# A Systematic Discussion of Fusion Techniques for Multi-Modal Affect Recognition Tasks

Florian Lingenfelser
Lab for Human Centered Multimedia
University of Augsburg
Universitätsstr. 6
Augsburg, Germany
lingenfelser@hcm-lab.de

Johannes Wagner
Lab for Human Centered Multimedia
University of Augsburg
Universitätsstr. 6
Augsburg, Germany
wagner@hcm-lab.de

Elisabeth André
Lab for Human Centered Multimedia
University of Augsburg
Universitätsstr. 6
Augsburg, Germany
andre@hcm-lab.de

## ABSTRACT

Recently, automatic emotion recognition has been established as a major research topic in the area of human computer interaction (HCI). Since humans express emotions through various channels, a user's emotional state can naturally be perceived by combining emotional cues derived from all available modalities. Yet most effort has been put into single-channel emotion recognition, while only a few studies with focus on the fusion of multiple channels have been published. Even though most of these studies apply rather simple fusion strategies – such as the sum or product rule – some of the reported results show promising improvements compared to the single channels. Such results encourage investigations if there is further potential for enhancement if more sophisticated methods are incorporated. Therefore we apply a wide variety of possible fusion techniques such as feature fusion, decision level combination rules, meta-classification or hybrid-fusion. We carry out a systematic comparison of a total of 16 fusion methods on different corpora and compare results using a novel visualization technique. We find that multi-modal fusion is in almost any case at least on par with single channel classification, though homogeneous results within corpora point to interchangeability between concrete fusion schemes.

## Categories and Subject Descriptors

I.5.1 [**Computing Methodologies**]: Pattern Recognition

## General Terms

Algorithms

## Keywords

Emotion Recognition, Multi-Modal Fusion

*ICMI'11,* November 14–18, 2011, Alicante, Spain.

## 1. INTRODUCTION

In human interaction, social signals are generally expressed through multiple available modalities. Emotions in particular are illustrated by a combination of vocal behaviour, facial expressions, gestures and postures. One main goal of human computer interaction (HCI) is to make information about a users emotional state available to a machine via automatic emotion recognition and classification. A generic approach to this problem is to choose one type of signal, train the computer to extract and recognize preassigned features and cues from it, and finally associating made observations with predefined emotional classes. But as humans tend to base and refine their predictions on emotional states on more than one modality, a machine should do so too if possible.

This means fusing multi-modal observations at some point of the prediction-process. Generally said this effort can be done at different levels, mainly the feature level by merging cues from all modalities into one classification scheme, or at decision level by combining outputs of several classifiers (one can think of other levels of fusion and we will present some later on). If we however consider the great amount of meanwhile established and further possible ensemble based strategies, the question arises if there exist generally advisable ones or if the success of a strategy is based on the observed problem. The *No Free Lunch Theorem* [15] has proven for supervised machine learning that there is no universally applicable classification scheme for all given classification tasks. When observing all possible problems, solutions perform on an equal level on average. Studies like [3, 9, 8, 4] examine rather basic fusion strategies and sometimes try to give advise on which scheme dominates others. Results are not consistent throughout mentioned experiments, so suspicion that the *No Free Lunch Theorem* holds for combination rules as well as for the underlying classification methods seems reasonable.

In the field of emotion recognition fusion has been mainly applied to audio-visual data. Authors in [16] cite 18 studies dealing with audio-visual fusion. The authors distinguish between feature-, decision- and meta-level fusion, where the latter describes approaches, which use a 2nd-level classifier to combine predictions of the single channels. While none of the mentioned studies uses methods of all three kinds, it is also difficult to compare the results between the studies as they differ greatly in their methodology, as well as, the underlying databases. We will try to enrich the ongoing discussion with a comprehensive comparison of various established and novel fusion strategies, ranging from feature fusion and elaborated decision level combination rules to meta-level and hybrid-fusion. These will be applied to different corpora for emotion recognition in order to directly compare relative recognition-success on different classification-problems. Thus, on the one hand we investigate the general potential of multi-modal fusion compared to single channel emotion recognition and on the other hand give clear hints on benefits of certain fusion schemes or even their interchangeability for future studies. To our best knowledge this is the most comprehensive empirical fusion study applied on audio-visual emotion recognition.

## 2. APPLIED FUSION TECHNIQUES

When confronted with multi-modal fusion for audio-visual emotion recognition, a vast amount of eligible fusion strategies come into consideration. In the following, possible fusion techniques for combining available modalities are presented and discussed in detail. For the sake of clear arrangement, possible methods can be differentiated by the levels on which they are executed. Fusion at decision level can be further sub-divided into class-label combination strategies, algebraic combination rules and specialist selection methods.

### 2.1 Feature Level Fusion

A very straightforward way to fuse all observed channels is to merge all calculated features into a single and high dimensional feature set. One classifier is then trained for the task of classification. The accumulated features contain a bigger amount of information than a single modality. Thus, increased classification accuracy can theoretically be expected. The eventually occurring *Curse of Dimensionality* has to be accounted for on small datasets. If the available observations are not proportional to the amount of features covered by a sole classifier, the classification results become non-meaningful. As a second, it has to be mentioned that a growing feature vector may stress computational resources. However, appliance of feature selection techniques may relieve both problems.

### 2.2 Decision Level Fusion

Decision level fusion sums up combination rules[1] for the outputs of several classification models. Instead of using all available features for a sole classifier, the available feature set is divided into subgroups (e.g. one classifier per modality) and the partitions are used to form classifiers. The assembly of these classifiers is called an ensemble. The outcomes of these slim classifier models are taken into account for the final decision making process.

#### 2.2.1 Class-Label Combination

Voting could be considered the most generic approach to decision level fusion, because it simply combines class labels gained from $T$ classifiers by summing up decisions. The ensemble decision for an observed sample $x$ is chosen to be the class $\omega_n$ 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 (thus it is not capable of producing definite decisions in many practical applications and is therefore often replaced by the weighted variant).

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

In Weighted Majority Voting each vote is associated with a pre-calculated weight (in our case weights are determined by evaluations of classifiers on training data) of the ensemble member. Ties are not likely to happen this way, which makes the weighted variant more suited for most classification problems.

Another way of combining the class labels generated by ensemble members is to construct a lookup-table. This method is introduced by [5] as Behavior Knowledge Space (BKS). During training the table counts combinations of labelling outputs together with the true class and occurrences of this composition. Test samples then are compared to that table and the true class for which the currently observed labelling combination was recorded most often gets chosen as ensemble decision.

#### 2.2.2 Algebraic Combination Rules

Algebraic combiners for continuous outputs mathematically compute the ensemble decision from probabilities for each class over all classifiers. The Maximum Rule and Minimum Rule respectively choose the maximum or minimum support generated by $T$ ensemble members. The ensemble decision for an observed sample $x$ is chosen to be the class $\omega_n$ for which support $\mu_n(x)$ is largest.

$$\mu_n(x) = max, min_{t=1..T}\{s_{t,n}(x)\}$$

The Sum Rule simply sums up the support given to each class $\omega_n$ in order to generate total support $\mu_n$ for each class. By averaging the support ($\frac{1}{T}$ serves as normalization factor) given to each class $\omega_n$ we obtain the Mean Rule. When additionally adding classifier weights $w_t$, the Weighted Average method calculates total support $\mu_n$ for class $n$ as:

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

By multiplying the support given to each class $\omega_n$, the Product Rule determines total support $\mu_n$ for class $n$ as:

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

The following two combination rules make more extensive use of continuous outputs of ensemble classifiers. Given sample $x$, the decision profile $DP(x)$ for $T$ ensemble members contains the probability distributions among $N$ classes:

$$DP(x) = \begin{array}{c|c|c} s_{1,1}(x) & ... & s_{1,N}(x) \\ \hline ... & ... & ... \\ \hline s_{T,1}(x) & ... & s_{T,N}(x) \end{array}$$

Decision template $DT_n$ can then be defined for each class $\omega_n$ as respective decision profile during training, averaged by the cardinality of observed class. Given an unlabelled test-sample $x$, we first construct $DP(x)$ from ensemble members and then calculate similarity (as Squared Euclidean distance) $S$ between $DP(x)$ and the decision template $DT_n$ for

[1] 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 $\omega_n$ is chosen, $d_{t,n} = 0$ otherwise. 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 $s_{t,n} \in [1,0]$.

each class $\omega_n$. Finally the most similar class is chosen as ensemble decision.

Further utilisation of decision templates is based on on the Dempster-Shafer theory of evidence [12]. It can be applied to decision making by interpreting the classifiers outputs as a measure of evidence. Instead of similarities, proximities and resulting beliefs (evidence) is calculated. This represents the belief in one classifier correctly classifying observed instance into respective classes. Following Dempsters rule of combination, these beliefs can be multiplied throughout the ensemble in order to obtain the final decision.

### 2.2.3 Specialist Selection

In contrary to the fusion schemes described so far, the Cascading Specialists [6] method does not focus on merging outputs from all ensemble members, but on selecting specialists for each class and bringing them in a reasonable order. In a preparation step, experts for every class of the classification problem (based on evaluation of training data) are chosen. Next, classes are rank ordered, from worst classified class across all classifiers to the best one. The algorithm for classification works as follows: First class in the sequence is chosen and the corresponding expert 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.

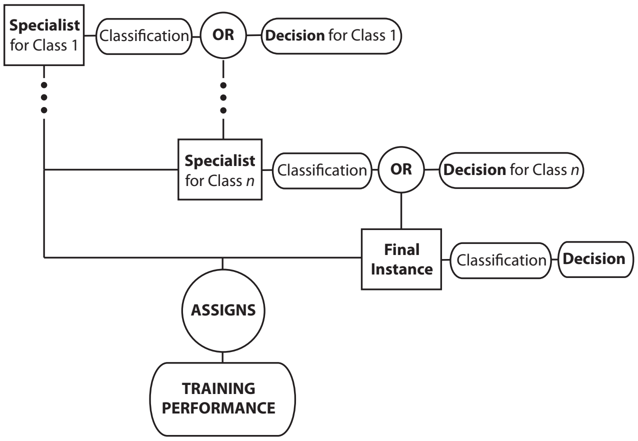

**Figure 1: Cascading Specialists Scheme**

## 2.3 Meta Level Fusion

In meta level fusion, the outputs of several ensemble classifiers are not fused by predefined combination rules. Instead their results are used as input for one or more meta classification models, that generate the final ensemble decision. This process is lent from meta-classification and conforming to notations used by Wolpert and Macready, ensemble classifiers correspond to level-0 base classifiers, the meta classifiers fusing their results equate to level-1 meta generalisers.

In Stacked Generalisation – as proposed by [13] – a level-1 classifier tries to learn the probability distribution among level-0 ensemble classifiers together with the true class that lead to this combination. When asked to classify an unknown sample $x$, the method first collects probability estimates of all ensemble members that consecutively form the basis for the level-1 classifier's final prediction.

Another approach to meta-classification is Grading [11], where the goal of level-1 classifiers is to correct potentially false decisions of level-0 ensemble members. During training every base classifier is complemented by a meta classifier with same training data but a graded label – a boolean value stating correct or incorrect prediction of the ensemble classifier. At classification time our implementation fuses ensemble predictions as every member adds the probability of correctness (generated by it's grading classifier) to the final support of the class it predicted. As usual, the class with highest support is chosen as final ensemble decision.

## 2.4 Hybrid Fusion

In this study we use the term hybrid fusion to characterise fusion techniques that incorporate classifiers with merged features into used ensembles and therefore combine decision and feature level fusion. Of course this approach is applicable to most ensemble combination rules discussed so far, but we decided to develop a refined fusion scheme with two variants in order to explore the capabilities of hybrid fusion.

The One Versus Rest approach trains $N$ classifiers on every available feature-set (excluding merged features), each specialised in recognising one of $N$ classes. This breakdown on several two-class classification problems is done by relabelling. Additionally the ensemble is completed by one multi-class classification model trained on the merged feature set.

Given test-sample $x$, variant one multiplies probabilities gained from classifiers trained on recognising class $n$ with the associated probability generated by the multi-class classification model. This is done for classes $1..N$ and the class with the highest accumulated support gets chosen as ensemble decision. Variant two chooses among the two-class classification models the most promising one for every class. The specialist's probability is then summed with the respective probability from the multi-class classification model.

# 3. AFFECTIVE CORPORA

To draw a comparison of the presented fusion techniques we use two different corpora - the DaFEx and CALLAS corpus. These corpora have been chosen as they both contain audio-visual recordings of Italians. They differ, however, in the number of expressed emotional states and their level of naturalness. In the past most studies dealing with the recognition of emotions were based on recordings from professional actors. Lately, we can observe an increasing trend towards more natural data sets. It turned out that findings derived from acted data must not necessarily be transferred to spontaneous emotions [14]. Using corpora of both kinds allows us to investigate to what extent the choice of the fusion technique depends also on the naturalness of the observed data.

The DaFEx corpus [1] contains recordings of 8 professional Italian actors (4 male and 4 female) expressing 6 basic emotions and neutral. It was initially constructed as a benchmark for the evaluation of facial expressivity of Embodied

Conversional Agents, but is well suited for the evaluation of emotion recognition systems, too (see e. g. [10]). It consists of 1008 short videos clips, where each clip corresponds to one the basic emotions happiness, surprise, fear, sadness, anger and disgust, or neutral. The facial expressions are available at 3 intensity levels (low, medium and high), but for our purpose are combined to a single class. Finally, we select only those samples, where the actors were also uttering a sentence, resulting in 84 samples per subject equally distributed among the 7 classes.

The CALLAS expressivity corpus [2] was constructed within the European Integrated Project CALLAS. Designed for examination of cultural differences it was actually recorded in three countries, Germany, Italy and Greece. In contrast to the DaFEx corpus participants in the CALLAS corpus have no special acting abilities and were asked to perform expressions in the three broad categories positive, neutral and negative. For this study we only consider the Italian subcorpus consisting of 1539 samples of 13 persons (7 female and 6 male) equally distributed among the three classes. Specific emotions were elicited by a mood induction technique. For this study samples were consequently labelled corresponding to the state of the stimuli sentences.

# 4. METHODOLOGY OF EXPERIMENTS

Extraction of descriptive features is a necessary step to convert the raw signals into the compact form required for classification. From the audio channel we extract acoustic features related to the paralinguistic message of speech, i. e. "how" something is said. MFCCs and spectral features as well as prosodic features from pitch, energy, duration, voicing and voice quality total to the amount of 1316 features calculated by EmoVoice[2]. For video analysis we use SHORE, a library for facial emotion detection developed by Fraunhofer IIS[3] [7]. For each face detected, SHORE is able to extract a set of features including amongst others the position of the face, the eyes, nose and mouth, as well as information whether the eyes or the mouth are open or closed. Collected for each frame, these values build a series of 24 short-term features. Together with the calculation of 11 long-term measurements, we obtain a facial feature set with 264 entries.

Both resulting feature sets are then reduced by correlation based feature selection followed by a sequential forward search – a simple but popular wrapper approach that uses a classifier to determine significant features. This and all other classification tasks within this study is done via Naive Bayes. This classification model is a probabilistic classifier based on the Bayes' theorem, which makes the (unrealistic) assumption that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Hence, the probability for the occurrence of a certain category given a set of observations can be estimated as the product of the individual attributes. We chose this rather simple classification scheme due to its successful application in earlier emotion recognition tasks [14] and its fast computation, which allows us to run our experiments in a reasonable amount of time. After feature selection 35 audio and 40 video features remained for the DaFEx corpus, while on the CALLAS corpus 64 audio and 45 video features were chosen.

For evaluation of made experiments we agreed on a realistic, user independent approach: Leave-One-Person-Out. We consecutively draw samples belonging to one subject out of the available corpus. Remaining samples are used for training of classification models which then are tested against the isolated samples. The procedure is repeated until every person was once evaluated.

The described analysis was implemented and run with the Social Signal Interpretation (SSI) framework. SSI offers tools to record, analyse and recognize human behaviour in real-time, such as gestures, mimics, head nods, and emotional speech. In particularly it supports the machine learning pipeline in its full length and suits the fusion of multimodal information at different stages including early and late fusion. SSI is written in C++ and source code is available under LGPL[4].

# 5. RESULTS

Results for DaFEx and CALLAS corpora are summarised in Table 1. For the DaFEx data we observe an improvement of up to 7% and 10% compared to classification results on the audio and video channel. Here, BKS and One Versus Rest turn out to give the best performance closely followed by Feature Fusion, Mean Rule, Sum Rule, Weighted Average, Product Rule, Grading and One Versus Rest-Specialists. All in all, remarkable are results established across all fusion levels on the acted affective corpus.

In case of the more natural CALLAS corpus no improvement is achieved compared to the audio channel. However, on the video channel an enhancement of up to 8% is observed. Except for Grading and Min Rule results of all other fusion strategies lie within 3%.

Across corpora, simple fusion techniques like Feature Fusion as well as Mean, Sum and Product Rule perform on a very stable basis. More elaborate strategies seem to be more reliant on the structure of observed data. For example, the Cascading Specialist method generates the desired flattening effect among classes on the CALLAS corpus and therefore lists among the best fusion approaches. In contrary, needed specialist selection seems to be harder on the DaFEx corpus and it ranges among worst combination rules. Sophisticated ensemble strategies bare the potential to outperform more simple ones, but success is not guaranteed. Differences in accuracy tend to be rather small among all considered fusion techniques.

# 6. DISCUSSION

In Figure 2 recognition results are visualized per sample, as we compare the prediction for each sample with its real label. If the sample was correctly classified, it is marked with a white square, otherwise with a black one. Each column represents one sample of the data set and each row stands for the used classification method. The first row, for instance, visualizes classification results obtained for the single audio channel. We can for example infer from the DaFEx pattern on top of Figure 2 that the first two samples were correctly classified by audio, video and most fusion schemes, while sample three and four were obviously misclassified.

[2] http://hcm-lab.de/EmoVoice.html

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

[4] http://hcm-lab.de/ssi.html

|  | DaFEx |  |  |  |  |  |  |  | CALLAS |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | anger | disgust | fear | happiness | neutral | sad | surprise | **average** | positive | neutral | negative | **average** |
| *Single Modalities* |  |  |  |  |  |  |  |  |  |  |  |  |
| **Audio** | 0.39 | 0.32 | 0.43 | 0.21 | 0.86 | 0.67 | 0.25 | **0.45** | 0.59 | 0.64 | 0.61 | **0.61** |
| **Video** | 0.57 | 0.34 | 0.11 | 0.82 | 0.72 | 0.59 | 0.22 | **0.48** | 0.60 | 0.50 | 0.48 | **0.53** |
| *Feature Level Fusion* |  |  |  |  |  |  |  |  |  |  |  |  |
| **FeatureFusion** | 0.54 | 0.36 | 0.36 | 0.79 | 0.77 | 0.70 | 0.26 | **0.54** | 0.57 | 0.59 | 0.62 | **0.59** |
| *Decision Level Fusion* |  |  |  |  |  |  |  |  |  |  |  |  |
| **WeightedMajorityVoting** | 0.57 | 0.34 | 0.11 | 0.82 | 0.72 | 0.59 | 0.22 | **0.48** | 0.59 | 0.64 | 0.61 | **0.61** |
| **BKS** | 0.53 | 0.45 | 0.30 | 0.84 | 0.85 | 0.51 | 0.35 | **0.55** | 0.62 | 0.62 | 0.56 | **0.60** |
| **MaxRule** | 0.48 | 0.31 | 0.22 | 0.80 | 0.84 | 0.69 | 0.16 | **0.50** | 0.62 | 0.55 | 0.64 | **0.60** |
| **MinRule** | 0.44 | 0.39 | 0.41 | 0.44 | 0.73 | 0.59 | 0.39 | **0.48** | 0.56 | 0.61 | 0.55 | **0.57** |
| **MeanRule** | 0.52 | 0.38 | 0.36 | 0.79 | 0.78 | 0.71 | 0.26 | **0.54** | 0.59 | 0.58 | 0.59 | **0.59** |
| **SumRule** | 0.52 | 0.38 | 0.36 | 0.79 | 0.78 | 0.71 | 0.26 | **0.54** | 0.59 | 0.58 | 0.59 | **0.59** |
| **WeightedAverage** | 0.58 | 0.41 | 0.28 | 0.83 | 0.77 | 0.66 | 0.23 | **0.54** | 0.61 | 0.58 | 0.58 | **0.59** |
| **ProductRule** | 0.50 | 0.39 | 0.38 | 0.79 | 0.77 | 0.70 | 0.27 | **0.54** | 0.59 | 0.58 | 0.59 | **0.59** |
| **DecisionTemplate** | 0.51 | 0.41 | 0.30 | 0.67 | 0.81 | 0.61 | 0.22 | **0.50** | 0.57 | 0.60 | 0.59 | **0.59** |
| **DempsterShafer** | 0.48 | 0.41 | 0.31 | 0.67 | 0.81 | 0.59 | 0.25 | **0.50** | 0.56 | 0.62 | 0.59 | **0.59** |
| **CascadingSpecialists** | 0.35 | 0.38 | 0.44 | 0.53 | 0.90 | 0.66 | 0.27 | **0.50** | 0.60 | 0.63 | 0.61 | **0.61** |
| *Meta Level Fusion* |  |  |  |  |  |  |  |  |  |  |  |  |
| **StackedGeneralisation** | 0.53 | 0.40 | 0.39 | 0.72 | 0.74 | 0.61 | 0.28 | **0.52** | 0.59 | 0.57 | 0.64 | **0.60** |
| **Grading** | 0.60 | 0.44 | 0.18 | 0.80 | 0.89 | 0.64 | 0.23 | **0.54** | 0.67 | 0.50 | 0.49 | **0.55** |
| *Hybrid Fusion* |  |  |  |  |  |  |  |  |  |  |  |  |
| **OneVersusRest** | 0.53 | 0.34 | 0.36 | 0.83 | 0.79 | 0.71 | 0.25 | **0.55** | 0.59 | 0.59 | 0.60 | **0.59** |
| **OneVersusRest-Specialists** | 0.59 | 0.31 | 0.40 | 0.82 | 0.76 | 0.70 | 0.21 | **0.54** | 0.60 | 0.58 | 0.63 | **0.60** |

**Table 1: Recognition results for DaFEx and CALLAS corpora**

A clear characteristic shown by this visualisation on both corpora is the behaviour of fusion schemes in relation to single modalities. Depending on the outcomes of audio and video, there is a clear trend of forming white and black vertical columns within the picture: If both modalities classify correctly, most fusion approaches do so too; if both channels misinterpret the sample, most fusion strategies fail. Especially algebraic combiners like the sum or product rule amplify consistent (correct or incorrect) ensemble decisions because of their inherent combination rules. An exception to this trend is shown by some decision level approaches (BKS, decision template and dempster shafer) and the meta-learners of stacked generalisation and grading: These approaches are meant to learn the behaviour of available modalities. The logical connection between observed ensemble members' decisions and the actual true-class enables the phenomenon of "error learning" and therefore the potential of generating correct predictions though both modalities classify incorrect. This is visualized by white break-ups in black columns caused by consistent miss-classification of audio and video. Figure 2 unfortunately show that this desirable "error-learning" is also the reason for these ensemble methods to predict wrong classes even though both modalities chose the correct one – an undesirable characteristic that is not likely to be manifested by more simple fusion schemes.

Both modalities perform on an equal level in the DaFEx corpus (with a slightly better video channel), the more realistic CALLAS corpus clearly shows better results on prosodic observations and therefore the audio channel outperforms the facial modality (though former studies like [17] have shown a rather contrary behaviour). These results may be caused by the appliance of mood-inducing sentences for CALLAS sample generation: As the DaFEx corpus features professional actors, vocal and facial expressions are be expressed on an equally convincing level, while the unexperienced CALLAS probands focus strongly on expressing the given sentences verbally. Sample-wise recognition outcomes for the two modalities are therefore clearly more consistent on the DaFEx corpus than on CALLAS data. Some fusion strategies seem to benefit from consistency of modalities, leading to improved fusion results compared to the single channel-classifiers on the DaFEx corpus. But then again do recognition accuracies on the CALLAS corpus show the potential of combination rules to handle disagreeing modalities in away so that at least comparable results to the best channel can be received. However, it is difficult to recognize from the graphs in Figure 2, what exactly causes a potential gain in recognition accuracy for combination rules compared to the two single channels. To carve out those samples, which are misclassified in one of the single channels, but correctly recognized after fusion, we have included Figure 3.

Figure 3 compare sample-wise the results obtained from the superior modality with the results for the fusion techniques. For each corpus, the topmost row again shows sample-wise classification results for respective best modality. Within the following block we mark for each fusion strategy those samples with a white square, which receive a false recognition on the observed channel, but a correct one if fused with a further modality. The numbers behind listed methods gives the amount of relative improvement. Within the lower block, samples which are correctly predicted from the video channel, but misclassified after fusion with the audio channel are marked black. The numbers behind the method now illustrates the amount of impairment. For example did Feature Fusion correct 13% of the misses in the video channel on the DaFEx corpus. On the other hand it failed in 7% of the cases in which the facial modality did classify correctly. This leads to an overall improvement of 6% compared to the video channel.

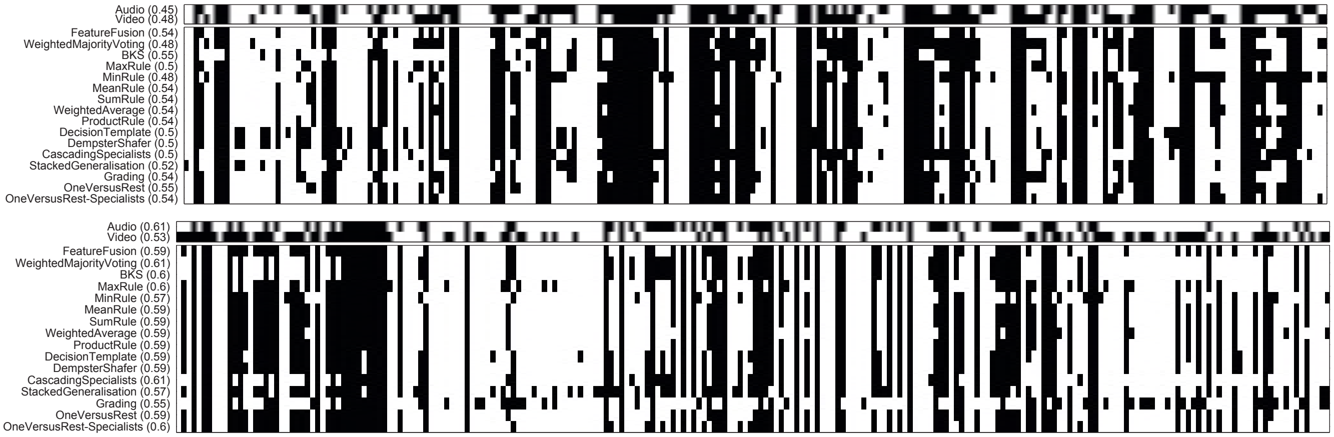

**Figure 2: Visualization of predictions for excerpts of the DaFEx and CALLAS samples**

Again, a clear trend to vertical columns in the patterns is obvious. This means that gains and losses of different fusion schemes are – most generally spoken – produced on the same samples. Outlying improvements achieved by elaborate ensemble combination techniques like the mentioned "error learning" are subsequently lost by outlying misclassification, caused by just this technique. The same regularities become apparent if we compare the amount of improvement with the amount of impairment between the different fusion methods. Obviously does a high amount of positive correction on miss-classified samples come along with a high number of errors on samples that were already correctly classified. For instance, on DaFEx we improve correct classification on 16% of the samples with stacked generalisation, but at the same time also lose a correct prediction in 11%, while for the weighted average rule a smaller improvement of 10% is accompanied by a lesser impairment of 5% (note that these numbers describe absolute, sample-wise values and do not directly transfer into class-wise recognition rates presented in Table 5). The difference between both values is more or less the same for all tested methods. On the CALLAS corpus rates for improvement are mostly smaller than generated impairments (as the audio channel clearly performs better than the facial modality), resulting at best in equal recognition results for some ensemble approaches as for the single modality classification with vocal features.

Made observations imply to a certain degree an interchangeability of presented fusion schemes. Sample-wise investigations show a common recognition-behaviour and gains in some specialised areas are paid for in others. Fusion techniques, on the other hand, have the potential to outperform single-channel classification on multi-modal datasets. Results are either considerably improved (DaFEx) or at least in line with the dominating modality (CALLAS). This characteristic is especially desirable, whenever the best modality is not known in advance. But is there a generally advisable fusion scheme? Performance of different strategies lie within a close range, sophisticated methods do not necessarily outperform simple combination rules. Without prior knowledge, appliance of feature fusion simple algebraic combiners seems reasonable, as they perform stable across different datasets despite simple mechanisms.

# 7. CONCLUSION

In this study we performed a comprehensive comparison of fusion techniques for multi-modal affect recognition tasks. Experiments were run on two Italian emotion corpora featuring vocal and facial modalities: The acted DaFex corpus and the more natural CALLAS expressivity corpus. Implemented fusion rules included feature, decision, meta and hybrid level strategies and results were discussed using novel visualisations. We found interesting, common characteristics among ensemble methods as well as uncommon effects like "error learning". Though a certain degree of interchangeability between tested fusion approaches can be suspected – and therefore no fusion scheme can be advised in general – are most approaches capable of enhancing single channel classification or are at least on par with the superior modality, if inconsistency of observed modalities prevent further improvements. Inconsistencies of fusion schemes performances across corpora make no strategy advisable in general, the applied fusion method should be chosen based on the underlying classification problem.

# 8. ACKNOWLEDGMENTS

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

# 9. REFERENCES

[1] A. Battocchi, F. Pianesi, and D. Goren-Bar. Dafex: Database of facial expressions. In M. T. Maybury, O. Stock, and W. Wahlster, editors, *INTETAIN*, volume 3814 of *Lecture Notes in Computer Science*, pages 303–306. Springer, 2005.

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

[3] R. Duin and D. Tax. Experiments with classifier combining rules. In *Lecture Notes in Computer Science*, volume 1857, pages 16–29. Springer, 2000.

[4] G. Fumera and F. Roli. A theoretical and experimental analysis of linear combiners for multiple classifier

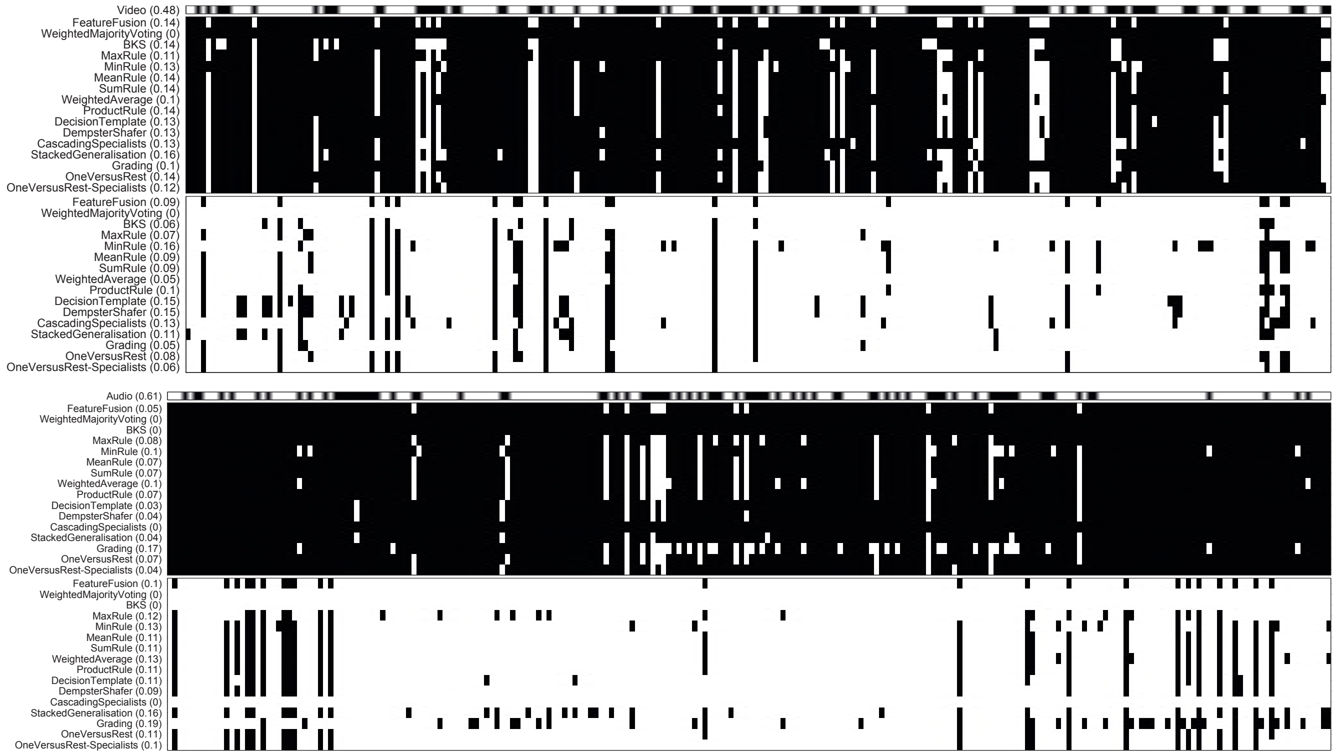

**Figure 3: Visualization of recognition improvements and impairs of the DaFEx and CALLAS samples**

systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):942–956, 2005.

[5] Y. S. Huang and C. Y. Suen. Behavior-knowledge space method for combination of multiple classifiers. *Proc. of IEEE Computer Vision and Pattern Rcog.*, 20:347–352, 1993.

[6] J. Kim and F. Lingenfelser. Ensemble approaches to parametric decision fusion for bimodal emotion recognition. In *Int. Conf. on Bio-inspired Systems and Signal Processing (Biosignals 2010)*, 2010.

[7] C. Küblbeck and A. Ernst. Face detection and tracking in video sequences using the modifiedcensus transformation. *Image Vision Comput.*, 24:564–572, June 2006.

[8] L. I. Kuncheva. Switching between selection and fusion in combining classifiers: An experiment. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 32(2):146–156, 2002.

[9] L. I. Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):281–286, 2002.

[10] A. Rabie, B. Wrede, T. Vogt, and M. Hanheide. Evaluation and discussion of multi-modal emotion recognition. In *Proceedings of the 2009 Second International Conference on Computer and Electrical Engineering - Volume 01*, ICCEE '09, pages 598–602, Washington, DC, USA, 2009. IEEE Computer Society.

[11] A. K. Seewald and J. Fuernkranz. An evaluation of grading classifiers. In *Advances in In Intelligent Data Analysis, 4th International Conference, IDA 2001, Proceedings*, volume 10, pages 271–289. Springer, 2001.

[12] G. Shafer. *A Mathematical Theory of Evidence*. NJ: Princeton Univ. Press, Princeton, 1976.

[13] K. M. Ting and I. H. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:115–124, 1999.

[14] T. Vogt and E. André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *IEEE International Conference on Multimedia & Expo (ICME 2005)*, 2005.

[15] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1996.

[16] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, January 2009.

[17] 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. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 137–143, 2004.