BIMODAL FUSION OF EMOTIONAL DATA IN AN AUTOMOTIVE ENVIRONMENT

S. Hoch, F. Althoff

BMW Group Research and Technology
Department of Human-Machine Interaction
Hanauerstr. 46, 80992 Munich, Germany
{Stefan.Hoch, Frank.Althoff}@bmw.de

G. McGlaun, G. Rigoll

Munich University of Technology (TUM)
Institute for Human-Machine Communication
Arcisstr. 21, 80290 Munich, Germany
{mcglaun, rigoll}@ei.tum.de

ABSTRACT

In this work, we present a flexible bimodal approach to person dependent emotion recognition in an automotive environment by adapting an acoustic and a visual monomodal recognizer and combining the individual results on an abstract decision level. The reference database consists of 840 acted audiovisual examples of seven different speakers, expressing the three emotions *positive* (joy), *negative* (anger, irritation) and *neutral*. Concerning the acoustic modul, we calculate the statistics of commonly known low-level features. Facial expressions are evaluated by a SVM classification of gabor-filtered face regions. At the subsequent integration stage, both monomodal decisions are fused by a weighted linear combination. An evaluation of the recorded examples yields an average recognition rate of $90,7\%$ for the fusion approach. This adds up to a performance gain of nearly 4% compared to the best monomodal recognizer. The system is currently used to improve the usability for automotive infotainment interfaces.

1. INTRODUCTION

The conceptual design of both powerful and intuitive user interfaces has evolved to an important factor in the development of interactive systems. Confronted with increasing functional complexity and extensive learning periods users become frustated more and more. Thus, research in the field of human-machine-interaction is looking for various possibilities to make the dialog between the user and system flexible, natural and error-robust. Inspired by the example of human communication, information is interpreted among various sources. In this context, analysing the emotional state of the user plays a key role.

Concerning an automotive environment, the driver has to cope with different tasks like steering the car, controlling the speed and operating several assistence and information systems. Each of these tasks can have a significant influence on the current emotional state of the user. By adapting dialog strategies, an automatic emotion recognition modul could be used to reduce the mental workload of the driver and avoid dangerous situations.

In this work, we focus on multimodal emotion recognition. Two state-of-the-art technologies for classifying spoken utterances and facial expressions are adapted to the specific boundary conditions in the automotive environment and combined by a late-semantic fusion approch. As we are not aware of any public available databases containing emotional affected material in the automotive domain, we have decided to collect our own database. Furthermore, we have decided not to differ between the common six discrete MPEG4 basis emotions (anger, joy, surprise, fear, sad, disgust). Concerning the automotive environment we propose a more suitable class separation, discriminating between the classes positive (joy), negative (anger, irritation) and neutral.

1.1. Related work

A large community of researchers has dealt with the problem of emotion recognition analysing different information channels. Pantic et. al [1] give an exhaustive overview about the common methods on this topic. Most work has been done on classifying the prosody of speech [2, 3], lately also combined with verbal information [4]. The static classification of statistical measurements of low-level acoustic features has proven to be the state-of-the-art approach from this research sector.

Another field of emotional information derived from the class of physiological signals, e.g., galvanic skin response, heart rate and body temperature [5]. This modality has to deal with the disadvantage that most of its signals can not be measured without direct body contact. Thus, these sources are improper in most possible environments.

A further quite intensely researched topic in emotion recognition is the analysis of facial expressions. There are two common approaches, one exploring the movement of fiducial points called facial action units (FACS) [6], the other classifying the output of filtered face regions [7, 8] without locating specific facial features. The latter approach has a significant advantage as it does not depend on a perfect location of the FACS, which still are hard to detect automatically. In comparison to monomodal systems, multimodal approaches, that make use of multiple sensor information have only recently been explored [5, 9]. Thus, our work focusses on a domain-specific integration of various recognition results.

2. METHODOLOGY

The following section gives a short description of the pattern recognition processes in our work. We first focus on the acoustic and visual feature extraction and then explain the applied pattern classification techniques. Finally, we present a decision level approach as a method of sensor fusion.

2.1. Feature Extraction

2.1.1. Acoustic Features

Extracting the prosodic parameters *pitch*, *power*, *formants* and *duration of voiced segments* has turned out to be a solid way of recognizing emotion from human speech [2, 3, 4]. We use the Snack

Sound Toolkit [10] to preprocess the audio signal, and afterwards extract 51 features consisting of different statistical measurements (e.g., mean, median, max, min, max-min, range) derived from the course of these acoustic low-level parameters and their first derivations. We ranked the acoustic features by measuring the expected *information gain* which displays the mutual information between the class **Y** and an attribute **X**, and is calculated according to the following equation.

$$H(\mathbf{Y}, \mathbf{X}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) \quad (1)$$

According to the calculated values, the size of the feature vector can be reduced in order to save computation time. In section 3, a variation of the number of acoustic features based on their information gain is discussed, comparing the recognition results of the full and the reduced feature sets.

2.1.2. Visual Features

Our approach to emotion recognition from facial expressions is based on an adapted implementation of the algorithm originally proposed by Movellan and Bartlett [7, 8]. All visual algorithms are realized in C^{++}, using the Intel Open Source Computer Vision Library (OpenCV) [11]. In a first step, the human face is located by a frontal face detector based on the object recognition algorithm of Viola & Jones [12] which uses an AdaBoost [13] feature selection process to create a robust object representation out of a large number of Haar features. The detected face area is scaled to a square region (70x70 pixels), and converted into a grayscale image in order to normalize the analyzed pattern. After this preprocessing, the detected region of interest is filtered with a set of gabor wavelets [8, 14]. This filterbank is made up of filters (see equation 2, [14]) with three different spatial frequencies and six different orientations which influence the wave vector $\vec{k}$. In combination with the variance σ of the gaussian envelope, the wave vector affects the shape of a gabor filter. Figure 2.3 shows the real parts of the gabor set (a), an example face (b), and the resulting magnitude images after the filtering process (c).

$$p(x) = \frac{|k|^2}{\sigma^2} \exp\left(-\frac{|k|^2 |x|^2}{2\sigma^2}\right) \left(\exp(i\vec{k}\vec{x}) - \exp\left(-\frac{\sigma^2}{2}\right)\right) \quad (2)$$

Filtering the 70x70 pixels face region with these 18 gabor filters results in 88200 different magnitude coefficients. For a more compact representation and reduced computation time, feature selection is performed, using the AdaBoost algorithm which detects the features with the greatest discriminative power regarding the training data. In our work, the number of selected features averages 118 across all test persons, reducing the length of the input vector by a factor of 700.

2.2. Machine Learning

2.2.1. Prosody Classification

We have examined two established pattern recognition techniques according to their performance in classifying the prosodic feature vectors calculated from the acoustic sequences. The first method is a support vector machine (SVM) with a linear kernel function and a one-against-all class separation approach resulting in three SVMs for our problem. The second classifier we have tested in the acoustic emotion recognition modul is an artificial neural network

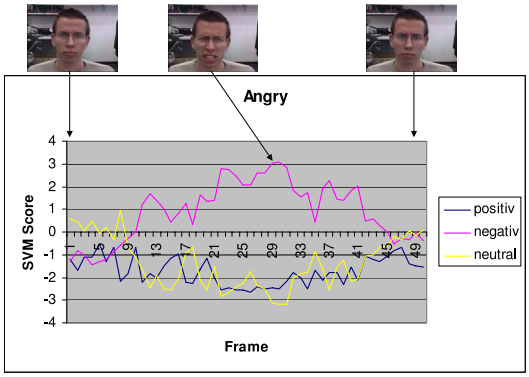


Fig. 1. Course of the SVM output-values for an example sequence of a negative facial expression

(ANN). It consists of an input layer with 51 neurons, one hidden layer with 27 neurons, and an output layer with 3 neurons, one for each analyzed class. No matter which classifier is used, the acoustic modul produces a three-dimensional output vector (x_1, x_2, x_3) on the basis of the standardized feature vectors. Afterwards, the output is transferred into a probability distribution (ρ_1, ρ_2, ρ_3) by the following soft-max function.

$$\rho_n = \frac{\exp(x_n)}{\sum_{i=1}^{3} \exp(x_i)}, \; n \in \{1, 2, 3\} \quad (3)$$

2.2.2. Facial Expression Classification

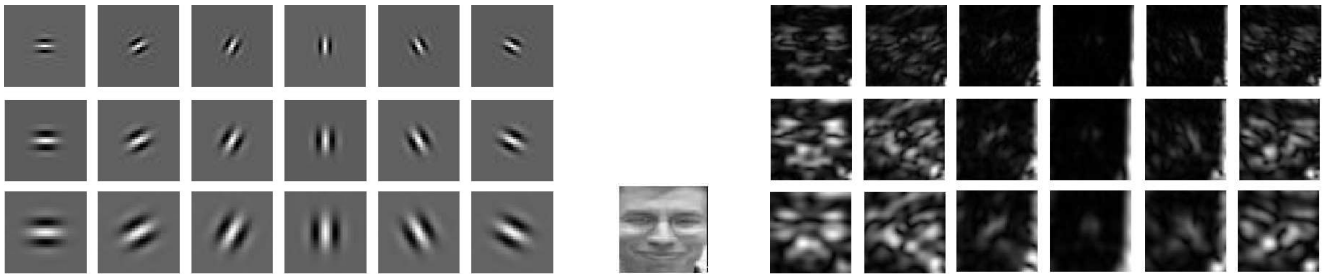
The visual part of our emotion recognition system uses the same, previously described type of SVMs. For each image of a sequence, the face region is detected, preprocessed, and the selected features are calculated. Then the SVM classifier produces an estimation of the current facial expression, resulting in a course of classification scores for a complete video sequence (see figure 1). To obtain the results of a facial expression, we integrate the class scores for the whole sequence, and calculate the average score for each emotion class. Finally, we transform these mean values to a probability distribution, using the softmax function (see equation 3).

2.3. Decision-based Fusion Approach

Both monomodal emotion recognition systems provide an output vector containing the individual confidence measurements of the monomodal classification process. Our sensor fusion approach combines these two monomodal results to a multimodal decision. The acoustic confidence measurements $\rho_{ac,n}$ and the corresponding visual results $\rho_{vis,n}$ are merged to fusion confidences $\rho_{fus,n}$ by the following weighted linear combination.

$$\rho_{fus,n} = \eta \cdot \rho_{ac,n} + (1 - \eta) \cdot \rho_{vis,n} \quad (4)$$

An important factor of this approach is the parameter η, called the *linear fusion coefficient* (LFC). The LFC is bound to the interval $[0, 1]$, and controls the influence of each modality on the fusion result. Therefore, it offers a high potential for an adaptive readjustment which will be discussed later in this paper. Using a



(a) (b) (c)

Fig. 2. (a) Real parts of the gabor filters, (b) localized and normalized face, (c) magnitude of the filter operations

LFC value of 0.5, results in an equal weighted fusion which corresponds to an average of the monomodal scores. The influence of the acoustic analysis can be increased by raising the LFC, decreasing the visual influence vice versa.

3. EVALUATION

We have recorded an audiovisual data collection with seven non-professional actors. Thereby, we have put some constraints on our material. All data have been recorded in a standing car with cost-efficient sensor equipment, a standard webcam, and an array microphone (see figure 3(b)), trying to match industrial standards for in-car systems. At first, our test persons were asked to make only facial expressions. For all persons, 50 images have been captured for each emotion class in order to build the person dependent facial expression models. Then they were shown different sentences from one to ten words which they should act in an emotional manner. These sentences included emotional sensitive phrases and examples without any connection to a specific emotional temper. The recordings have resulted in a set of 840 audiovisual sequences of seven different speakers, five male Germans, one female German, and one male American. For the multimodal system, evaluation data for each person has randomly been separated in two-thirds for training and one-thirds for testing purpose. For the acoustic feature tests, the person dependent instances have randomly been subdivided into ten folds, and have been exploited in a cross-validation. In every iteration, nine folds have been used to train the classifiers and, afterwards, the system has been tested with the remaining samples. The final recognition rate for each person has then been calculated by averaging the results over these ten iterations.

3.1. Acoustic Feature Reduction

Concerning the evaluation of the acoustic features, the calculated feature vectors have been enlarged bit by bit, starting with the acoustic feature with the greatest information gain, and then adding the next best feature. Every iteration, a person dependent 10-fold-cross-validation has been performed. Figure 3(a) shows the mean results across all seven speakers in a course for both test classifiers, ANN and SVM. The recognition rate for both classifiers increases in the beginning until a vector length of 16 features (80,7% for ANN) which is only a small difference compared to the maximum recognition rate of 81,8% with 30 prosodic features (ANN). After this point both classification curves reach a saturation level around 80% recognition rate. Therefore, adding more prosodic features to the feature vector did not result in significant gain in recognition performance for our problem. The ten first selected features are all statistics derived from the course of pitch and power. This result approves the high significance of these low-level features for acoustic emotion recognition.

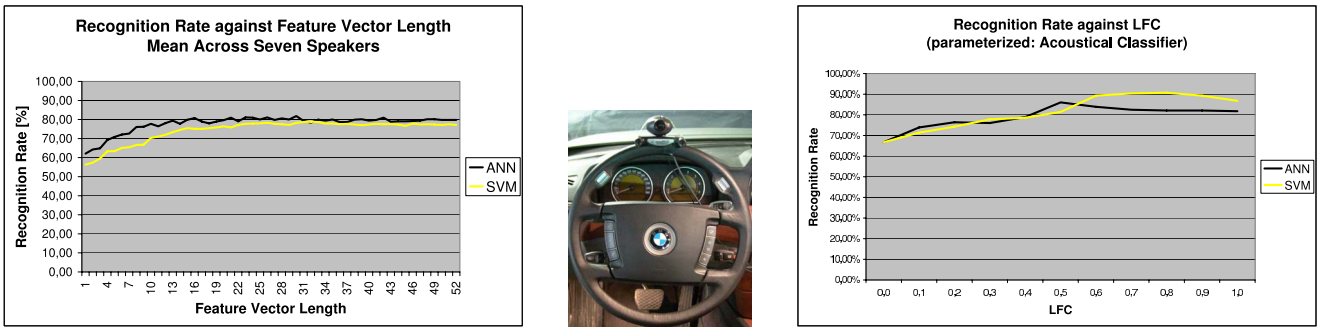
3.2. Multimodal Fusion

In the evaluation session of our multimodal approach, we analyzed in which way both information channels supply useful information to solve the problem of emotion recognition in an automotive environment. In several test sets, we explored the influence of different weightings of the modalities in the fusion process, ranging from pure facial expression analysis ($\eta = 0.0$) to pure prosodic classification ($\eta = 1.0$) without any feature selection. Figure 3(c) shows the results of this evaluation in a diagram comparing the mean recognition rate over all seven speakers against the used LFC. The diagram is parameterized by the different classification schemes, ANN and SVM, which were used for the prosodic classification. In both evaluation configurations the prosodic classification achieved the better monomodal recognition performance, $81,7\%$ (ANN) and $86,8\%$ (SVM), resp., compared to $66,8\%$ for the SVM classification of the facial expressions. Both diagrams show a gain in recognition performance in between the monomodal configurations. The fusion of the bimodal recognition confidences results in a gain of up to $4,3\%$ and $3,9\%$, respectively .

4. DISCUSSION

The results presented in the last section demonstrate the potential of using multimodal information for this research topic. The bimodal information channels contain both redundant and complementary emotional signals which can be used for a more robust recognition. In an automotive environment, this is a particularly important system property, since the single information channels can be influenced by strong noise (e.g., acoustic channel in a convertible). Moreover, the experiments showed that the test persons differed in the degree of emotional signals in the analyzed information channels. Thus, our multimodal approach can provide essential input for adapting a target application to the individual user.

At first, the performance of the facial expression analysis has been below our expectations, as we had already obtained much more satisfying results in a laboratory environment. Many errors arise from the fact that the system classifies the expression as neutral in comparison to the person dependent training images. The training data contained distinctive facial expressions which were not always similar to the ones in the acted audiovisual emotion se-



(a) (b) (c)

Fig. 3. (a) Acoustic recognition rate against the number of used acoustic features, sorted by the information gain across all speakers, (b) sensor placement in the car, (c) Recognition rate against LFC, parameterized by the used acoustic classifier

quences. Additionally, the performance and the accuracy of the frontal face detector lacked because of the poorer image quality in comparison to the laboratory environment, resulting in a less accurate matching between the selected feature points for the gabor filtering. Nevertheless, we are optimistic that these visual problems can be solved either by a better camera equipment or a retraining of the object detector. Additional experiments showed that creating an illumination invariant visual representation, for example, using NIR cameras, could be another solution for this problem.

Furthermore, our weighted linear combination shows a lot of potential to be expanded to a more complex system, e.g., by an adaptive readjustment of the LFC. While we only checked preadjusted LFC values, the LFC could be changed dynamically depending on several context parameters, like the SNR of the data channels, user profiles, or different emotion classes. The fusion can also be extended by combining even more data sources, e.g., by adding word phrase semantics, physiological signals or other context information.

5. CONCLUSIONS AND FUTURE WORK

We have presented a bimodal approach to automatic emotion recognition in an automotive environment that is based on two state-of-the-art techniques for acoustic and facial expression analysis. Sensor fusion has been executed on decision level, where the confidence measurements for the three discrete emotion classes positive, negative and neutral of each individual recognizer have been combined through a weighted linear combination. The multimodal system has yielded a maximum recognition rate of $90,7\%$, and has performed about 4% better compared to the best monomodal recognizer (acoustic classifier) of our system.

We are currently working on a person independent realization of our system, as for a real application, it is hardly acceptable to ask the user to train person dependent classification models. Furthermore, we try to optimize both monomodal recognizers in terms of signal and preprocessing quality which could increase the overall system performance. In the near future, we think about evaluating other fusion techniques, like fusion on feature extraction level, e.g., based on multistream HMMs, which could be more powerful in registering the dynamics of an emotional expression, and, further on, could deliver a possibility for an automatic emotion spotting process.

6. REFERENCES

[1] M. Pantic et al., "Towards an affect-sensitive multimodal human-computer interaction," in *Proc. of the IEEE*, 2003.

[2] F. Dellaert et al., "Recognizing emotions in speech," in *Proc. ICSLP '96*, 1996.

[3] P.-Y. Oudeyer, "The production and recognition of emotions in speech: Features and algorithms," in *Int. Journal of Human Computer Studies*, 2003.

[4] B. Schuller et al., "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture," in *Proc. of the ICASSP 2004, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2004.

[5] C. L. Lisetti et al., "Developing multimodal intelligent affective interfaces for tele-home health care," in *Int. Journal of Human-Computer Studies Special Issue on Applications of Affective Computing in HCI*, 2003.

[6] I. Cohen et al., "Facial expression recognition from video sequences: Temporal and static modeling," in *Computer Vision and Image Understanding*, 2003.

[7] G. Littlewort et al., "Dynamics of facial expression extracted automatically from video," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2004.

[8] J. R. Movellan and M. S. Bartlett, "The next generation of automatic facial expression measurement," in *What the Face Reveals*. 2003, Oxford University Press.

[9] M. Song et al., "Audio-visual based emotion recognition a new approach," in *Proc.of the 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, 2003.

[10] K. Sjolander et al., "Web-based educational tools for speech technology," in *Proc. of Matisse 99*, 1999.

[11] Intel®, "Intel open source computer vision library," .

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[13] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Int. Conf. on Mach. Learning*, 1996.

[14] F. Jiao et al., "Face alignment using statistical models and wavelet features," in *Proc. of the 2003 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, 2003.