A Multimodal Human-Robot-Dialog Applying Emotional Feedbacks*

Alexander Bannat[1], Jürgen Blume[1], Jürgen T. Geiger[1], Tobias Rehrl[1], Frank Wallhoff[1,4], Christoph Mayer[2], Bernd Radig[2], Stefan Sosnowski[3], and Kolja Kühnlenz[3]

[1] Human-Machine Communication, Department of Electrical Engineering and Information Technologies, Technische Universität München, Munich, Germany
{bannat,blume,geiger,rehrl,wallhoff}@tum.de
[2] Image Understanding and Knowledge-Based Systems, Department of Informatics, Technische Universität München, Munich, Germany
{mayerc,radig}@in.tum.de
[3] Institute of Automatic Control Engineering, Department of Electrical Engineering and Information Technologies, Technische Universität München, Munich, Germany
{sosnowski,koku}@tum.de
[4] Jade University of Applied Sciences, Oldenburg, Germany

Abstract. This paper presents a system for human-robot communication situated in an ambient assisted living scenario, where the robot performs an order-and-serve-procedure. The interaction is based on different modalities that extract information from the auditory and the visual channel in order to obtain an intuitive and natural dialog. The required interaction dialog structure is represented in first-order logic, which allows to split a complex task into simpler subtasks. The different communication modalities are utilized to conclude these subtasks by determining information about the human interaction partner. The system works in real-time and robust and utilizes emotional feedback to enrich the communication process.

1 Introduction

Despite of many advantages of computers (data storage and quick data processing), human-computer interaction still lacks a natural way of communication and further human-interaction techniques like gestures and mimics. Traditional human-machine interaction relies on a fixed set of operations performable on the machine in a certain way. These operations are static and force the human to adapt to the machine interface.

Instead, we aim at granting machines the ability to adapt to typical human behavior [1]. The idea is to equip machines with higher flexibility in the interaction process. To accomplish this objective, the different steps of interaction have to be performed autonomously and robustly and machines need to infer information from typical human interaction channels, such as gaze, facial expressions, head gestures or speech.

We tackle this challenge in a two-fold approach. Firstly, we present an abstraction of human-machine dialogs by using first order logic. By applying equivalence rules,

* All authors contributed equally.

S.S. Ge et al. (Eds.): ICSR 2010, LNAI 6414, pp. 1–10, 2010.

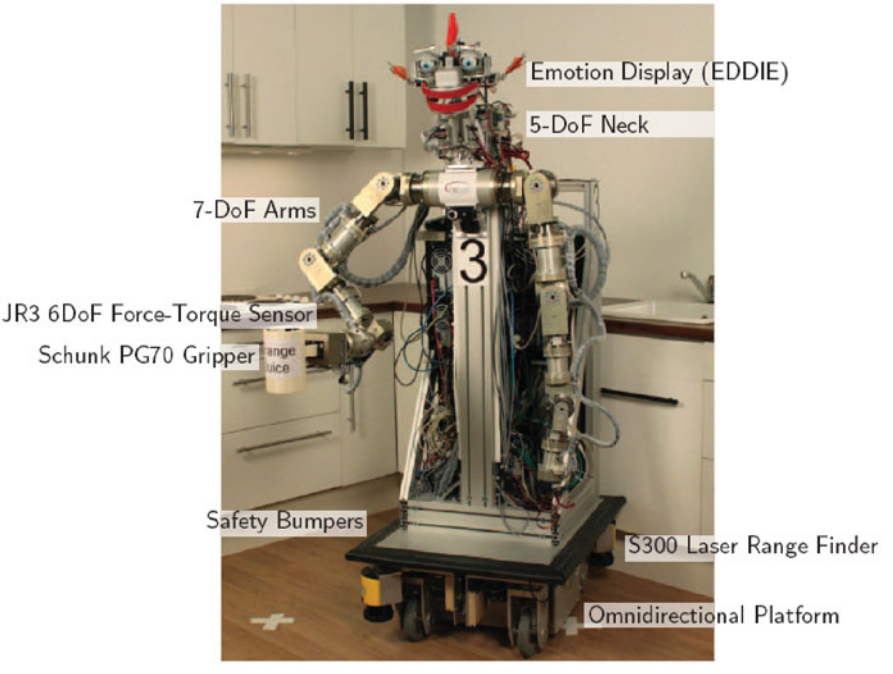

Fig. 1. Robot hardware overview

missing information is determined and complex tasks can be split into subtasks. This allows for a intuitive representation of dialogs as well as high flexibility during runtime. Secondly, we integrate multiple modalities, like emotional feedback, in the dialog to better approximate the natural human communication with a robot (see Figure 1).

1.1 Related Work

In conversational systems, Spoken Language Understanding (SLU) aims at extracting speech concepts from spontaneous speech and their relations. With increased complexity in spoken dialog systems, a richer set of features is needed to perform SLU, ranging from a-priori knowledge over long dependency dialog history to system belief [2].

A rough separation of systems for multimodal fusion leads into two classes: early fusion systems and late fusion systems [3]. Systems that integrate several modalities already on feature level are called early fusion systems. This approach is very popular for systems that integrate modalities like speech and pen input, which are closely correlated to each other. In [4] an extensive overview of several early fusion systems is given.

In recent years, late fusion systems have become more popular. For each modality, those systems involve separate recognition modules, which send their recognition

results to a multimodal fusion component. This component generates a combined interpretation of the input. Examples for late fusion systems are the robot systems presented in [5] and [6]. The integration of vision and speech enables both of these robots to communicate with humans.

We apply Hidden Markov Models (HMMs) [7] to the recognition of head gestures as well as to person identification. HMMs represent a sequence of feature vectors using statistical processes. They have been proven highly applicable to both, forming an abstraction of sequential processes and determining the probability that a specific feature vector sequence is generated by a modeled process.

Referring to the survey of Pantic et al. [8], the computational task of facial expression recognition is usually subdivided into three subordinate challenges: face detection, feature extraction, and facial expression classification. After the position and shape of the face in the image are detected in the first step, descriptive features are extracted in the second step. In the third step, high-level information from these features is derived by a classifier. Due to the generality of this approach, we apply it to recognize head gestures and person identification.

Models rely on a priori knowledge to represent the image content via a small number of model parameters. This representation of the image content facilitates and accelerates the subsequent interpretation task. Cootes et al. [9] introduce modeling shapes with Active Contours which use a statistics-based approach to represent human faces. Further enhancements extended the idea and provided shape models with texture information [10]. However, both models rely on the structure of the face image rather than the structure of the real-world face. Therefore, information such as position or orientation in 3D space is not explicitly considered but has to be calculated from the model parameters. Since this mapping is again not provided by the model, it is error-prone and renders them difficult for extracting such information.

Recent research considers modeling faces in 3D space [11,12]. In contrast to 2D models these models directly provide information about position and orientation of the face.

Face recognition is an important topic for security relevant systems. In early 1990 a revolutionary face recognition algorithm was developed by Turk and Pentland, called Eigenfaces [13]. When HMMs are applied for classification, typically we use continuous HMMs with one Gaussian per state, due to the large number of states and the limited training data for each face. However it turned out that discrete modeling techniques in conjunction with HMMs are efficient for large vocabulary speech recognition (LVCSR) and even in the field of handwriting applications [14,15]. Thus, it seems promising to test this technology also for our face recognizer. The advantage of discrete systems is the higher computation speed and smaller model size compared to continuous HMMs.

1.2 Organization of the Paper

The remainder of this paper is structured as follows: In the subsequent Section, a short overview of the scenario is given. Section 3 gives a general system overview and focuses on the following topics: *Dialog Manager* (3.1), *Communication Backbone* (3.2), *Person Identification* (3.3), *Head Gestures* (3.4) and *Robotic Head* (3.5).

The paper presents a conclusion and closes with an outlook over future work in Section 4.

2 Scenario Description

Our scenario is located in an ambient assisted living environment where the robot takes the role of a waiter. In general, the ordering dialog is initiated when a new guest is entering the scene. Afterwards the robot approaches the human by utilizing sophisticated navigation and path planning strategies combined with online collision avoidance and orients itself towards the guest, providing a comfortable communication situation. After the adjustment of the ”eyes” of the robotic head, the face based verification is conducted (for more information see Section 3.3). If the person is known, the guest is approached directly with his name, otherwise a more formal kind of greeting is applied. The robot takes the first initiative by asking the dialog partner what kind of drink is desired. The human customer can ask for different drinks ranging from coffee to water. The order is conducted via speech and can be confirmed by the user either by speech or head gestures (nodding or shaking). Based on positive or negative dialog feedback, the emotions of the robotic head are set correspondingly. For example, if the speech recognition module fails to recognize a name of a drink, the robot informs the user of this failure, which is done both verbally and non-verbally by the speech output and the facial expression of sadness.

Additionally, to signal the readiness for speech input, the head can arrange the ”ears” showing a listening behavior. As soon as the ordering process is accomplished, the robot starts to fetch the corresponding drink from a bartender and returns the drink to the customer. With a handover the cup is delivered to the thirsty guest. In the ordering dialog the robot can choose from a large set of possible phrases (english or german), where for each step in the ordering dialog, several sets of phrases are available, e.g. the greeting sequence can start with ”Hello.”, ”Hi.”, ”How are you?” etc. For gaining a more natural and intuitive dialog, the robotic head is also equipped with so called ”idle motions” to simulate a vivid behavior.

3 System Overview

We equipped a robot with visual and auditory interaction capabilities. The complete platform is mobile and therefore able to move in the laboratory. An in-eye camera system infers visual information about the human interactant to identify the person from its face and recognize head gesture. A mounted robot head signals the robot state back to the human via the display of facial expressions. In addition, a system of microphones and speakers is utilized to communicate with the human via natural language. Therefore, our system provides a bidirectional communication based on the visual and auditory channel.

3.1 Dialog Manager

A dialog manager keeps track of the ongoing communication to estimate when human user or machine response is expected by the dialog partners. The complete dialog

structure is represented by first-order logic. Tasks to solve are represented by predicates with variables. These variables represent information which is to be determined during the dialog. Equivalence rules on these predicates are specified to navigate through the dialog by splitting a task into several subtasks. Evaluating predicate truth values and binding variables models real-world interaction.

We demonstrate this with help of an example that consists of the task of serving a human a drink. This task is represented by the predicate $orderDialog(A,B)$ with A being the person to be served and B being the drink. The dialog manager inspects the equivalences of the system and applies a rule replacing the $orderDialog(A,B)$ predicate with $isIdentified(A)$ & $isOrdered(B)$ & $isConfirmed(B)$. In the next step, the dialog manager will determine the truth value $isIdentified(A)$ by inspecting the value of A. At this point, interaction with the environment is required to determine what person is interacting with the robot, setting the value of A. At the same time, the predicate $isIdentified(A)$ evaluates to `true`.

In a similar manner, $isOrdered(B)$ is evaluated by assigning some name of a drink from the speech recognition system to the variable B. In the last step, the value of $isConfirmed(B)$ is determined from either speech recognition or head gesture recognition.

Note that this representation allows to integrate different sensory modules that determine information about the environment. The information determined by the modules is considered in a very similar manner in the dialog structure, simplifying the integration of multiple modalities.

We will now introduce the modules considered in the example dialog (see Figure 2).

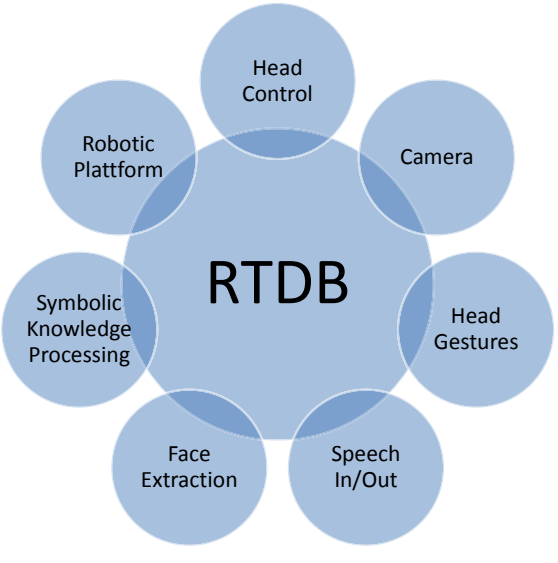

Fig. 2. Schematic System Overview

3.2 Communication Backbone

For the described setup and the desired functionalities of the system, a lot of modules are required, orchestrated with a suitable communication backbone basing on the Real-time Database (RTDB) introduced in [16]. The RTDB is based on data-objects, which are written by input modules (called writers) to be accessible by multiple processing modules (called readers) without blocking effects, exploiting a shared memory characteristic. For example, a camera image can be stored in the object buffer allowing all image processing modules to analyze this image in parallel. In the following we will present a short overview of the involved software modules basing on the RTDB. These software modules can be roughly subdivided into two entities:

Module Controlling. A generic control container is supplied by every processing module for delivering an inter-process communication. Other modules can write commands into a container and will receive in turn a result object containing information about the processed command. This can be used for basic commands like "stop" or "start processing" to use the available computational power efficiently. Furthermore, module specific commands can be exchanged via this interface, e.g. the speech recognition module can receive a grammar as input and return the answer of the user as a result.

Image and Audio Processing. The RTDB manages different image and audio processing modules covering simple video-writers and audio-recording up to sophisticated gesture detectors and the integration of a commercial speech synthesize and recognition framework.

3.3 Person Identification

We determine the identity of the human interaction partner from a frontal face image, which is easy to obtain in the introduced dialog situation. The face position within the image is determined by applying the algorithm of Viola et al.[17] and is extracted and then scaled to a size of $64x96$ pixels. The subsequent feature extraction consists of a rectangular windowed 2D Discrete Cosine Transform (DCT). The sampling window is moved in the vertical direction first, then in the horizontal direction. With each displacement the window is moved not by a full window size but rather one fourth of the window size so that an overlap of 75% to the previous window arises. To preserve the 2D structure of the images, a special marker is inserted at the beginning of each row.

For the classification process the following maximum likelihood decision is used:

$$\delta^* = argmax_{M \in DB} P(X|M) \tag{1}$$

In this formula, X is the feature vector sequence of an unknown image and M represents one HMM of an individual contained in the database. The system recognizes the image as belonging to the individual whose corresponding model δ yields the highest production probability $P(X|M)$. To solve this equation, the values of $P(X|M)$ for all models have to be computed.

We evaluated the system on the FERET database that is provided by the Army Research Laboratories. Comparing the results of the discrete and continuous systems the discrete systems outperforms the continuous systems with a recognition rate of 98.13%.

This implies that only 6 individuals of the 321 tested are not recognized correctly. The computation time while classifying the discrete models was just about half the time of the continuous ones. This is important, because in real-world scenarios the robot has to adapt to new persons quickly.

3.4 Head Gesture Recognition

Models based on the analysis of face images impose knowledge about the object of interest and reduce the large amount of image data to a small number of expressive model parameters. We utilize a rigid, 3D model of human faces in our system, because it inherently considers position and orientation of the face in space. The 3D model is fit onto the face in the image to determine corresponding model parameters. The small amount of model parameters guarantees a short calculation time which in turn provides real-time capability. Five model parameters are considered to train a classifier for the recognition of head gestures. The data vector d_i extracted from a single image I_i is composed of the in-plane transition of the face and the three rotation angles (pitch, yaw and roll). However, we do not utilize the absolute values of the five parameters but temporal parameter changes. Due to their time-sensitive nature, we apply continuous HMMs for classification.

In total, fourteen different persons constitute the model for classifying head gestures. We record two sequences per person and head gesture (nodding, shaking, neutral). The model is tracked through these short image sequences consisting of roughly $n = 12$ frames I_i, $1 \leq i \leq n$ and the model parameters are exploited to train a classifier. Per training image sequence we create one set of data vectors $(d_1, ..., d_{12})$ of fixed size. Each of these sets form one observation to train the HMM. Note that therefore the HMM determines the head gesture for a sequence of images rather than for a single image. In total, we present $14 \times 3 \times 2 = 74$ observations to the HMM. The only parameter given manually is the number of states J. We train different HMMs to correctly determine this parameter. Inspection of the training errors shows that the best parameterization is $J = 5$ [18].

3.5 Robotic Head

In this system a robotic head with 23 DOF is used for intuitive, natural communication feedback. In order to achieve dynamic, continuous, and realistic emotional state transitions, the 2D emotional state-space based on the circumplex model of affect is directly mapped to joint space corresponding to particular action units of the facial action coding system (FACS). It is also possible to display the six basic emotions according to Fridlund, Ekman and Oster [19]. The head is largely developed and manufactured in a rapid-prototyping process. Miniature off-the-shelf mechatronics are used for the face, providing high functionality while extremely low-cost. The active vision system consists of two firewire cameras, which are integrated into the eyes (2DOF each), forming a stereo pair with a baseline of 12cm and 30fps at a resolution of 640x480 pixels, and a 5DOF neck.

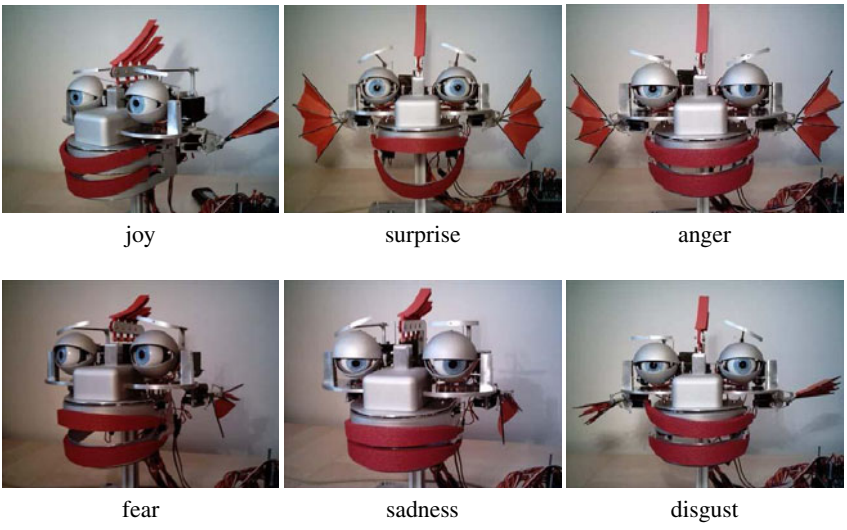

Fig. 3. Robotic head displaying the basic facial expressions

3.6 Facial Design

Facial action coding system. In order to get an objective distinction among traceable changes in facial expressions, the FACS based on face muscles [20] is applied. For this purpose FACS defines smallest movable units, the so called action units. The limitation on visually distinguishable changes in faces leads to the definition of 44 action units [21]. The current construction covers 13 of the total 21 action units needed for the basic emotions. For an even better match, the integration of a flexible skin would be required.

3.7 Emotion Models

Discrete emotional states. The existence of universal emotions, being represented and interpreted equally in the whole world, is said to be assured. Fridlund, Ekman and Oster affirm in their literature research in 1987, that six basic emotions are clearly identified in a multitude of different cultural groups. Joy, surprise, fear, sadness, anger and disgust, therefore, are considered universal [19]. Therefore these emotions are highly suitable for this scenario. Figure 3 shows the display of these emotions by the robotic head.

4 Conclusions and Outlook

In this paper we present a system that realizes a simple dialog between a human and a robot. An early version of the proposed system has successfully been shown as a live demonstrator at the *1. International Workshop on Cognition for Technical Systems*. Furthermore, a user study has been conducted to show the degree to which humans can understand the facial expressions of the robot [22]. More experiments are scheduled. Two different communication channels are regarded: The machine receives simple

commands and asks for confirmation via spoken language. Furthermore, head gestures are recognized via model-based image understanding techniques and classification with HMMs. The system operates without manual control and all important algorithms base on objective machine learning techniques instead of subjective manual design. For obtaining a more natural dialog, the system is capable of identifying its counterpart with an algorithm for person identification. Furthermore, the robots head is able to express different kinds of emotions for user feedback and reflect internal system states towards the human.

Objective of this first system is the implementation of an integrated framework for the recognition of user feedback resulting in a mostly natural human-robot dialog system. As such, there are now several ways to further improve the overall process and its performance. The first major step is to further exploit the existing implementation of a face recognition/verification module to track the user. The second major research topic is the expansion and improvement of the head gesture and facial expression classification system to include hesitation or confusion.

Future work also focuses on increasing the robustness with respect to real-life scenarios (lighting conditions, multiple points of view, etc.) and integrating facial expressions into the classification process. Thus the system might achieve an overall higher level of comfort and acceptance.

Acknowledgement

This ongoing work is supported by the DFG excellence initiative research cluster *Cognition for Technical Systems CoTeSys*, see `www.cotesys.org` and [1] for further details. The authors would also like to thank all project partners for the many fruitful discussions. This work is supported in part by Institute for Advanced Study (IAS), Technische Universitaet Muenchen, Munich, Germany.

References

1. Brščić, D., Eggers, M., Rohrmüller, F., Kourakos, O., Sosnowski, S., Althoff, D., Lawitzky, M., Mörtl, A., Rambow, M., Koropouli, V., Medina Hernández, J.R., Zang, X., Wang, W., Wollherr, D., Kühnlenz, K., Mayer, C., Kruse, T., Kirsch, A., Blume, J., Bannat, A., Rehrl, T., Wallhoff, F., Lorenz, T., Basili, P., Lenz, C., Röder, T., Panin, G., Maier, W., Hirche, S., Buss, M., Beetz, M., Radig, B., Schubö, A., Glasauer, S., Knoll, A., Steinbach, E.: Multi Joint Action in CoTeSys - setup and challenges. Technical Report CoTeSys-TR-10-01, CoTeSys Cluster of Excelence: Technische Universität München & Ludwig-Maximilians-Universität München, Munich, Germany (June 2010)
2. Raymond, C., Riccardi, G.: Generative and discriminative algorithms for spoken language understanding. In: Proceedings of the Interspeech Conference, Antwerp, Belgium (2007)
3. Sharma, R., Pavlovic, V.I., Huang, T.S.: Toward multimodal human-computer interface. Proceedings of the IEEE 86, 853–869 (1998)
4. Oviatt, S.: Multimodal interfaces. In: The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, pp. 286–304 (2003)
5. Stiefelhagen, R., Ekenel, H., Fügen, C., Gieselmann, P., Holzapfel, H., Kraft, F., Nickel, K., Voit, M., Waibel, A.: Enabling multimodal human-robot interaction for the karlsruhe humanoid robot. IEEE Transactions on Robotics 23, 840–851 (2007)

6. Fransen, B., Morariu, V., Martinson, E., Blisard, S., Marge, M., Thomas, S., Schultz, A., Perzanowski, D.: Using vision, acoustics, and natural language for disambiguation. In: HRI 2007: Proceeding of the ACM/IEEE International Conference on Human-Robot Interaction, pp. 73–80. ACM Press, New York (2007)
7. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77 (1989)
8. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1424–1445 (2000)
9. Cootes, T.F., Taylor, C.J.: Active shape models – smart snakes. In: Proceedings of the 3rd British Machine Vision Conference, pp. 266–275. Springer, Heidelberg (1992)
10. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
11. Ahlberg, J.: Candide-3 – an updated parameterized face. Technical Report LiTH-ISY-R-2326, Linköping University, Sweden (2001)
12. Blanz, V., Scherbaum, K., Seidel, H.P.: Fitting a morphable model to 3d scans of faces. In: Proceedings of International Conference on Computer Vision (2007)
13. Turk, M., Pentland, A.: Face Recognition using Eigenfaces. In: Conference on Computer Vision and Pattern Recognition, pp. 586–591 (1991)
14. Rigoll, G., Kosmala, A., Rottland, J., Neukirchen, C.: A Comparison Between Continuous and Discrete Density Hidden Markov Models for Cursive Handwriting Recognition. In: International Conference on Pattern Recognition (ICPR), Vienna, Austria, August 1996, vol. 2, pp. 205–209 (1996)
15. Neukirchen, C., Rigoll, G.: Advanced Training Methods and New Network Topologies for Hybrid MMI-Connectionist/HMM Speech Recognition Systems. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Munich, Germany, April 1997, pp. 3257–3260 (1997)
16. Goebl, M., Färber, G.: A real-time-capable hard- and software architecture for joint image and knowledge processing in cognitive automobiles. In: Intelligent Vehicles Symposium, pp. 737–740 (June 2007)
17. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision (2004)
18. Gast, J., Bannat, A., Rehrl, T., Rigoll, G., Wallhoff, F., Mayer, C., Radig, B.: Did I Get It Right: Head Gestures Analysis for Human-Machine Interactions. In: Human-Computer Interaction. Novel Interaction Methods and Techniques, pp. 170–177.
19. Altarriba, J., Basnight, D.M., Canary, T.M.: Emotion representation and perception across cultures. Online Readings in Psychology and Culture (2003)
20. Ekman, P., Friesen, W.V.: Facial Action Coding Consulting. Psychologist Press, San Diego (1977)
21. e-learning-Kurs "'about facs"' (2003), `http://www.uni-saarland.de/fak5/orga/Kurs/home.htm`
22. Mayer, K.K.C., Sosnowski, S., Radig, B.: Towards robotic facial mimicry: system development and evaluation. In: 19th IEEE International Symposium in Robot and Human Interactive Communication, Speccial Session on Cognition for Interactive Robots (2010)