# Visual-Based Emotion Detection for Natural Man-Machine Interaction

Samuel Strupp, Norbert Schmitz, and Karsten Berns

Faculty of Computer Science, Robotics Research Lab,
University of Kaiserslautern, D-67653 Kaiserslautern, Germany

**Abstract.** The demand for humanoid robots as service robots for everyday life has increased during the last years. The processing power of the hardware and the development of complex software applications allows the realization of "natural" human-robot interaction. One of the important topics of natural interaction is the detection of emotions which enables the robot to react appropriate to the emotional state of the communication partner. Humanoid robots designed for natural interaction require a short response time and a reliable detection. In this paper we introduce a emotion detection system realized with a combination of a haar-cascade classifier and a contrast filter. The detected feature points are then used to estimate the emotional state using the so called action units. Final experiments with the humanoid robot ROMAN show the performance of the proposed approach.

## 1 INTRODUCTION

Humanoid robots are - as it can be seen in many movies - of great interest. Either as entertainment robots, enduring workers or for the care of elderly people. All these possible scenarios have one thing in common: the robot is an accepted member of society and therefore it must behave as a human would do. To reach this goal it is necessary that a robot can interact and communicate in a natural way.

The communication between humans is not limited to speech. Moreover it is a complex combination of speech, gestures, mimics and body pose. Robots interacting with people therefore have to be able to use these communication skills. This paper focuses on the aspect of emotion detection using face images which influences the behavior of a robot according to the situation of the human interaction partner.

The detection of emotions is often divided into two steps. The first step is to get information about the current expression of the face. This includes the detection of features like eyes, eyebrows or mouth. Different approaches for feature detection like feature point tracking ([13], [16], [18]), optic flow fields ([12], [1], [11] ), neural networks ([4], [6], [7] ), deformable contours ([9], [3], [15] ) or difference imaging ([1], [6], [7]) have been presented in many publications.

Beside the feature detection algorithms several approach use specific properties of the human face. Kawato for example [10] uses the blinking of the eyes

to detect them. Depending on the frame rate and resolution of the camera the detection can fail. Another approach of Soderstrom uses the nostrils to detect the region of the mouth [14]. To recognize the nostrils in a picture you must have a viewer perspective under the tip of the nose which is generally not realizable for a humanoid robot. The detection of features is a well know research topic therefore this paper focuses on the emotion interpretation and integration into the humanoid robot.

The second step in the detection process is to interpret the feature information and associate them to an emotion. Canzler [2] proposes a system which interprets gestures including the detection of facial expressions. The disadvantage of his system is the slow processing and the robustness of the system. The system by Wimmer and Fischer [8] for an emotional sales agent uses a classifier based on the optical flow within the face to interpret an emotion. The disadvantage is that an emotion can only be interpreted if the facial expression of the emotion starts from a neutral expression. The initialization must be done manually when the person looks neutral which leads to an unnatural behavior for humanoid robots.

In the following a concept for feature detection and emotion interpretation is introduced. Final experiments performed with the humanoid robot ROMAN will show the performance and integration of the system.

## 2 FEATURE DETECTION

To interpret emotions, it is necessary to have simple information about the current state of the face. So the first step is to find exact positions of some feature points. Therefore detectable and useful feature points for the interpretation of emotions have to be selected. In coherency to the emotion interpretation in section 3, 10 significant features points have been defined: 2 points for each eyebrow, the pupils, the mouth corners as well as the upper and the lower lip. Fig. 1 shows the location of these points as red crosses.

The goal of the feature detection is to localize these 10 points in the face region. Our emotion interpreter and the feature detection assume that all faces are frontal to the camera. Although this is an strong limitation it can be assumed that a person is facing a robot during a conversation. The face region itself can be found with the frontal face classifier of OpenCV [1]. For each feature point a search region within the face region has been defined to reduce the search space. In Fig. 1 the search regions are drafted. Two different techniques are used to find the feature points in this region. The eyebrows are extracted using an adaptive contrast filter which uses the min and max gray values in the eyebrow region to detect the borders. This filtering results in an exact representation of the eyebrow as a polygon. Using these polygons two points on the eyebrows are localized at a certain distance from the pupils. Fig. 1 shows a sketch of the eyebrow search regions. The other feature points are detected with self-trained classifiers based

[1] see http://www.intel.com/technology/computing/opencv/index.htm for details

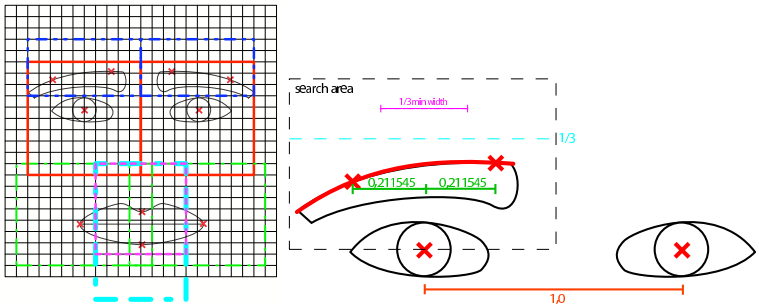


**Fig. 1.** (left) Feature points (red crosses) and their search regions. Blue: eyebrows; Red: pupils; Green: mouth corners; Cyan: lower lip; Magenta: upper lip. (right) Sketch of the search for the right eyebrow.

on the technique of Viola and Jones [17] implemented in OpenCV. The trained classifiers are shown in Fig. 2.

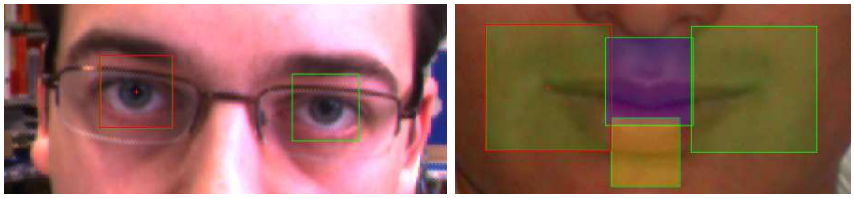

**Fig. 2.** (left) Pupil classifier region for the right and left pupil. The center of the region is located in the center of the pupil. (right) Mouth classifier regions. [green] right and left mouth corner regions [blue] upper lip region [yellow] lower lip region

The detection of all these mentioned feature points is realized in a detection pipeline. This pipeline has the following steps:

1. Detection of the face region with the frontal-face classifier of OpenCV.
2. Localization of the pupils in their search areas with the pupil classifier. The pupil classifier provides very few false positives. Therefore the first detection is taken and the rest discarded.
3. The eyebrow detection is performed using the adaptive contrast filter.
4. The mouth corners are classified using a heuristic which searches the minimal vertical distance from the center point of the search area.
5. Detection of the lower lip point. If more than one area is classified the lowest match with minimal distance to the vertical axis in the center of the search area is chosen.
6. Finally the upper lip point is searched. If multiple detections occur the point with minimal distance to the vertical axis is chosen.

With the help of the detected points the emotion interpretation is performed as described in the following section. The detection rates of the specified regions in out test data set are: left pupil 86.2%, right pupil 87.2%, inner left brow 74.4%, outer left brow 63.6%, inner right brow 78.5%, outer right brow 63.6%, left mouth corner 83.6%, right mouth corner 80.0%, upper lip 59.0% and lower lip 49.7%.

## 3 EMOTION INTERPRETATION

To interpret an emotion with the detected feature points the "Facial Action Coding System" (FACS) [5] is used. This system defines so called "Action Units" for every muscle in the face which influences the facial expression. The general idea is to conclude from the positions of the feature points to the activity of some action units. The goal is to interpret the emotion of a facial expression activities of the action units. All detectable action units corresponding to the feature points described in 2 are listed in Tab. 1.

**Table 1.** Detectable action units with feature points from section 2

| AU | Description | AU | Description | AU | Description |
|---|---|---|---|---|---|
| 1 | Inner Brow Raiser | 10 | Upper Lip Raiser | 20 | Lip Stretcher |
| 2 | Outer Brow Raiser | 12 | Lip Corner Puller | 24 | Lip Presser |
| 4 | Brow Lowerer | 15 | Lip Corner Depressor | 26 | Jaw Drop |

To interpret the activity of an action unit a neutral and a maximum distance from the eye line has been defined. The eye line is the straight line from the right pupil center to the left pupil center. If the distance between the feature point and the eye line is less than or equal to the neutral distance, the activity of the corresponding action unit is 0. If the distance is greater than or equal to the maximum distance the activity is 1. If the distance lies between the neutral and maximum distance then the activity value is linearly interpolated. For the calculation of the activities only vertical or horizontal distances from the eye line or the normal of the eye line which intersects the middle between the two pupils are used. In Fig. 3 the feature points with their corresponding action units and activation directions are illustrated.

The distance to the main axis is a simple point to straight line distance. To be independent from the size of the face we build the ratio of the distance between feature point and eye line and the distance between the pupils. So the distance between the pupils is 1 on all faces. Since only frontal faces are considered this unit definition is usable. With this method we measure the neutral and maximum distances from different faces for all action units. We simply choose the average values for each action unit distance. In Tab. 2 you can see the results. With this data it is possible to associate a point position to an activity value of an action unit.

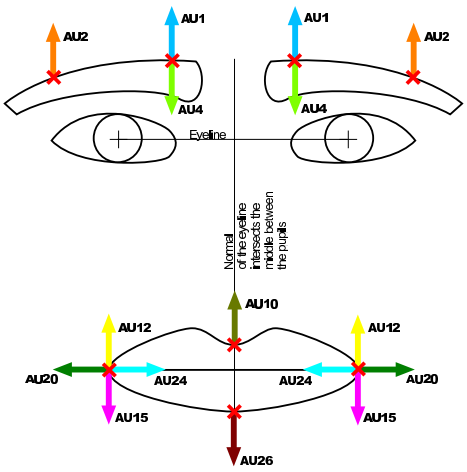


**Fig. 3.** Detectable action units. The arrow shows the direction a feature point must move to to activate this action unit.

**Table 2.** Neutral and maximum distances for each action unit extracted from example images of different persons. The reference distance with the value 1 is between the pupils. The action units with * use the vertical distance. If the maximum distance is smaller than the neutral distance then the action unit gets active if the feature point gets closer to the reference line.

| AU | neutral | maximum | AU | neutral | maximum | AU | neutral | maximum |
|---|---|---|---|---|---|---|---|---|
| 1 | 0,268 | 0,443 | 10 | 0,982 | 0,875 | 20* | 0,396 | 0,513 |
| 2 | 0,284 | 0,446 | 12 | 1,056 | 0,902 | 24* | 0,396 | 0,276 |
| 4 | 0,268 | 0,148 | 15 | 1,056 | 1,111 | 26 | 1,223 | 1,714 |

Based on the activities of the action units the displayed emotion has to be estimated. The emotional state is divided into the 6 basic emotions anger, disgust, fear, happiness, sadness and surprise. If none of these emotions is expressed the emotion should be classified as neutral. Each of the presented emotions has an activity between 0 and 1. Starting with the original emotion formulas from Ekman shown in Tab. 3 an improvement step has been performed. Using example images of faces showing a specific emotion various combinations and weights have manually been optimized to gain better results. The resulting formulas are also presented in Tab. 3. Using these activity values a emotion has been defined as active when the activity is higher than a specified threshold $t_{emo} = 0.26$, which has been extracted from the test data set. Additionally an emotion is defined as being dominant if the emotion has the highest activity of all 6 emotions.

To test the emotion interpretation and the emotion detection two tests on the same data set have been performed. The data set has multiple examples of different faces for each of the six basic emotions and the neutral expression. In the first test the feature points have been annotated manually. This test allows a

**Table 3.** Original and adapted action unit combinations to express the six basic emotions. $L$ and $R$ indicate the action unit on the left and right side of the face.

| Emotion | Original Definition | Adapted Definition |
|---|---|---|
| Fear | $1+2+4+5+20+25$ | $(1L+1R+2L+2R+20L+20R)/6$ |
| Surprise | $1+2+5+26$ | $(1L+1R+2L+2R+26+26)/6$ |
| Anger | $4+5+7+24$ | $(4L+4R+24L+24R)/4$ |
| Sadness | $1+4+15$ | $(4L+4R+15L+15R)/4$ |
| Disgust | $4+9+10+17$ | $(4L+4R+10)/3$ |
| Happiness | $6+12+25$ | $(12L+12R)/2$ |

testing of the emotion interpreter without feature detection errors. In the second test we use the feature detection described in section 2 to get the positions of the feature points. Fig. 4 shows the detection rates of the first and second test.

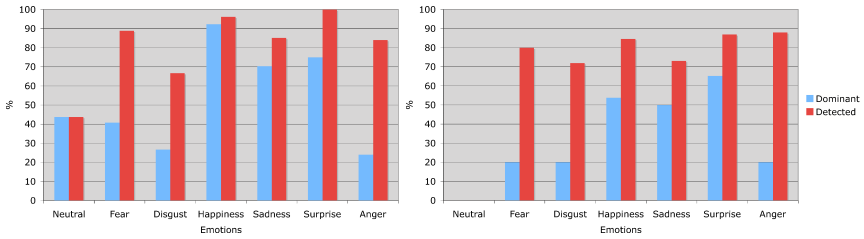


**Fig. 4.** (left) Emotion detection rates of our emotion interpreter. The feature points are marked manually (right) Emotion detection rates of our emotion detection. The feature points are detected automatically

As expected the detection rate decreases with the automatic detection of the feature points. The missing neutral facial expression is a result of the feature detection process. Minor localization errors of a single point have a great influence on the activity of an action unit which leads to an unwanted activity of an emotion. These tests show that the emotions happiness, surprise and sadness can be detected correctly in over 70% of the cases. In about 50 % of the cases they are even dominant.

### 3.1 Integration

For interaction purposes the robot (Fig. 5) must have a description of the communication partner. Therefore it is necessary to generate a description for each person in the environment. The description includes all information which are necessary to react properly. These information include body pose, gaze direction, emotional state and many more. The object descriptions generated in the perception system are then transferred to the motives and the habits of interaction. Further details concerning the control architecture will be given in future publications.

The control of the robot itself uses these models of the communication partner to adapt and modify the behavior of the robot. The emotional state of a communication partner for example influences the way the robot responses. The development of these models and the correct reaction of the robot are a current research topic which will be addressed in near future.

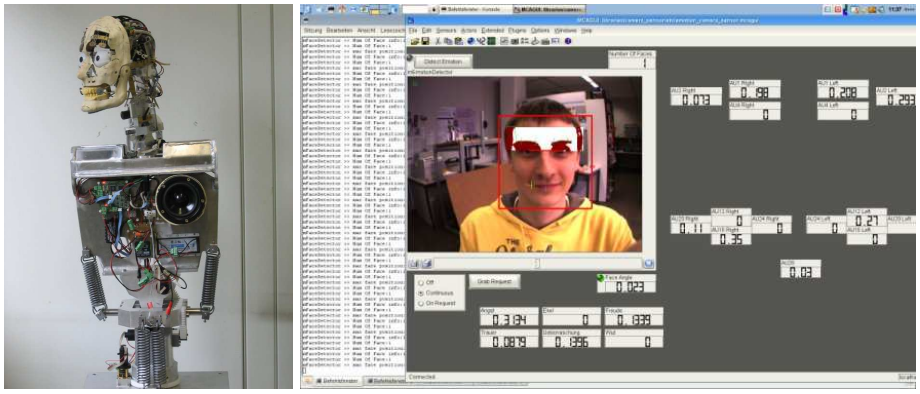

**Fig. 5.** (left) The humanoid robot ROMAN without clothes and silicone skin (right) Screen shot of the graphical user interface for the emotion detection

## 4 CONCLUSIONS AND FUTURE WORK

This paper presents a camera-based system for the detection of emotions of a human interaction partner. This system is divided into two main parts the feature detection and the emotion interpretation. In the first part important features in the face are localized with the help of a haar cascade classifier and a contrast filter. Based on the exact position of these features a probability for each of the six basic emotions fear, happiness, sadness, disgust, anger and surprise can be assigned. Final test experiments with reference images show correct detection rates of about 60% for the emotions happiness, sadness and surprise.

For the future it is necessary to detect the feature points with a better accuracy to improve the results of the emotion interpreter. Another aspect is to detect more feature points to describe a larger set of action units. One possible candidate is the nose wrinkler which is influenced by action unit 9. The detection of this action unit will improve the results for the emotion disgust.

Another important sector for development are the distance values of table 2. These values are average values from a few persons. A model of each person with personal distance values will increase the robustness of the system. To realize this task an identification of the person is required. Even better would be a system that can measure these values directly from the observation of the face.

## References


1. M. Bartlett. *Face Image Analysis by Unsupervised Learning and Redundancy Reduction*. PhD thesis, University of California, San Diego, 1998.
2. Ulrich Canzler. Automatische erfassung und analyse der menschlichen mimik. In *Bildverarbeitung fr Medizin 2001. Algorithmen, Systeme, Anwendungen*, 2001.
3. Greg I. Chiou and Jenq-Neng Hwang. Lipreading by using snakes, principal component analysis, and hidden markov models to recognize color motion video. unknown, 1996.
4. M.N. Dailey and G.W. Cottrell. Pca = gabor for expression recognition. Technical Report CS1999-0629, University of California, San Diego, October 26 1999.
5. P. Ekman and W. Friesen. *Facial Action Coding System*. Consulting psychologist Press, Inc, 1978.
6. Beat Fasel and Juergen Luettin. Recognition of asymmetric facial action unit activities and intensities. In *Proceedings of International Conference on Pattern Recognition (ICPR 2000), Barcelona, Spain*, 2000.
7. W. Fellenz, J. Taylor, N. Tsapatsoulis, and S. Kollias. *Comparing Template-based, Feature-based and Supervised Classification of Facial Expressions from Static Images*. Computational Intelligence and Applications, World Scientific and Engineering Society Press, 1999, 1999.
8. S. Fischer, S. Dring, M. Wimmer, and A. Krummheuer. Experiences with an emotional sales agent. In *ADS*, 2004.
9. M. Kass, A. Witkin, and D. Terzopoulos. Snakes:active contour models. *International Journal of Computer Vision*, pages pp. 321–331, 1988.
10. Shinjiro Kawato and Nobuji Tetsutani. Detection and tracking of eyes for gaze-camera control. In *VI*, 2002.
11. Jenn-Jier James Lien. Automatic recognition of facial expressions using hidden markov models and estimation of expression intensity. Technical Report CMU-R1-TR-31, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, April 1998.
12. K. Mase and A. Pentland. Recognition of facial expressions from optical flow. IEICE Transactions (Special Issue on Computer Vision and its Applications), 1991.
13. M. Rosenblum, Y. Yacoob, and L.S. Davis. Human expression recognition from motion using a radial basis function network architecture. *IEEE transactions on neural networks*, 7:1121–1138, 1996.
14. Ulrik Soderstrom and Haibo Li. Customizing lip video into animation for wireless emotional communication. Technical Report DML-TR-2004:06, Department of Applied Physics and Electronics, Umea University, Sweden, 2004.
15. D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *Pattern Analysis and Machine Intelligence PAMI*, 15(6):596–579, June 1993.
16. Ying-Li Tian, Takeo Kanade, and Jeffery Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001.
17. P. Viola and M. Jones. Robust real-time object detection. In *Second International Workshop on Statistical and Computational Theories of Vision*, Vancouver, Canada, July 13 2001.
18. M. Wang, Y.Iwai, and M. Yachida. Expression recognition from time-sequential facial images by use of expression change model. In *FG '98: Proceedings of the 3rd. International Conference on Face and Gesture Recognition*, page 324, Washington, DC, USA, 1998. IEEE Computer Society.