Proceedings Second International Symposium on New Frontiers in Human-Robot Interaction, K. Dautenhahn, J. Saunders (Eds.),
at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Mirror my emotions! Combining facial expression analysis and synthesis on a robot

Stefan Sosnowski[1] and **Christoph Mayer**[2] and **Kolja Kühnlenz**[3] and **Bernd Radig**[4]

Abstract. Everyday human communication relies on a large number of different communication mechanisms like spoken language, facial expressions, body pose or gestures. Facial expressions are one of the main communication mechanisms and pass large amounts of information between human dialogue partners [22]. Therefore, the analysis and the synthesis of facial expressions are important steps towards an intuitive human-machine interaction and form valuable research targets. We present a system that tackles both challenges. It relies on a fully automated, model-based, real-time capable approach to distinguish universal facial expressions and their intensities from camera images. Facial expression synthesis is conducted via the robot head EDDIE, a flexible low-cost emotion-display with 23 degrees of freedom. Static facial expressions at continuous intensities are included, as well as smooth transitions based on the circumplex model of affect. Miniature off-the-shelf mechatronic components are used to provide high functionality at low cost. Evaluations conducted in a user-study show that emotions displayed by EDDIE are recognizable by humans very well. By combining facial expression recognition and display on the robot, a first demonstration is presented in which the robot mirrors the human's emotions, as a basis for further research in the field of emotional closed loop systems.

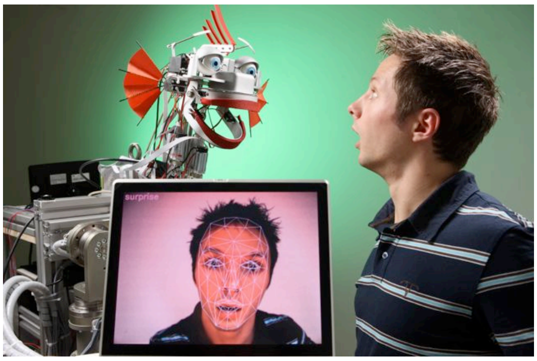

Figure 1. Demonstration setup with EDDIE (foto: Kurt Fuchs).

1 Introduction

Robots will be an integrated part of future living or working environments making fast and efficient human-robot interaction important. Traditional human-human interaction passes important information on many different communication channels like the auditory channel for spoken language or the visual channel for facial expressions. In contrast, traditional human-machine interaction via mouse and keyboard is often considered slow and non-intuitive. Today, dedicated hardware often derives the emotional state from blood pressure, perspiration, brain waves, heart rate, skin temperature, etc [5, 14, 13]. Humans, however, interpret emotion via video and audio information that is acquired without interfering with other persons. Fortunately, this information can be acquired by computers as well using general purpose hardware. For a comprehensive overview on applications in emotion recognition, we refer to Lisetti et al. [10]. Yet, recent research aims at enabling machines to utilize communication channels more natural to human beings, such as gestures or facial expressions. Knowledge about human behavior, intention, and emotion is necessary as well to construct convenient interaction mechanisms.

We present a robot system with combined facial expression recognition and synthesis, that aims at providing communication via facial actions between the robot and the human and therefore partially filling the gap between human-human and human-machine interaction. Facial expressions form an important aspect of human-human communication and pass important information about the interaction partners' emotional states, which robots should be able to utilize to understand and eventually influence the human interaction partner. The proposed system recognizes facial expressions from camera images, processes the information encoded in the facial expression and reacts with facial expressions of the robotic head EDDIE . In a preliminary version, the facial expression recognized from the human is merely mirrored on the robotic head, see Figure 1. However, since facial expressions are strong communication signals in human-human interaction, more complex reactions according to psychological insights will be integrated in the future.

[1] Institute of Automatic Control Engineering, Department of Electrical Engineering and Information Technologies, Technische Universität München, Munich, Germany email: sosnowski@tum.de
[2] Chair for Image Understanding and Knowledge-Based Systems, Computer Science Department, Technische Universität München, Munich, Germany email: mayerc@in.tum.de
[3] Institute for Advanced Study (IAS) and Institute of Automatic Control Engineering, Department of Electrical Engineering and Information Technologies, Technische Universität München, Munich, Germany email: koku@tum.de
[4] Chair for Image Understanding and Knowledge-Based Systems, Computer Science Department, Technische Universität München, Munich, Germany email: radig@in.tum.de

1.1 Related Work

Integrating human-inspired forms of communication in robotic systems will allow for a human-machine interaction with increased efficiency, security and robustness and ease of use [16]. The work presented in this paper presents a fusion of two research areas: facial expression recognition and facial expression synthesis. In this paper, we present a combination of two systems that have already be inspected in the past independent of each other. Please note, that the main contribution of this work is not to present a novel way of facial expression analysis or synthesis but rather to introduce a system that combines both approaches in a closed loop. Both are based on the Facial Action Coding System (FACS), which precisely describes the muscle activity within a human face that appear during the display of facial expressions [3]. Ekman and Friesen find six universal facial expressions that are expressed and interpreted independent of the cultural background, age or country of origin all over the world, see Figure 2 [2].

1.1.1 Facial Expression Recognition

Pantic et al. [12] propose a three-step approach for facial expression recognition which we also follow in our work. Firstly, the position and shape of the face visible in the image is determined. Secondly, features descriptive for facial expressions are extracted. Finally, in the third step a classifier is applied to identify the expression. Some approaches infer the facial expression by exploiting a face model. Kotsia et al. take this approach by fitting a face model to the example images showing either a neutral face or a strongly displayed facial expression and extracting the model parameters [8] to apply rules stated by Ekman and Friesen [3].Similar to our approach, Gokturk et al. rely on person-specific information of a neutral face and track a three-dimensional model in image sequences [4] . In contrast, they utilize non-public image data for building the face model, training the classifier and evaluating their approach. De la Torre et al. [9] extract temporal segmentation of facial behavior from image data to assist professional FACS coders. They demonstrate their approach on image data that has not been taken with a computer vision application in mind and therefore face challenging context conditions. However, they rel y on person-specific active appearance models (AAM) which renders the approach not applicable to previously unseen data. In contrast, we utilize publicly available data throughout

Figure 2. The six universal facial expressions as they occur in the Cohn-Kanade Facial Expression Database.

Figure 3. The face model is fitted to each image in order to estimate the facial expression currently visible.

the whole approach. The accuracy of model fitting is of great importance to model-based approaches since the accuracy of facial expression recognition depends directly on the model's capability to reduce the image data and reflect its content. Pure image data without prior data reduction is also applied for facial expression classification. Littlewort et al take this approach [27]. They utilize AdaBoost and Support Vector Machines to determine the facial expression from convolutions of image data with Gabor energy filters and their approach obtains very high recognition results. Local binary patterns are applied in recent work by Shan et al. [28]. These approach suffer from large changes in face appearance which is usually not provided by standard databases.Similar, Michel et al extract the motion of 22 feature points within the face between a neutral image and an image that displays a strong facial expression to train a Support Vector Machine (SVM). [11].

1.1.2 Facial Expression Synthesis

Facial expression synthesis is utilized in virtual agents and robots for social interaction. In virtual systems there are two major methods: One is an abstraction based approach, using blend shapes, form interpolation or deformation based morphing. The other is muscle based, being separable in physical mass-spring systems and muscle group based approaches like FACS [26]. In robotic systems, mostly FACS is used (e.g. Kismet [16] , Saya [25]).

2 Facial Expression Recognition

We rely on model-based techniques to determine the exact location of facial components such as eyes or eye brows in the image. Geometric model form an abstraction of real-world objects and contain knowledge about their properties, such as position, shape or texture in a model parameter $\mathbf{p}$, see Figure 3. The Candide-3 face model consists of 116 anatomical landmarks and has been specifically designed for facial expression recognition [8]. Its parameter vector $\mathbf{p} = (r_x, r_y, r_z, s, t_x, t_y, \sigma, \alpha)^T$ describes the affine transformation $(r_x, r_y, r_z, s, t_x, t_y)$ and the deformation (σ, α) [1]. The 79 deformation parameters indicate the shape of facial components such as the mouth width or the roundness of the eyes and partially refer to the Facial Action Coding System, see Figure 4. Model-fitting is the computational challenge of determining correct model parameterizations for single images without prior knowledge of the image content. We rely on the work of Wimmer et al. to tackle this challenge [15].

2.1 Feature Extraction

Two aspects generally characterize facial expressions: they turn the face into a distinctive state and the involved muscles show a distinctive motion. Our approach considers both aspects by extracting static and person-adapted features and assembles them in a feature vector $\mathbf{x}$. Since we consider the transformation parameters not related to the facial expression, they are not considered in the feature assembly. The static features are assembled by calculating the model parameters for a single image and extracting the deformation parameters σ, α. To calculate the person-adapted features, we acquire a neutral reference image of the person visible in the image sequence and determine the model parameter change $\Delta\sigma, \Delta\alpha$ from the reference image to the inspected image.

2.2 Classification

The aim of this step is to infer the facial expression currently visible from the extracted features. For training purpose, the Cohn-Kanade Facial Expression Database [6] and MMI Face Database [7] serve as training and testing data. They provide the facial expression label for every image sequence. To determine the expression intensity, we manually annotate the facial expression starting, peak and ending image index in every sequence and model the facial expression intensity to increase or decrease linearly. The intensity is defined to range between 0 (neutral image) and 1 (peak expression image). We apply decision trees and Support Vector Machines (SMVs) to map the feature vector $\mathbf{x}$ on the facial expression visible in a single image. Since the databases do not provide the point of transition between a neutral face and a displayed facial expression within a single sequence, we decide to use all images with an intensity less than 0.33 as neutral face representatives. Once the facial expression is determined, a model tree is trained to infer its intensity from $\mathbf{x}$. Model trees are flow-chart-like tree structures, in which internal nodes denote a test on a feature to branch the flow of the computation into one of the two subordinate nodes [20]. Furthermore, the leaf nodes contain line segments. Advantages of Model Trees over other means of classification are the low run-time and fast generation of the tree from a previously acquired set of training data.

2.3 Experimental Evaluation

Every sequence in the image databases starts with a neutral face and develops into the apex expression. In contrast to the image sequences of the Cohn-Kanade Facial Expression Database, the MMI Face Database image sequences also show decreasing facial expression intensity until a neutral face is reached again. We fit the face model to each image and compute the static and person-adapted features for each image individually.

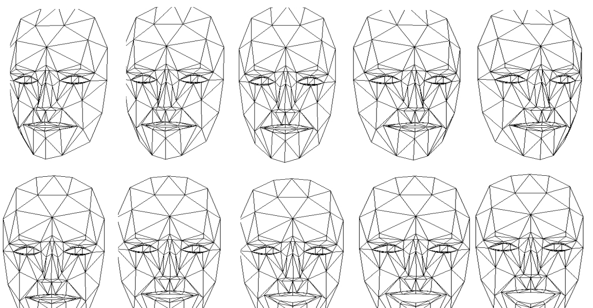

Figure 4. The face model reflects face properties such as gaze direction or facial component shapes.

We provide the classification algorithm with different sets of features. The classifier C1 is trained on the static features only and therefore it is applicable to single images, without requiring a comparative image depicting a neutral facial configuration. The classifier C2 is presented the person-specific features only and adapts to the person visible in the image. The classifier C3 includes both static and person-specific features, with the intention of combining the "best of both worlds". The classifiers are evaluated by a cross-validation of ten independent subsets of the training data, where nine sets are used to train a classifier and the remaining set is used for testing purposes. The intensity estimation is evaluated by correlation between the value calculated by the model tree and the image groundtruth, see Table 1.

	C1	C2	C3
decision tree	83.6% / 91.8%	84.7% / 91.8%	86.1% / 92.0%
SVM	87.9% / 93.1%	88.4% / 92.7%	91.4% / 87.8%

database	Cohn-Kanade	MMI
intensity correlation	92.8%	93.7%

Table 1. Recognition accuracies for different feature sets and classifiers. The numbers are given for Cohn-Kanade Facial Expression Database and MMI Face Database, respectively.

3 Facial Expression Synthesis

According to Mori's theory of the uncanny valley, a robotic interaction partner is considered to be visually most unpleasant, if it is designed to look very similar to a human, but not entirely human. Therefore, in order to design a comfortable robotic interaction partner, a reasonable degree of familiarity should be achieved, balancing well between a machine-like and a human-like appearance which guided the design process of EDDIE [18].

A distinctive feature of this robot head is that it includes not only human-like facial features, but also animal-like features. The ears can be tilted as well as folded/unfolded and a crown with four feathers is included. For further details on the actual design of the head, please refer to [21]. EDDIE has a total of 23 degrees of freedom. To acquire video data from a human interaction partner, IEEE1394 Point-Grey Dragonfly cameras are integrated in the eyes, covered with natural looking eyeballs. The mounting in the eyes ensures full frontal images of the interaction partner, with the partner normally fixating the gaze on the eyes of the robot. To compensate for movements, the eyes are movable in two degrees of freedom (pan/tilt). Video data is streamed with a resolution of 640x480 at 30 frames per second. The head can be controlled via an embedded control architecture over either a TCP/IP connection or an interface to the RealTime DataBase (RTDB) develop ed by Matthias Göbl [24]. The RealTime DataBase is a shared memory architecture, implementing a publisher and subscriber concept. In this demonstration setup the robot is controlled over RTDB, which is also used to pass the video-stream captured from EDDIE's eye cameras to the expression recognition software component.

3.1 Expression Generation

The emotional state of the display can be controlled in two ways, either referring to the discrete basic emotions found by Ekman et al.

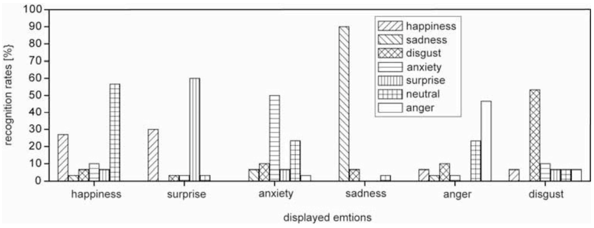

Figure 5. Results of the evaluation study

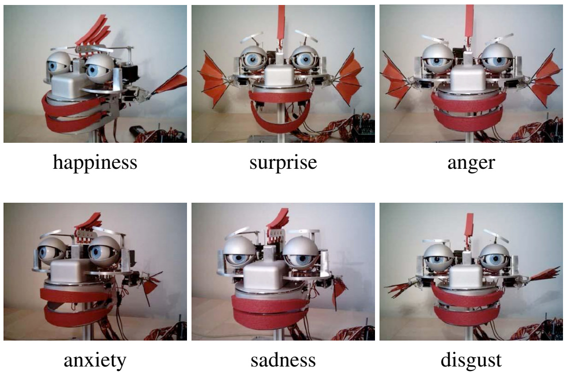

Figure 6. EDDIE displaying the basic facial expressions proposed by Ekman et al. [2].

or referring to the circumplex model of affect proposed by Russel et al.[2, 17]. Each state of the discrete emotions was modeled according to the FACS description, transitions between states are animated by linear interpolation of the respective motor commands of the start - / end-state. The expression of the six emotional states can be seen in Figure 6. A second way to generate the emotional state of the robot and the according facial expressions is the circumplex model of affect. In this model, the state-space is spanned by a valence (pleasure) and an arousal (exitement) axis. Every mood-describing expression can be defined according to its pleasure- and excitement-component. With a state-space to joint-space mapping that we developed, arbitrary facial expression in this emotional space can be generated believable and with smooth transitions [21].

3.2 Experimental evaluation

A study with 30 participants (researchers of Technische Universität and the University of the Armed Forces, München, 11 female, 19 male, age-mean 30 years) has been conducted. Six basic emotions have been presented to each individual. The subjects' ratings of the displayed emotions were rated with a multiple-choice test with a seven-item-scale. The result of this study can be seen in Figure 5. On the abscissa the six displayed emotions are shown. For each displayed emotion the amount of assumed emotions by the subjects is presented. For example, 90% of the participants agree in seeing a sad face if a sad face is displayed. For the displayed emotions anger and anxiety about 50% of the answers were correct, the remaining 50% are shared between the other emotions. This evaluation study proves that the generated facial expressions can be correctly identified by untrained users and matched to the corresponding emotional state of the robot to a reasonable degree [23].

4 Demonstration setup

In this setup we are using the RealTime DataBase as an interface between the recognition and the synthesis components, providing a fast and reliable data exchange. Images captured from the robot head's eye cameras are passed to the recognition module via the RealTime DataBase. The face model is fitted to faces found in the images. The facial expression and it's intensity are extracted from the face model and passed to back to the RTDB. This information about the emotional class and the intensity is then pushed to the synthesis component on the robotic head. Although all six basic facial expressions are recognized and mirrored on both databases, a preliminary demonstrator version includes only the strongest facial expressions, laughing and surprise from live images. The display of the facial expression is computed as presented in Section 3.1. In order to account for the intensity, the shown facial expression results from a linear interpolation between the emotional class and the neutral expression. The human interaction partner perceives the emotional state displayed by the facial expression, closing the loop, as can be seen in Figure 1 and 7.

5 Conclusion and Future Work

In this paper an approach is presented that recognizes six universal facial expressions together with their intensity and mirrors them on a robotic head, as a first step towards an emotional closed loop system. We integrate the Candide-3 face model and exploit its parameter vector to extract feature values. Accuracy on independent training data peaks at 91%. Our approach runs in real-time on standard hardware. The drawback of our approach is the inability to recognize very subtle changes due to the lack of texture information. Furthermore, the recognition accuracy is limited by the training data provided causing the facial expression recognition to fail if presented face movement not part of the training data.

Future work includes evaluation on and improvement of our approach on data obtained from more realistic real-world scenarios and extension to the live mirroring of all six basic expressions. So far the system components have been evaluated independently of each other. Although the recognition rates of facial expressions obtained by the system and the inspection of facial expression recognizability by humans indicate that the synthesized expression can be recognized with the one displayed by the human, there is no scientific proof, yet. Therefore, future work will include evaluating the complete analysis-synthesis loop on human test subjects. A test will be set up that asks human probands to match a synthesized facial expression with the original human face.

Real-time capabilities are important for future use in emotional closed loop systems, where delays in information processing should be kept to a minimum. While the mirroring in future experiments could be extended to more sophisticated versions, the final goal is to influence the human interaction partner in it's perceived emotional state with the induction of feelings through the emotional display. This might improve the acceptance or performance rating of the robot in human-machine-interaction.

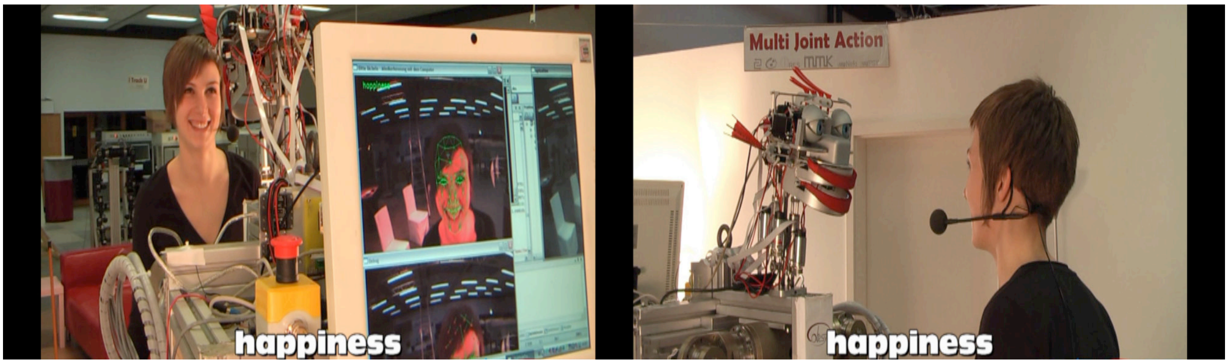

Figure 7. Facial expression recognition and synthesis demonstration

ACKNOWLEDGEMENTS

This work is supported in part within the DFG excellence initiative research cluster Cognition for Technical Systems CoTeSys, see www.cotesys.org, and by the Institute for Advanced Study (IAS), Technische Universität München, see also www.tum-ias.de.

REFERENCES

[1] J. Ahlberg. Candide-3 – an updated parameterized face. Technical Report LiTH-ISY-R-2326, Linköping University, Sweden, 2001.

[2] P. Ekman. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation 1971*, University of Nebraska Press.

[3] P. Ekman. Facial expressions. *Handbook of Cognition and Emotion*, New York, 1999. John Wiley & Sons Ltd.

[4] S. B. Gokturk, J.-Y. Bouguet, C. Tomasi, and B. Girod. Model-based face tracking for view-independent facial expression recognition. *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[5] C. S. Ikehara, D. N. Chin, and M. E. Crosby. A model for integrating an adaptive information filter utilizing biosensor data to assess cognitive load. In *User Modeling*, Springer Berlin / Heidelberg, 2003.

[6] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *International Conference on Automatic Face and Gesture Recognition*, 2000.

[7] M. Pantic and M. F. Valstar and R. Rademaker and L. Maat Web-based Database for Facial Expression Analysis. Proc. IEEE Int'l Conf. Multmedia and Expo (ICME'05), 2005

[8] I. Kotsia and I. Pitaa. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *Trans. On Image Processing*, 16(1), 2007.

[9] F. D. la Torre Frade, J. Campoy, Z. Ambadar, and J. Cohn. Temporal Segmentation of Facial Behavior. In *International Conference on Computer Vision*, 2007.

[10] C. L. Lisetti and D. J. Schiano. Automatic facial expression interpretation: Where human interaction, articial intelligence and cognitive science intersect. *Pragmatics and Cognition, Special Issue on Facial Information Processing and Multidisciplinary Perspective*, 1999.

[11] P. Michel and R. E. Kaliouby. Real time facial expression recognition in video using support vector machines. In *Fifth International Conference on Multimodal Interfaces*, 2003.

[12] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE PAMI*, 2000.

[13] E. M. Sheldon. *Virtual agent interactions*. PhD thesis, 2001. Major Professor-Linda Malone.

[14] R. M. Vick and C. S. Ikehara. Methodological issues of real time data acquisition from multiple sources of physiological data. *Annual Hawaii International Conference on System Sciences*, 2003.

[15] M. Wimmer, F. Stulp, S. Pietzsch, and B. Radig. Learning local objective functions for robust face model fitting. *IEEE PAMI*, 2008.

[16] C.L.Breazeal. Designing sociable robots. *IEEE The MIT Press* 2002

[17] J.A.Russell A circumplex model of effect *Journal of Personality and Social Psychology* 1980

[18] J.Reichard Robots: Fact, Fiction and Prediction *Penguine Books* 1978

[19] A.Mehrabian and J.A.Russell An Approach to Environmental Psychology *MIT Press* 1974

[20] J. R. Quinlan Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on Artifcial Intelligence*, pages 343-348, 1992.

[21] S.Sosnowski and A.Bittermann and K.Kühnlenz and Martin Buss Design and Evaluation of Emotion-Display EDDIE In *Int. Conference on Intelligent Robots and Systems* , 2006

[22] A. Mehrabian Communication without Words, Psychology Today, volume 2, pages 53-56, 1968

[23] K. Kühnlenz and S. Sosnowski and M. Buss Evaluating Emotion Expressing Robots in Affective Space In *Advanced Robotic Systems and I-Tech* , Vienna, Austria , 2007

[24] M. Goebl and G. Färber, "A real-time-capable hard- and software architecture for joint image and knowledge processing in cognitive automobiles," *Intelligent Vehicles Symposium*, pp. 737 – 740, June 2007.

[25] T. Hashimoto and S. Hitramatsu and T. Tsujiand H. Kobayashi Development of the Face Robot SAYA for Rich Facial Expressions SICE-ICASE 2006. International Joint Conference

[26] Z. Deng and J. Noh Data-Driven 3D Facial Animation Computer Facial Animation: A Survey. Springer London, 2007.

[27] G. Littlewort and M. Stewart Bartlett and I. Fasel and J. Susskind and J. Movellan Dynamics of facial expression extracted automatically from video Image and Vision Computing, vol. 24, pp. 615–625, 2006

[28] C. Shan and S. Gong and P. W. McOwan Facial Expression Recognition based on Local Binary Patterns: A comprehensive Study Image and Vision Computing, vol. 27, pp. 803-816, 2009