Multiple Parallel Vision-based Recognition in a Real-Time Framework for Human-Robot-Interaction Scenarios

Tobias Rehrl*, Alexander Bannat*, Jürgen Gast*, Frank Wallhoff*, Gerhard Rigoll*
Christoph Mayer†, Zadid Riaz†, Bernd Radig†
Stefan Sosnowski‡, Kolja Kühnlenz‡
* *Human-Machine Communication, Department of Electrical Engineering and Information Technologies, Technische Universität München, Munich, Germany*
†*Chair for Image Understanding and Knowledge-Based Systems, Computer Science Department, Technische Universität München, Munich, Germany*
‡*Institute of Automatic Control Engineering,Department of Electrical Engineering and Information Technologies, Technische Universität München, Munich, Germany*

***Abstract*—Everyday human communication relies on a large number of different communication mechanisms like spoken language, facial expressions, body pose and gestures, allowing humans to pass large amounts of information in short time. In contrast, traditional human-machine communication is often unintuitive and requires specifically trained personal. In this paper, we present a real-time capable framework that recognizes traditional visual human communication signals in order to establish a more intuitive human-machine interaction. Humans rely on the interaction partner's face for identification, which helps them to adapt to the interaction partner and utilize context information. Head gestures (head nodding and head shaking) are a convenient way to show agreement or disagreement. Facial expressions give evidence about the interaction partners' emotional state and hand gestures are a fast way of passing simple commands. The recognition of all interaction queues is performed in parallel, enabled by a shared memory implementation.**[1]

***Keywords*-real-time image processing, gesture recognition, human-robot interaction, facial expressions**

I. 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. Yet, recent research aims at enabling machines to utilize communication channels more familiar to human beings.

The excellence research cluster *Cognition for Technical Systems* CoTeSys attempts to equip technical systems with a high degree of cognition, thus facilitating more intelligent and useful reactions of these technical systems in human-machine interaction scenarios [1]. The cluster of excellence focuses on two main research fields: ambient living and advanced robotics [2], [3]. To provide intuitive communication, a cognitive system should be able to apprehend the human in a multimodal manner. We present a robot system that aims at providing such communication between robots and humans to partially fill the gap between human-human and human-machine interaction.

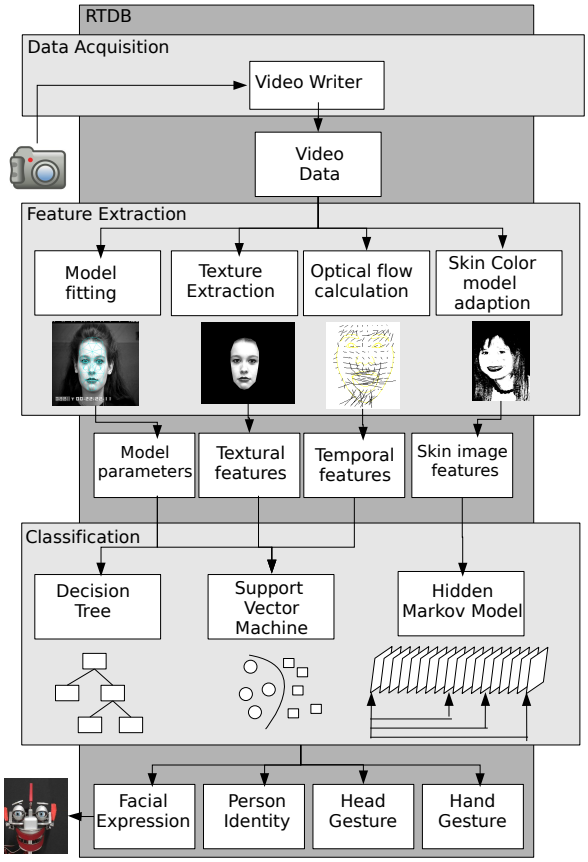

Figure 1. The system processes the data in parallel, transferring information between modules via the RTDB architecture.

The rest of this paper is organized as follows: A brief overview of related literature is presented in Section II. Afterwards, in Section III, we have a closer look on the

[1]All authors contributed equally to the work presented in this paper.

From: Proc. ACHI 2010, St. Maarten, Netherlands Antilles

features provided to the classifiers detailed in Section IV. Section V provides information about facial expression synthesis. Finally, evaluation results are presented in Section VI. The paper closes with a summary and an outlook over the next planned working steps.

II. Related Work

This section reviews related work in the fields of face image analysis, facial expression synthesis and hand gesture recognition. Due to the widespread nature of the presented system, the different modules are inspected separately. Integrating human-inspired forms of communication in robotic systems will allow for a human-machine interaction with increased efficiency, security, robustness and ease of use [15].

A. Face Recognition

A large number of techniques have been published for face recognition, including feature based recognition, face geometry based recognition, classifier design and model based methods. In [16] the authors give a comprehensive survey of face recognition and some commercially available face recognition software. Subspace projection methods like Principal Components Analysis (PCA) were adopted by M. Turk and A. Pentland, introducing the famous idea of Eigenfaces [17]. Active Appearance Model (AAMs), introduced by Cootes et. al, model shape and texture of human faces. [18]. Edward et al propose AAMs and weighted distance classifier for face recognition. [19]. Blanz et al. use state-of-the-art morphable models from laser scanner data for face recognition by synthesizing 3D faces. The proposed approach combines the presented methods, by relying on a 3D shape model in combination with a PCA-based texture extraction [20].

B. Head and Hand Gestures

Head gestures are pleasant means to show agreement or disagreement [4]. Nonetheless, it is also imaginable to use head gestures for controlling operations, like document browsing [5]. A classical paper utilizing hand gesture recognition resembles partially our approach and has overlap with respect to the classification method basing on Hidden Markov Models [6]. Nonetheless, our area of application is situated in a typical living environment and thus the detection of the human hand is more intricate. Robust hand detection in cluttered surroundings is also a problem for the approach of [8]. In [9] a remote control is introduced that utilizes ten predefined hand gestural commands to control a device selected via a pointing gesture.

C. Facial Expression Recognition

Some approaches infer the facial expression from rules stated by Ekman and Friesen [27]. 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 [28]. De la Torre et al. extract temporal segmentation of facial behavior from image data to assist professional FACS coders [23]. 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 rely on person-specific AAMs which renders the approach not applicable to previously unseen data. In contrast, we utilize publicly available data throughout the whole approach.

D. 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 [35]. In robotic systems, mostly FACS is used (e.g. Kismet [15] , Saya [34])

III. Feature Extraction

As can be seen in Figure 1, the RTDB allows for a parallel processing of the video data by multiple modules. First, features are extracted from the raw image data. Second, high level information is obtained from these features. The raw image data, the extracted image features and the determined high level information is stored and buffered in the RTDB. This section details the first of these processing steps, the extraction of descriptive image features.

A. Framework

We utilize the head of the service robot EDDIE for two purposes: First, we utilize the robot head's two movable firewire cameras to obtain image data. Second, the robot head displays facial expressions and therefore provides information passing from the system to the human [10]. The Real-Time DataBase (RTDB) serves as technical communication platform between the various modules, since it is capable of dealing with a large amount of input streams, having diverse features (i.e. data rate, packet losses, etc.) [11]. The RTDB is based on the concept of shared memory and serves as both, communication platform and data buffer.

B. Skin Color Features

We apply an implementation of the object detection approach published by Viola et al. to detect human faces in the video images obtained from the RTDB memory [12]. Since this information is required by multiple modules, it is also made available to them in the RTDB. Similar to the approach mentioned in [13], we adapt a skin color model with regard to the obtained face image. Please refer to [31] for further details on our approach. To constrain the hand gesture recognition, we define a region of gesture action ("roga"), which is located on the right side of the interaction partner's head. For the feature extraction process, we rely on features that are invariant to scale, translation, and rotation as well that. Hu moments are chosen from a set of possible candidates, since their computation is very fast, and therefore

they are advantageous for real-time constraints [14]. They are calculated from the roga in the skin color image.

C. Face Pose and Shape

Geometric models form an abstraction of real-world objects and contain knowledge about their properties, such as position, shape or texture and represent them in a parameter vector. Model-fitting is the computational challenge of determining correct model parameterizations for single images without prior knowledge of the image content. We integrate the Candide-III face model and rely on the work of Wimmer et al. for model fitting [33], [22]. To obtain descriptive information about head gestures, the temporal changes of the in-plane transition of the face and the temporal change of the three rotation angles (pitch, yaw and roll) are extracted. Facial expression are generally characterized by two important aspects: 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. The static features are assembled by calculating the model parameters for a single image whereas the person-adapted features are calculated as the model parameter change between a neutral reference image of the person visible in the image sequence and the current image.

D. Face Texture

We extract texture from the face region by mapping it to a reference shape. The 3D model points of the face model are projected into the image and delauny triangulation is applied. Texture warping between the triangles is performed using affine transformation. The extracted data is reduced using PCA to obtain a parameter vector. The large amount of data is reduced to roughly 15%. In addition, temporal features of the facial changes are also calculated. The local motion of predefined feature points is observed using optical flow and the relative location of these points is connected to the structure of the face model. To determine robust descriptors, again PCA is applied. This reduces the number of descriptors by enforcing robustness towards outliers as well. Please note the difference between person-adapted features, which are calculated from the model parameters, and the temporal features, which are directly calculated from the image data.

IV. Classification

In this sections, we will delineate our parallel classification approach. Note, that no knowledge of the exact starting point of a single gesture is required. We exploit the shared memory characteristics of the RTDB to process the same obtained feature vectors multiple times in parallel. Note, that with this approach, these classifiers share a certain amount of features elements, however, the temporal length of the investigated feature vector sequence varies for each classification process.

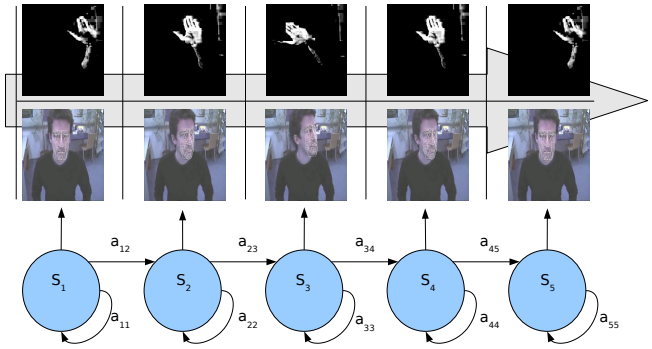

Figure 2. We apply left-right Hidden Markov Models for gesture recognition. The Model states reflect the current execution state of a single gesture, here demonstrated with the example gestures "head shaking" and "waving".

A. Person Identity

Real-time capability is important, and therefore, a small number of feature points is considered only. We use reduced descriptors to trade off between accuracy and run time performance. The static, textural and temporal features, which are extracted as described in Section III-C and Section III-D, are used for classification. We combine them into a single feature vector. In order to fuse the features, we normalize them to avoid the dominance of one type of feature over the other. Single image information is considered by the structural and textural features, whereas image sequence information is considered by the temporal features. Our approach achieves real-time performance and provides robustness against facial expressions in real-world scenarios.

B. Head and Hand Gestures

We apply continuous Hidden Markov Models as described in [7] with a left-right structure for the classification process as visualized in Figure 2. The HMM is presented with sequences of varying length by applying different sliding window sizes on the data buffered by the RTDB. To reduce the computational demand of this parallel classification process, the amount of applied windows is limited to three. These window sizes have correspondences to the time interval, where a gesture is completely performed under normal circumstances in the everyday life. Due to the fact that dynamic hand gestures have shorter duration in time, the maximum window size is smaller than the window size for the head gestures. The head gesture classification is based on the temporal face pose change, which is calculated as described in Section III-C. Three different head gestures are evaluated by our approach: Head nodding, head shaking and a neutral head position. For the hand gestures, we evaluated five different gestures: grasp, lay, no (waving index finger), up, and wave and their calculation is based on the skin color image features, see Section III-B

C. Facial Expression

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 and MMI Face Database 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. We apply decision trees and SVMs to map the feature vector on the facial expression visible in a single image. The classification is provided the static and person-adapted features, see Section III-C. 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 one third maximum intensity as neutral face representatives. A model tree is trained to infer the facial expression intensity [26]. 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 and the final intensity is calculated from line segments in the leaf nodes. Figure 3 depicts some example images that are captured from a real-world application of the system.

V. 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 [21].

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. 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 movements, the eyes are movable in two degrees of freedom (pan/tilt). Video data is streamed with a resolution of 640x480 with 30 frames per second. For further details on the actual design of the head, please refer to [10]. The head can be controlled via an embedded control architecture over either a TCP/IP connection or an interface to the RTDB [11]. 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.

The emotional state of the display can be controlled in two ways, either referring to the discrete basic emotions found by Ekman et al. or referring to the circumplex model of affect proposed by Russel et al. [32], [29]. 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 4. 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 pleasure and an arousal axis. 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 [30]. Evaluation studies prove that the generated facial expressions can be correctly identified by untrained users and matched to the corresponding emotional state of the robot [10]. In the preliminary setup the robot head is merely mirroring the facial expression recognized from the human interaction partner as described in Section IV-C by imitating the activation of facial action units.

VI. Evaluation

This section presents experimental evaluations of the components integrated into the complete system. Stratified cross-validation is applied to determine recognition accuracies.

A. Person Identity

We experiment on subjects of the Cohn-Kanade Facial Expression Database. The model is fit onto the single images

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

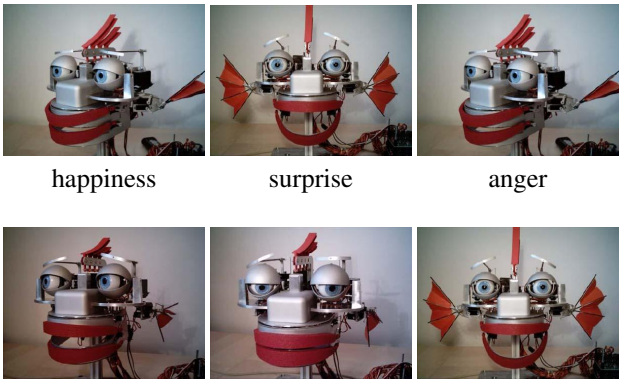

Figure 4. EDDIE displaying the basic facial expressions proposed by Ekman et al. [32].

and static, textural and temporal features are extracted. The recognition results are obtained in the presence of strong facial expressions but restricted to frontal face views only. We used binary decision trees with 10-fold cross validation to avoid over-fitting. In a preliminary experiment, only the texture information is considered for person identification. A successful recognition rate of 92.88% is obtained. However, applying the complete feature set, the recognition rate is recorded to be 99.52% which demonstrates that considering facial expression during face recognition is more useful than neutralizing them.

B. Head and Hand Gestures

In total 24 image sequences of 14 persons were used for the evaluation process. Six-fold cross-validation is applied to calculate the recognition accuracies presented in Table I. For the hand gestures, we evaluated five different gestures: grasp, lay, no (waving index finger), up, and wave. For the performance evaluation a Hidden Markov Model with 13 states is used.

Classified As	Sequence Label		
	Shaking	Neutral	Nodding
Shaking	99%	1%	0%
Neutral	9%	83%	8%
Nodding	3%	11%	86%

Table I
ARITHMETIC AVERAGE OVER THE RESULT OF THE HEAD GESTURE RECOGNITION.

Classified As	Sequence Label				
	Grasp	Lay	No	Up	Wave
Grasp	71%	12%	0%	0%	17%
Lay	0%	100%	0%	0%	0%
No	0%	0%	100%	0%	0%
Up	0%	0%	0%	100%	0%
Wave	0%	0%	0%	0%	100%

Table II
ARITHMETIC AVERAGE OVER THE RESULTS OF THE HAND GESTURE RECOGNITION.

C. Facial Expression Recognition

The Cohn-Kanade Facial Expression Database serves as data to train and test classifiers that recognize the degree of each universal facial expression. Each sequence starts with a neutral face and develops into the apex expression. In addition, the MMI Face Database image sequences also show decreasing facial expressions. 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-adapted features only and adapts to the person visible in the image. The classifier C3 includes both static and person-adapted features, with the intention of combining the "best of both worlds". The classifiers are evaluated by a 10-fold cross-validation on both databases as presented in Table III.

	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%

Table III
RECOGNITION ACCURACIES FOR DIFFERENT FEATURE SETS AND CLASSIFIERS. THE NUMBERS ARE GIVEN FOR COHN-KANADE FACIAL EXPRESSION DATABASE AND MMI FACE DATABASE, RESPECTIVELY.

VII. CONCLUSION AND FUTURE WORK

We introduced a real-time capable framework for human-machine interaction considering person identity, facial expressions, head and hand gestures. The system has not reached its ultimate state, nonetheless, we are sanguine that the system will develop towards a holistic framework for human-robot interaction.

For the subsequent steps, we anticipate and tackle the following challenges: First, integration of additional depth information to improve hand segmentation and gesture classification as well. A second point for improvement will be extension of the gesture vocabulary. Third, a fusion of results originating from different classification methods is imaginable.

ACKNOWLEDGMENT

This ongoing work is supported by the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see www.cotesys.org for further details and information. This system has been implemented with partial support of the Technische Universität München – Institute for Advanced Study, funded by the German Excellence Initiative.

REFERENCES

[1] D. Vernon, G. Metta, and G. Sandini, "A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents," *Evolutionary Computation, IEEE Transactions on*, vol. 11, no. 2, pp. 151–180, 2007.

[2] M. Beetz, F. Stulp, B. Radig, J. Bandouch, N. Blodow, M. Dolha, A. Fedrizzi, D. Jain, U. Klank, I. Kresse, A. Maldonado, Z. Marton, L. Mösenlechner, F. Ruiz, R. Bogdan Rusu, and M. Tenorth, "The assistive kitchen — a demonstration scenario for cognitive technical systems," in *IEEE 17th International Symposium on Robot and Human Interactive Communication (RO-MAN), Muenchen, Germany*, 2008, Invited paper.

[3] C. Lenz, S. Nair, M. Rickert, A. Knoll, W. Rösel, A. Bannat, J. Gast, and F. Wallhoff, "Joint Actions for Humans and Industrial Robots: A Hybrid Assembly Concept," in *Proc. 17th IEEE International Symposium on Robot and Human Interactive Communication*, Munich, Germany, 2008.

[4] C. M., Y. Yacoob, and L. Davis, "Recognition of head gestures using hidden markov models," in *In Proceeding of ICPR*, 1996, pp. 461–465.

[5] L.-P. Morency and T. Darrell, "Head gesture recognition in intelligent interfaces: the role of context in improving recognition," in *Proceedings of the 11th international conference on Intelligent user interfaces*, 2006, pp. 32–38.

[6] T. S., J. Weaver, and A. Pentland, "Real-time american sign language recognition from video using hidden markov models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1371–1375, 1998.

[7] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE 77*, 1989.

[8] M. Kölsch and M. Turk, "Fast 2d hand tracking with flocks of features and multi-cue integration," in *In IEEE Workshop on Real-Time Vision for Human-Computer Interaction (at CVPR)*, 2004, p. 158.

[9] J. Do, J. Jung, S.H. Jung, H. Jang, and Z. Bien, "Advanced soft remote control system using hand gesture," in *5th Mexican International Conference on Artificial Intelligence, Apizaco, Mexico*, November 2006, pp. 215–220.

[10] S. Sosnowski, K. Kuhnlenz, and M. Buss, "EDDIE - An Emotion-Display with Dynamic Intuitive Expressions," in *The 15th IEEE International Symposium on Robot and Human Interactive Communication. ROMAN 2006*, University of Hertfordshire, Hatfield, United Kingdom, 6-8 September 2006, pp. 569–574.

[11] 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.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Conference on Computer Vision and Pattern Recognition1*, 2001, pp. 511–518.

[13] K. Nickel and R. Stiefelhagen, "Visual recognition of pointing gestures for human-robot interaction," 2007, pp. 1875–1884, Elsevier.

[14] M. K. Hu, "Visual pattern recognition by moment invariants," *IRE Transaction on Information Theory*, pp. 179–187, 1963.

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

[16] W. Zhao, R. Chellapa, A. Rosenfeld and P.J. Philips, Face Recognition: A Literature Survey In *ACM Computing Surveys* Vol. 35, No. 4, December 2003, pp. 399-458.

[17] Turk M. A., Pentland A. P., Eigenfaces for Recognition *Journal of Cognitive Neuroscience 3 (1)* 1991, pp 71-86.

[18] Cootes T. F., Edwards G. J., Taylor C. J. Active Appearance Models. In *Proceedings of European Conference on Computer Vision* Vol. 2, pp. 484-498, Springer, 1998.

[19] G. J. Edwards, T. F. Cootes and C. J. Taylor, Face Recognition using Active Appearance Models *in Proceeding of European Conference on Computer Vision* 1998 vol. 2, pp-581-695, Springer 1998.

[20] V. Blanz, T. Vetter Face Recognition Based on Fitting a 3D Morphable Model. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol.25 no. 9, pp 1063 - 1074, 2003.

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

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

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

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

[25] D. R. Carney and C.R.Colvin The Circumplex Structure of emotive social behaviour, *Northeastern University* 2005

[26] R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993

[27] P. Ekman and W. Friesen. *The Facial Action Coding System: A Technique for The Measurement of Facial Movement*. Consulting Psychologists Press, San Francisco, 1978.

[28] 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.

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

[30] 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*

[31] C. Mayer and M. Wimmer and B. Radig Adjusted Pixel Features for Facial Component Classification Image and Vision Computing Journal, 2009

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

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

[34] 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

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