Towards Multimodal Detection and Classification of Emotional Patterns in Human-Machine Interaction – Results of a Baseline Study

Gregor McGlaun, Frank Althoff, Manfred Lang, and Gerhard Rigoll

Institute for Human-Machine Communication
Technical University of Munich
Arcisstr. 16, 80290 Munich, Germany
phone: +49 89 289-28541

{mcglaun, althoff, lang, rigoll}@ei.tum.de

ABSTRACT

The study described in this contribution represents the first stage towards a new integral approach in which multimodal (i.e. visual, acoustical, and tactile) information is used for evaluating emotional patterns. One aim was to collect data material of real user emotions provoked in reproducible scenarios. We gathered phrases the subjects uttered spontaneously during the test and compared them to their counterparts spoken in an acted emotional state. Moreover, we investigated the relevance and the quota of visual and acoustical features regarding the classification of emotions. Subjects had to evaluate emotional patterns provided by still pictures, sound tracks, video tracks, and audiovisual clips. For this, they had to use a special three-dimensional classification scheme. Our study showed that natural emotions regarding the visual and the acoustical features significantly differ from acted ones regarding parameters of the voice (f_0, amplitude) and visual features (eye movement, blink frequency, gaze retention periods). As a further result, we obtained that visual and acoustical features complementarily, in some cases even redundantly contribute to the classification of emotional patterns. Often, the visual information serves as a pre-classifier. The multimodal combination of both sources showed the highest recognition rate of emotional patterns.

Keywords: Emotion, behavior, multimodal, features, classification, human-machine interaction, human error

1. INTRODUCTION

Many systems are capable of evaluating speech and even gesture commands besides the classical tactile input paradigms so far, but a highly reliable automatic recognition and evaluation of emotional patterns in user behavior is still one of the great goals in human-machine interaction. The benefit of emotional information appears in numerous domains. Considering the automotive environment in which the driver interacts with audio and communication devices while going by car, it is very important to distinguish between her/his different emotional states. When the driver is negatively affected, for example, the length of the system replies could be varied, i.e. shortened. Thus, the driver's workload is kept down, and the error robustness of the system is enhanced [1]. In high security environments (e.g. power stations or oil platforms), emotions can negatively influence the concentration of the user and cause dangerous situations. By a respective monitoring, the system can intervene, prohibit the user from further interacting and start respective measures of security. Emotions can also be used as a basis to adapt sales strategies in e-commerce [2]. In medicine, recognizing emotional outbursts of patients or of the elderly helps to improve monitoring anxiety or pain states. Further applications employ emotions to control the behavior of animated interactive agents [3]. Particularly, in some tactile input devices, like touch-screen or mouse, emotional patterns (based on features, like the frequency and intensity of pushing) are evaluated [4].

2. FORMAL DESCRIPTION

Our baseline study was motivated and based on numerous theories and trials that have been developed and carried out to classify and evaluate emotions. Before our test is described in detail, we take a glance at former research work in this topic:

Overview and Categorization

Approaches to detect and evaluate emotions can largely be categorized in *invasive* and *non-invasive* techniques. In invasive investigations, trials have been carried out by applying small amounts of stimulating substances (e.g.

adrenaline) to the subject (Schlachter's and Singer's experiment, 1962). In other methods, various sensors and detectors are attached to the human body, e.g. in electro-cardiographic (blood pressure) or electro-myographic evaluations (regarding galvanic skin response). But in many target application domains, the user cannot be expected to attach any kinds of sensors or detectors to his body (e.g. in the automobile environment). Therefore, our study is based on non-invasive methods, in which mainly facial expressions, voice intonation, gestures, or body temperature (via infrared detection) are investigated. Most of approaches in this area have in common that they were strictly monomodal, i.e. they focused exclusively on one source of information. Based on the investigations of Tomkins (1962), Ekman and Friesen (1971) developed a facial feedback theory. It includes the *Facial Action Coding System* (FACS), which is a comprehensive, anatomically based system for measuring all visually discernible facial movements [5]. FACS comprises 44 unique *action units* (AUs) by which all visually distinguishable mimic expressions, head and eye positions can be described. In speech based methods, mostly the f_0 and the pitch are analyzed as characteristic features of voice [6]. The speech signal is normalized to a defined neutral state. In newer adaptive systems, this neutral state is steadily leamed and updated by means of a dedicated user model, which is created upon different knowledge bases (e.g. comprising a dialog history).

Natural versus Imitated Emotions

A large set of investigations has been done on the basis of artificially played emotions, but even actors often could only approximately imitate emotional states. With emotional patterns occurring spontaneously, one aspect of the baseline study was to develop reproducible scenarios in which real emotions were systematically aroused. In the scope of the study, we focused on the six *basic emotions* proposed by Ekman, Friesen, and Ellsworth [7], which are *anger, disgust, fear, happiness, sadness,* and *surprise.* On the basis of this data material, an effective quantitative analysis of the emotional features could be carried out.

Primary and Secondary Emotions

According to Leventhal [8], emotions can largely be divided into *primary* and *secondary emotions* regarding their formation phase. Primary emotions are the ones that humans spontaneously express after an event (e.g. reactions, like reflexes). Emotions that proceeded by larger cognitive processes are subsumed under secondary emotions. In human-machine interaction, the user often does not utter any phrases during cognitive processes, whereas image based methods still lack of accuracy in feature extraction. Depending on the actual scenario, either the image- or the speech-based feature extraction will be more effective. Thus, a system containing a multimodal combination, including facial actions, speech, and tactile features (if implicated by the particular application domain) is likely to promise a significant increase of the robustness of emotion recognition and a more flexible way to adapt to different types of scenarios.

Multimodal Classification of Emotional Patterns

With human beings using both their sense of seeing and hearing, another motivation was to investigate, how both senses work together in the classification of emotions. Thus, a further goal of our study was to delimit the most important visual and acoustical features and to analyze the way that they contribute to the categorization of emotional patterns. In the past, there have been many empirical approaches using different kinds of classifications. R. Woodworth (1938) applied a one-dimensional classification [9], and P. Lang (1995) made a similar trial on the basis of a two-dimensional scheme [10]. As emotional patterns vary depending on the individual and cultural differences, there are no defined stereotypes for emotional patterns. Thus, in our study, we introduced a new method of categorization. Test persons had to characterize emotional patterns, using three unique dimensions (valence, potency, and activity). The model contained a continuous semantic differential scale [11], so the appropriate rating along one dimension is done by choosing a numerical value. The ranges of the dimensions were, as follows: negative to positive (evaluation), weak to powerful (potency), and calm to excited (activity).

3. EXPERIMENTAL SETUP

The first step how to effectively use different kinds of information streams for recognizing and evaluating emotional patterns is to investigate particular methods and strategies human beings apply. In the study, we focused on visual and acoustic features. We use the results as a prototypical basis on which a respective multimodal approach for classifying and evaluating emotional patterns is developed.

Test Environment

The user study was carried out at the usability laboratory of the institute. In the observation room, the experimental subjects were supposed to deal with the experimental setup. They sat at a table in front of two remotely controlled video cameras. One camera recorded the whole upper part of the body. The second exclusively focused on the face. For later analysis, during the test, we permanently recorded the subjects' actions by video and audio equipment. We prepared a special test platform consisting of a table microphone, a keyboard with a mouse, and a screen on which the single test tasks were displayed. These devices were connected to a computer in a separate control room. Here, the test supervisor monitored the whole trial. He started the single tasks of the test by operating a semi-automatic tool, which read and displayed the

content of the task to the participant. Additionally, he could observe the subject's behavior via a one-way mirror.

Test Procedure

The study consisted of three main parts. In the first part, we developed a set of reproducible scenarios aiming at the arousal of spontaneous, natural patterns of emotions. To get unaffected results, from the beginning, the test subjects had no idea that the study actually dealt with the analysis of emotions. They were rather lead to believe that the trial is concerned with gathering some empirical data for the evaluation of a commercial multimodal system that can be operated by classical tactile input modalities (mouse, keyboard), but also via natural speech (which was important to decrease command-like expressions). To vindicate the camera setup, we told the participants that the system also comprised a gaze-tracker for which training data is needed. From the subjects' point of view, the particular arousal scenarios were masked in means of single tasks. Hence, according to a predefined run chart, the supervisor systematically tampered with the actions of the subjects.

There was a total of six different scenarios which we expected to cover the whole set of basic emotions.

Scenario 1 (E-Questionnaire): In the first task, we prepared an electronic questionnaire. With the participant filling in the entry mask, the test supervisor remotely interfered by adding single letters, or even erasing complete lines.

Scenario 2 (Video Clip): The test participant was asked to watch a short scene of a video. To have the subject's concentration on the scene, (s)he was told that her/his eye movements were recorded and that (s)he would be asked some questions about the scene afterwards. During the clip, we suddenly increased the volume of the video to a very high level.

Scenario 3 (Mathematics): In a mathematical test, the subject had to orally solve ten very simple mathematical problems (addition and subtraction of integers, respectively). The answer had to be given as quickly as possible. But instead of the actual result, in each case, the test supervisor told the participant that only one of the ten tasks had been solved correctly. Moreover, the test supervisor provoked the subject by different comments and questions.

Scenario 4 (O/G-Stimulation): For gathering data material of natural facial expressions, we applied olfactory and gustatory stimulants. The subject was told that a company would make an ascertainment on the perception of different smells and tastes. (S)he got the task to categorize samples of different smells and tastes, using a scale with a bipartite semantic differential (pleasant, good, neutral, bad, disgusting). In particular, the drinks were multivitamin juice, water, lemonade, wine, and salty water. For the smells, we chose perfume, coffee, garlic, vinegar, and an acidic cleaner.

Scenario 5 (Advertisement): The goal of this scenario was to get distinctive facial expressions of happiness. Thus, we presented a short clip with a funny advertisement. To distract from the actual aim of the study, the subjects should change the screen size of the video player by different speech commands, and they were asked some questions on the clip afterwards.

Scenario 6 (Computer game): In the last task, the subjects had to participate in a computer-game competition (a TETRIS-like game was applied). To increase their motivation, we told the test persons that they would be awarded by a gift of money, if obtaining a certain minimum score. For controlling, the subjects could use speech as well as tactile input. While they were playing, the test supervisor once more manipulated the subject's input. In particular, the supervisor firstly deteriorated the result, but in the end, let every of the subjects win the game.

At the end of the first part, the participants were told about the actual aim of the study and enlightened that their actions have been sabotaged. The subjects had to fill in a questionnaire, in which we asked them to describe the emotional states they felt during the particular scenarios.

In the second part of the test, a cross-examination between natural and artificially played emotions was carried out. During the first part of the trial, the test supervisor had selected a set of phrases the participant used, when emotionally affected. The test supervisor showed each of the recorded video clips to the subject. We asked the participant to characterize her/his own emotional state shown in the video. On the one hand, (s)he had to assign one of the basic emotions to the clip; on the other hand, the three-dimensional classification scheme described above (see 2.) should be applied.

To get a basis of comparison, the subject had to reproduce each of the shown phrases in a particular emotional way. First, to get a defined neutral state, (s)he just had to read the phrase from a prompter. Then we asked the subject to speak the phrase in the different basic emotional states.

The last part of the test was concerned with the contribution of visual and acoustic features in emotional patterns. First, in random order, a total of ten pictures was presented to the subject for five seconds. Each picture showed the facial expression of a person in a particular emotional state. Moreover, we evaluated the power of acoustic information on the classification of emotions. The test subjects just heard the sound of five video cuts containing scenarios with persons showing different natural emotional states. The language in these clips was German. Subsequently, in random order, the same scenarios in terms of a video clip were presented. This time, there was just the visual information, but no sound was played with the sample. Finally, the clip with both the visual and the acoustical information was presented for a cross-evaluation. The subject had to classify the pictures as well as the sound files, the visual and audiovisual sce-

narios by means of the 3D-model and also by assigning one of the basic emotional states.

4. RESULTS

A total of 18 subjects participated in the study. 16 of them were graduate students, and 2 of them were PhD candidates. 17 test participants were male, and one was female. The average age of the subjects was 25a (ages ranging from 21a to 31a). As the intensity and the way of emotional expressions depend on patterns that are closely connected with cultural behavior, we acquired test subjects from 14 different nations. The participants came from Germany (5x), Lebanon, Vietnam, Israel, Turkey, Korea, Bangladesh, Indonesia, Thailand, Japan, Tunisia, Croatia, India, and Malta. Four test persons of the non-German group were also able to speak and understand the German language. Thus totally, one half of the test subjects could not speak and understand German.

In the first part of the test, a huge set of natural emotional reactions could be ascertained in each of the described scenarios.

While filling in the questionnaires (scenario 1), 78% firstly thought that they caused the errors while typing. 44% of the subjects defined themselves to be in neutral state during the whole task. Five subjects pointed out that their emotional states changed rapidly after the errors occurred increasingly often. They switched from neutral to surprise and finally to anger. Two of them even kept themselves amused with seeing the lines disappear.

In scenario 2, most subjects showed surprise (39%) or anger (28%), when the volume was increased for the first time. But when the same effect recurred, half of the subjects did not react at all.

When the participants were informed about their low performance in the mathematical test (scenario 3), 11 subjects made objections and started a discussion with the test supervisor, whereas 7 accepted the results without any complaints, but each of them showed a respective facial expression (anger or surprise). 44% of the participants stated that they were surprised when they heard about the result, 22% rated their emotional state as angry. According to FACS, in this scenario, we could state mostly combinations of AUs 4, 5 or 9 in combination with AUs 25-27. Moreover, the visual and the acoustical information were strongly complementary. 13 of 18 test persons showed a cognitive process (up to 7 s), until it came to a verbal reaction. Thus, before vocal features could be realized, visual parameters of an emotional pattern could be detected.

The scenario in which facial expressions were aroused by olfactory and gustatory stimulation furnished interesting results concerning a black-and-white decision between positive and negative emotions. Earlier studies (e.g. [12]) point out that there is a close association between tastes or smells and a respective emotional stimulus of facial expressions. The facial actions could be detected, as follows: For tastes or smells that were rated “pleasant” or “good,” the most frequent AUs were 1, 11, and 18, whereas if rated “bad” or “disgusting,” subjects showed AUs 2, 4, 9, 10, 15, and 20. With neutral tastes like water, also AU 18 and 23 was ascertained. The subjects rated their individual emotional states happy (50%), when tasting the sweet drink (multivitamin) and sniffing the perfume (66%). 28% pointed out that after a short state of surprise, they felt disgust when smelling the container with the acidic cleaner.

In scenario 5, we get strong emotional reactions regarding patterns of happiness. Facial actions were redundantly linked with vocal expressions. In the subjective rating, 55% pointed out that they were in neutral state, and only 22% specified that they experienced happiness.

During the computer game (scenario 6), significant utterances, like interjections (“oh”, “shit”) with respective changes of the facial expressions appeared. In this scenario, the visual and the acoustical expressions were nearly simultaneous and strongly redundant.

Concerning the second part of the study, we compared the natural emotional patterns to the corresponding imitated emotional states of the participants. The null hypothesis H_0^{voc} that characteristic vocal features do not differ from natural to imitated emotional patterns could be rejected (at $p < .01$). Considering all subjects and normalizing to the reference values of the neutral state, the contour of f_0 showed variations up to 32%, the intensity ranged 52% at maximum, as well as the speech rate (max. 42%), and the volume (max. 62%). In figure1, clear differences in pitch, intensity ranges and f_0 can be seen in an example of the test.

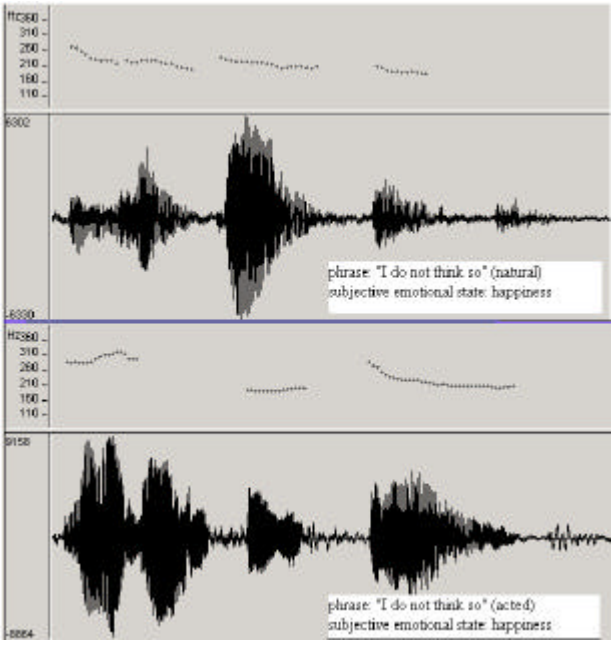

Figure 1: f_0 and waveforms of natural and respective acted phrase (female subject)

Just as well, the respective null hypothesis H_0^{vis} that features of facial actions do not differ from natural to imitated emotions could be denied (at $p < .001$). With 89% of the subjects, identical AUs of imitated emotions were less intense then real ones. Moreover, 78% of the test subjects had a lower blink frequency, when they simulated patterns of fear.
Concerning the classification of the single still pictures in the third part of the study, the test persons often had big difficulty in making a definite characterization of the shown emotion. (4 of the 10 presented pictures can be seen in figure 2a-d).

Figure 2a: test picture P1

Figure 2b: test picture P2

Figure 2c: test picture P3

Figure 2d: test picture P4

Regarding the classification results, there is a large variance of the ratings in P3 and P4 (see table 1, * is the actual state). In case of P3, the subjects pointed out that the half-frontal view made it difficult to classify (33%), and that a picture sequence (89%) or some vocal information (50%) would have helped to reach a definitive decision. P4 was even associated with contradictory types of emotions (happiness versus fear and disgust). All test participants mentioned that the closed eyes of the person in the picture aggravated a rating of the emotional pattern, and that at least, some context information (78%) or previous knowledge of the scene (56%) is needed.

	P1	P2	P3	P4
neutral	16*	0	1	0
happiness	0	0	0	9*
sadness	1	16*	2	0
anger	0	0	7	0
fear	0	1	2	5
disgust	0	1	2	4
surprise	1	0	4*	0

Table 1: Subjective rating of the pictures P1 to P4

Finally, we evaluated the contribution of visual and acoustical information in the classification of emotions. As a general result, we found out that the subjects rather used the acoustical information in cases of strongly extroverted primary emotions, like anger (89%), even if there were perceptible facial expressions. On the other hand, they used visual information for the detection of emotions strongly based on cognitive processes (83% of the subjects mentioned sadness as an example). Moreover, 78% stated that the facial information often served as a pre-classifier, and the final decision was made upon characteristics of the voice. As the language of the video clips was German, expectedly, the group that understood the German language (9 subjects) showed a better classification result (at 43%) listening to the audio files than those who did not have the semantic information. But 94% of all test persons pointed out that, for example, interjections helped them with the classification. 13 of 18 subjects did not have an identical classification for the visual, the acoustical, and the combination of both streams, respectively. In nearly all scenarios (except for the sadness state), the clips providing audiovisual information had the best recognition rate (see figure 3).

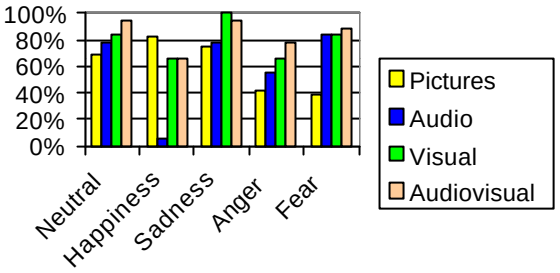

Figure 3: comparison between the recognition rates of emotions when using different streams of information

The results clearly prove that a human being obviously uses both its visual and acoustical sense for the classification of emotional patterns. Thus, exploiting the multimodal input stream will give promising results with regard to a more reliable and error-robust detection and classification of emotional patterns, respectively.
In the overall evaluation of the scenarios and the subjects, we obtained the quantitative 3D-classification scheme of emotional patterns, as follows:

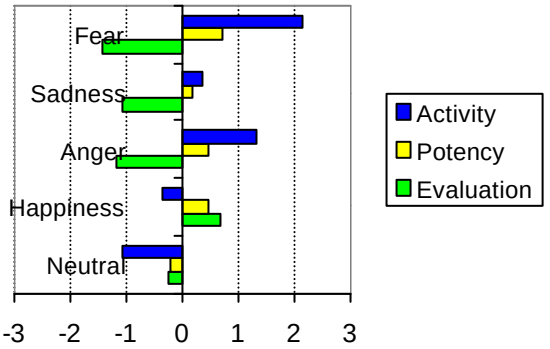

Figure 4: Distinction of the emotional states in the 3D-classification scheme (-3 and 3 are the scale limits)

The bias of the neutral state results from the fact that many of the subjects rated this mood as a strong state of relaxation. Normalizing the other states to neutral, the scheme provides a useful distinction between the particular states.
Based on the results of the questionnaires, we could provide a table with the most important qualitative features of emotions in acoustic and visual information streams (see table 2).

	Acoustic Features	**Visual Features**
Evaluation	progression of f_0	movements of eyes, mouth, and eyebrows
Potency	changes in pitch, amplitude	duration of facial expression
Activity	degree of verbosity	gaze retention period, blink frequency

Table 2: Important features for classifying emotions in the particular evaluation dimensions

Not surprisingly, all features are differential variables, i.e. factors that are obtained by observation of picture sequences or changes in voice. This also confirms the comparatively bad recognition rates of the still pictures and the respective subjective statements discussed above.

5. CONCLUSIONS AND FURTHER WORK

The study showed that, regarding the recognition of emotional patterns, it will be very promising to use a multimodal detection and classification scheme. The subjects used the information in a complementary or a redundant way, which improved the recognition rate in comparison to the monomodal classification. Evaluation strategies of emotional states will be carried out on the basis of natural emotions, as they significantly differ from acted ones (visual and acoustic features). The applied 3D-scheme proved to be a usable non-deterministic way of classifying emotions. Based on the results of the study, we currently develop an automatic multimodal emotion recognition system to be implemented in an existing system architecture [13]. The respective results of the monomodal emotion recognizers will be merged via a Late Semantic Fusion and weighted due to the particular situation. To effectively manage the different information sources, FIFO queues based on a slot-and-filler mechanism will be applied. The classification process will be influenced by a dedicated user model containing additional context parameters, the dialog history and diverse status variables. With this approach, redundant and complementary combinations can be effectively evaluated to increase the recognition robustness of user emotions. The system is to be used in applications of medical technology as well as in virtual reality and automotive environments.

6. REFERENCES

[1] Project FERMUS: Error Robust Multimodal Speech Dialogues, cooperation between the Technical University of Munich, the BMW Group, the Siemems-VDO AG, and the DaimlerChrylser AG, website: www.fermus.de

[2] F.N. Egger: "Affective Design of E-Commerce User Interfaces: How to Maximise Perceived Trustworthiness", Proceedings of The International Conference on Affective Human Factors Design, Asean Academic Press, London, 2001

[3] E. André et al.: "Exploiting Models of Personality and Emotions to Control the Behavior of Animated Interactive Agents," Proceedings of the workshop on "Achieving Human-Like Behavior in Interactive Animated Agents" in conjunction with the Fourth International Conference on Autonomous Agents, Barcelona, 2000

[4] Project BlueEyes: "The Emotion Mouse," IBM Almaden computer science research, website: www.almaden.ibm.com/cs/blueeyes

[5] P. Ekman et al.: Final Report to NSF of the Planning Workshop on Facial Expression Understanding, National Science Foundation, Human Interaction Lab, UCSF, CA 94143, 1993

[6] B. Schuller et al: "Automatic Emotion Recognition by the Speech Signal," Proceedings of the SCI 2002, 6th World Multi-conference on Systemics, Cybernetics, and Informatics, Orlando 2002

[7] P. Ekman et al. "Emotions in the Human Face," Cambridge University Press, London, 1982

[8] H. Leventhal: "Perceptual-motor procession model of emotion", in "Perception of Emotion in self and others," Plenum Press, New York, 1979

[9] R. Woodworth: "Experimental psychology," Henry Holt, New York, 1938

[10] P. Lang: "The Emotion Probe: Studies of Motivation and Attention," American Psychologist, 1995 pp. 372-385

[11] Jakob Nielsen, Usability Engineering, Morgan Kaufmann Publishers, Inc., San Francisco, California, 1999, pp. 93-114

[12] J. O'Doherty et al.: "Representation of Pleasant and Aversive Taste ion the Human Brain," Journal of Neurophysiology 85, 2001, pp. 1315-132

[13] Gregor McGlaun et al.: "A new approach for Integrating Multimodal Input via Late Semantic Fusion," VDI/VDE-Proceedings of USEWARE 2002, Darmstadt, 2002