Multimodal Emotion Classification in Naturalistic User Behavior

Steffen Walter[1], Stefan Scherer[2], Martin Schels[2], Michael Glodek[2], David Hrabal[1], Miriam Schmidt[2], Ronald Böck[3], Kerstin Limbrecht[1], Harald C. Traue[1], and Friedhelm Schwenker[2]

[1] Medical Psychology, University of Ulm, Germany
[2] Institute of Neural Information Processing, University of Ulm, Germany
[3] Chair of Cognitive Systems, Otto von Guericke University Magdeburg, Germany
steffen.walter@uni-ulm.de

Abstract. The design of intelligent personalized interactive systems, having knowledge about the user's state, his desires, needs and wishes, currently poses a great challenge to computer scientists. In this study we propose an information fusion approach combining acoustic, and bio-physiological data, comprising multiple sensors, to classify emotional states. For this purpose a multimodal corpus has been created, where subjects undergo a controlled emotion eliciting experiment, passing several octants of the valence arousal dominance space. The temporal and decision level fusion of the multiple modalities outperforms the single modality classifiers and shows promising results.

1 Introduction

In the future, technological cognitive systems will more and more find their way into human's everyday life by supporting us with helpful information and mediating intentions in a social and technical environment. These systems need to be adaptive towards the individual user states such that the system is able to appropriately react to the user's needs and wishes. Requirements for this visionary goal are a robust and reliable automatic classification of emotional states. Since emotional user behavior is a complex, multimodal and dynamic process, it is important to consider a variety of channels such as prosody, mimics, gestures and bio-physiologic data [8,12]. For the analysis of naturalistic user behavior we therefore suspect multimodal approaches to outperform methods using a single modality with respect to classification performance [9].

In the presented study, we investigate a Wizard-of-Oz design aiming at inducing emotions according to the valance, arousal and dominance (PAD) model [15] in the subjects by simulating typical situations found in human computer interaction. Among the manifold possibilities to induce emotions we chose, e.g. delays, misunderstandings, ignoring commands, time pressure, but also positive feedbacks. For the automatic classification analysis we chose two octants of the PAD space, namely "positive valence, low arousal, high dominance" vs. "negative valence, high arousal, low dominance". The modalities of choice include

J.A. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2011, LNCS 6763, pp. 603–611, 2011.

respiration, heart rate, electromyography (M. corrugator), skin conductance, as well as audio based prosody, energy, voice quality and standard speech features, and manually annotated analysis of the facial expression due to the Facial Action Coding System (FACS) [5].

2 Experimental Design

To study the emotional behavior of humans in the different octants of the PAD space, a certain number of participants has been recorded in a Wizard-of-Oz (WOz) experiment. This method allows to simulate the technical system in a human-computer interaction by a psychologist in a separate room, which is connected via cameras and microphones. A group of 10 volunteers participated in the experiment. This group is divided equally in four subsets: younger female (mean of age = 27.6), older female (m = 46.4), younger male (m = 25.2) and older male (m = 56.2). In each gender group, the threshold of the age is 40 years. In Sect. 2.1 the intrinsic setting of the experiment will be illustrated, with which we expect to induce the demanded emotional states[1].

2.1 Course of the Experimental Sequences

The experiment is composed of two rounds: the first is divided into five so-called experimental sequences (ES). In each ES (ES-1_1-ES-5_1), the subject plays a game of Concentration (also known as Memory) of varying difficulty. The second round comprises six ES (ES-1_2-ES-6_2). The two rounds are conjoined by a center part, in which images of the International Affective Picture System (IAPS) [11] are shown to the subject. Figure 1 shows the schematic configuration of the course of the experiment.

Table 1 lists the characteristics of the different experimental sequences. The first column denotes the ES code, the second column the expected emotional states of the subject (e.g. "PAD +−+" represents positive valence, low arousal and high dominance). In the third column, the size of the Concentration matrix is shown, increasing from ES-1 to ES-6. The fourth column specifies the difficulty of the game. The last column lists the quality of the feedback given to the subject.

3 Feature Extraction

3.1 Audio Features and HMM Sequence Encoding

Prior to the audio processing it is necessary to perform a speech/non-speech detection to isolate the portions of the sequences in which the subject utters commands, indicating the position of the next memory card to be flipped in a two dimensional grid. The commands comprise the uttering of a letter indicating

[1] The study was carried out according to the ethical guidelines of Helsinki (ethic committee of the university of Ulm: C4 245/08-UBB/se).

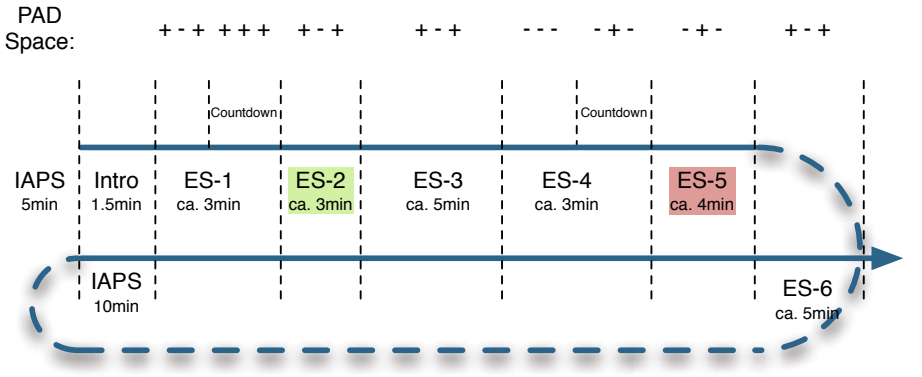

Fig. 1. Experimental design, including the expected position in the PAD space. After a five minute initialization with IAPS images, the experiment starts with a short introduction, followed by the first round of Concentration (ES-1,...,ES-5). The second round of Concentration contains six games (ES-1,...,ES-6) and starts after ten minutes presentation of IAPS images. The experiment is closed by standardized questionnaires. The expected octant of the PAD space is coded as plus and minus symbols in the order pleasure, arousal and dominance.

Table 1. Characteristics of the experimental sequences. For some sequences the arousal was increased (*) by introducing a countdown.

ES	Expected state	Matrix size	Difficulty	Feedback
ES-1 *	PAD +−+	4 × 4	low	slightly positive feedback (e.g. "you have successfully solved the first task.")
ES-2	PAD +−+	4 × 4	low	positive feedback (e.g. "great!")
ES-3	PAD +−+	4 × 5	low	neutral feedback (e.g. "your performance is mediocre.")
ES-4 *	PAD − − −	4 × 5	high	slightly negative feedback (e.g. "your performance is declining")
ES-5	PAD −+−	5 × 6	very high	negative feedback (e.g. "you are doing very poorly.")
ES-6	PAD +−+	5 × 6	low	positive feedback (e.g. "your performance is improving.")

the row of the board and a number indicating the column. Hence, the acoustic information is very limited and traditional features such as Mel frequency cepstral coefficients (MFCC) might not perform sufficiently well. However, since the commands are quite frequent and very similar in their structure and length, features such as the rate of speech, the length of the uttered letter or number, or the fundamental frequency could contain information on the subject state in the given task.

In order to investigate the distinction between positive and negative subject states, we computed seven diverse feature sets for the commands. Each of the

sets is used to train an independent classifier. The decision for all the classifier is combined in a fusion step by simply multiplying the outputs of the respective classifier for each utterance. In the following we briefly introduce the feature sets:

1. The MFCC and ΔMFCC[2] features are extracted from 32ms sized windows with a sample rate of 100Hz. The features are design with respect to the biological perceptual variations in the human ear in order to capture the phonetically important characteristics of speech [3].
2. The biologically inspired Modulation spectrum (ModSpec) based features reflect the variations in the frequencies of the signal and are extracted from 200ms sized windows with a sample rate of 50Hz [10,6].
3. Perceptual linear predictive (PLP) features are an extension to the well known linear predictive coding features with respect to human perceptual capabilities comprising similar filtering as for MFCC and equal loudness curves [7,14]. The features are extracted from 32ms of speech and have a sample rate of 100Hz.
4. Statistics of the fundamental frequency (f_0) are calculated from each of the commands to form one feature vector [14]. The statistics comprise minimum, maximum, span, mean, and standard deviation as well as the same statistics on the differentiation of the f_0 signal.
5. Corresponding to the computation of the f_0, statistics of the energy of the signal are computed for each command.
6. Based on the voiced parts of the commands, we calculate statistics on the durations and pauses in between the latter and the number of the command using the f_0 signal. These features form the seventh set of audio based features, namely the speech rate and syllable duration features.

For the feature sets MFCC, ΔMFCC, ModSpec and PLP we calculate the distance matrices using a hidden Markov model encoding in order to compensate varying command lengths [17]. These encodings are then used as the actual features for the classification.

3.2 Facial Expressions Utilizing FACS Coding

The Facial Action Coding System (FACS) introduced in [5] systematically describes facial expressions. Each facial expression is decomposed by experts into so-called Action Units (AU), which encode contractions or relaxations of facial muscles (e.g. AU1: inner brow raiser, AU12: lip corner puller). These encodings could directly be useful as a high level feature input to the automatic classification.

3.3 Bio-physiological Feature Extraction

In this study we are provided with a variety of bio-physiological signals of the subject. From these signals sampled at 512 Hz, features were extracted for classification on the basis of a five second time window.

[2] The derivative of the Mel frequency cepstral coefficients.

One of the most prominent physiological features is the skin conductance level (scl), measuring the transpiration of a person. The gradient of this channel was calculated and thereupon the minimum, maximum and average gradients in the 5 second window are used for classification. A further intuitive feature for emotion recognition is the heart rate of a subject. In order to determine this characteristic, the blood volume pulse was recorded, which reflects the well known QRS structure of a heart beat. For every 5 second time window the minimum, maximum and the average rate of occurrences of R artifacts was calculated. For the EMG channels of the zygomaticus and the corrugator, we have first calculated the result of the moving average algorithm with window size 20 for each point of the segment to acquire a value for the tension of the muscle. Again, the minimum, maximum and average values of these signals were passed to classification.

4 Methods for Classification

Within this study two kinds of classifiers, namely Multilayer Perceptrons (MLP) and a Support Vector Machine (SVM) are used. This section gives a brief introduction to those classifiers. For a detailed description we recommend the reader to follow the citations.

5 Multilayer Perceptron

The multilayer perceptron (MLP), a universal function approximator, maps an input $\mathbf{x}$ to an output $\mathbf{y}$ processing the input via weighted ($\mathbf{w}$) connections through potentially multiple hidden layers [1]. In general, each perceptron calculates a weighted sum of all incoming variables plus an additional bias and then creates the output by mapping the value using a differentiable function called activation function. A simple perceptron without a hidden layer and a sigmoid activation function is therefore given by

$$y{=}\sigma(\sum_i x_i w_i + w_0). \tag{1}$$

An MLP is trained using error back-propagation and a gradient based weight adaptation after presenting labeled examples to the network and calculating an error function.

5.1 Probabilistic Outputs for Support Vector Machines

In recent years, SVM have found a broad acceptance in the Machine Learning Community. The success might have its origin in the intuitive generalization approach using a maximum margin between the classes and the capability to reformulate the SVM such that kernels can be used to transfer the data points implicitly into a higher dimensioned feature space. The SVM output in its original formulation renders crisp class decisions depending on which side of the

margin the data point is located [19]. In many applications, especially in information fusion tasks, it can be advantageous to know the degree of belief that the output belongs to a certain class. Therefore, Platt [13] proposed an extension to SVM in which the distance of a point $\hat{\mathbf{x}}$ to the hyperplane is mapped onto a probability $p(\mathtt{t}|\mathbf{y} = y(\hat{\mathbf{x}}))$ using a sigmoid function where the distance to the hyperplane is given by $y(\hat{\mathbf{x}}) = \mathbf{w}^{\mathrm{T}}\hat{\mathbf{x}} + b$ and $\mathtt{t} \in \{-1, +1\}$. In order to fit the posterior the logistic sigmoid

$$p(\mathtt{t}|\mathtt{y} = y(\hat{\mathbf{x}}))=\frac{1}{1 + \exp(-ay(\hat{\mathbf{x}}) + d)} \tag{2}$$

having the parameters a and d is optimized using the cross-entropy error function.

5.2 Information Fusion

Combining information from multiple sources is a powerful means to stabilize classifiers but also to correct decisions of individual classifiers. In principle, there are two distinct ways of information fusion considering classification: firstly one can combine on a lower level of the general classification architecture, called *early* fusion , e.g. on a feature level. On the other hand, fusion is often considered on a decision level: the outputs of individual classifiers are combined in order to correct a more precise new decision. This approach is called *late* fusion or multiple classifier system (MCS) [9].

Generally information fusion is promising, when the combined sources show independence given the true class label, which is commonly called diversity in the MCS community [2]. Also, in case of an application using sequential data, e.g. physiological data or audio data, which are both used in this study, the temporal accumulation of individual decisions is advantageous. Thus, if the sequence of classifications have only a small bias to the correct class, an over-all correct classification can be accomplished [4]. Classifier fusion can be conducted using static combination rules as described in [18], but also using more complex trainable mappings [16].

6 Classification Experiments

Due to the heavily varying subject dependent bio-physiological data, we do not expect to be able to generalize over multiple subjects at once. Therefore, we conduct personalized experiments. As described above, we possess recordings of each individual undergoing the positive and negative experimental sequences twice (compare Section 2.1). Hence, in order to provide statistically salient results we conduct four cross-evaluation runs, i.e. training on round one and testing on the second, vice versa and folding the sequences of the two rounds.

For every audio feature, SVM with probabilistic outputs were trained and its results were combined on a decision basis using the averaging rule. These audio based results on the single utterance level led to an accuracy of 74.3 % with

Table 2. Confusion matrix of audio based utterance level late fusion

\	ES-2	ES-5
ES-2	0.477	0.523
ES-5	0.083	0.917

Table 3. Confusion matrix of bio-physiological five second window based classification using early fusion

\	ES-2	ES-5
ES-2	0.420	0.580
ES-5	0.474	0.526

Table 4. Confusion matrix of audio based sequence level fusion computed with temporal fusion

\	ES-2	ES-5
ES-2	0.556	0.444
ES-5	0.028	0.972

Table 5. Confusion matrix of bio-physiological sequence level classification computed with temporal fusion

\	ES-2	ES-5
ES-2	0.861	0.139
ES-5	0.417	0.583

Table 6. Confusion matrix after combining both modalities

\	ES-2	ES-5
ES-2	0.583	0.417
ES-5	0.028	0.972

an F_1 value of 0.600 for the positive class and 0.812 for the negative one over all the subjects and evaluation runs. The confusion matrix for the test data set is seen in Table 2. The bio-physiological data was combined using early fusion, i.e. concatenation of the features, and classified using a MLP. The results on the five second window classification led to an accuracy of 49.2 % and an F_1 value of 0.348 for the positive class and 0.584 for the negative class over all the subject and evaluation runs. The confusion matrix for this experiment is found in Table 3.

In the following temporal fusion is conducted using an average over the particular sequence. The results of this temporal fusion are shown in the confusion matrices seen in Table 4 and 5. For the audio classification, this procedure resulted in an over all accuracy of 76.4 % and an F_1 value of 0.702 for the positive class and 0.805 for the negative class. In case of the bio-physiological data, we achieved an accuracy of 72.2 % and an F_1 value of 0.756 for the positive sequences and 0.677 for the negative class.

These temporally combined decisions for both modalities audio and bio-physiology are further combined in a decision fusion step and yield the following results: an overall accuracy of 77.8% is achieved with an F_1 score of 0.712 for the positive class and 0.805 for the negative one.

Additionally, a certified FACS-coder observed all changes in facial musculature as described in the FACS manual. Conducted Wilcoxon rank sum tests showed statistically significant differences of the distribution of the occurring action units AU1 ($p = .042$), AU10 ($p = .041$), AU14 ($p = .017$), and AU28 ($p = 0.020$) when comparing ES-2 and ES-5. Moreover, the observed mean values were significantly lower in ES-2 than in ES-5.

7 Summary and Discussion

In the presented study information fusion techniques are applied in order to improve results for multimodal emotion classification. We conducted an utterance based classification using seven distinct audio feature sets yielding a high accuracy of 74.3 %. On the other hand using bio-physiological data only 49.2 % of the five second long clips were correctly classified, suggesting that emotion recognition in the utilized channels requires more information than the one that is present in these short windows.

In a further step a temporal fusion of both modalities was conducted, improving both results significantly. Especially the bio-physiological approach improved to an accuracy of 72.2 %.

Observing the confusion characteristics seen in Table 4 and 5, reveals opposing error behavior. Hence, a decision fusion of both could yield much improved results. The decision fusion, however, only slightly improved the overall result, which indicates that more effort needs to be put into this particular step.

The results considering the facial expression analysis have not yet incorporated into the information fusion framework. The presented tests nonetheless reveal a significant differences between the considered classes, which makes the video modality a promising source for classification in this application.

Acknowledgment

This research was supported in part by grants from the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). The work of Martin Schels is supported by a scholarship of the Carl Zeiss Foundation. Miriam Schmidt is supported by a scholarship of the graduate school Mathematical Analysis of Evolution, Information and Complexity of the University of Ulm.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. Springer, Heidelberg (2006)
2. Brown, G., Kuncheva, L.I.: "Good" and "Bad" diversity in majority vote ensembles. In: El Gayar, N., Kittler, J., Roli, F. (eds.) MCS 2010. LNCS, vol. 5997, pp. 124–133. Springer, Heidelberg (2010)
3. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing 28(4), 357–366 (1980)
4. Dietrich, C., Schwenker, F., Palm, G.: Classification of time series utilizing temporal and decision fusion. In: Kittler, J., Roli, F. (eds.) MCS 2001. LNCS, vol. 2096, pp. 378–387. Springer, Heidelberg (2001)
5. Ekman, P., Friesen, W.V.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)

6. Hermansky, H.: The modulation spectrum in automatic recognition of speech. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 140–147. IEEE, Los Alamitos (1997)
7. Hermansky, H., Hanson, B., Wakita, H.: Perceptually based linear predictive analysis of speech. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1985), vol. 10, pp. 509–512 (1985)
8. Kim, J., André, E.: Emotion recognition based on physiological changes in music listening. IEEE Trans. Pattern Anal. Mach. Intell. 30, 2067–2083 (2008), http://portal.acm.org/citation.cfm?id=1477073.1477535
9. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley Interscience, Hoboken (2004)
10. Maganti, H.K., Scherer, S., Palm, G.: A novel feature for emotion recognition in voice based applications. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 710–711. Springer, Heidelberg (2007)
11. Peter, J., Lang, M.M.B., Cuthbert, B.N.: International affective picture system (iaps): Affective ratings of pictures and instruction manual. Tech. rep., NIMH Center for the Study of Emotion & Attention, University of Florida (2008)
12. Picard, R.W.: Affective Computing. MIT Press, Cambridge (2000)
13. Platt, J.: Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in Large Margin Classifiers, 61–74 (1999)
14. Rabiner, L.R., Schafer, R.W.: Digital processing of speech signals. Prentice-Hall Signal Processing Series. Prentice-Hall, Englewood Cliffs (1978)
15. Russell, J.: A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178 (1980)
16. Schwenker, F., Dietrich, C., Thiel, C., Palm, G.: Learning of decision fusion mappings for pattern recognition. International Journal on Artificial Intelligence and Machine Learning (AIML) 6, 17–21 (2006)
17. Smyth, P.: Clustering sequences with hidden markov models. Advances in Neural Information Processing Systems 9, 648–654 (1997)
18. Tax, D.M.J., van Breukelen, M., Duin, R.P.W., Kittler, J.: Combining multiple classifiers by averaging or by multiplying. Pattern Recognition 33(9), 1475–1485 (2000)
19. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York (1995)