Fusion for Audio-Visual Laughter Detection

Boris Reuderink

September 13, 2007

University of Twente
Enschede - The Netherlands

Abstract

Laughter is a highly variable signal, and can express a spectrum of emotions. This makes the automatic detection of laughter a challenging but interesting task. We perform automatic laughter detection using audio-visual data from the AMI Meeting Corpus. Audio-visual laughter detection is performed by combining (fusing) the results of a separate audio and video classifier on the decision level. The video-classifier uses features based on the principal components of 20 tracked facial points, for audio we use the commonly used PLP and RASTA-PLP features. Our results indicate that RASTA-PLP features outperform PLP features for laughter detection in audio. We compared hidden Markov models (HMMs), Gaussian mixture models (GMMs) and support vector machines (SVM) based classifiers, and found that RASTA-PLP combined with a GMM resulted in the best performance for the audio modality. The video features classified using a SVM resulted in the best single-modality performance. Fusion on the decision-level resulted in laughter detection with a significantly better performance than single-modality classification.

Acknowledgements

I would like to thank my supervisors, Maja Pantic, Mannes Poel, Khiet Truong and Ronald Poppe for supervising me, supporting me and pointing me in the right directions when needed. In addition to my supervisors, I would like to thank the HMI department of the University of Twente for supporting my trip to Londen. I would also like to thank Michel Valstar, who helped me tremendously during my stay in London, and Stavros Petridis, for helping with the creation of the video-features. I would like to thank Luuk Peters and Roald Dijkstra for proofreading my drafts. And last but not least I would like to thank Sanne Beukers for supporting me while I was working on my thesis.

Contents

Chapter 1

Introduction

Laughter is important. Someones mental state and emotions are conveyed in paralinguistic cues, such as laughter, a trembling voice and coughs. Because laughter occurs frequently in spontaneous speech, it is an interesting research subject. Laughter is not limited to positive emotions; negative feelings and attitudes such as sadness and contempt can be expressed with laughter [35]. This spectrum of expressed emotions combined with the high variability of laughter makes the automatic detection of laughter a challenging but interesting task.

Automatic laughter detection can be used for example in meetings where laughter can provide cues to semantically meaningful events. Another application of laughter detection is the detection of non-speech for automatic speech recognition. Laughter can possibly be used as a feedback mechanism in Human Computer Interaction interfaces.

Earlier work on laughter detection has mainly focused on laughter detection in audio only. Currently the focus starts to shift for laughter detection in audio to audio-visual detection of laughter because additional visual information can possibly improve the detection of laughter. Research investigating audio-visual laughter detection was suggested by Truong et al. [38]. In this thesis we investigate fusion for audio-visual laughter detection. We will investigate if fusion of the audio and video modality can improve the performance of automatic laughter detection. Fusion will be performed on the decision level, which means that audio and video are classified separately, and the results are fused to make a final classification. We will evaluate different feature sets and different classification algorithms in order to find strong audio and video classifiers, and fuse those results to create a audio-visual classifier which hopefully outperform both the audio and the video classifiers.

The rest of this thesis is organized as follows. In the next chapter we describe earlier work on laughter detection in audio, detection of facial expressions in video, and work on audio-visual emotion recognition. Then we describe the methodology we use to evaluate the performance of decision-level fusion. This includes a description of the data set, the machine-learning techniques we use, and the performance measure we use. The results are presented in the next chapter, followed by the conclusions in the last chapter.

Chapter 2

Literature

2.1 Laughter

Although laughter occurs frequently in conversations, we do not seem to know a lot about laughter. Research on the acoustic properties of laughter often contradicts other research, and the used terminology differs from work to work. To increase the confusion even more, smiles and laughter are often not discussed together while they seem to be related. Therefore we will describe different terminologies to describe laughter, the relation between speech, laughter and smiles before we describe the variability of the laughter signal.

Laughter is usually analyzed on three levels. Bachorowski [5] defines the following three levels: bouts, calls and segments. Bouts are entire laugh episodes that occur during one exhalation. Calls are discrete acoustic events that together form a bout. Bouts start with long calls, followed by calls that were about half as long. A call is voiced, unvoiced, mixed, or is made up by glottal pulses, fry registers and glottal whistles. The mouth can be open or closed during the production of a call. Calls can be subdivided in segments. Segments are temporally delimited spectrogram components. A similar division in three level was suggested by Trouvain [36]: phrases, syllables and segments. Phrases are comparable to bouts. Syllables are defined as interpulse intervals, and form phrases when combined. Segments can be vowels or consonants. The consonantal segment in a laugh is often seen as an interval or pause.

Smiling and laughter often occurs together, and seem to be different forms of the same event. Laughter often shows a facial expression similar to smiling combined with an involuntarily exhalation, sometimes followed by uncontrolled inhalations and exhalations. This involuntarily breathing is not present during smiling. Laughter and smiles could be extremes of a smile-laugh continuum, but there are some indications that there is a more complex relation between laughter, smiling and even speech than we would expect. Aubergé and Cathiard demonstrated that a genuine smile includes a specific manipulation of the prosody of the speech [4], which cannot be attributed to the facial deformation of a smile; not only laughter, but also smiling is audible. Like smiling, laughter does occur during speech, and does so very often according to Trovain [35]. In the KielCorpus of Spontaneous Speech, 60% of all labeled laughs are instances that overlap speech. Simultaneous production of speech and laughter is not simply laughter imposed on articulation, and there is no prototypical pattern for speech-laughs. In a later work, Trouvain reports [36] that laughter is a mix of laughter interspersed with speech-laughs and smiled speech. Smiling and laughter seem to be different categories rather than extremes of a continuum.

Laughter is a highly variable signal [5, 40]. Voiced laughter shows much more source-related variability than is associated with speech, and the individual identity and sex are conveyed in laugh acoustics. The variability between individuals greatly exceeded the variability within an individual. Laughter seems to be better conceptualized as a repertoire of sounds, which makes it difficult to detect it automatically. Kipper and Todt [27] report that the successive syllables of laughter, which appear similar, show dynamic changes of acoustic parameters, and are in fact different. For example, the fundamental frequency, the amplitude and the duration of a syllable varies during a laughter bout. Laughter seems to be a very variable signal, both on phrase and syllable level.

The automatic recognition of of laughter seems to be a very challenging problem. The laughter signal is highly variable on multiple levels, and can be described best as a group of sounds. Laughter and smile seem to be different categories, and should not be regarded as different manifestations of the same event.

2.2 Laughter detection in audio

Automatic laughter detection has been studied several times, in the context of meetings, for audio indexing and to detect affective states. We will describe a few studies on automatic laughter detection, and summarize some characteristics of these studies. An overview of automatic laughter detection can be found in Table 2.1.

Campbell et al. [8] developed a system to classify a laugh in different categories. They constructed a corpus containing four affective classes of laughter: A hearty laugh, an amused laugh, a satirical laugh and a social laugh. A training set of 3000 hand labeled laughs was used to train hidden Markov models (HMMs). The HMMs recognized the affective class correctly in 75% of the test cases.

Automatic laughter detection can be used in audio indexing applications. For example, Lockerd and Mueller [28] performed laughter detection using their affective indexing camcorder. Laughter was detected using HMMs. One HMM was trained on 40 laughter examples, the other HMM was trained on speech. The classifier correctly identified in 88% of the test segments. Misleading segments were sounds such as coughs, and sounds produced by cars and trains.

Arias et al. [2] performed audio indexing using a Gaussian mixture models (GMMs) and support vector machines (SVMs) on spectral features. Each frame is classified and then smoothed using a smoothing function to merge small parts. The accuracy of their laughter detection is very high (97.26% with GMMs and 97.12% with a SVM). However, their data set contains 1 minute of laughter for every 180 minutes of audio. Only prediction non-laughter would result in a baseline accuracy of 99.4%, which makes it unclear how well their laughter-detection really performs.

Automatic laughter detection is frequently studied in the context of meetings. Kennedy and Ellis [25] detected multiple laughing participants in the ICSI Meeting database. Using a SVM on one second windows of Mel-Frequency Cepstrum Coefficients (MFCCs) features, a equal error rate (EER) of 13% was obtained.

The same data set was used by Truong and Van Leeuwen [37]. Using Gaussian mixture models (GMM) on Perceptual Linear Predictive Analysis (PLP) features, they also obtained an EER of 13%. The data set contained examples in which both speech and laughter were present, and some inaudible laughs. After removing these difficult instances, the performance

Study	Dataset	Performance	Remarks
Truong (2007) [38]	ICSI-Bmr, clean set	EER_{gmm}: 6.3, EER_{svm}: 2.6, EER_{fused}: 2.9	EER_{fused} was tested on different corpus than EER_{svm}
Arias (2005) [2]	Broadcast	A: 97%	MFCCs with GMMs and SVM, 1 minute laughter, 180 minutes non-laughter
Campbell (2005) [8]	ESP	A: 75%	HMMs to classify a laugh into 4 categories
Ito [23] (2005)	Audio visual laughter	Audio: (95% R, 60% P)	
Truong (2005) [37]	ICSI-Bmr	EER: 13.4%, EER_{clean}: 7.1%	PLP, GMM, EER_{clean} on set with unclear samples removed
Kennedy (2004) [25]	ICSI-Bmr	EER: 13%	MFCCs + SVM
Lockerd (2002) [28]	Single person, 40 laughs	A: 88%	HMMs

Table 2.1: Automatic laughter recognition in audio.

increased, resulting in a EER of 7.1%. Different audio features were tested and resulted in PLP outperforming pitch and energy, pitch and voicing and modulation spectrum features.

In a more recent work, Truong and Van Leeuwen [38] used the cleaned ICSI meeting data set to train GMM and SVM classifiers. For the SVM classifier the frame level features were transformed to a fixed length using a Generalized Linear Discriminant Sequence (GLDS) kernel. The SVM classifier performed better than the GMM classifier in most cases. The best feature set appeared to be the PLP feature set. The scores of different classifiers based on different features were fused using a linear combination of the scores or fused using a SVM or a MLP trained on the scores. Fusion based on GMM- and SVM-classifiers increases the discriminative power, as does fusion between classifiers based on spectral features and classifiers based on prosodic information.

When we compare the results of these studies, GMMs and SVMs seem to be used most for automatic laughter recognition. Spectral features seem to outperform prosodic features. An EER of 12–13% seems to be usual. Removing unclear examples improves the classification performance enormously. This suggests that the performance largely depends on the difficulty of the chosen data set.

2.3 Facial expressions

The detection of facial expressions in video is a popular area of research. Therefore, we will only describe a few studies that are related to fusion and the Patras-Pantic particle filtering tracking scheme [33] which we will use to extract video features.

Valstar et al. [39] conducted a study to automatically differentiate between posed and spontaneous brow actions. Timing is a critical factor for the interpretation of facial behavior. The facial expressions are labeled according to the Facial Action Coding System (FACS) [16] action units (AUs). The SVM based AU detectors detect temporal segment (neutral, onset, apex, offset) of the atomic AUs based on a sequence of 20 tracked facial points. These points were tracked using the Patras-Pantic particle filtering tracking scheme. Using the detected three brow AUs (AU1, AU2, AU4), mid-level features based on intensity, duration, trajectory, symmetry and co-occurrence with other muscle actions were created to determine the spontaneousness of an instance. This resulted in a classification with an accuracy of

90.7%.

Gunes and Piccardi [18] compare fusion of facial expressions and affective body gestures at the feature and decision level. For both modalities, single expressive frames are manually selected, which are classified into six emotions. The body-modality classifier was able to classify frames with a 100% accuracy. Using feature-level fusion, which combines the feature vectors of both modality into a single multi-modal feature vector, again an accuracy of 100% was obtained. Decision level fusion was performed using different rules to combine the scores of the classifiers for both modalities, which resulted in an accuracy of only 91%. Clearly the used fusion rules are not well chosen for their problem.

Pantic et al. [32] used the head, face and shoulder modalities to differentiate between spontaneous and posed smiles. The tracking of the facial expressions was performed using the Patras-Pantic particle filtering tracking scheme [33]. For mid- and high-level fusion, frames are classified, and filtered to create neutral-onset-apex-offset-neutral sequences. Mid-level fusion is performed by transforming features into symbols such as temporal aspects of AUs, and the head and shoulder actions. For these symbols, mid-level features such as morphology, speed, symmetry, the duration of apex-overlap of modalities are calculated, and the order of the different actions are computed. Low level fusion (recall: 93%, precision: 89%) yields better results than mid-level (recall: 79%, precision: 79%) and high-level fusion (recall: 93%, precision: 63%). The head modality is the most important modality for the recognition for this data set, although the difference is not significant. The fusion of these modalities improves the performance significantly.

2.4 Audio-visual fusion

Most work on audio-visual fusion has focused on the detection of emotion in audio-visual data [49, 47, 44, 18, 45, 48, 46, 21, 41, 17, 7]. Some other audio-visual studies are conducted on cry detection [31], movie classification [42], tracking [6, 3], speech recognition [13] and laughter detection [23]. These studies all try to exploit the complementary nature of audio-visual data. Decision level fusion is usually performed using the product, or a (weighted) sum of the predictions of single-modality classifiers, or using hand-crafted rules for classification. Other commonly used fusion techniques include mid-level fusion using multi-stream hidden Markov model (MHMM), and feature level fusion. We will describe some studies in more detail and make some general observations. A overview of these studies can be found in Table 2.2.

Zeng et al. [48] used a sparse network of Winnow (SNoW) classifier to detect 11 affective states in audio-visual data. Fusion was performed using voting on frame-level to obtain a class for each instance. For a second, person-independent test, fusion was performed by using a weighted summation of component HMMs. In a following study [46], Zeng et al. performed automatic emotion recognition of positive and negative emotions in a realistic conversation setting. The facial expressions were encoded using FACS. Video features were based on the facial texture; prosodic features were used for audio classification. Fusion was regarded as a multi-class classification problem, with the outputs of the different component HMMs as features. An AdaBoost learning scheme performed best of the tested classifiers.

Another study that compared feature-level fusion and decision level fusion for automatic emotion recognition was conducted by Busso et al. [7]. Video texture and prosodic audio features were classified using SVMs. The confusion matrices of the audio and video modalities show that pairs of emotions that are confused in one modality can be easily classified using

Study	Dataset	Performance	Remarks
Zajdel [44] (2007)	Posed, 2 emotions	A: ≈ 45%, V: ≈ 67%, MF: ≈ 78%	Dynamic Baysian Network
Zeng [46] (2007)	AAI, 2 emotions	A: 70%, V: 86% DF: 90%	Adaboost on component HMMs
Zeng [48] (2007)	Posed, 11 emotional states	A: 66%, V: 39%, DF: 72%	SNoW, MHMM
Pal [31] (2006)	Unknown, 5 cry types	A: 74%, V: 64%, DF: 75%	Rule based fusion using confusion matrices
Zeng [45] (2006)	AAI, 2 emotions	A: 70, V: 86% DF: 90%	Adaboost on component HMMs
Asoh [3] (2005)	Speech, 2 states and location	MF: 85%	Particle filter
Hoch [21] (2005)	Posed, 3 emotions	A: 82%, V: 67%, DF: 87%	SVM, weighted-sum fusion
Ito [23] (2005)	Spontaneous, laughter	A: (95% R, 60% P), V: (71% R, 52% P)%, DF: (71% R, 74% P)	Rule based fusion
Wang [41] (2005)	Posed, 6 emotions	A: 66%, V: 49%, FF1: 70%, FF2: 82%	FF1: FLDA classfier, FF2: Rule-based voting
Xu [42] (2005)	Movies, horror vs comedy	DF: (R=97%, P=91%)	Voting, rule based fusion
Busso [7] (2004)	Posed, single person, 4 emotions	A: 71%, V: 85%, FF: 89%, DF: 89%	DL: product fusion
Go [17] (2003)	6 emotions	A: 93, V: 93%, DF: 97%	Rule based
Dupont [13] (2000)	M2VTS, 10 words, noisy	A: 52% V: 60%, FF: 70%, MF: 80%, DF: 82%	Fusion using MHMMs

Table 2.2: Audio-visual fusion

the other modality. Different decision-level fusion rules were tested, the best results (89%) were obtained using the product of the prediction of both modalities. Feature-level fusion resulted in an accuracy of 89%. In this experiment, feature-level fusion and decision-level fusion had a similar performance.

Dupont and Luettin [13] used both acoustic and visual speech data for automatic speech recognition. A MHMM is used to combine the audio and video modalities on the feature level, decision level and mid-level. The fused system performs better than systems based on single modality in the condition of noise. Both mid-level and decision-level fusion perform better than feature-level fusion. Without the addition of noise to the features, audio-classification alone is sufficient for almost perfect classification.

A quite different approach for fusion was taken by Asoh et al. [3]. Particle filtering was used to track the location of human speech events. The audio modality consisted of a microphone array, the video-modality consisted of a monocular camera. Audio-visual tracking was performed by modeling the position and the type of the signal as a hidden state. The noisy observations are used to estimate the hidden state using particle filtering. This approach provides a simple method to compute the probability of a location and occurrence of a speech event.

Ito et al. [23] focused on the detection of a smiling face and the utterance of laughter sound in natural dialogues. A database was created with Japanese, English and Chinese subjects. Video features consist of the lip lengths, the lip angles and the mean intensities of the cheek areas. Frame level classification of the video features is performed using a

perceptron, resulting in a recall of 71%, and a precision of 52%. Laughter sound detection is performed on MFCC and delta-MFCC features, using two GMMs, one for laughter, and one for other sounds. Using a moving-average filter the frame-by-frame sequences are smoothed. A recall of 96% and a precision of 60% was obtained with 16 Gaussian mixtures. The audio and video channels are combined using hand-crafted rules. The combined system obtained a recall of 71% and a precision of 74%. Ito et al. do not report if fusion significantly increases the performance of their detector.

Xu et al. [42] performed affective content analysis of comedy and horror movies using audio emotional events, such as laughing and horror sounds. The audio is classified using a left-to-right HMM with four states. After classification, the predictions are filtered using sliding window majority-voting. Short horror sounds were too short to detect using a HMM, they were detected by finding large amplitude changes. The audio features consist of MFCCs, with delta and acceleration features to accentuate the temporal characteristics of the signal. The recall and precision are over 90% for horror sounds and canned laughter.

The performance of decision-level fusion seems to be similar to the performance of feature-level fusion. The fusion of audio and video seems to boost the classification performance in these studies with about 4%. However, most work does not report the significance of this gain in performance. Fusion seems to work best when the individual modalities both have a low performance, for example due to noise in the audio-visual speech recognition of Dupont [13]. When single classifiers have a high performance, the performance gain obtained by fusion of the modalities is low, and sometimes fusion even degrades the performance, as observed in the work of Gunes [18].

Chapter 3

Methodology

Fusion of audio and video can be performed on different levels. We perform fusion on the decision-level where the audio and video modality are classified separately. When the classifiers for both modalities have classified the instance, their results are fused into a final multi-modal prediction. See Figure 3.1 for a schematic overview. An alternative approach is fusion on feature-level, where the audio and video features are merged into a single, fused feature set. A classifier classifies the fused features of a single instance. We have chosen to evaluate decision-level fusion instead of feature-level fusion for two reasons. The first first reason is that decision-level fusion allows the use of different classifiers for the different modalities. The different results for the different classifiers helps us understand the nature of the audio-visual signal, and it possibly results in a better performance. The second reason is that we use a very small data set. The feature-level fusion approach has a higher dimensionality, which requires a lager data set to learn a classifier [1]. We therefore use decision-level fusion.

In the next subsections, we will describe the preprocessing we applied to our data set, the features we used and the design we used to evaluate our fusion techniques.

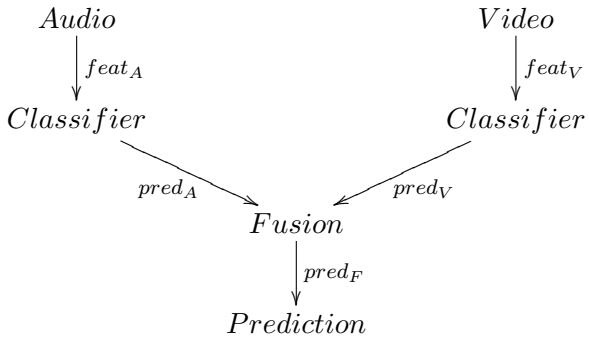

Figure 3.1: Decision-level fusion.

3.1 Dataset

In order to measure the classification performance of different fusion techniques, we need a corpus containing both laughter and non-laughter examples to use for training and testing. We created a corpus based on the AMI Meeting Corpus [29]. In the following sections,

we will describe the AMI Meeting Corpus, the segmentation process used to select examples (instances), and the details regarding the construction of the corpus based on the segmentation data.

3.1.1 AMI Meeting Corpus

The AMI Meeting Corpus consists of 100 hours of meeting recordings, stored in different signals that are synchronized to a common time line. The meetings are recorded in English, mostly spoken by non-native speakers.

For each meeting, there are multiple audio and video recordings. We used seven non-scenario meetings recorded in the IDIAP-room (IB4001, IB4002, IB4003, IB4004, IB4005, IB4010, IB4011). These meetings contain a fair amount of spontaneous laughter. In the first five meetings, the four participants plan an office move. In the last two meeting four people discuss the selection of films to show for a fictitious movie club. We removed two participants, one displayed extremely asymmetrical facial expressions (IB4005.2), the other displayed a strong nervous tick in muscles around the mouth (IB4003.3, IB4003.4). Both participants were removed because their unusual expressions would have a huge impact on our results due to the small size of our dataset. The remaining 10 participants are displayed in Figure 3.3.

We used the close-up video recording (DivX AVI codec 5.2.1, 2300 Kbps, 720×576 pixels, 25 frames per second) and the headset audio recording (16 KHz WAV file) of each participant for our corpus. In total we have used 17 hours of raw audio-visual meeting data to construct our corpus.

3.1.2 Segmentation

The seven meetings we selected from the AMI Meeting Corpus were segmented into laughter and smile segments. The presence of laughter was determined using the definition for audible laughter of Vettin and Todt [40]:

> Vocalizations compromising several vocal elements must consist mainly of expiratory elements; inspiratory elements might occur at the end of vocalisations; expiratory elements must be shorter than 600ms and successive elements have to be similar in their acoustic structure; single-element vocalizations must be expiratory with a vowel-like acoustic structure, or, when noisy, the element must begin with a distinct onset.

For smiles we used a definition based on visual information. We define a visible smile as the visible contraction of the Zygomatic Major (FACS Action Unit 12). The activation of AU12 pulls the lip corners towards the cheekbones [14]. We define the start of the smile as the moment the corners of the mouth start to move, the end is defined as the moment the corners of the mouth return to a neutral position.

Using these definitions, the 17 hours of audio-visual meeting recordings were segmented into 2049 smiles and 960 laughs. Due to the spontaneous nature of these meetings, speech, chewing and occlusions sometimes co-occur with the smile and laugh segments.

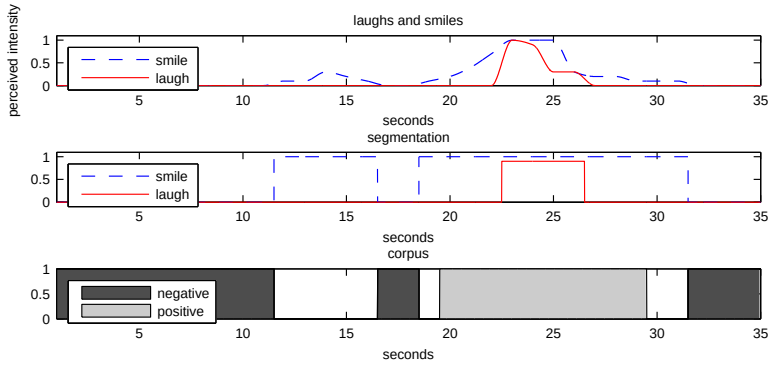

Figure 3.2: Segmentation of the data and extraction of instances for the corpus. On top the typical observed intensity of the smiles and laughs is shown. Based on the observations, a segmentation is made, as shown in the middle diagram. The laughs are padded with 3 seconds on both sides to form the positive instances. The negative instances are created from the remaining non-smile space.

3.1.3 Corpus

The final corpus is built using this segmentation data. The laughter instances are created by padding each laughter segment with 3 seconds on each side to capture the onset and offset of a visual laughter event (see Figure 3.2). A preliminary experiment showed that these onset and offset segments increased the performance of the classifier. Laughter segments that overlapped after padding are merged into a single laughter instance. This effectively merges separate laughter calls to a instance containing a single laughter bout. The non-laughter instances are created from the audio-visual data that remains after removing all the laughter and smile segments; the smile segments are not used during this research. The length of the non-laughter instance is taken from a random Gaussian distribution with a mean and standard deviation equal to the mean and standard deviation of the laughter segments. Due to time constraint we have based our corpus on selected 60 randomly selected laughter and 120 randomly selected non-laughter instances, in which the 20 facial points needed for tracking are visible. Of these 180 instances, 59% contains speech of the visible participant. Almost all instances contain background speech. Together these instances consist 25 minutes of audio-visual data.

3.2 Features

This section outlines the features we have used for the audio and video modalities. For audio we use features that are commonly used for the detection of laughter in audio. For video we used features based on the location of 20 facial points.

3.2.1 Audio features

In order to detect laughter in audio, the audio signal has to be transformed to useful features for classification algorithms. Spectral or cepstral audio features, such as Mel-Frequency Cepstrum Coefficients (MFCC) [25] and Perceptual Linear Predictive (PLP) Analysis [19], have been used successfully for automatic speech recognition and laughter detection. We decided

Figure 3.3: Laughter examples for each individual in the corpus.

to use PLP features, with the same settings as used by Truong and van Leeuwen [38] for automatic laughter detection, and RASTA-PLP features with similar settings.

PLP and RASTA-PLP can be understood best as a sequence of transformations. The first transformation is Linear Predictive Coding (LPC). LPC encodes speech based on the assumption that speech is comparable with a buzzer at the end of tube; the formants of the speech are removed, and encoded with the intensity and frequency of the remaining buzz. PLP adds a transformation of the short term spectrum to LPC encoded audio, in order to mimic human hearing. We used these PLP features for audio classification. In addition to the PLP audio features, we derived RASTA-PLP [20] features. RASTA-PLP adds filtering capabilities for channel distortions to PLP, and yield significantly better results for speech recognition tasks than PLP in noisy environments [13]. A visualisation of PLP and RASTA-PLP features can be found in Appendix B.

For PLP-features we used the same settings as were used by Truong and Van Leeuwen [38] for laughter detection (see Table 3.1). The 13 cepstral coefficients are calculated (12 model order, 1 gain) over a window of 32 ms with a step-size of 16 ms. Combined with the temporal derivative (calculated by convolving with a simple linear-slope filter over 5 audio frames) this results in a 26 dimensional feature vector per audio frame. The RASTA-PLP features are created using the same settings. We normalize these 26-dimensional feature vectors to a mean $\mu = 0$ and a standard deviation $\sigma = 1$ using z-normalisation.

	PLP	RASTA-PLP
Sampling frequency:	16 kHz	16 kHz
Window size:	32 ms	32 ms
Window step-size:	16 ms	16 ms
Model order:	12	12
Delta window:	5 frames	5 frames
Log-RASTA filtering:	false	true

Table 3.1: Settings used for the PLP and RASTA-PLP features

3.2.2 Video features

The video channel was transformed into sequences of 20 two-dimensional facial points located on key features of the human face. These point sequences are subsequently transformed into orthogonal features using a Principal Component Analysis (PCA).

The points were tracked as follows. The points were manually assigned at the first frame of an instance movie and tracked using a tracking scheme based on particle filtering with factorized likelihoods [33]. We track the brows (2 points each), the eyes (4 points each), the nose (3 points), the mouth (4 points) and chin (1 point). This tracking configuration has been used successfully [39] for the detection of the atomic action units of the FACS. This results in a compact representation of the facial movement in a movie using 20 (x, y) tuples per frame (see Figure 3.4).

After tracking, we performed a PCA on the 20 points per video-frame. A PCA linearly transforms a set of correlated variables in a set of uncorrelated variables [24]. The principal components are ordered so that the first few retain most of the variance of the original variables. Therefore a PCA can be used as a dimension-reduction technique for features [1], however we chose to keep all the dimensions because we do not know in advance which

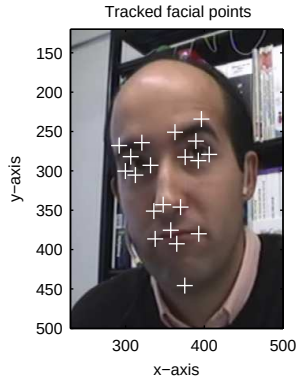

Figure 3.4: The tracked facial points

principal components are useful for laughter detection. We have chosen to use PCA over manually defined features because PCA can detect factors – such as differences in head shape – that are otherwise difficult to detect and remove from the features.

For each frame in the videos we defined a 40-dimensional shape vector by concatenating all the Cartesian (x, y) coordinates. Using a PCA we extracted 40 principal components (eigenvectors) for all the frames in the data set. The original shape vectors can be reconstructed by adding a linear combination of these eigenvectors, to the mean of the shape vectors:

$$x = \overline{x} + bP^T \tag{3.1}$$

Here x is the original shape vector, $\overline{x}$ is the mean of the shape vectors, b is a vector of weights and P matrix of the eigenvectors.

An analysis of the eigenvectors revealed that the first five principal components encode the head pose, including translation, rotation and scale. The other components encode interpersonal differences, facial expressions, corrections for the linear approximations of movements and less obvious factors of the facial configuration. See Figure 3.5 for a visualisation of the first 12 principal components, for more information please refer to Appendix A.

The matrix of eigenvectors serves as a parametric model for the tracked facial points. The Active Shape Model developed by Cootes et al. [10] used a similar technique to create a model for shapes. The main difference is that Cootes at al. removed global linear transformations from the model by aligning the shapes before the PCA is applied. We did not align the shapes because the head modality seems to contain valuable cues for laughter detection we want to include in the model.

We use the input for this model (the weight vector b) as feature vector for the video-data. For unseen data, this feature vector can be calculated using Equation 3.2.

$$b = (x - \overline{x})P \tag{3.2}$$

In order to capture temporal aspects of this model, the first order derivative for each weight is added to each frame. The derivative is calculated with $\Delta t = 4$ frames on a moving average of the weights with a window length of 2 frames. Facial activity (onset-apex-offset) can last from a 0.25 seconds (for example a blink) to several minutes [16]. With a $\Delta t = 4$ frames even the fastest facial activity is captured in the derivative of the features. We normalize this 80-dimensional feature vector to a mean $\mu = 0$ and a standard deviation $\sigma = 1$ using

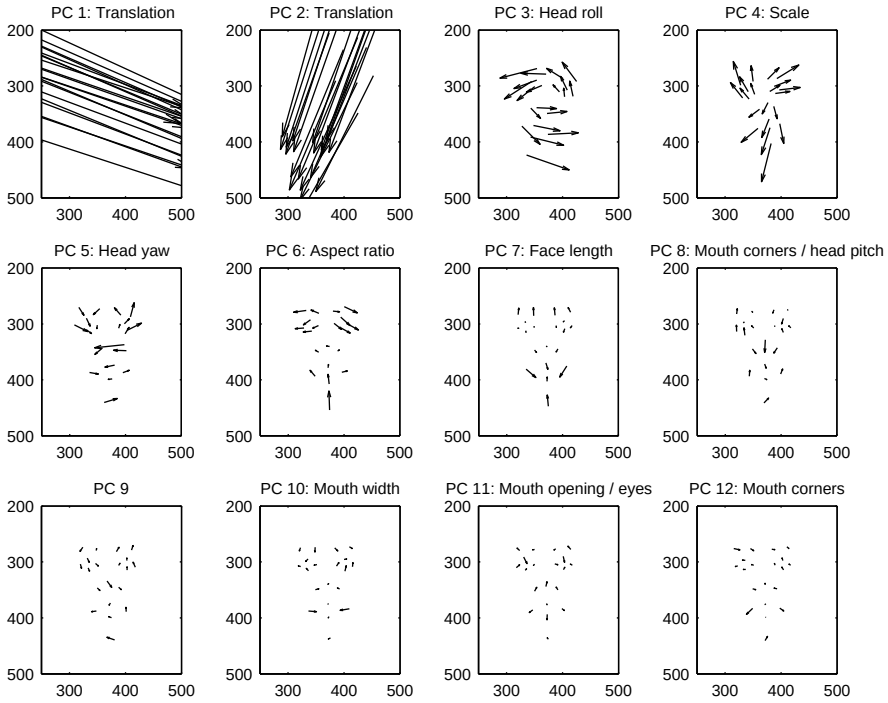

Figure 3.5: A visualisation of the influence of the first 12 principal components. The arrows point from -3σ to 3σ, where σ is the standard deviation.

z-normalisation. This results in a normalized 80-dimensional feature vector per frame which we use for classification (Appendix B).

3.3 Test setup

3.3.1 Classifiers

We selected Gaussian mixture models (GMMs), hidden Markov models (HMMs) and support vector machines(SVMs) as machine learning techniques to be used for classification. GMMs and HMMs are frequently used in speech recognition and speaker identification, and have been used before for laughter recognition [2, 38, 23, 8, 28, 26, 42]. SVMs have been used for laughter detection in [25, 2, 34, 38].

HMMs and GMMs are generative models. Therefore, a different model has to be trained for each class. After training using the EM algorithm [11, 43], the log-likelihood for both class-models is computed and compared for each instance. Using these log-likelihoods the final output is computed as the logarithm of the ration between the probability of the positive and the negative model (Eq. 3.3).

$$score(I) = log(\frac{P_{pos}(I)}{P_{neg}(I)}) = logP_{pos}(I) - logP_{neg}(I) \tag{3.3}$$

We use HMMs that model the generated output using a mixture of Gaussian distributions. For the HMMs classifiers we used two different topologies (Figure 3.6). The first is commonly used in speech recognition, and contains only forward connections. The advantage of this left-right HMM model is that less parameters have to be learned, and the left-right architecture seems to fit sequential nature of speech. An ergodic HMM allows state transitions from every state to every state. This topology is more flexible, but more variables have to be learned. Kevin Murphy's HMM Toolbox [30] was used to implement the GMM and the HMM classification.

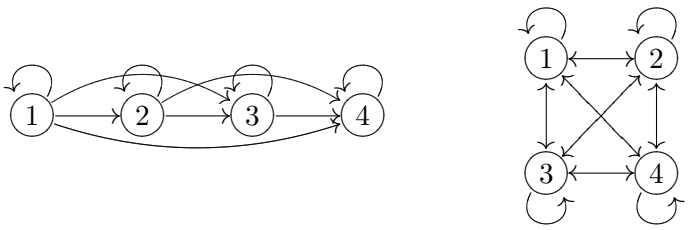

Figure 3.6: left-right HMM (left) and an ergodic HMM (right)

SVMs expect a fixed-length feature vector, but our data consists of sequences with a variable length. Therefore we use a sliding window to create features for the SVM. During training the class of windowed sections of the instances are learned. During classification a probability estimate for the different windows of an instance is calculated. The final score of an instance is the mean of its window-scores, a median could be used as well. We use Radial Basis Function (RBF) kernel SVMs, which are trained using LIBSVM [9].

3.3.2 Fusion

Fusion is performed on the decision-level, which means that the output of an audio and a video classifier are used as input for the final fused prediction. For each instance we classify, we generate two numbers, representing the probability of laughter in the audio and the video modality. Fusion SVMs are trained on these numbers using same train, validation and test sets as used for the single modality classifiers (see Section 3.3.3). The output of these SVMs is a multi-modal prediction based on high-level fusion. As an alternative to this learned fusion, we test fusion using a weighted-sum (Equation 3.4) of the predictions to fuse the scores of the single-modality classifiers.

$$s_{fused} = \alpha * s_{video} + (1 - \alpha) * s_{audio} \tag{3.4}$$

3.3.3 Cross validation scheme

In order to compare different fusion techniques, we need to be able to measure the generalisation performance of a classifier. We decided to use a preprocessed data set, so the preprocessing is done once for the whole data set. We have chosen to exclude the preprocessing from the cross-validation loop in order to measure the generalisation error of the fusion without the additional generalisation error of the preprocessing. The preprocessing consists of feature-extraction, and z-normalisation which transforms the data to a mean $\mu = 0$ and $\sigma = 1$. Using this setup we measure the generalisation error of the classification, and not the combined generalisation error of preprocessing and classification.

Because we have a small data set we use a cross-validation scheme to create multiple train, validation and test sets (see Figure 3.7).

Algorithm 1: The used cross-validation scheme.

for *K in [1..10]* **do**
  $S_{train} = S - S_K$;
  for *L in [1..3]* **do**
    $S_{validation} = S_{K_L}$;
    $S_{test} = S_K - S_{K_L}$;
    C = trainer.learn(S_{train}, $S_{validation}$);
    $S_{test}.performance$ = trainer.test(C, S_{test});
  end
end

The preprocessed data set is divided into K=10 subsets. During each of the K folds, 1 subset is set aside. The other 9 subsets are used for training. The remaining subset is used to create a validation and a test set for three folds. One third is used as validation set and the remaining two thirds as test set (see Algorithm 1). Different model-parameters are used to train classifiers on the train set. The classifier with the best performance on the validation-set is selected, and tested on the test set. This results in performance measurements for $10 \times 3 = 30$ different folds of the data set.

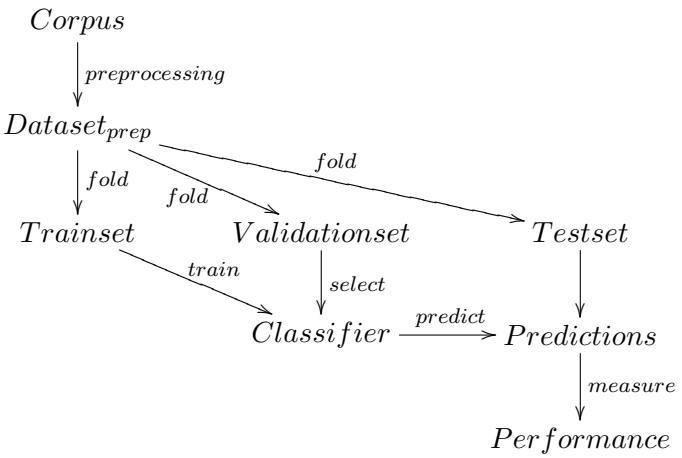

Figure 3.7: A train, validation and test set are used to measure the generalisation performance of a classifier

3.3.4 Model-parameter estimation

Most machine learning techniques have model-parameters (for example, the number of states and the number of Gaussian mixtures for a HMM, the C and γ parameters for a SVM with a RBF kernel) that influence their performance. We find good parameters by performing a multi-resolution grid-search [22] in the model-parameter space in which we search for the parameters that result in the best performance on the validation set after training. For a SVM with a RBF-kernel, we test different values for the $log(C)$ and $log(\gamma)$. The parameters that result in the highest AUC-ROC (see section 3.3.5) form the center of a smaller grid, whose values are again tested on the validation set. The best scoring classifier is the final classifier.

For generative models, such as HMMs and GMMs, we perform the same grid-based parameter search. Because we need a model for both the positive and the negative instances, the grid-search is performed for both classes individually. The performance measure during this search is the log-likelihood of the model on the validation set. For GMMs, we estimate the best number of Gaussian mixtures for our data set. For HMMs, we search the best values for the number of states, the number of Gaussian mixtures and a Boolean that determines if the HMM is fully connected or not.

3.3.5 Performance measure

In order to calculate the generalisation performance of a classifier, we need to select a suitable measure for the performance. We have chosen to use and the Area Under Curve of the Receiver Operating Characteristic (AUC-ROC) [15] as primary and the Equal Error Rate (EER) as secondary performance measure.

$$accuracy = \frac{TP + TN}{P + N} \tag{3.5}$$

$$recall = \frac{TP}{P} \tag{3.6}$$

$$precision = \frac{TP}{TP + FP} \tag{3.7}$$

The most commonly used measure in previous work is the accuracy (Equation 3.5), or the recall and precision pair (Equation 3.6 and Equation 3.7). The accuracy measure is not suitable to measure the performance for a two-class problem, because a very high accuracy can be obtained by predicting the most frequent class for problems with a high class skew. The combination of recall and precision is more descriptive. Recall expresses the fraction of detected positive instances, precision describes the fraction of the detected instances that is a real positive. Those measures can be calculated using the values found in the confusion matrix (Fig. 3.8).

prediction \ class	positive	negative
positive	TP	FP
negative	FN	TN
all	P	N

Figure 3.8: A confusion matrix, where the columns represent the real class, and the rows represent the prediction of a classifier. The cells contain the true positives (TP), false positives (FP), the false negatives (FN) and true negatives (TN).

Most classifiers can be modified to output a probability of a class instead of a binary decision. A trade-off for the cost of different errors – FP versus FN – can be made by thresholding this probabilistic output. This trade-off can be visualized in a receiver operating characteristic (ROC), in which the true-positive rate is plotted against the false-positive rate for different thresholds (see Figure 3.9). A recall-precision pair corresponds to a single point on the ROC. One of the advantages of the ROC over other thresholded plots is its invariancy to class-skew [15]. Because we do not know in advance which costs are associated with the different errors, we cannot define a single point of interest on the ROC. Therefore we measure the performance using the area under the ROC curve (AUC-ROC). The AUC-ROC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. In addition to the AUC-ROC performance, we will report the EER for a classifier. The EER is single a point on the ROC, defined as the point for which the false-positive rate equals the false-negative rate.

We will use a paired two-tailed t-test to compare the AUC-ROCs of the cross-validation folds. This K-fold cross-validated paired t-test suffers from the problem that the train sets overlap, which results in an elevated probability of detecting a difference between classifiers when no such difference exists (type I error) [12]. As a solution for this problem the 5×2 cross-validated paired t-test has been developed, which has an acceptable type I error. Because this method uses only half of the data for training during a fold it is unsuitable for our data set. Therefore we use the K-fold cross-validated paired t-test to compare the AUC-ROC values for different classifiers, and note the possibility of a type I error.

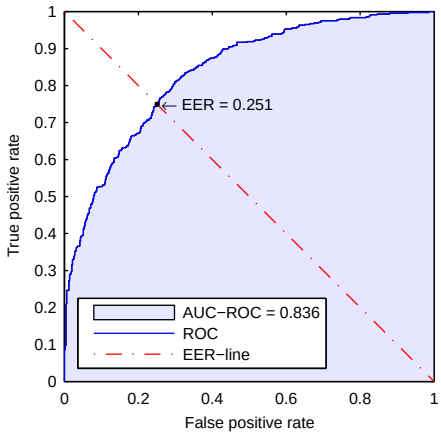

Figure 3.9: The ROC, the ROC-AUC and the EER for a classifier. The probabilistic output of the classifier is thresholded to generate the ROC-curve. Points on the curve define the relation between the true positive rate and the false positive rate. The area under the ROC (the AUC-ROC) is our primary performance measure. The EER for a classifier is the error-rate in the intersection of the ROC with the EER-line from $(0, 1)$ to $(1, 0)$.

Chapter 4

Results

In this chapter we will describe the results of our experiments. We will start with the results for the single-modality classifiers. The best single-modality classifiers are used to construct a fused classifier, which we will compare to the best performing single modality classifier.

4.1 Single-modality classifiers

We will start with the audio classifiers. We have trained different classifiers on the the two sets of audio features. Figure 4.1 shows a ROC-plot for the audio features . The figure shows that all the trained classifiers have similar performance for PLP features. The only real differences are in the area with a very low threshold (high recall, low precision) and the are with a high threshold (low recall, high precision). In those areas the generative models (GMMs and HMMs) seem to perform better. When we look at Table 4.1 we see that the number of Gaussian mixtures for the positive and negative model seems to be proportional to the amount of train data. We expect that more train data would increase the number of mixtures and possibly the performance of our GMM and HMM classifiers. This is supported by the work of Truong et al. [38], where models with 1024 Gaussian mixtures were trained using more than 3000 instances.

Classifier	Features	Positive model		Negative model		AUC-ROC	EER
		#states	#mix.	#states	#mix.		
GMM	PLP	-	21.0 (3.2)	-	48.8 (3.1)	0.794 (0.169)	0.331
GMM*	RASTA	-	16.9 (2.8)	-	35.6 (5.9)	0.825 (0.143)	0.258
GMM	Video	-	3.0 (0.7)	-	3.3 (0.6)	0.871 (0.129)	0.208
HMM	PLP	11.0 Erg. (0)	2.1 (0.5)	18.5 (1.1) Erg.	2.5 (0.9)	0.791 (0.160)	0.333
HMM	RASTA	11.6 Erg. (1.9)	2.1 (0.4)	21.3 (1.9) Erg.	2.0 (0)	0.822 (0.135)	0.242
HMM	Video	2.5 LR (0.5)	4.0 (0)	1.2 (0.4) Erg.	3.0 (0)	0.844 (0.129)	0.258

Classifier	Features	Window	Step	$log_2(C)$	$log_2(\gamma)$	AUC-ROC	EER
SVM	PLP	1.12 s	0.64 s	-8.9 (3.7)	-22 (2.6)	0.775 (0.173)	0.315
SVM	RASTA	1.12 s	0.64 s	-9.8 (4.1)	-21.7 (3.2)	0.621 (0.157)	0.400
SVM*	Video	1.20 s	0.60 s	1.3 (5)	-18 (0)	0.916 (0.114)	0.133

Table 4.1: Results of the different classifiers trained on different features. For the model-parameters and the performance measure, the mean value is diplayed with the standard deviation displayed between parenthesis. The classifiers marked with an asterisk are the best performing classifiers for the audio and video modality.

The results for the RASTA-PLP features are remarkebly different. The ROC is not as smooth as for PLP features, and the SVM-performance is degraded dramatically. However, RASTA-PLP features result in a slightly better performance than the PLP features for the generative models. The filtering that RASTA-PLP adds to PLP seems to smoothen the signal (Appendix B). This results in features that can be modeled using fewer mixtures (see Table 4.1), which allows for the training of more states, or training with a higher accuracy. RASTA-PLP was developed with speech recognition in mind, with explains why the generative models that are commonly used in speech recognition perform better with RASTA-PLP features than with PLP features. While the distribution of the values of the features is simplified, the performance for SVMs degrades. SVM-classifiers trained on RASTA-PLP features generally have a lower C-parameter, which indicates a smoother hyper-plane. Therefore we assume that the smoother RASTA-PLP signal allows for more overfitting, which can explain the degraded performance for SVMs on RASTA-PLP features.

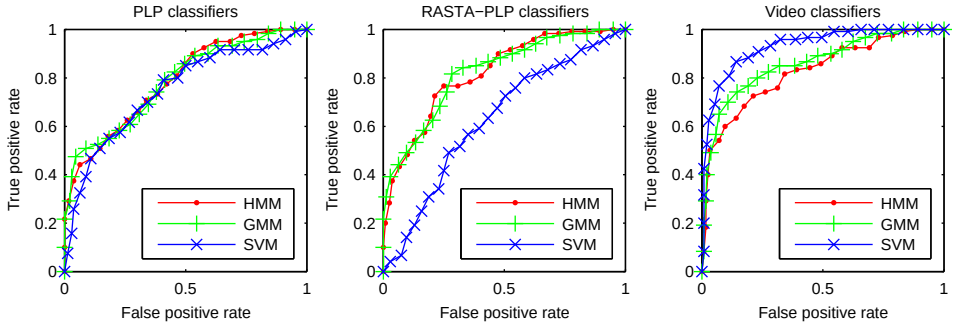

Figure 4.1: The ROC for the PLP features (left), the RASTA-PLP features (mid) and the video features (right).

When we compare the results of the different classifiers trained using PLP and RASTA-PLP features, we observe that the SVM-based classifiers have the worst performance. The difference in performance for the generative models is not as clear. Using a paired samples t-test, we find that the RASTA-PLP features have a significantly higher AUC-ROC ($t(59) = 2.15, p < 0.05$) than the PLP features. We conclude that the combination of a GMM or HMM classifier with RASTA-PLP features results in best performance for laughter detection in audio using our data set.

For the video features we evaluated the same classifiers using different model-parameters. These ROC-plots can be found in Figure 4.1. The ROC-plot shows that classifiers trained on the video modality have a better performance than classifiers trained on the audio modality. When we look at the average model for the HMM-classifier trained on the video features, we notice that the model for the positive instances is a left-right (LR) HMM, while the model for the positive instance for audio is an ergodic HMM (see Table 4.1). The visual laugh seems to display a sequential order, that is not modeled in the audio HMMs. Another difference is that the video modality is modeled using fewer Gaussian mixtures. This can be the result of the higher dimensionality of the video features. The best result for the video modality was obtained using a SVM-classifier. This can be the result of the more sequential pattern of visual laughter, that can be detected more reliable inside of a sliding window than the variable audio signal. The video-SVM classifier has the best single-modality performance.

4.2 High level fusion

For fusion we have selected the best performing single-modality classifiers. For video we use the Video-SVM (VSVM) classifier. We use the GMM or HMM classifiers trained on RASTA-PLP features to classify audio. The results for decision-level fusion can be found in Table 4.2.

Fusion	Features	Compared to Video-SVM	AUC-ROC	EER
RBF-SVM	Video-SVM + RASTA-GMM	$t(29) = 2.45, p < 0.05$	0.928 (0.107)	0.142
RBF-SVM	Video-SVM + RASTA-HMM	$t(29) = 1.93, p = 0.06$	0.928 (0.104)	0.142
Linear-SVM	Video-SVM + RASTA-GMM	$t(29) = 1.51, p = 0.14$	0.925 (0.109)	0.140
Linear-SVM	Video-SVM + RASTA-HMM	$t(29) = 1.78, p = 0.09$	0.927 (0.104)	0.142
W-sum, $\alpha = 0.57$	Video-SVM + RASTA-GMM	$t(29) = 2.69, p < 0.05$	0.928 (0.107)	0.142
W-sum, $\alpha = 0.55$	Video-SVM + RASTA-HMM	$t(29) = 2.38, p < 0.05$	0.930 (0.101)	0.142

Table 4.2: Results of the decision-level fusion. The t-test is a paired samples t-test on the AUC-ROC of the Video-SVM classifier and the specified fusion classifier. The mean value of the ROC-AUC is displayed with the standard deviation displayed between parenthesis.

The fused classifiers have a higher mean AUC-ROC than all the single-modality classifiers. In the case of SVM-fusion, the combination of the Video-SVM classifier and the RASTA-GMM classifiers outperforms the Video-SVM classifier significantly. Inspection of the trained SVM-classifiers reveals that the separating-hyperplane is nearly linear; therefore we tried to replicate the fusion using a linear SVM. Table 4.2 shows that the performance of the linear fusion SVM is slightly worse than fusion using a RBF-kernel.

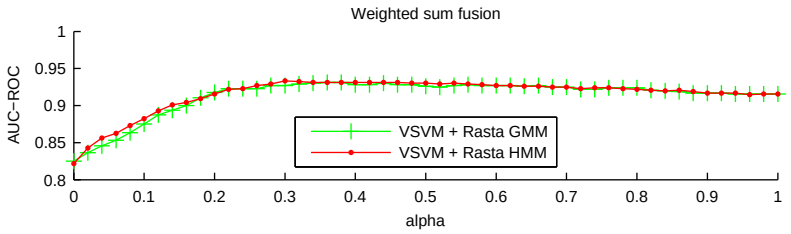

Figure 4.2: The mean AUC-ROC performance of weighted sum fusion, displayed as a function of the α parameter.

In addition to implicit fusion using a SVM, we used a weighted-sum rule (Equation 3.4) to combine the output of the audio and video classifiers. The influence of both modalities is determined using the α parameter. See Figure 4.2 for plot of the AUC-ROC performance as a function of α. For a low values of α, the audio-modality is dominant, the video modality is dominant for values near 1. The highest mean AUC-ROC values are in the region with a more dominant audio-classifier. However, for a significant improvement over the Video-SVM classifier fusion with $\alpha = 0.57$ or $\alpha = 0.55$ is needed for the RASTA-GMM and the RASTA-HMM classifier respectively (see Table 4.2). This indicates a dominant video-classifier as we would expect from the results of the single-modality classification.

When we compare the ROCs of the fusion classifiers with the video-SVM classifiers, we can see that the EER-point of the fused classifiers often has a lower performance than the video-classifier (see Figure 4.3). Most of the performance-gain is obtained in the direct vicinity of the EER point, where the error-rates are not equal. For these thresholds it is easier to exploit the complementary nature of both modalities. For the threshold with an equal error rate, the

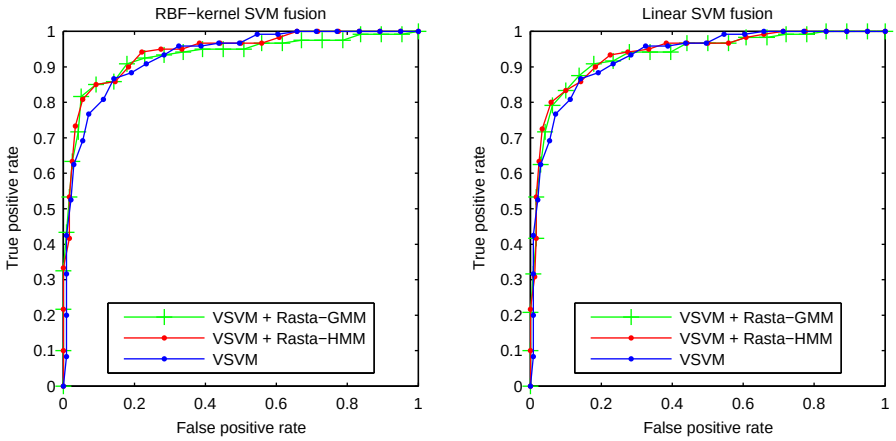

Figure 4.3: The ROC for the high-level fusion using a RBF-kernel SVM (left) and using a linear SVM (right).

hyperplane needs to separate instances for which both modalities are uncertain. This may explain why the EER is slightly higher for the fused classifiers.

Chapter 5

Conclusions

Our goal was to perform automatic laughter detection by fusing audio and video signals on the decision level. We built audio and video-classifiers, and demonstrated that the fusion of the best classifiers significantly outperformed both single-modality classifiers. The best classifier were the following. For audio, the GMM classifier trained on RASTA-PLP features performed best, resulting in a AUC-ROC of 0.825. A mean of 16.9 Gaussian mixtures was used to model laughter, non-laughter was modeled using 35.6 Gaussian mixtures. The best video-classifier was a SVM-classifier with an AUC-ROC of 0.916, trained on windows of 1.20 seconds using a $C = 2.46$ and a $\gamma = 3.8 \times 10^{-6}$. The best audio-visual classifier was constructed by training a SVM on the output of these two classifiers, resulting in a AUC-ROC performance of 0.928.

During the fusion we evaluated different feature-sets. For laughter-detection in audio, we obtained significantly better results with RASTA-PLP features than with PLP features. RASTA-PLP features have not been used before for laughter detection as far as we know. For laughter detection in video we successfully used features based on the PCA of 20 tracked facial points. The performance of the video classifiers was very close to the fused classifiers, which is a promising result for laughter detection in video. However, during this research we excluded instances that contain smiles. It is likely that our video-classifier also classifies smiles as laughter.

The audio and video modalities show some striking differences. The HMM classifiers trained on audio were all ergodic (fully connected) instead of the left-right HMMs that are commonly used for speech recognition. This indicates that there was no strict sequential pattern for laughter that could be exploited for recognition, which seems to support the claim that laughter is a group of sounds [36]. In video such a sequential sequence of states was found. It seems that visual laughter has a more sequential nature than audible laughter. The GMM and HMM classifiers are useful for classification of audio. For laughter detection in video a SVM trained on sliding windows outperforms the other classifiers.

For future work we recommend an investigation of fusion on feature-level. We have demonstrated that decision-level fusion improves the performance, but it is not clear how this relates to other fusion techniques. For low-level fusion, a dimensionreduction technique is most likely needed. During this research we have not performed dimension reduction on the features. A comparison between our best classifier and a low-level fusion classifier both trained on a reduced feature set would therefore be very interesting. Another limitation of this research is that we only perform classification of segmented instances. The extension of these classifiers to a laughter detection system that automatically finds laughter segments in streams forms

an interesting challenge. For example, the predictions of the Video-SVM could be used to segment a stream in candidate laughs, which could be further refined using the log-likelihoods of the generative models.

Bibliography

[1] E. Alpaydin. *Introduction To Machine Learning*. MIT Press, 2004.

[2] J.A. Arias, J. Pinquier, and R. André-Obrecht. Evaluation of Classification Techniques for Audio Indexing. 2005.

[3] H. Asoh, I. Hara, F. Asano, and K. Yamamoto. Tracking human speech events using a particle filter. *Acoustics, Speech, and Signal Processing, 2005. (ICASSP'05). Proceedings of IEEE International Conference on*, 2, 2005.

[4] V. Aubergé and M. Cathiard. Can we hear the prosody of smile? *Speech Communication*, 40(1-2):87–97, 2003.

[5] J.A. Bachorowski, M.J. Smoski, and M.J. Owren. The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, 110:1581–1597, 2001.

[6] M.J. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):828–836, 2003.

[7] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211, 2004.

[8] N. Campbell, H. Kashioka, and R. Ohara. No Laughing Matter. *Proceedings of the Ninth European Conference on Speech Communication and Technology (Interspeech)*, pages 465–468, 2005.

[9] C.C. Chang and C.J. Lin. LIBSVM: a library for support vector machines. `http://www.csie.ntu.edu.tw/cjlin/libsvm`, 2001.

[10] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[11] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[12] T.G. Dietterich. Approximate Statistical Test For Comparing Supervised Classification Learning Algorithms, 1998.

[13] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *Multimedia, IEEE Transactions on*, 2(3):141–151, 2000.

[14] P. Ekman and W.V. Friesen. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, 6(4):238–252, 1982.

[15] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[16] W.V. Friesen, J.C. Hager, and A.H. Face. *Facial Action Coding System*. A Human Face, 2002.

[17] H.J. Go, K.C. Kwak, D.J. Lee, and M.G. Chun. Emotion recognition from the facial image and speech signal. *SICE 2003 Annual Conference*, 3, 2003.

[18] H. Gunes and M. Piccardi. Fusing face and body display for Bi-modal emotion recognition: Single frame analysis and multi-frame post integration. *Lecture notes in computer science*, pages 102–111, 2005.

[19] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87:1738–1752, 1990.

[20] H. Hermansky, N. Morgan, A. Bayya, P. Kohn, U.S.W.A. Technologies, and CO Boulder. RASTA-PLP speech analysis technique. *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, 1, 1992.

[21] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll. Bimodal fusion of emotional data in an automotive environment. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2, 2005.

[22] C.W. Hsu, C.C. Chang, and C.J. Lin. A practical guide to support vector classification. *National Taiwan University, Tech. Rep., July*, 2003.

[23] A. Ito, X. Wang, M. Suzuki, and S. Makino. Smile and Laughter Recognition using Speech Processing and Face Recognition from Conversation Video. *Cyberworlds, 2005. International Conference on*, pages 437–444, 2005.

[24] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[25] L. Kennedy and D. Ellis. Laughter detection in meetings. *Proc. NIST Meeting Recognition Workshop*, 2004.

[26] D. Kimber and L. Wilcox. Acoustic segmentation for audio browsers. *Proc. Interface Conference*, 1996.

[27] S. Kipper and D. Todt. The Role of Rhythm and Pitch in the Evaluation of Human Laughter. *Journal of Nonverbal Behavior*, 27(4):255–272, 2003.

[28] A. Lockerd and F.L. Mueller. Leveraging Affective Feedback Camcorder. *Proc. of CHI'02*, 2002.

[29] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, et al. The AMI Meeting Corpus. *Proceedings of Measuring Behavior*, 2005.

[30] K. Murphy. Hidden Markov Model (HMM) Toolbox for Matlab. `http://www.ai.mit.edu/murphyk/Software/HMM/hmm.html`.

[31] P. Pal, A.N. Iyer, and R.E. Yantorno. Emotion detection from infant facial expressions and cries. *Proc. Int'l Conf. Acoustics, Speech & Signal Processing*, 2, 2006.

[32] Pantic, M. A multimodal approach to automatic recognition of posed vs. spontaneous smiles. 2007.

[33] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 97–102, 2004.

[34] T. Pohle, E. Pampalk, and G. Widmer. Evaluation of frequently used audio features for classification of music into perceptual categories. *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing*, 2005.

[35] J. Trouvain. Phonetic Aspects of Speech-Laughs. *Proc. Confer. on Orality & Gestuality (Orage)*, pages 634–639, 2001.

[36] J. Trouvain. Segmenting Phonetic Units in Laughter. *Proc. 15th International Conference of the Phonetic Sciences, Barcelona, Spain*, pages 2793–2796, 2003.

[37] K.P. Truong and D.A. van Leeuwen. Automatic detection of laughter. *Proc. Interspeech Euro. Conf*, pages 485–488, 2005.

[38] K.P. Truong and D.A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007.

[39] M. Valstar, M. Pantic, Z. Ambadar, and JF Cohn. Spontaneous vs. posed facial behavior. *ACM International Conference on Multimodal Interfaces,(Banff, Canada, 2006)*, 2006.

[40] J. Vettin and D. Todt. Laughter in Conversation: Features of Occurrence and Acoustic Structure. *Journal of Nonverbal Behavior*, 28(2):93–115, 2004.

[41] Y. Wang and L. Guan. Recognizing human emotion from audiovisual information. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2, 2005.

[42] M. Xu, L.-T. Chia, and J. Jin. Affective Content Analysis in Comedy and Horror Videos by Audio Emotional Event Detection. *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 622–625, 2005.

[43] G. Xuan, W. Zhang, and P. Chai. EM algorithms of Gaussian mixture model and hidden Markov model. *Image Processing, 2001. Proceedings. 2001 International Conference on*, 1, 2001.

[44] W. Zajdel, D. Krijnders, T. Andringa, and D.M. Gavrila. Cassandra: Audio-video sensor fusion for aggression detection, 2007.

[45] Z. Zeng, Y. Hu, Y. Fu, T.S. Huang, GI Roisman, and Z. Wen. Audio-visual emotion recognition in adult attachment interview. *Proceedings of the 8th international conference on Multimodal interfaces*, pages 139–145, 2006.

[46] Z. Zeng, Y. Hu, G.I. Roisman, Z. Wen, Y. Fu, and T.S. Huang. Audio-visual spontaneous emotion recognition.

[47] Z. Zeng, J. Tu, M. Liu, and T.S. Huang. Multi-stream confidence analysis for audio-visual affect recognition. *Lecture notes in computer science*, pages 964–971.

[48] Z. Zeng, J. Tu, M. Liu, TS Huang, B. Pianfetti, D. Roth, and S. Levinson. Audio-Visual Affect Recognition. *Multimedia, IEEE Transactions on*, 9(2):424–428, 2007.

[49] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T.S. Huang, D. Roth, and S. Levinson. Bimodal HCI-related affect recognition. *Proceedings of the 6th international conference on Multimodal interfaces*, pages 137–143, 2004.

Appendix A

Principal components for the video features

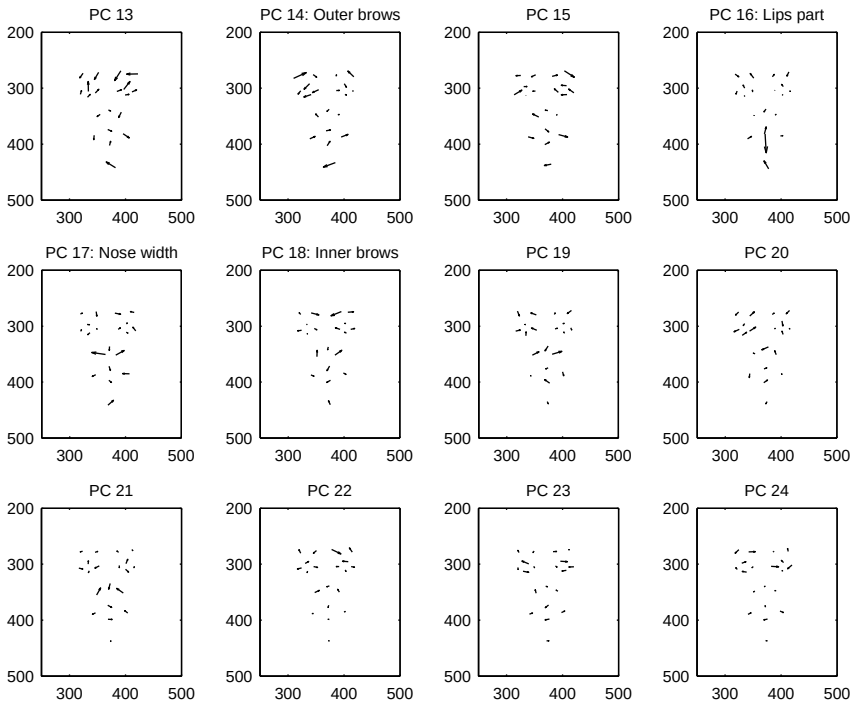

Figure A.1: A visualisation of the influence of principal components 13 to 24. The arrows point from -6σ to 6σ, where σ is the standard deviation.

PC	Description
1	Translation top-left to lower-right, small rotation
2	Translation top-right to lower-left, small rotation
3	Head roll counter clockwise
4	Scale
5	Head yaw
6	Aspect ratio
7	Brow raise, mouthcorners lower, mouth gets smaller, chin raises
8	Mouth corners raise, eyes get smaller
9	Horizontal bending (2nd order rotation?)
10	Mouth size, eye distance, eye size
11	Mouth size, eye size
12	Horizontal S-bend (3rd order rotation?)
13	Waving displacement (high order rotation?)
14	Outer brow movement, horizontal chin movement
15	Nose width, brow movement
16	Mouth open/closed
17	Nose width
18	Nose widht, eye and brows distance
19	High order rotation component
20	High order rotation component
21	High order rotation component
22	Eye size, brows rotation
23	Eye rotation
24	Tracker noise
25	Mouth corner movement, outer eye movement
26	High order rotation
27	Asymmetrical brow movement
28	Left eye rotation
29	Asymmetrical eye size
30	Mouth rotation
31	Eye shape change
32	Mouth skew
33	Left brow rotation
34	Left brow movement, eye movement
35	Left eye skew
36	Eye shape change
37	Right eye shape change
38	Nose rotation
39	Asymmetrical eye open/closed
40	Asymmetrical eye shape change

Table A.1: Principal components for the video frames

Appendix B

Normalized features

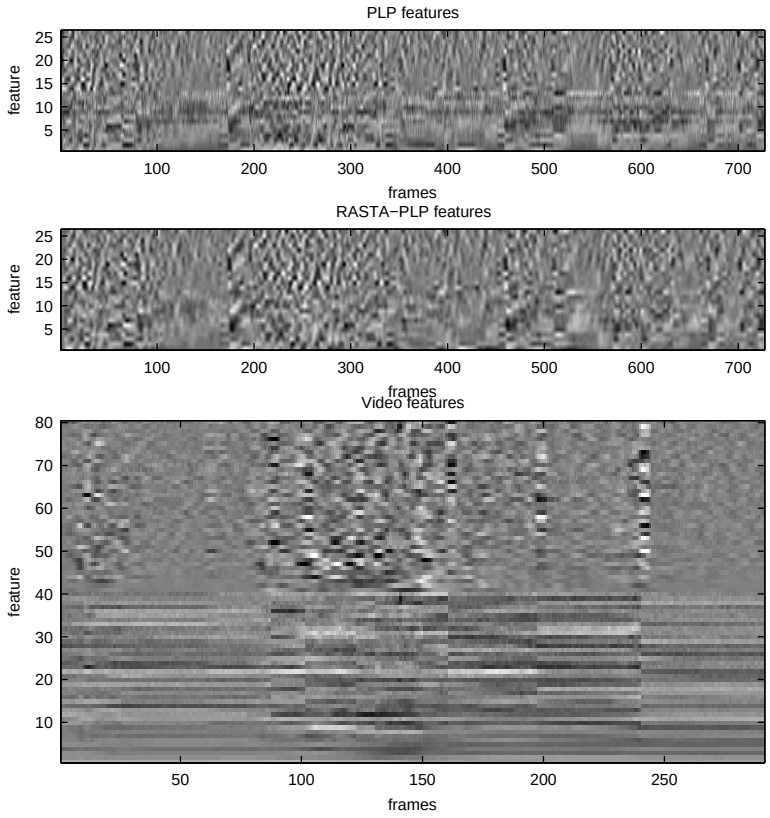

Figure B.1: A visualisation of the raw, normalised features for one instance. Time proceeds from left to right. It can been seen that the RASTA-PLP audio features are more smooth than the PLP audio features. For each feature-set, the high-numbered features are the temporal derivative of the lower numbered features.