in heterogeneous video collections. In *Proc. Interpeech*, pages 2925–2928, Antwerpen, 2007.

M. Huijbregts, D. van Leeuwen, and F. de Jong. Speech overlap detection in a two-pass speaker diarization system. In *Proc. Interpseech 2009*, 2009a.

M. Huijbregts, D. van Leeuwen, and F. de Jong. The majority wins: a method for combining speaker diarization systems. In *Proc. Interspeech*. ISCA, September 2009b.

D. Imseng. Novel initialization methods for speaker diarization. Technical report, Ecole Polytechnique Federale de Laussane, 2009. Nominated for the IBM/EPFL outstanding master thesis award.

D. Imseng and G. Friedland. Robust speaker diarization for short speech recordings. In *Proc. 11th Biannual IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009a.

D. Imseng and G. Friedland. Tuning-robust initialization methods for speaker diarization. *IEEE Transactions on Acoustics, Speech, and Language Processing*, 2009b. submitted.

O. Jesorsky, K. Kirchberg, and R. Frischholz. Robust Face Detection using the Hausdorff Distance. In *Proc. of the 3rd Intl. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, pages 90–95, 2001.

H. Jin, Q. Liu, H. Lu, and X. Tong. Face detection using improved LBP under bayesian framework. In *Proc. of the Third International Conference on Image and Graphics (ICIG)*, pages 306–309, 2004.

C. I. Jones and A. L. Abott. Color Face Recognition by Hypercomplex Gabor Analysis. In *IEEE Intl Conf. on Automatic Face and Gesture Recognition (AFGR)*, pages 126–131, 2006.

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. J. H. M. Kester. Designing networked adaptive interactive hybrid systems. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 516–521, 2008.

S. Kombrink, L. Burget, . Matejka, M. Karafiat, and H. Hermansky. Posterior-based out of vocabularyword detection in telephone speech. In *Proc. Interspeech 2009*, pages 80–83, 2009.

S. Lefèvre and J.-M. Odobez. Structure and appearance features for robust 3d facial actions tracking. In *International Conference on Multimedia and Expo (ICME)*, june 2009.

W. Li, J. Dines, M. Magimai.-Doss, and H. Bourlard. Non-linear mapping for multichannel speech separation and robust overlapping speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

J. Luettin and G. Maitre. Evaluation protocol for the extended M2VTS database (XM2VTSDB). Idiap Communication 98-05, Idiap, 2000.

H. Luo. Optimization design of cascaded classifiers. In *Computer Vision and Pattern Recognition*, pages 480–485, 2005.

S. Lyu. Infomax Boosting. In *Computer Vision and Pattern Recognition*, pages 533–538, 2005.

K. Messer, J. Matas, J. Kittler, J. Lüttin, and G. Maitre. XM2VTSDB: The Extended M2VTS Database. In *Intl Conf. Audio- and Video-based Biometric Person Authentication (AVBPA)*, volume 43, pages 72–77, 1999.

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.

D. Novick, B. Hansen, and K. Ward. Coordinating Turn-Taking with Gaze. In *Proc. ICSLP*, 1996.

T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29:51–59, 1996.

J. M. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multiple distant microphones: mixing acoustic features and inter-channel time differences. In *Proc. Interspeech*. ISCA, 2006.

S. H. K. Parthasarathi, M. Magimai.-Doss, H. Bourlard, and D. Gatica-Perez. Investigating privacy-sensitive features for speech detection in multiparty conversations. In *Proceedings of Interspeech*, 2009.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

S. Petridis and M. Pantic. Audiovisual discrimination between voiced laughter, unvoiced laughter and speech. *IEEE Transactions on Multimedia*, submitted.

P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The Feret evaluation methodology for face recognition algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 22(10):1090–1104, 2000.

R. Poppe. Evaluating example-based pose estimation: Experiments on the HumanEva sets. In *Proceedings of the Workshop on Evaluation of Articulated Human Motion and Pose Estimation at the Conference on Computer Vision and Pattern Recognition (CVPR-EHuM)*, Minneapolis, MN, June 2007.

R. Poppe. *Discriminative Vision-Based Recovery and Recognition of Human Motion*. Phd thesis, University of Twente, 2009. ISBN: 978-90-365-2810-8.

R. Poppe and M. Poel. Discriminative human action recognition using pairwise CSP classifiers. In J. Cohn, T. Huang, M. Pantic, and N. Sebe, editors, *IEEE International Conference on Automatic Face and Gesture Recognition (FGR'08)*, pages 1–8, Los Alamitos, 2008. IEEE Computer Society Press.

D. Povey. Improvements to fMPE for discriminative training of features. In *Proc. of Interspeech2005*, pages 2977–2980, Lisbon, Portugal, Sep 2005.

D. Povey. *Discriminative Training for Large Vocabulary Speech RecognitionK*. PhD thesis, Cambridge University Engineering Department, Mar. 2003.

D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiat, P. Schwarz, A. Rastrow, R. Rose, , and S. Thomas. Subspace gaussian mixture models for speech recognition. In *submitted to ICASSP 2010*, 2010.

L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989.

S. Reiter, B. Schuller, and G. Rigoll. Hidden conditional random fields for meeting segmentation. In *Proceedings of the 8th ICME*, 2007.

E. Ricci and J. Odobez. Real-time simultaneous head tracking and pose estimation. In *IEEE International Conference on Image Processing (ICIP)*, november 2009.

R. Rienks and D. Heylen. Automatic dominance detection in meetings using easily obtainable features. In *Proceedings of the 2nd Workshop on MLMI*, 2006.

Y. Rodriguez. *Face Detection and Verification using Local Binary Patterns*. PhD Thesis, École Polytechnique Fédérale de Lausanne, 2006.

R. Russell, P. Sinha, I. Biedermann, and M. Nederhouser. Is Pigmentation Important For Face Recognition ? Evidence From Contrast Negation. *Perception*, 35:749–759, 2006.

M. Sadeghi, S. Khoshrou, and J. Kittler. SVM-Based Selection of Colour Space Experts for Face Authentication. In *Intl Conf. on Biometrics (ICB)*, volume 4642 of *Lecture Notes in Computer Science*, pages 907–916. Springer, 2007.

F. Samaria and S. Young. HMM-based Architecture for Face Identification. *Image and Vision Computing*, 12(8):537–543, Oct. 1994.

B. Schiele. *Object Recognition using Multidimensional Receptive Field Histograms*. PhD thesis, I.N.P.Grenoble, 1997.

G. Schwartz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

L. Sigal and M. J. Black. HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, Department of Computer Science, Providence, RI, September 2006.

P. Sinha, B. Balas, Y. Ostrovsky, and R. Russel. Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of the IEEE, Special Issue on Biometrics: Algorithms and Applications*, 94(11):1948–1962, 2006.

J. Sochman and J. Matas. WaldBoost-Learning for time constrained sequential detection. In *Computer Vision and Pattern Recognition*, pages 150–156, 2005.

J. Sun, J. Rehg, and A. Bobick. Automatic cascade training with perturbation bias. In *Computer Vision and Pattern Recognition*, pages 276–283, 2004.

Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proc. of the Annual Meeting of the ACL*, volume 44, 2006.

J. Tejedor, D. Wang, S. King, J. Frankel, and J. Colas. A post erior probability-based system hybridisation and combination for spoken term detection. In *Proc. Interspeech 2009*, 2009.

L. Torres, J. Y. Reutter, and L. Lorente. The Importance of the Color Information in Face Recognition. In *IEEE Intl Conf. on Image Processing (ICIP)*, volume 3, pages 627–631, 1999.

M. Turk and A. Pentland. Face Recognition Using Eigenfaces. In *IEEE Intl. Conf on Computer Vision and Pattern Recognition (CVPR)*, pages 586–591, 1991.

V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1989.

R. Vertegaal, R. Slagter, G. Van der Veer, and A. Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proc. of ACM SIGCHI*, 2001.

D. Vijayasenan, F. Valente, and H. Bourlard. Mutual information based channel selection for speaker diarization of meetings data. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2009a.

D. Vijayasenan, F. Valente, and H. Bourlard. KL realignment for speaker diarization with multiple feature streams. In *10th Annual Conference of the International Speech Communication Association*, 2009b.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 511–518, 2001.

F. Wallhoff, M. Zobl, and G. Rigoll. Action segmentation and recognition in meeting room scenarios. In *Proceedings of the 11th ICIP*, 2004.

H. Wang, P. Li, and T. Zhang. Histogram Features-Based Fisher Linear Discriminant for Face Detection. In *Asian Conference on Computer Vision*, pages 521–530, 2006.

L. Welling, S. Kanthak, and H. Ney. Improved methods for vocal tract normalization. In *Proc. ICASSP 1999*, volume 2, pages 764–764, Phoenix, AZ, USA, Mar. 1999. ISBN 0-7803-5041-3.

S. Yan, S. Shan, X. Chen, and W. Gao. Locally Assembled Binary (LAB) Feature with Feature-centric Cascade for Fast and Accurate Face Detection. In *Computer Vision and Pattern Recognition*, 2008.

M.-H. Yang, D. Roth, and N. Ahuja. A SNoW-based face detector. In *Advances in Neural Information Processing Systems*, pages 855–861, 2000.

B. Zhang, S. Matsoukas, and R. Schwartz. Recent progress on the discriminative region-dependent transform for speech feature extraction. In *Proc. of Interspeech2006*, pages 2977–2980, Pittsburgh, PA, USA, Sep 2006a.

D. Zhang, D. Gatica-Perez, S. Bengio, and D. Roy. Learning influence among interacting markov chains. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1577–1584. MIT Press, 2006b.

H. Zhang, W. Gao, X. Chen, and D. Zhao. Learning Informative Features for Spatial Histogram-Based Object Detection. In *Proc. of International Joint Conference on Neural Networks, Montreal, Canada*, pages 1806–1811, 2005.

L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Li. Face detection based on Multi-Block LBP representation. In *2nd Intl. Conf. on Biometrics (ICB)*, pages 11–18, 2007.

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 PETS-ICVS*, pages 32–36, 2003.