Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data

Johannes Wagner, Florian Lingenfelser, Elisabeth André, Jonghwa Kim, *Senior Member, IEEE*

Abstract—The study at hand aims at the development of a multimodal, ensemble based system for emotion recognition. Special attention is given to a problem often neglected: missing data in one or more modalities. In off-line evaluation the issue can be easily solved by excluding those parts of the corpus where one or more channels are corrupted or not suitable for evaluation. In real applications, however, we cannot neglect the challenge of missing data and have to find adequate ways to handle it. To address this, we do not expect examined data to be completely available at all time in our experiments. The presented system solves the problem at the multimodal fusion stage, so various ensemble techniques – covering established ones as well as rather novel emotion specific approaches – will be explained and enriched with strategies on how to compensate temporarily unavailable modalities. We will compare and discuss advantages and drawbacks of fusion categories and extensive evaluation of mentioned techniques is carried out on the CALLAS Expressivity Corpus, featuring facial, vocal and gestural modalities.

Index Terms—Ensemble Based Systems, Decision Level Fusion, Multimodal Emotion Recognition, Missing Data

1 Introduction

EMOTIONAL sensitivity in machines is believed to be a key element towards more human-like computer interaction. Due to the complex nature of human emotions, automatic emotion recognition still remains a challenging task since many years. One difficulty a machine has to face is the fact that humans express their emotions rarely exclusively, but use several channels such as speech and mimics. Studies that have focused on the fusion of multiple channels, however, often start from too optimistic assumptions, e. g. that all data from the different modalities is available at all time. As long as a system is only evaluated on off-line data this assumption can be easily ensured by examining given samples beforehand and excluding parts where one or more channels are corrupted or not suitable for evaluation. In real, application-oriented on-line systems, however, we cannot neglect the issue of missing data and have to find adequate ways to handle it, so that robustness of recognition performance can be guaranteed.

Generally we can identify various causes for missing data: a sensor device can fail so that an according signal is no longer available. Even if a sensor device is running properly there is the possibility of desired information within a signal being no longer accessible, e. g. a tracked object disappears from the view of a camera. We can also think of a situation in which the desired information theoretically is at hand but practically corrupted to some extent, e. g. a speech signal that is overlaid by noise. Finally, not only technical problems can be responsible for one or more modalities to become useless. If a subject simply does not generate observable material, no meaningful data can be recorded, e. g. the gesture modality will not contribute relevant information while monitoring a momentarily motionless user. A system capable of handling missing data must therefore dynamically decide, which channels are available and to what extent the present signals can be trusted. For the case that data is partially missing a couple of treatments have been suggested in literature: Multiple imputation predicts missing values using existing values from previous samples [23]. In data marginalization unreliable features are marginalized to reduce their effect during the decision process [11]. However, consequences of imperfect data on the performance of fusion algorithms have not been systematically explored.

The realization of handling missing data within our aspired multimodal system is engineered within the fusion of available modalities. We explore various standard fusion schemes that are amongst others comprehensively described in [24] as well as more sophisticated and emotion specific fusion techniques. All presented strategies are enriched with strategies on how to treat unavailable modalities.

For final evaluation, every on-line system can be pre-trained with data gathered from multiple subjects and afterwards classification tasks are carried out on a user not known during the training phase. An alternative to this user-independent approach is training the system with data recorded from the subject that is going to use the system afterwards. This approach presumably adapts the system strongly to a single subject and therefore better classification results on this person can be expected at the cost of universality.

- *The authors are with the Department of Human-Centered Multimedia, University of Augsburg, Germany.*
E-mail: <name>@informatik.uni-augsburg.de

Both methods have their advantages and should be compared in terms of recognition accuracy.

The reported processing and classification methods, as well as, all fusion based approaches used in our experiments have been developed with Smart Sensor Integration (SSI) [33], a framework for multimodal signal processing developed at our lab aiming to support the building of on-line recognition systems. SSI supports the design and evaluation of machine learning pipelines by offering tailored tools for signal processing, feature extraction, and pattern recognition, as well as, tools to apply them off-line (training phase) and on-line (real-time recognition).

2 Related Work

Recently many studies in multimodal affect recognition have been done by exploiting synergistic combination of different modalities. Most of the previous works focus on fusion of audiovisual information for automatic emotion recognition, e.g. combining speech with facial expression. De Silva et al. [28] and Chen et al. [10] proposed a rule-based decision-level fusion method for a combined analysis of speech and facial expressions. Huang and colleagues [18] used boosting techniques to automatically determine adaptive weights for audio and visual features. In the work of Busso et al. [5], an emotion-specific comparison of feature-level and decision-level fusion has been reported by using an audiovisual database containing four emotions, sadness, anger, happiness, and neutral state, deliberately posed by an actress. They observed for their corpus that feature-level fusion was most suitable for differentiating anger and neutral state while decision-level fusion performed better for happiness and sadness. They concluded that the best fusion method depends on the application. Interestingly, in addition to speech and facial expression, the thermal distribution of infra-red images is also integrated to a multimodal recognition system [34] by considering the fact that infra-red images are hardly affected by lighting conditions, which is one of the main problems in facial image analysis.

Humans use several modalities jointly in a complementary manner [10]. For decision-level fusion, however, multiple uni-modal classifiers are trained for each modality separately and those decisions are fused by using specific weighting rules. This means that such kind of fusion method is necessarily based on the assumption of conditional independence between modalities. To address this problem, a number of model-level fusion methods have been proposed, originally in the research field of speaker identification, that are capable to exploit cross-correlations between modalities. Song et al. [29], for example, proposed tripled HMM that models correlations between upper face, lower face, and prosodic dynamic behaviours. By relaxing the general requirement of synchronized segmentation for audiovisual streams, Zeng et al. [35] proposed a multi-stream fused HMM which provides the possibility of optimal combination among multiple streams from audio and visual channels. For the estimation of correlation levels between the streams, they used the maximum entropy and the maximum mutual information criterion. Sebe and colleagues [27] suggest the use of dynamic Bayesian Networks to model the interdependencies between audio and video data and handle imperfect data by probabilistic inference. Finally, various types of multimodal correlation models are based on extended artificial neural networks, e.g. [14] [6].

Although there are many studies in psychology supporting that the combined face and body approaches are the most informative for the analysis of human expressive behaviour [1], there is surprisingly very few effort reported on automatic emotion recognition by combining body gesture with other modalities such as facial expression and speech. For example, bimodal fusion methods at different levels for emotion recognition are presented by Balomenos et al. [2] and Gunes & Piccardi [15], using facial expression and body gesture. Kaliouby and Robinson [12] proposed a vision-based computational model to infer acted mental states from head movements and facial expressions. Castellano et al. [9] presented a multimodal approach for the recognition of eight emotions that integrates information from facial expressions, body gestures, and speech. They showed a recognition improvement of more than 10% compared to the most successful uni-modal system and the superiority of feature-level fusion to decision-level fusion. All these approaches are based on visual analysis of expressive gestures and dealt with mapping different gesture shapes to relevant emotions. In our experiment, however, we use three-axis accelerometer instead of visual information in order to extract non-proportional movement properties such as relative amplitude, speed, and movement fluidity, under the assumption that distinct emotions are closely associated with different qualities of body movement rather than gesture shapes.

All the studies reviewed above have shown that the performance of automatic emotion recognition systems can be improved by employing multimodal fusion. Some of them highlight the benefits of fusion mechanisms in situations with noisy features or missing values of features, for example, see [27]. Nevertheless, surprisingly few fusion approaches explicitly address the problem of non-available information. Most of them are based on the assumption that all data is available at all time. This precondition is not realistic in practical environments. In order to guarantee consistent classification, ways of handling missing sensory input have to be thought of.

3 The CALLAS Expressivity Corpus

For training and evaluation of classification systems the choice of an adequate corpus is substantial. Only significant empirical data enables meaningful statements about the performance of investigated recognition techniques. The CALLAS Expressivity Corpus [8] constitutes all desired demands for our aims. It was constructed within the European Integrated Project CALLAS and contains affective behaviour, incorporating vocal utterances, facial expressions and gesture expressivity in three primary emotion classes. The gesture stream is not available for all made observations, so applicable ways of handling missing data must be incorporated into experiments on this corpus.

3.1 Data Generation

As the present corpus was originally designed for examination of cultural differences between emotion expressions of persons from different European countries it was initially made with subjects from Greece, Italy and Germany. This work bases solely on data collected from German participants, as we do not aim at dealing with cultural differences in detail. The German sub-corpus contains 21 persons (10 female and 11 male) and almost 5h of recorded interaction[1], which is sufficient for our investigations.

During the experiment participants were asked to perform expressive sentences through voice, face and body language. During a session 120 emotion inducing sentences were successively displayed to the participants. The sentences, which were inspired by the Velten mood induction technique [31], can be divided according to their semantic content in three broad categories, namely *positive*, *negative* and *neutral*. After he or she had read a sentence silently the projection was blanked out and the sentence got expressed in their own words and with whatever gesture or voice they felt to be fitting. It should be noted that recorded persons had no acting background and it was left to their discretion to what extent they expressed the emotions. This, of course, leads to a broader diversity among observed expressions as it would under a more restrictive setup. However, it comes closer to what a system must expect under a realistic setting. When in some situations subjects were not using gestures to accompany their speech at all this just reflects what happens in real life and hence renders a situation an emotion recognition system must deal with. For our experiments no samples were removed from the data set.

While users were performing the sentences their actions were captured with two cameras one steering at the the proband's face and one at the whole body. In the following progress we only analyse videos captured from the face and refer to it as the facial modality. Voice was captured from a microphone hanging above the users head. Gestures were tracked in three different modes: either free handed by the body-camera, a Humanware™ data-glove on one hand or two Nintendo Wii™ remote controls, one in each of the user's hands. For purpose of the study at hand, we only use data gathered by the Wii™ controller for movement and acceleration tracking.

3.2 Modelling Emotions

In order to deal with multimodal emotion recognition, a concept of discrete emotion modelling has to be chosen. The term emotion itself is a very abstract concept, describing a vast amount of human feelings. These feelings are too numerous to use them directly for recognition tasks, so they have to be integrated into quantifiable categories of emotions. A discrete emotion model is necessary to define target emotions, so that the recognition system is able to understand the problem to be solved. Moreover, the emotion model supports a convergent labelling process. Such procedures narrow the amount of identifiable feelings and group the wide field of possible individual emotions into a small amount of discrete emotion-classes. One possibility is to have all recorded emotional experiences labelled by external specialists and subsumed under predefined expressions like love, hate, sadness, surprise, etc. However, this approach could restricting, as many blended feelings and emotions cannot adequately be described by the chosen categories. Selection of some particular expressions can not be expected to cover a broad range of emotional states and could suffer from randomness. Another way of categorising emotions is to attach the experienced stimuli to continuous scales. Lang [21] proposes arousal and valence as measurements. These scales describe multiple aspects of an emotion, the combination of stimuli's alignments on these scales defines single emotions. More precisely the valence scale describes the pleasantness of a given emotion. A positive valence value indicates an enjoyable emotion such as joy or pleasure. Negative values are associated with unpleasant emotions like sadness and fear. This designation is complemented by the arousal scale which measures the agitation level of an emotion. Combina-

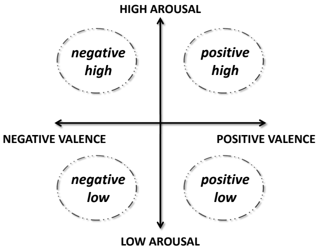

Fig. 1. Arousal - Valence based Emotion Model

[1] Since user sessions were continuously captured this includes intermediate parts where users were reading sentences or changing devices.

tion of the two scales forms four emotion-quadrants, representing a four–class classification problem to be dealt with by a emotion recognition system.

3.3 Segmentation and Annotation

As explained before, users were asked to utter expressive sentences and accompany them with whatever gesture or voice they felt to be fitting. As a natural consequence of this experimental design the voice channel became the dominant modality, which usually triggered on- and offset of a performance. Only in very few cases mimics or gestures were observed before or after an utterance. Consequently, we decided to segment the streams according to the speech signal and used the beginning and ending of each utterance of a sentence as borders. In this way, we ended up with 2513 segments[2], each containing according snapshots of available audio, video and acceleration streams. In average these segments have a length of $2.9s$ with a standard deviation of $0.9s$. Figure 3.3 shows a histogram of the distribution.

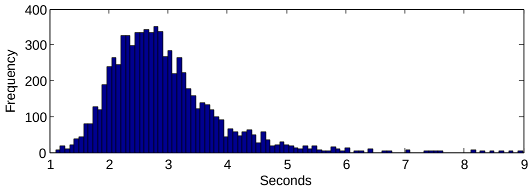

Fig. 2. Histogram of segment lengths.

When annotating a multimodal corpus, we must decide which of the available modalities should serve as source throughout the labelling process. In case of the corpus at hand, judgements could be based either exclusively on the audio or video channel, or by presenting both modalities simultaneously. We did decide in favour of the audio channel as it is – in consequence of the chosen segmentation procedure – the only channel that is always available throughout the whole corpus. Labels assigned to the speech chunks are then consequently applied to the other modalities. In doing so, we actually neglect a problem that often comes across in social communication, namely the problem of blended and masked emotions, which lead to ambiguous expressions across modalities. For instance, if we consider a situation where we are forced to talk with calm voice, while at the same time we use mimics to express our anger about something. However, since such a situation is unlikely to occur in the present corpus, we assume that emotions are more or less homogeneously expressed across the three modalities.

Though used mood inducing sentences are more categorised along the valence axis (positive, neutral and negative), it becomes obvious that an arousal categorisation is also needed. Especially when looking at negative sentences, there are samples tending to a depressed and sad mood (*negative-low*), while others are expressed in an aroused and angry way (*negative-high*). Nearly all neutral and part of the positive observations share a calm and optimistic sub-tone (*positive-low*) in contrast to fewer positive examples bearing clear hints of joy and laughter (*positive-high*). Based on these impressions we refrained from including neutral moods as an own class and label calm and non-negative emotions as *positive-low*. For annotating recorded samples, three experts were asked to independently label made observations in terms of high or low arousal as well as positive or negative valence. The term experts is used to denote that the raters have some sort of knowledge about emotions and their recognition that goes beyond every-day experience. During annotation phase audio segments were replayed in chronological order to each expert independently. Raters could loop an utterance as often as needed and even jump forth and back in order to repeat older segments and re-assign labels.

Final combination of differing annotations is done via majority decision, as three assessments are given to each orientation of valence or arousal respectively, decisions are definite. For instance, if the 1st rater assigns label *low* and *positive*, the 2nd *low* and *negative*, and the 3rd *high* and *negative*, the segment is finally labelled as *low and negative*. Applying the described voting we end up with 1145 samples labelled as *positive-low*, 527 as *positive-high*, 608 as *negative-low*, and 233 as *negative-high*.

To report inter-rater reliability we calculate the kappa value according to Fleiss [13]. Fleiss' Kappa value is a common way to measure the agreement over multiple raters. It is expressed as a number between 0 and 1, where 1 indicates a perfect agreement. Applied to the decision of our raters, we get for valence a kappa value of 0.84, which indicates an almost perfect agreement. For arousal, on the other side, we measure a notably lower value of 0.38, which implies a fair agreement. This drift obviously follows from the fact that expressed sentences were selected to be either negative, neutral or positive. Thus, it should be easier for the raters to agree on the valence of an utterance than its level of arousal. The kappa value for all four classes amounts to 0.52, which expresses a moderate agreement. The obvious disagreement of the raters regarding the labels that were assigned to the samples makes already clear that we must not reckon a 100% accuracy in classification.

4 Methodology

In this section we introduce feature extraction methods for available modalities and describe the classification scheme and feature selection strategies we applied to the CALLAS corpus. As we mainly focus on building on-line recognition systems, we rely solely

[2] 7 segments were skipped either because subjects refused to perform or due to technical problems

on methods that can be computed within a feasible amount of time. In particular, feature extraction and classification should be applicable in (near) real-time and must not require manual tuning at runtime.

4.1 Classification Model

Naive Bayes (NB) is a simple but efficient classification scheme. It is based on the Bayes Theorem which states:

$$P(E_i|f_1,\ldots,f_n) = \frac{P(E_i)\prod_{j=1}^{n} P(f_j|E_i)}{P(f_1,\ldots,f_n)}$$

In other words, this means that the probability of the emotion E_i given an observed feature vector $(f_1,\ldots,f_n)$ of dimension n depends on the *a-priori* probability $P(E_i)$ of the emotion, multiplied by the product of the probability of each feature f_i given the emotion, divided by the *a-priori* probability of the feature vector. As classification result, the emotion E_i from a set of N emotions $E_1,\ldots,E_N$ that maximises the equation is chosen. This is simplifying in so far (and hence the name *Naive* Bayes), as the Bayes Theorem assumes the features to be independent from each other. Parameters for the probability distributions $P(E_i)$ and $P(f_j|E_i)$ are gained from the annotated training material.

Some of the fusion algorithms tested in this paper combine results from no less than 12 internal classification models. To run experiments in a reasonable amount of time it was more important for us to rely on a fast classification scheme than one that gives highest classification rates. Hence, we have chosen Naive Bayes as it is extremely fast in training and testing, even for high-dimensional feature vectors and large training databases. Using a more sophisticated, but at the same time much slower classifier, such as Support Vector Machine (SVM), would certainly improve recognition accuracies. However, since the main focus of this study is the comparison of fusion strategies, the underlying classification model is of less importance.

4.2 Feature Extraction

The extraction of descriptive features brings the raw signals into the compact form required by the classifier. Often features are computed after a preprocessing phase during which additional properties of the signals are carved out and unwanted aspects are suppressed. Depending on whether the features are extracted on a small running window of fixed size or for longer chunks of variable length, we denote them as short- or long-term feature. While in offline analysis the whole signal is available from the beginning and processing can fall back on global statistics, such as global mean and standard-deviation, or perform zero-phase filtering by processing the input data in both the forward and reverse directions, such treatments are not possible in on-line processing. In our experiments signals are processed in small blocks with a fixed window size[3]. This implies that only the information of the current and previously seen blocks are available to our algorithms. Table 1 offers a summary of the applied processing methods. A detailed description is given in the following paragraphs.

4.2.1 Speech Features

CALLAS corpus includes mono audio recordings from a single USB microphone (Samson C01U) placed near the subject's head. The audio stream was captured at a sample rate of 16 kHz and quantised by 16 bit PCM. Recording quality and noise level are similar among all sessions.

The list of proposed features suited for emotion recognition from speech is long. A co-operation of different sites under the name CEICES (Combining Efforts for Improving Automatic Classification of Emotional user States) has carried out experiments based on a pool of more than 4000 features including acoustic and linguistic feature types [3]. Results for the individual groups, as well as a combined set, have led to the following assumptions: Among acoustic features duration and energy seem to be most relevant, while voice quality showed less impact. Yet, no single group outperformed the pool of all acoustic features.

In our experiments we restricted the set of features to those that can be extracted in real-time and in a fully automatic manner. For example, no features have been included that require information on the spoken words or grammatical context, as such information is difficult to get without manual annotation. Hence, we only compute acoustic features related to the paralinguistic message of speech, i. e. we analyse ”how” something is said. In our previous studies [32] MFCC and spectral features turned out to be good candidates. In addition we also compute features from pitch, energy, duration, voicing and voice quality and use feature selection to reduce the full set of 1316 features to the most relevant ones. In a previous study the feature set was evaluated on the Berlin Database of Emotional Speech [4] that is commonly used in offline research (7 emotion classes, 10 professional actors) and achieved an average recognition accuracy of 80%. On the FAU Aibo Emotion Corpus [30] as part of the INTERSPEECH Emotion Challenge 2009 [25], we were able to slightly exceed the baseline given by the organizers for a 5 class problem (anger, emphatic, neutral, positive and rest) [32]. Both corpora have been intensively studied in the past by many researchers working on emotion recognition from speech and serve as a kind of benchmark in this area.

[3] At each processing step a small portion of the signal equal to the window size is processed. Afterwards the window is moved by a certain number of samples, which is defined by the frame shift.

TABLE 1
Overview of pre-processing steps and feature extraction methods applied in our experiments.

modality	channels	pre-processing	short-term feature	long-term feature	total
voice	mono audio, 16kHz	pre-emphasis filter	pitch, energy, MFCCs, spectral, voice quality	mean, median, maximum, minimum, variance, median, lower/upper quartile, absolute/quartile range	1316
face	RGB video, 720x576, 25fps	conversion to gray image	bounding box of face, position of eyes, mouth and nose, opening of mouth, facial expression happy/angry/sad/surprised	mean, energy, standard deviation, minimum, maximum, range, position minimum/maximum, number crossings/peaks, length	264
gesture	3 axes acceleration, 50fps	remove trend	acceleration and 1st derivative, velocity, position	power, fluidity, volume, mean, minimum, maximum, position minimum/maximum, length	75

4.2.2 Facial Features

As mentioned before we use only video recordings of the subject's face. The according camera was placed in a distance of about two meters and captured frames include a close-up of shoulder and head. The resolution of the video is 720x576 pixels at 25 fps. Videos are stored in uncompressed 24-bit RGB format. Video processing is provided by SHORE, a library for facial emotion detection developed by Fraunhofer IIS [4][20]. In the first place, SHORE offers a robust tracking of in-plane rotated faces up to sixty degrees. For each face that is found, SHORE reports the bounding box of the face, as well as position of the left/right eye and the nose tip. These features measure head movement. In addition the left and right corner of the mouth and its degree of opening is reported. Most important, SHORE also calculates scores for four facial expressions – namely happy, angry, sad and surprised. These scores are also extracted for each frame and used in addition to the geometric features. In total, for each segment a series of 24 short-term features is derived by joining features extracted for each frame in the clip. Finally, we extract 11 long-term measurements, leading to an overall feature set with 264 entries.

4.2.3 Gestural Features

The acceleration sensors included in the Wii™ remote control include a three-axis linear accelerometer, which measures the force exerted by a set of small proof masses inside of it with respect to its enclosure. To categorize gestural style we can rely on expressivity parameters, e.g. how fast a gesture is done, how much space one uses to perform a gesture etc. In a previous study we tested the feasibility of this approach on a training set containing 1260 samples by 7 subjects [26]. Using a ten-fold cross-validation of a Nearest Neighbor classifier, we obtained recognition results of at least 94% for power, speed and spatial extent.

The features we extract from the acceleration signal go back to a set of expressivity parameters originally defined by Hartman et. al for expressive gesture synthesis for embodied Conversational Agents [17]. Caridakis and collegues have applied similar features to measure gesture expressivity in hand tracking from video images [7]. They propose a set of six expressivity parameters, namely overall activation, spatial extent, temporal, fluidity, power/energy and repetitivity. Overall activation, e. g. is considered as the overall quantity of movement, while fluidity differentiates smooth/graceful from sudden/jerky gestures.

Since we do not have direct access to the 2D position in space, but instead measure the second derivation, i. e. acceleration, we apply a couple of changes to their algorithm. First, we eliminate the influence of gravity by removing the linear trend from each of the three acceleration axis. Next, we build the first derivative and use cumulative trapezoidal numerical integration to deduce velocity and position. This is done for each axis separately. Finally we calculate the power from each signal, as well as fluidity from position. We also extract another 7 statistical features from the acceleration signal and add the length of the gesture. Overall, we obtain 75 features from the three acceleration axes.

4.3 Feature Selection

Especially on small corpora a large number of features can lead to a problem known as ”curse of dimensionality”. This term was introduced by Richard Bellman in 1961 as mathematical problem and in machine learning it describes the exponentially rising need for numbers of samples for a sufficient description of a high dimensional feature space. In short this means that the more features are given to a classification model, the more samples are needed to train it. In most cases however, not all of the features add useful information to the classification problem, while some may carry redundant information. This fact is related to the challenge of finding that subset of features, which tweaks the best recognition performance.

[4] http://www.iis.fraunhofer.de/en/bf/bv/ks/gpe/demo/

To select a set of most relevant features in a feasible amount of time we apply a combination of two selection approaches. First, we select the best 150 features according to correlation-based feature subset selection (CFS, [16]). CFS aims at finding a subset of features where the correlation of each feature with the class is maximised, while the correlation of the features among each other is low. This strategy is especially beneficial for the Naive Bayes classifier which performs badly when features are highly correlated since it assumes features to be independent for simplification reasons.

Afterwards, sequential forward selection (SFS) is applied on this subset and stopped after 100 iterations (i. e. after 100 features have been selected). SFS is a simple, but popular selection method. Like other wrapper approaches it uses a classifier to measure the performance gain of different feature subsets. SFS starts from an empty subset and adds at each step the feature that brings the highest performance gain. In order to avoid over-fitting cross-validation should be used to evaluate the feature sets. Finally, the subset, which by then gives the best performance, is selected.

4.4 Recognizing Missing Data

As described earlier, the CALLAS corpus is well suited for exploring the problem of missing data. The gesture modality is partially missing or no movements of hands were executed during a sample. The facial modality is missing at points in time, when SHORE lost track of a recorded person and therefore no meaningful facial features could be extracted. As the recordings of samples start and end concurrently with a mood inducing sentence, the vocal modality is the only source that is always available. Handling of missing data is modelled within the multimodal fusion process. Recognition of missing data has to be executed beforehand. Therefore we keep track of when SHORE looses the bounding box around an observed face – this happening marks recorded data as missing until the face is recognised again. We furthermore introduce a threshold for minimum energy within a signal recorded from the Wii™ controller and whenever energy falls below this mark, we assume that no gesture was performed at all during the recorded phrase.

5 Multi Sensor Data Fusion

The examined CALLAS dataset consists of up to three modalities - video, speech and gesture data. As a special challenge channels are not always accessible, so dedicated ways of fusing all available data channels have to be thought of. In order to discuss preconditions and advantages of decision level fusion in a multi sensor environment, we begin with a description of differences between the possible levels on which fusion can be executed and follow up with precise reviews on possible decision level fusion methods and inherited strategies meant to deal with missing data.

5.1 Feature Level Fusion

Feature level fusion is a common and straightforward way to fuse all recorded observation-channels. All desired features are merged into a single high dimensional feature set. One single classifier is then trained for the task of classification. As the fused data contains a bigger amount of information than single modalities, an increase in classification accuracy can theoretically be expected. In practice these classifiers yield reliable classification results. But this very accessible approach to data fusion comes along with a couple of major problems: First drawback is the eventually occurring 'curse of dimensionality' on small datasets (see Section 4.3). If the available data is not ample, the classification results become non-meaningful. As a second, it has to be mentioned that a growing feature vector may stress computational resources for training and evaluation of the classification model. In some examinations these obstacles may be not of interest due to a fair availability of time and resources, other ones may refuse the feature level approach solely because of these reasons and consider decision level approaches to data fusion instead.

A very crucial shortcoming of the feature fusion approach can be observed in particular on the CALLAS dataset. The single classifier trained on the whole feature set is by default not capable of handling the problem of missing data. Furthermore, the feature level fusion approach does not give the opportunity to employ any strategies like elaborating the fusion strategy for affective tasks. The behaviour of feature level fusion is strictly determined by the selected feature-set and underlying classification model.

5.2 Decision Level Fusion

Contrary to feature level fusion and its reliance on a single classifier that deals with a high dimensional feature vector, decision level fusion focusses on the usage of small classifiers and their combination. Instead, the available feature set is divided into subgroups and the partitions are used to form several small classification models (of course these classifiers can also be generated by sub-sampling training data or the usage of different classification models). The assembly of these classifiers is called an ensemble. Outcomes of these slim classifier models are taken into account for the final decision making process. The term decision level fusion sums up a variety of methods designed in order to merge the decisions of ensemble members into one single ensemble decision.

5.2.1 Ensemble Based System

Classification models used in creating the underlying ensemble for all discussed decision making algorithms stem from multimodal emotion observations using facial, vocal and gesture recordings. Our implemented system forms an ensemble by providing features from each listed channel with a classification model. Neither must they provide perfect performance on some given problem, nor do their outputs need to resemble each other. It is preferable that the chosen classifiers make mistakes, at best on different instances. A base idea of ensemble based systems is to reduce the total error rate of classification by strategically combining the members of the ensemble and their errors. Therefore the single classifiers need to be diverse from one another.

5.2.2 Benefits of Decision Level Fusion

Ensemble based systems and corresponding decision level fusion offer some significant advantages over the use of a single classifier. A few important ones should be listed here.

Training Efficiency: In many applications a vast amount of data is gathered and computational efficiency can greatly suffer from training and evaluation of a single classifier with huge datasets. Partitioning of data, training of independent classifiers with different subsets and combination for a final decision often proves to be more practical, time-saving and yields at least competitive results in most cases[5]. Given the contrary case that too little data is available, various re-sampling techniques can be used to form overlapping sub-samples of the dataset. Each of the resulting sub-sets can be applied for training of classifiers, which then are capable of decision making via combination.

Divide and Conquer: Another classification problem can arise if the underlying dataset and the corresponding feature distribution is too complex for a sole classifier to learn. Classification accuracy seriously suffers if the needed decision boundary cannot be found by the used classification model. This undesirable phenomenon can be counteracted by an appropriate set of classifiers. Using a divide-and-conquer approach, the feature space is divided into several (perhaps overlapping) distributions that are easier to learn. Each of these partitions is then handled by one classifier. Adapted combination of the gained classifiers and their simplified decision boundaries adequately simulates the original, complex boundary.

Field Performance: The training and testing of classification models typically takes place on data gained from some kind of laboratory environment. Statements about generalized classification performance experienced in field testing – whenever previously unknown samples appear – are difficult to estimate. The risk of performing below average in the field is much higher for a single classifier than for an assemblage of classifier models. Some by chance poorly trained classifiers within a set are a much less of a menace than a single classifier performing poorly.

Concerned with systems for multi sensor data fusion and real-time applications, these can be implemented in a way that they resist the breakdown of one or more attached sensors. If the classifiers involved in decision making each represent the observations of an associated sensory device, the absence of a single contribution to the final decision is unlikely to result in a drastic quality fall-off for overall classification accuracy - especially if the sensory malfunction is recognized and the corresponding classifier's (most likely counter-productive) contribution is accordingly rated.

5.2.3 Decision Level Fusion Techniques

Having established an ensemble based system, the diverse decisions of its members have to be merged into a single ensemble decision. For this purpose we can choose from various established fusion strategies. Among the strategies we find all forms of algebraic combinations of the classifiers continuous outputs, ranking methods as well as varying voting schemes and other ways of class label combination. One feature of some discussed methods is the appliance of weights to the ensemble members. Based on prior knowledge - like for example gained by evaluation of training performance - single classifiers can be associated with a certain weight. This way their importance within the ensemble is reflected. Note the immense importance of not taking any knowledge of data to be classified into account for the calculation of mentioned weights. Otherwise unrealistic prior knowledge is hypothesized and regarded experimental results can no longer be rated as significant. We apply described fusion schemes to modalities given by the CALLAS corpus. As our system handles temporarily[6] missing modal-

[5] Several smaller classifiers may not save training time when using a classification model of linear time complexity (e.g. Naive Bayes - $O(n)$), as the training will consume as much time as for an overarching classifier. But as time complexity rises (e.g. Support Vector Machines classification scheme (SVM) - $O(n^2)$ or worse) this behaviour changes in favour of small classifiers, using only a section of the original, high dimensional feature vector or training data.

[6] Temporarily means that we adjust the fusion scheme per input sample (not as in most fusion approaches for the whole corpus), i. e. for each modality and each sample, we decide whether to include the information to the fusion process or not. A the moment this decision is exclusive. If we had some way of estimating the degree of corruption within the sample's modality, we could simply assign proper weights instead.

ities in the fusion step, all described fusion methods are enriched with strategies on how to handle missing data streams.

For the explanation of reviewed algorithms the following annotations are used: The decision of ensemble member t for class n is denoted as $d_{t,n} \in \{0,1\}$, with $t = 1..T$ and $n = 1..N$ and $d_{t,n} = 1$ if class ω_n is chosen, $d_{t,n} = 0$ otherwise. Respectively the support given to each class n (i.e. the calculated probability for the observed sample to belong to single classes) by classifier t is described as $s_{t,n} \in [1..0]$.

Weighted Majority Voting
Majority Voting simply sums up decisions of T classifiers. The ensemble decision for an observed sample x is chosen to be the class ω_n which received the most votes (decisions) v_n. A definite decision is only guaranteed if an odd number of ensemble members handle a two-class problem. In Weighted Majority Voting each vote is associated with the pre-calculated weight of the ensemble member. The ensemble decision for an observed sample x is chosen to be the class ω_n which received the most weighted votes v_n. Ties are not likely to happen this way, which makes the weighted variant more suited for practical application.

$$v_n(x) = \sum_{t=1}^{T} w_t d_{t,n}(x)$$

Handling Missing Data:
Ensemble member t containing training data from modalities not featured in an observed sample is not included in the poll.

Weighted Average
In contrast to Weighted Majority Voting, the Weighted Average strategy applies weights not to class labels, but to continuous outputs of ensemble members. By summing up the weighted support given to each class ω_n, total weighted support μ_n for class n is calculated as:

$$\mu_n(x) = \sum_{t=1}^{T} w_t s_{t,n}(x)$$

The ensemble decision for an observed sample x is chosen to be the class ω_n for which support $\mu_n(x)$ is largest.

Handling Missing Data:
Ensemble member t containing training data from modalities not featured in an observed sample is weighted with a value of zero.

Maximum Rule, Minimum Rule, Median Rule
These strategies choose the maximum, minimum or median support generated by T ensemble members. The ensemble decision for an observed sample x is chosen to be the class ω_n for which support $\mu_n(x)$ is largest.

Handling Missing Data:
Ensemble member t containing training data from modalities not featured in an observed sample gives support s_t of value zero to each of the n classes.

Sum Rule, Mean Rule
The Sum Rule simply sums up the support given to each class ω_n in order to generate total support μ_n for each class. The ensemble decision for an observed sample x is chosen to be the class ω_n for which support $\mu_n(x)$ is largest.
By averaging the support ($\frac{1}{T}$ serves as normalization factor) given to each class ω_n, the Mean Rule calculates total support μ_n for class n as:

$$\mu_n(x) = \frac{1}{T} \sum_{t=1}^{T} s_{t,n}(x)$$

Handling Missing Data:
Ensemble member t containing training data from modalities not featured in an observed sample gives support s_t of value zero to each of the n classes.

Product Rule
By multiplying the support given to each class ω_n, total support μ_n for class n is calculated as:

$$\mu_n(x) = \frac{1}{T} \prod_{t=1}^{T} s_{t,n}(x)$$

Note that this fusion strategy reacts very sensitive to pessimistic ensemble members, as a support of value zero virtually nullifies the chance of a class to become the final decision.

Handling Missing Data:
Ensemble member t containing training data from modalities not featured in an observed sample gives support s_t of value 1 to each of the n classes.

Cascading Specialists
The Cascading Specialists method [22] does not focus on merging outputs from all ensemble members, but on selecting experts for each class and bringing them in a reasonable order. Based on evaluation of training data, experts for every class of the classification problem are chosen. Next, classes are rank ordered, from worst classified class across the ensemble's members to the best one. Given these preparations, classification works as follows: First class in the sequence is chosen and the corresponding specialist is asked to classify the sample. If the output matches the currently observed class, this classification is chosen as ensemble decision. If not, the sample is passed on to the next weaker

class and corresponding expert whilst repeating the strategy. Whenever the case occurs that none of the experts classifies its connected class, the classifier with the best overall performance on the training data is selected as final instance and is asked to label the sample. This strategy aims at a flattening effect among class accuracies that will – at best – improve overall classification performance.

Handling Missing Data:
The concept of choosing experts for certain classes has to be broadened, so that an expert unable to handle the given sample (because of missing data) can be adequately replaced. Instead of selecting one single ensemble member for expert and final classification tasks, ordered lists containing all classifiers - ranked by their qualification for the given task - replace sole classifiers. If in the classification step missing data is detected and the most qualified ensemble member is trained with data of that type, we move down in the prepared list to find the next best classifier that is able to handle the observed sample.

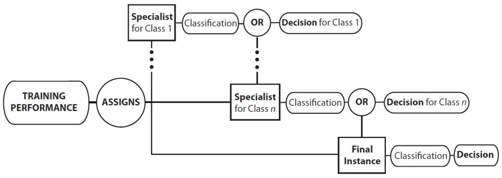

Fig. 3. Cascading Specialists Scheme

Arousal - Valence Combination
The generic approach to classification in a multi-class environment is to train classifiers and corresponding ensembles to categorise among the available classes, but the structure the chosen emotion model – consisting of two scales for valence and arousal – more emotion-adapted techniques for finding an ensemble decision can be applied [19]: Two ensembles are trained to recognise the observed emotion's axial alignment. Resulting outputs are logically combined for final decision, as decisions on valence and arousal orientations explicitly describe one of specified classes. Because of the mapping to orientation in the emotion model, this strategy cannot be generalised for common classification problems.

Handling Missing Data:
In our implementation, Weighted Majority Voting is used to generate the ensembles' decisions on axial alignments. Therefore the handling of missing data described for respective fusion scheme is adopted.

Arousal - Valence - Cross Axis Combination
The concept of a cross axis again is exclusive to the chosen emotion model and is meant to provide supplementary information to the arousal and valence ensembles. This axis can not be directly deduced from emotion theory but from a mathematical point of view it is a reasonable partition of the given 2D space. Just like arousal and valence axes, the cross axis divides the emotion model into two separate parts, each containing two emotion-quadrants. These parts contain the respective, complementary quadrants and therefore split the model in a diagonal way. According to arousal and valence, a proper ensemble for cross axis is constructed. A stepwise algorithm is used for combination of the sources of information[7]:
In *Step 1* each ensemble distributes its votes among the two quadrants that fit the recognised alignments in the emotion model. This step results in one of two possible outcomes:

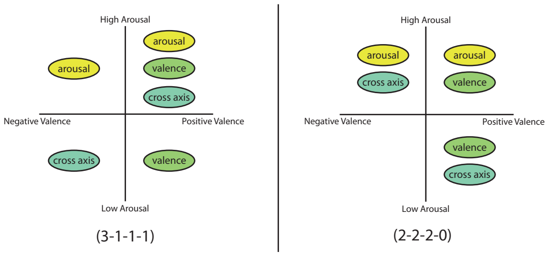

Fig. 4. Possible Vote Distributions after *Step 1*

If the ensembles agree on one emotion-quadrant, it receives three votes and can already be chosen as final decision, otherwise a voting tie occurs. No final decision can be chosen, instead the draw has to be dissolved and the algorithm moves on to *Step 2*. A direct classification ensemble designates exactly one vote to the class it predicts. Two situations can arise through this supplemental vote: If the ensemble chooses an emotion-quadrant that already holds two votes, the tie is resolved and the corresponding emotion is determined to be the final decision. Otherwise the ensemble chooses the emotion-quadrant that has not yet received any votes, the tie is not resolved and *Step 3* to be executed: The emotion-class that was originally determined by arousal and valence ensembles is chosen as final decision. In practice this case rarely occurs, but it is definitely needed to guarantee that no sample passes the decision process unclassified.

Handling Missing Data:
Again, Weighted Majority Voting and respective handling of missing data is used in all ensemble decisions.

6 Experiments

Presented experiments are done on the described CALLAS Expressivity corpus. Missing data is included in the facial and gestural modality, the vocal

[7] An additional ensemble for direct classification – as established for generic approaches – is needed for *Step 2* of the combination strategy

modality is always accessible, as samples represent one spoken sentence. In detail, audio signals are constantly present in all 2513 samples, facial features can successfully be extracted throughout 2251 observations (90%) and significant gestures are available for 569 samples (24%). In consequence of the experimental design the samples are more or less equally distributed among the 21 subjects, however, if we calculate the relative portion of samples per emotion we find a high variety between the users. In general, female users recorded in this experiment seem to perform more expressive than their male counterparts.

6.1 Evaluation Method

Choosing an adequate evaluation method is crucial for meaningful experiments. Among others, possibilities involve random drawing of samples for testing, percentaged subset drawing, k-fold splits or the leave-one-out strategy. As this work focusses strongly on the practical adaptability of presented methods and a good field performance under lifelike or even disadvantageous circumstances, the chosen evaluation method should reflect this intention. We agreed on a very realistic, user independent approach for evaluation of our experiments (Leave-One-Speaker-Out). As the employed corpus contains ten female and eleven male participants we consecutively draw samples belonging to one single subject out of the set. Remaining samples are used for training of classification models which then are tested against the isolated samples. Another important decision concerns the way recognition rates are presented. In consequence of the dominant presence of *positive-low* and a general imbalance in class-distribution, we base our studies on the class-wise recognition rate (sometimes referred to as unweighted average recall), which is the mean of the recognition rates observed for each class.

6.2 Discussion

Results shown in Table 2 are split into three parts: First, single channel performance is shown for each modality. Whenever missing data is found, the respective sample is not included for evaluation, so these results stem from different quantities of samples. Vocal modality clearly outperforms facial and gestural cues and establishes most balanced accuracies among observed classes, though positive emotions are recognised better than negative ones. The facial modality recognizes the *positive-high* emotion very well – presumably because it is well suited for detection of smiles and movements of the face associated with laughter – but lacks on other classes. Gestures are most often correctly classified during *positive-low* phases, the most calmly expressed class of observed emotional states with nearly no movement at all. *Negative-low* emotions were often expressed with despaired gestures, that could partially be separated from gestures with high arousal. These expressive movements were obviously often misinterpreted among each other, leading to very low accuracies on classes with highly aroused emotions.

TABLE 2
Results achieved for single modalities in comparison with decision level fusion and emotion adapted fusion.

	Recognition Results				
	p-low	p-high	n-low	n-high	**avg**
	Single Modalities				
Voice	0.61	0.50	0.49	0.43	**0.51**
Face	0.45	0.72	0.31	0.43	**0.48**
Gesture	0.57	0.30	0.44	0.36	**0.42**
	Generic Decision Level Fusion				
Cascading	0.49	0.63	0.47	0.40	**0.50**
MaxRule	0.55	0.65	0.36	0.46	**0.50**
MeanRule	0.57	0.66	0.41	0.37	**0.50**
MedianRule	0.58	0.67	0.41	0.38	**0.51**
MinRule	0.59	0.58	0.48	0.33	**0.49**
ProdRule	0.58	0.66	0.42	0.38	**0.51**
SumRule	0.57	0.66	0.41	0.37	**0.50**
WeightAvg	0.53	0.63	0.39	0.41	**0.49**
WeightMajVote	0.61	0.52	0.49	0.42	**0.51**
	Emotion Adapted Decision Level Fusion				
ArousalValence	0.56	0.46	0.55	0.51	**0.52**
CrossAxis	0.64	0.55	0.56	0.44	**0.55**

Second part of Table 2 shows results of generic fusion approaches that can be applied to any given classification problem. Theoretically do all decision level fusion strategies aim at exploiting mentioned differences in single channels in order to enhance combined performance. Practically, performances of facial and gestural modalities are too inferior to the audio channel to result in greater gains in overall recognition rates. If one compares the vocal modality to the fusion schemes, most approaches perform better on the *positive-high* class. This behaviour can be explained by the very good result of the facial modality in this area and the resulting influence on the ensemble. Unfortunately do bad results for remaining classes effect overall performance in a contrary way and these gains are lost again in other categories. This can be well observed when looking at the Product and Sum Rule – the ”standard” fusion schemes for merging classifier outputs – as recognition results stabilize around vocal performance with a better trend on the second class. Same estimations hold for other merging strategies like Mean or Weighted Average Rule. Behaviour of approaches that choose exactly one support value among ensemble members for each class (Max, Median and Min Rule) is harder to predict, but all of them resemble the just mentioned characteristics and their overall performances range from worst to best results within the generic fusion category. Weighted Majority Voting's inherent weighting method causes a strong reliance on the dominant modality, resulting

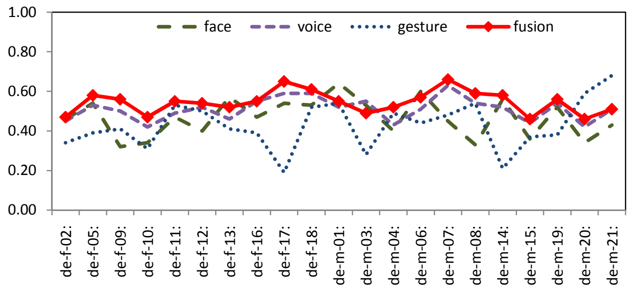

Fig. 5. Recognition performance per user of single-channel classification and the fusion approach (user independent).

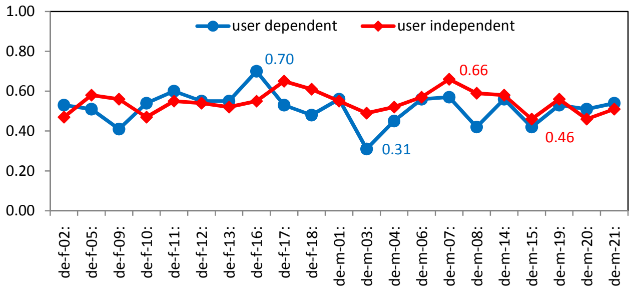

Fig. 6. Recognition rates for single-user dependent classification (blue line) and for user independent classification (red line).

in almost the same accuracies across all classes as the audio channel. The Cascading Specialists strategy generates acceptable results on negative classes – that are all in all more weakly categorised throughout the ensemble – but loses too much accuracy on the *positive-low* class in order to improve average accuracy. However, no strategy does generate drastically worse results than the best modality – actually they perform well compared to remaining modalities.

So in order to perceptibly enhance recognition rates compared to single channel classification on the dominant modality, we have to exploit deeper knowledge about the classification problem at hand with emotion adapted fusion strategies that employ more than a single generic ensemble. The combination of arousal and valence ensembles shows different characteristics than the generic approaches: The dominance of the *positive-high* class is gone and negative classes are well recognised. Overall these changes result in slightly superior accuracy than the best single channel. For further improvements we incorporate more available information from the 2D emotion model into combination strategies, leading to the additional cross axis ensemble. This fusion scheme exceeds the best modality on every observed class and therefore enhances average accuracy remarkably, however at the expense of a rising ensemble count.

Of course, observed results cannot be claimed to be universal, as they are highly corpus depended. Unfortunately, by the time this article was written there was – at least to our best knowledge – no publicly available emotion corpus, which could suite our needs in terms of modalities, naturalness and presence of missing data. Such a corpus could serve as a benchmark to compare presented methods with results achieved by other institutions and to prove generality. However, observations are in line with an earlier study based on a smaller emotional corpus, where emotion specific fusion approaches clearly outperformed generic ones [19].

6.3 Single-Channel Classification

To learn more about the contribution of the single modalities to the overall performance Figure 5 shows class-wise recognition rates for each channel and user (dashed lines). User names are encoded in the form `de-[f,m]-id`, where `f` denotes a female and `m` a male user. The line charts reveal high variability in all three modalities. The most extreme difference occurs in the gesture channel ranging from 19% (de-f-17) to 68% (de-m-21). For audio and video differences are slightly more stable, but still within a range of 21% and 31%, respectively. The results give no hints for a general preference of female and male users for a certain channel. In fact, it appears to be rather randomly distributed which are the strong channels of the single users. There are also no indications for a correlation between the channels, i. e. a low/high performance in one channel does not necessarily indicate a low/high performance in the other channels.

Figure 5 additionally reveals that the fusion approach (red line) improves the results achieved by the best single channel for 15 out of 21 subjects. For the remaining subjects (de-f-13, de-m-01, de-m-03, de-m-06, de-m-20, de-m-21) decision level fusion evens results, while it still outperforms the other two channels. Nevertheless, as fusion strategies consider all modalities, there is always a chance that a failure in one channel affects good performance of other ones. This is the trade-off for balancing weak modalities and corrupted data. After all, ensemble based strategies always guarantee results that are - if not superior - at least in line with the dominating modality. This characteristic is especially desirable, when the most trustworthy modality for a user is not known in advance.

6.4 Single-User Dependent Classification

Reasons for relying on user independent evaluation were already explained in 6.1. It may be still worthwhile to draw a comparison to single-user dependent classification. We surely would expect better classification performance from a system trained and tested with samples of the same user. In order to prove this assumption we select all samples of each user and investigate recognition rates for each user independently by 5-fold cross evaluation. Results are visualized in Figure 6.

Some remarkable facts have to be mentioned: As a first, one can observe a high variability in overall performance ranging from 31% to 70% in single-user dependent classification, whereas results in user independent classification are a bit more stable ranging from 46% to 66%. Obviously results of single-user dependent classification are not necessarily superior to the user independent approach chosen for our evaluations. In fact, for a noticeable amount of users (e. g. de-f-09 and de-m-03) overall classification rates prove to be significantly lower[8]. Finally we observe a common trend in user independent and single-user dependent evaluation, i. e. high performance in the single-user dependent classification implies high performance in the user independent case and the other way round.

Made observations suggest that success of single-user dependent as well as user independent emotion recognition always strongly correlates with the expressivity of investigated subjects. Unfortunately we have to reckon with a high variety in expressivity among users. Further, results suggest that a system trained on a considerable large database of pre-recorded users can yield similar results as a system trained on a limited set of personalized samples collected from the user.

7 Conclusion

In this work we report results of decision level fusion experiments carried out on the CALLAS Expressivity corpus, performing emotion recognition from three different modalities, namely voice, face and gesture. Classification accuracies of single modalities range from 42% to 51% while appropriately recognising and dealing with missing data in observed channels. By means of ensemble techniques results were raised up to 55%, including various generic fusion schemes as well as emotion adapted approaches like the combination of arousal, valence and cross axis. We see that generic approaches adapt to the most dominant modality in the ensemble while adopting some characteristics of other ensemble members – like the good performance on *positive-high* from the video channel. Exploiting the structure of the underlying emotion model leads to more elaborate fusion strategies – including the usage of mor than one ensemble – that are tailored for affective emotion recognition. These fusion schemes do not share above mentioned characteristics and are able to outperform single modalities and generic approaches by a significant rate, though bearing a higher complexity due to the generation of specialised ensembles.

Comparing recognition results of the single subjects we found a high variability within a range of more than 30%. This correlates with a high variety in expressivity among users. This variability also appears when observing results for the single channels. In fact, we could not identify a general trend that would suggest one channel to be more important than another. Though the vocal modality looks more stable overall, the facial channel is more suited for certain users. Even the gesture channel as the weakest of the three channels outperformed the others for some users. The uncertainty about the channel a user picks to express his or her emotion, however, is a strong argument for the benefit of fusion in our experiment as it reduces the risk to trust a weak modality.

As another interesting outcome we found that the results obtained for user independent classification also yielded similar and sometimes even better results than single-user dependent classification. Hence, an emotion recognition system trained with data of a large number of subjects must not necessarily achieve worse results than a personalized system trained with a small number of training samples from the user. Ideally the system would start with a pre-trained model, which is over time adapted to the user. However, we also observed that a low performance in the single-user dependent classification implies low performance in the user independent case, i. e. the success of an emotion recognition system always depends to a large part on the expressivity of the user.

For future work we also aim at improving the segmentation techniques applied to the CALLAS corpus. So far, recorded modalities have been segmented in a very straightforward way. Based on the user's speech, beginning and ending of a recorded sample coincide with the boundaries of the spoken stimuli sentence. Facial and gestural signals are simply observed and segmented over the according time span. This strategy suffers from two major problems: The hopefully expressed emotion could occur within a much shorter period somewhere within the spoken sentence and therefore information before and afterwards is not of great meaning for recognition. Furthermore significant hints from different modalities are not guaranteed to emerge at exactly the same time interval. Classification accuracy could be expected to improve, if modalities were segmented individually and the succession and corresponding delays between occurrences of emotional hints in different signals could be investigated more closely. However this approach gives room for a whole new set of hypotheses and experiments.

Acknowledgments

The work described in this paper is funded by the EU under research grant CALLAS (IST-34800) and CEEDS (FP7-ICT-2009-5).

[8] Note that the amount of training samples available for user independent classification is about twenty times higher compared to single-user dependent evaluation.

REFERENCES

[1] N. Ambady, R. Rosenthal, *Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis*, Psychological Bulletin 111 (2), pp. 256274, 1992.

[2] T. Balomenos, A. Raouzaiou, S. Ioannou, S. Drosopoulos, A. Karpouzis, S. Kollias, *Emotion analysis in manmachine interaction systems*, Workshop on Machine Learning for Multimodal Interaction, 2005.

[3] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, V. Aharonson, *Combining efforts for improving automatic classification of emotional user states*, Language Technologies, IS-LTC: Infornacijska Druzba (Information Society), pp. 240-245, 2006.

[4] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, *A Database of German Emotional Speech*, in Proceedings of INTERSPEECH, Lissabon, pp. 1517-1520, 2005.

[5] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan, *Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information*, in Proceedings of International Conference on Multimodal Interfaces ICMI, pp. 205-211, 2004.

[6] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Paouzaiou, K. Karpouzis, *Modeling Naturalistic Affective States via Facial and Vocal Expression Recognition*, in Proceedings of International Conference on Multimodal Interfaces ICMI, pp. 146-154, 2006.

[7] G. Caridakis, A. Raouzaiou, K. Karpouzis, S. Kollias, *Synthesizing Gesture Expressivity Based on Real Sequences*, Workshop on multimodal corpora, LREC, 2006.

[8] G. Caridakis, J. Wagner, A. Raouzaiou, Z. Curto, E. André, K. Karpouzis, *A multimodal corpus for gesture expressivity analysis Multimodal Corpora*, Advances in Capturing, Coding and Analyzing Multimodality, LREC, 2010.

[9] G. Castellano, L. Kessous, G. Caridakis, *Multimodal emotion recognition from expressive faces, body gestures and speech*, Affect and Emotion in Human-Computer Interaction, Lecture Notes in Computer Science, Springer, 2007

[10] L.S. Chen, T.S. Huang, T. Miyasato, R. Nakatsum, *Multimodal human emotion/expression recognition*, in Proceedings of International Conference on Automatic Face and Gesture Recognition FG, p. 366, 1998.

[11] C. Demiroglu, D.V. Anderson, M.A. Clements, *A Missing Data-based Feature Fusion Strategy for Noise-Robust Automatic Speech Recognition Using Noisy Sensors*, International Symposium on Circuits and Systems ISCAS, pp. 965-968, 2007.

[12] R. El Kaliouby, P. Robinson, *Generalization of a Vision-Based Computational Model of Mind-Reading*, In Proceedings of International Conference on Affective Computing and Intelligent Interfaces ACII, pp 582-589, 2005.

[13] J. L. Fleiss, B. Levin, M.C. Paik, *Statistical Methods for Rates and Proportions*, (3rd ed.), New York: John Wiley & Sons, 2003.

[14] F. Fragopanagos, J.G. Taylor, *Emotion recognition in human-computer interaction*, Neural Networks, 18(4): pp. 389-405, 2005.

[15] H. Gunes, M. Piccardi, T. Jan, *Face and body gesture recognition for a vision-based multimodal analyzer*, Pan-Sydney Area Workshop on Visual Information Processing 36, 2004.

[16] M. A. Hall, *Correlation-based feature subset selection for machine learning*, Masters thesis, University of Waikato, Hamilton, New Zealand, April 1998.

[17] B. Hartmann, M. Mancini, C. Pelachaud, *Implementing expressive gesture synthesis for embodied conversational agents*, Gesture Workshop, LNAI, Springer, 2005.

[18] L. Huang, L. Xin, L. Zhao, J. Tao, *Combining Audio and Video by Dominance in Bimodal Emotion Recognition* in Affective Computing and Intelligent Interaction ACII, pp. 729-730, 2007.

[19] J. Kim, F. Lingenfelser, *Ensemble Approaches To Parametric Decision Fusion For Bimodal Emotion Recognition*, in Proceedings of International Conference on Bio-inspired Systems and Signal Processing (Biosignals), pp. 460-463, 2010.

[20] C. Küblbeck, A. Ernst, *Face detection and tracking in video sequences using the modified census transformation*, Journal on Image and Vision Computing, vol. 24, issue 6, pp. 564-572, 2006.

[21] P. J. Lang, M. M. Bradley, B. N. Cuthbert, *Motivated attention: Affect, activation, and action.*, in Attention and orienting: Sensory and motivational processes, P. J. Lang, R. F. Simons, M. T. Balaban Eds. Mahwah, NJ: Erlbaum, pp. 97-135, 1997

[22] F. Lingenfelser, J. Wagner, T. Vogt, J. Kim, E. André, *Age and Gender Classification from Speech using Decision Level Fusion and Ensemble Based Techniques*, in Proceedings of INTERSPEECH, 2010.

[23] M. Pantic, L. Rothkrantz, *Toward an affect-sensitive multimodal human-computer interaction*, in Proceedings of the IEEE 91(9), pp. 1370-1390, 2003.

[24] R. Polikar, *Ensemble based systems in decision making*, IEEE Circuits and Systems Magazine, no. 3, pp. 21-45, 2006.

[25] B. Schuller, S. Steidl, A. Batliner, *The INTERSPEECH 2009 Emotion Challenge*, in Proceedings of INTERSPEECH, pp. 312-315, 2009.

[26] M. Rehm, N. Bee, E. Andé, *Wave Like an Egyptian Accelerometer Based Gesture Recognition for Culture Specific Interactions*, in Proceedings of HCI 2008 Culture, Creativity, Interaction, 2008.

[27] N. Sebe, I. Cohen, T. Gevers, T.S. Huang, *Multimodal approaches for emotion recognition: a survey*, in Proceedings of the SPIE, Volume 5670, pp. 56-67, 2004.

[28] L.C. De Silva, P.C. Ng, *Bimodal emotion recognition*, in Proceedings of International Conference on Automatic Face and Gesture Recognition FG, pp. 332-335, 2000.

[29] M. Song, J. Bu, C. Chen, N. Li, *Audio-visual based emotion recognition A new approach*, in Proceedings of International Conference on Computer Vision and Pattern Recognition CVPR, pp. 1020-1025, 2004.

[30] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, Logos Verlag, 2009.

[31] E. Velten, *A laboratory task for induction of mood states*, Behaviour Research and Therapy, 35:72-82, 1998.

[32] T. Vogt, E. André, *Exploring the benefits of discretization of acoustic features for speech emotion recognition*, in Proceedings of INTERSPEECH, pp. 328-331, 2009.

[33] J. Wagner, E. André, F. Jung, *Smart sensor integration: A framework for multimodal emotion recognition in real-time*, in Affective Computing and Intelligent Interaction ACII, 2009.

[34] Y. Yoshitomi, S. Kim, T. Kawano, T. Kitazoe, *Effect of sensor fusion for recognition of emotional states using voice, face image, and thermal image of face*, International Workshop on RobotHuman Interaction, 2000.

[35] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T.S. Huang, S. Levinson, *Audio-visual Emotion Recognition through Multi-stream Fused HMM for HCI Applications*, in Proceedings of International Conference on Computer Vision and Pattern Recognition CVPR, pp. 967-972, 2005.

Johannes Wagner graduated as a Master of Science in Informatics and Multimedia from the University of Augsburg, Germany, in 2007. Afterwards he joined the chair for Human Centered Multimedia of the same University. Among other projects, he has been been working on multimodal signal processing in the framework of CALLAS and is currently developing a general framework for the integration of multiple sensors into multimedia applications called Social Signal Interpretation (SSI).

Florian Lingenfelser received his M.Sc. degree in Informatics and Multimedia from the University of Augsburg, Germany, in 2009. In 2010 he joined the chair for Human Centered Multimedia of the same University as PhD student. He is currently contributing to multimodal data fusion within the CEEDS project and developing the Social Signal Interpretation framework.

Elisabeth André is full professor of Computer Science at Augsburg University and Chair of the Laboratory for Human-Centered Multimedia. Prior to that, she worked as a principal researcher at DFKI GmbH where she has been leading various academic and industrial projects in the area of intelligent user interfaces. In summer 2007 Elisabeth André was nominated Fellow of the Alcatel-Lucent Foundation for Communications Research. In 2010, she was elected a member of the prestigious German Academy of Sciences Leopoldina and the Academy of Europe. Her research interests include affective computing, intelligent multimedia interfaces, and embodied agents.

Jonghwa Kim received the BS and MS degree in electronic engineering from the Kyungwon University, Korea, in 1992 and 1994 respectively. He completed the diploma in sound engineering and received the Ph.D. degree in communication engineering from the Technical University of Berlin, Germany, in 1999 and 2003, respectively. In 2002, he joined the Institute of Computer Science at the University of Augsburg, Germany, and finished in 2010 his Habilitation (*Venia Legendi*) for the Computer Science.