VI International Conference on Computational Bioengineering
ICCB 2015
M. Cerrolaza and S.Oller (Eds)

A MULTIMODAL NONVERBAL HUMAN-ROBOT COMMUNICATION SYSTEM

S. SALEH[†*], M. SAHU[†], Z. ZAFAR[†] AND K. BERNS[†]

[†] Robotics Research Lab. - Dept. of Computer Science
University of Kaiserslautern
Kaiserslautern, Germany
web page: http://agrosy.cs.uni-kl.de
e-mail: {saleh, sahu, zafar, berns}@cs.uni-kl.de

[*] Dept. of Computer Science, University of Basrah
Basrah, Iraq

Key words: HRI, Facial Expression Recognition, Nonverbal Communication

Abstract. Socially interactive robot needs the same behaviors and capabilities of human to be accepted as a member in human society. The environment, in which this robot should operate, is the human daily life. The interaction capabilities of current robots are still limited due to complex inter-human interaction system. Humans usually use different types of verbal and nonverbal cues in their communication. Facial expression and head movement are good examples of nonverbal cues used in feedback. This paper presents a biological inspired system for Human-Robot Interaction (HRI). This system is based on the interactive model of inter-human communication proposed by Schramm. In this model, the robot and its interaction partner can be send and receive information at the same time. For example, if the robot is talking, it also perceive the feedback of the human via his/her nonverbal cues. In this work, we are focusing on recognizing the facial expression of human. The proposed facial expression recognition technique is based on machine learning. Multi SVMs have been used to recognize the six basic emotions in addition to the neutral expression. This technique uses only the depth information, acquired by Kinect, of human face.

1 INTRODUCTION

We are living in a world where technology is improving at a mind boggling speed. From just a wheel to an automatic driver-less cars, researchers have made human life much more effortless than it has ever been. The concept of a robot to do the vigorous tasks has been a thing in the past. The challenge is to build autonomous robots that are capable of interacting with humans. There has been few efforts already made to manufacture social service robots in the recent years. Such robots, geared with complex mechanical parts, should be able to interpret human emotions and provide assistance in real life environment.

Generally, the interaction between humans is not limited to verbal communication but also can be extended to non-verbal communication. Statistical research has claimed that, more than 65% of all the communication inferred from non-verbal communication [1]. From whole body gestures to facial expressions, non-verbal communication plays a significant role in

human daily life. Whether it is a work environment, while presenting a talk or it is a home environment while parenting a kid, non-verbal communication cannot be overlooked. Hence, a robot should interpret human verbal and non-verbal communication in order to interact with humans efficiently.

Humans use different modalities while interacting with each other. Speech is the basic modality used in daily life routine but as established before verbal content is not enough in efficient interaction. Humans also interact through facial expressions and express more feelings through facial gestures than any other body movement. Human facial expression is also the primary source of expressing emotions and feelings. In this context, there is a need of robust facial expression detector that can identify and recognize the internal state of human through Facial expression recognition.

In the paper, we present a perception and cognition model to realize the nonverbal communication between humans and robots. The perception part detects human faces, estimates head poses and recognize facial expressions. A proposed feature extraction method has been used in head pose estimation as well as in facial expression recognition. The cognition part fuses the detected information over time in order to recognize and interpret the partner's nonverbal cues as a feedback.

The remainder of this paper is organized as follows: In Section 2, a brief related work will be presented. Section 3 discusses the social communication and the model that has been used. Section 4 presents the proposed approach. Section 5 shows the experiments. In Section 6, conclusion and future work will be discussed.

2 RELATED WORK

Human facial expression recognition has been the center of attention in the field of human emotion and action recognition since last decade. Lots of researchers have attempted to address this challenge. However, due to variations in human appearances and facial deformations, nobody has solved this problem on a general scale.

Most of the researchers used different pattern recognition approaches that are primarily based on two dimensional facial features. Most common features in context to FER are geometric features such as the shapes of facial components (for example eyes, nose, mouth, eye brows etc.) and the position of facial points (lips, nose tip, chin, iris etc.) or the appearance features reporting the information of facial texture, wrinkles, frowns etc. Typical examples of geometric features based methods are Pantic et al. [2, 3, 4], Chang et al. [5] and Kotsia and Pitas [6]. Typical examples of appearance features based methods include Bartlett et al. [7, 8], who used gabor wavelets, Whitehill and Omlin [9], who used Haar features and Valstar et al. [10] who used temporal templates.

Both types of features has its own limitations. Using both features simultaneously can be the best choice for establishing a facial expression recognition system. Example of such hybrid use of features is Active Appearance Model (AAM) used by Lucey at al. [11] that extracts the essence of both features types. The scheme [11] builds upon active appearance models (AAMs) technique where AAM models fit shape and appearance components through a gradient descent search method. Support vector machine (SVM) is used for classification purpose. The critical part in AAMs approach lies in model fitting. Fitting a model perfectly on

a human face is challenging task and require highly efficient algorithms.

In context to human-robot interaction there have been few works reported in the literature. Cid et al. [12] presented appearance based approach. They applied gabor filters and the output edge image of this filters is used for detecting and extracting scale-invariant facial features. This feature extraction is robust and real time and they use Dynamic Bayesian Network to classify these features. They recognize universal facial expressions with more than 90% accuracy. The approach works well with static facial expressions but behaves poorly when dealing with dynamic facial expressions. Moreover, the approach only works with near-frontal faces.

Li and Hashimoto [13] presented a geometric approach. They detect facial components and then extract the positions of facial points such as height, width and angles. They used these features to recognize human emotion in HRI. Other than six basic facial expressions they also recognize some other expressions. Their approach is highly dependent on correct recognition of facial points and there can be a false recognition due to illumination variation and dynamic motion of head.

The current work uses the nonverbal cues as a back channel for communication that can be perceived by the robot to assess the human interactivity. A biological inspired interaction model has been used in order to design a human-like interaction behavior. This model is based on the interactive model presented by Schramm [14]. This interactive model will be discussed in the next section.

3 SOCIAL COMMUNICATION

Communication is the process in which humans exchange information and meanings using various technical or natural means. Many models have been proposed by scientists to describe inter-human communication. Most of these models are based on the mathematical linear model of Shannon and Weaver [15]. One of the important models was presented by Schramm [14]. He introduced an interactive model for communication. He believed that communication is a two way process between two individuals. He also used the concept of “interpreter” to analyze the meaning of messages. An important contribution Schramm made was to consider the fields of experience, or a common ground, of the sender and receiver. The sender encodes the message according to his/her field of experience. The receiver's field of experience guides decoding process. This model assumes that communication is circular and feedback is a central feature. Fig. 1 shows Shramm's interactive model.

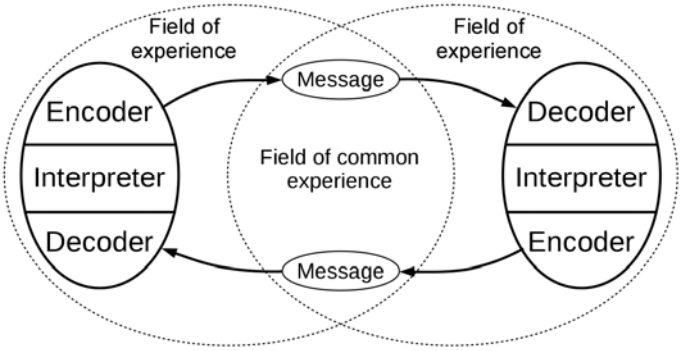

Figure 1: The interactive model of communication presented by Schramm [14].

As stated by Schramm, the communication between two individuals is highly depending on the overlap between the fields of experience of the two partners. In the present work, the interactive model of Schramm has been used to design a communication system between the robot and humans. In order for a natural, or smooth, interaction with a robot, it should be provided with some human skills of communication. One important skill of human communication is understanding of nonverbal cues. The robot should be able to decode and interpret the nonverbal cues of the interaction partner.

4 THE PROPOSED SYSTEM

The proposed system is based on the interactive model of communication by Schramm [14] in order to facilitate the interactivity process between human and robot. In this model, the robot and its interaction partner send and receive information at the same time. The robot sends information via speech and receive the feedback from the human via nonverbal cues (visual channel). This system consists of three main modules; human perception (decoder), cognition of nonverbal cues (interpreter), and the dialog system (encoder). In the following, the first two modules will be described in details. Fig. 2 depicts the proposed system.

4.1 Human Perception

The perception of human using only 2D images encounters serious problems due to its sensitivity to illumination and shadow. In addition to its cameras, a humanoid robot, usually, has multiple additional sensors such as those that provide depth information. These sensors can be used to enhance the human perception process. Thus, by doing so the interaction can be more efficient and more accurate.

This paper uses depth information provided by a Microsoft Kinect sensor in addition to RGB images to perceive the interaction partner in order to design a stable nonverbal communication system. To achieve good quality of human perception, following sub components are necessary.

4.1.1 Face Detection

One of the essential skills of robot interacting with humans is face detection. In computer vision terms, the face detection task is not easy, even though that humans can do it effortlessly [16]. The goal of face detection is to determine whether or not there is any face in a given image and returns the location and size of the face [17]. The challenges associated with face detection can be attributed to many variations in scales, locations, orientations, poses, facial expressions, lighting conditions, occlusions, etc. A fast and reliable face detection process is the most necessary condition for efficient human-robot interaction.

The present paper detects human face in two stages using RGB and depth images as in our previous work [18]. Experiments have shown that the use of depth information reduces the false positives tremendously. The location of the detected faces then passed to the next step to estimate the poses.

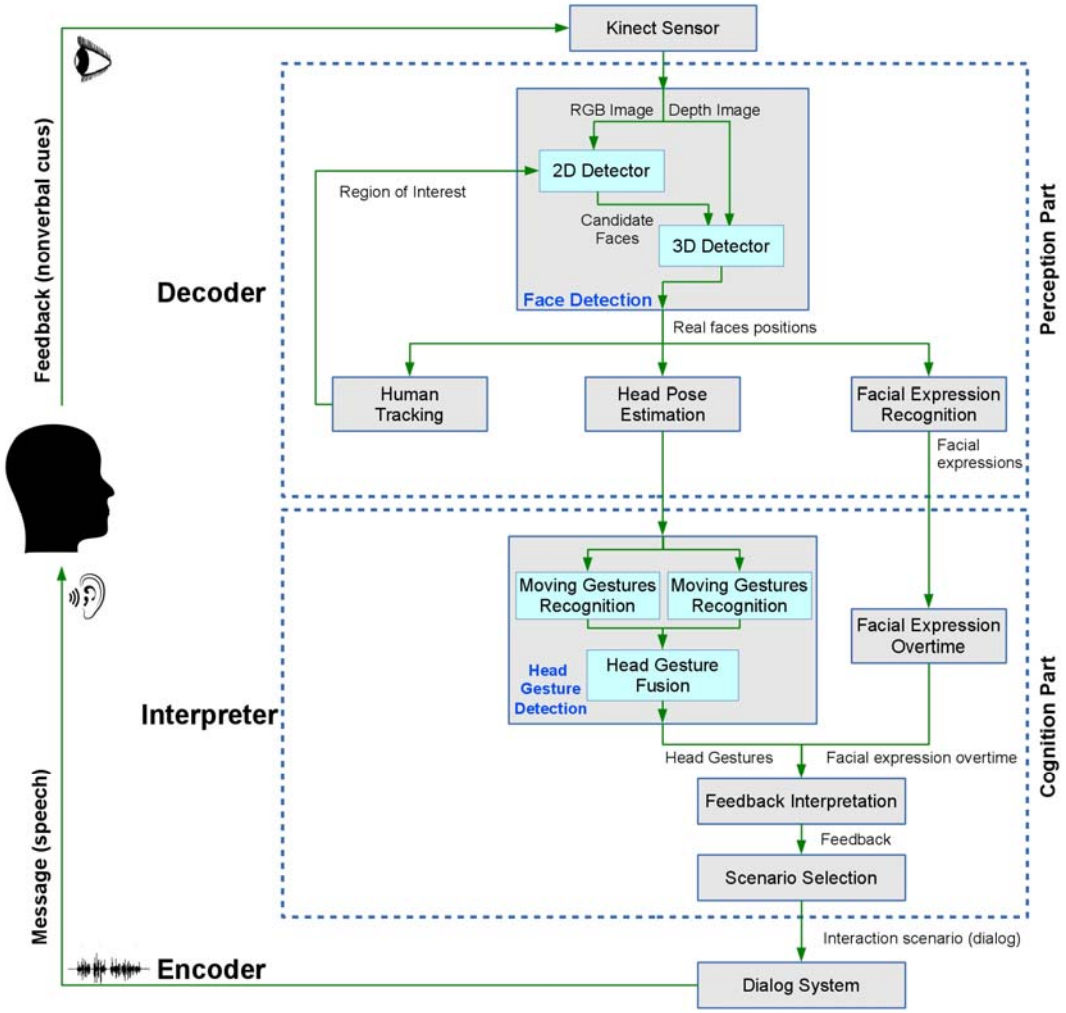

Figure 2: The proposed system overview. It consists of three modules: perception, cognition, and the dialog system. These three modules analogy to the decoder, interpreter, and encoder respectively.

4.1.2 Head Pose Estimation

Another nonverbal aspect, which this paper addresses, is a human head movement. Humans have the ability of interpreting these movements quickly and effortlessly, but it is regarded as a difficult challenge in computer vision and robotics. Detecting human head movement requires estimating head pose over the time. In order to build a robust human-robot interaction system, a robust and reliable head pose estimation algorithm is needed.

In addition to the sensitivity to illumination, the head pose estimation in 2D images suffer from the lack of features due to occlusion in some poses [19]. The present paper uses only depth information in head pose estimation. A Gaussian smoothing filter is used to remove the noise from each depth frame. Three linear Support Vector Machines (SVMs) for regression are trained to detect the pose angles *roll*, *pitch* and *yaw* [18]. The DMP (Direction-Magnitude Pattern) feature descriptor has been used in head pose estimation and facial expression recognition [20].

4.1.3 Facial Expression Recognition

Humans communicate effectively and are responsive to each other's emotional states. Facial expressions provide powerful cues to people's inner thoughts and emotions. Psychological studies have shown that facial expression is one of the most important modalities used in feedback. Smiles and head nods are the most frequent feedback gestures[21].

The proposed method is based on appearance analysis so the factors affecting the appearance of the face come naturally. The most annoying factors affecting the facial appearance are pose, lighting conditions and resolution. Therefore, some constraints were forced in this work, like we considered only nearly frontal faces. Moreover, the proposed work uses depth information for facial expression recognition, as 3D modality has advantage over 2D modality under the same feature extraction and classification algorithms [22]. The DMP feature descriptor, used in head pose estimation, has also been used in the facial expression recognition. This reduces the computation time for calculating features for each perception part in the whole system because it will be calculated once. Six SVMs with linear kernel function have been used to recognize the six basic expressions. Fig. 3 shows the facial expression recognition process.

SVM makes binary decisions, so six binary classifiers were created by using the one-versus-rest technique, which trains binary classifiers to discriminate one expression from all others. With regard to the parameter selection of SVM, we carried out several runs on the hyper-parameters and the parameter setting producing best accuracy was picked.

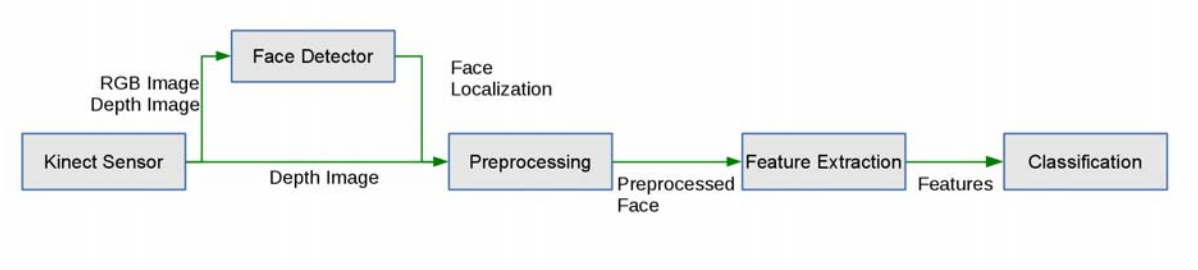

Figure 3: The proposed facial expression recognition system overview. The depth information of human face is preprocessed before extracting the features. The features are then used in the classification step.

4.1.5 Human Tracking

Human face tracking is important aspect in natural interaction. A Kalman filter is used for human face tracking. It also decreases the processing time of face detection. It predicts the position of the face in the next frame to reduce the search space. Depending on the face position and size in the next frame, a Region Of Interest (ROI) is determined to reduce the search space. The ROI is searched to detect the face in the next frame rather than the whole frame. This reduces detection time immensely.

4.2 Interpretation of Feedback

During inter-human conversations, people use visual channel to understand the feelings, emotions, and intentions. The interactive robots need also to look at the interaction partner to

interpret his/her movements and facial expressions as a feedback. Depending on the feedback, robot can select a suitable scenario for the conversation. The perception module of the present work plays a crucial role in the whole interaction process. In the perception module, the robot perceives a lot of information about the interaction partner. This information includes the human face, head pose over time, and facial expressions. The perceived information needs to be categorized and interpreted. The cognitive module, using the perceived information, detects and interprets the head gestures of the interaction partner and the accompanying facial expressions. The most significant cues considered in the present work are nodding, shaking, head up, head down, and head tilt as well as smiles. Detail information about head gestures can be found in [20].

5 EXPERIMENTS

5.1 Human Face Data

For experimental analysis, we considered a publicly available 3D face database Bosphorus[23], which is based on FACS coded Facial Action Units and emotion expressions, containing 105 subjects (60 men and 45 women) expressing multiple emotions in different poses and occlusion conditions.

In our experiments, we chose only those subjects which offer the basic seven expressions (anger, sad, happy, disgust, surprise, fear and neutral). We partitioned the data set into two subsets of variable size, trying to determine the train-test ratio for best recognition results.

5.2 Training

In order to obtain a high recognition rate, experiments were conducted with a combination of image resolution, image sub-region division, and the type of SVM as parameters. One additional parameter required for DMP features is the threshold set.

The lower resolution images were obtained by down-sampling the original images (faces) into 130x98, 114x86, 98x74, 82x62, 66x50, 50x38 or 34x26 image resolutions for different experiments. For calculation of DMP features, these low resolution images were divided into 3x3, 4x4, 5x5, 6x6, 7x7 or 8x8 sub regions. The descriptors are calculated by overlapping between adjacent regions. The descriptors produced by every combination of image resolution and sub-region size are fed into Linear SVM and Poly SVM. A number of possible thresholds were also taken into account for calculating the DMP features so as to find the best possible threshold values suitable for the facial expression recognition task.

5.3 Results

The results have shown that the face resolution, 82x62, produces best recognition rates among all considered face resolution. Fig. 4 shows that face sub-regions 6x6 is an optimal choice for our applications among all other sizes for face sub-regions. The threshold values for the DMP features that have been found empirically were (5, 10, and 30), which contributed in the best recognition rate. Fig. 5 shows the recognition rates for all of the expressions.

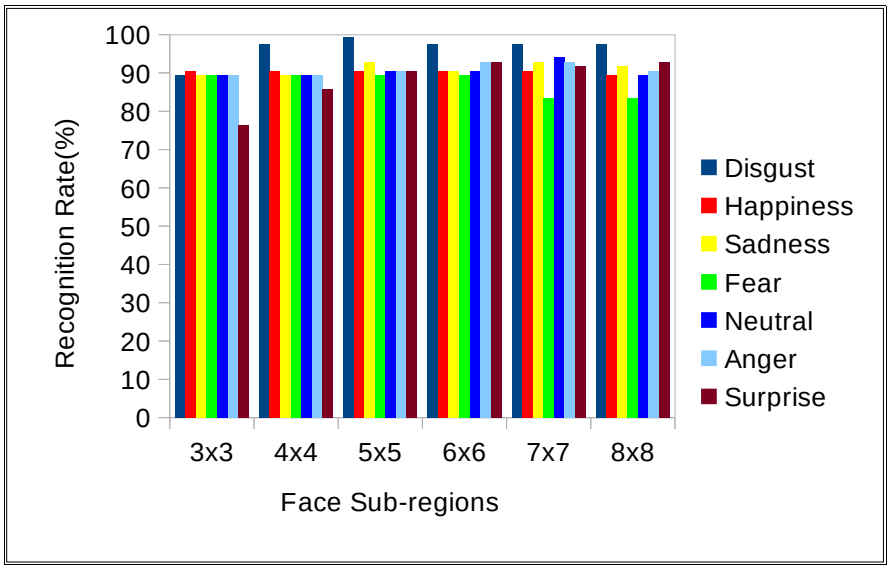

Figure 4: Recognition rate for different sizes for face sub-regions.

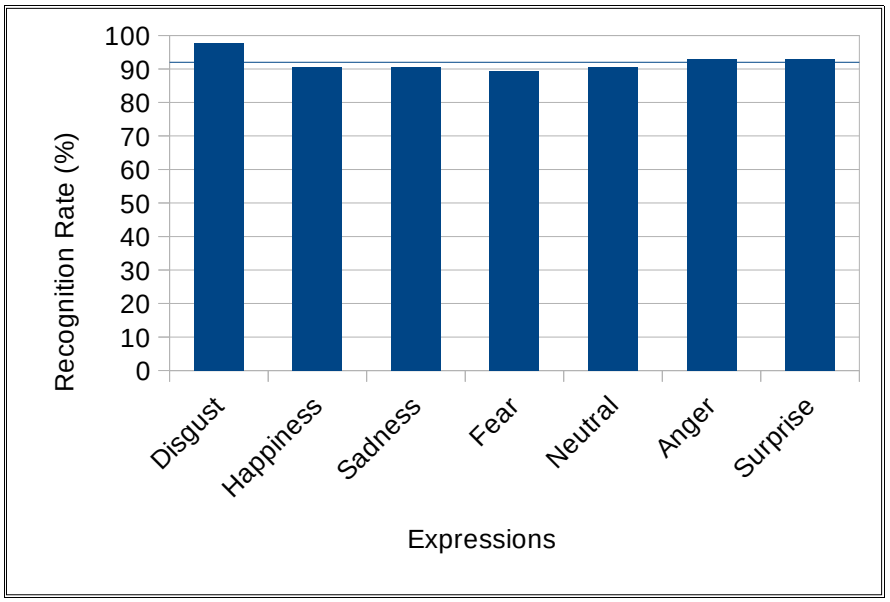

Figure 5: Recognition rate for all expressions

6 CONCLUSION AND FUTURE WORK

The paper presented an interaction model for communication with human nonverbally. The nonverbal communication used in this work includes head gestures and facial expressions. A combination of RGB images and depth information is used to improve the human perception process. The human perception process includes face detection, face tracking, head pose estimation, and facial expression recognition.

We proposed a method for facial expression recognition that is based on integrating depth and color images for describing the expression of a face image. We performed experiments on

a 3D face database (Bosphorus) for 6 basic expressions and demonstrated the effectiveness of the method on this database. A detailed analysis of performance is also carried out.

After the perception of the interaction partner, the cognition module should be able to interpret the human emotions and intentions. It combines the head gestures and facial expressions overtime and can be interpreted as a feedback in human-robot interaction.

As a future work, the system can be extended to include hand gestures as well as body postures in the nonverbal communication. Other nonverbal expressions like consciously performed gestures should be also recognized in order to improve the robot's interactive capabilities. Including a speech recognition system will improve the interaction with human tremendously. Another improvement could be stated is considering gestures for different cultures and learning new gestures from people.

REFERENCES

[1] Hogan, K. and Stubbs, R. *Can't get Through 8 Barriers to Communication*. Pelican Publishing, 2003.

[2] Pantic, M. and Bartlett, M.S. Machine Analysis of Facial Expressions, *Face Recognition*, K. Delac and M. Grgic, eds., pp. 377-416, I-Tech Education and Publishing, 2007.

[3] Pantic M. and Patras, I. Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments Form Face Profile Image Sequences, *IEEE Trans. Systems, Man, and Cybernetics* Part B, vol. 36, no. 2, pp. 433-449, 2006.

[4] Valstar, M. Pantic, M. Ambadar, Z. and Cohn, J.F. Spontaneous versus Posed Facial Behavior: Automatic Analysis of Brow Actions, *Proc. Eight Int'l Conf. Multimodal Interfaces* (ICMI '06), pp. 162-170, 2006.

[5] Chang, Y. Hu, C. and Turk, M. Probabilistic Expression Analysis on Manifolds, *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition* (CVPR '04), vol. 2, pp. 520-527, 2004.

[6] Kotsia, I. and Pitas, I. Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines, *IEEE Trans. Image Processing*, vol. 16, no. 1, pp. 172-187, 2007.

[7] Bartlett, M.S. Littlewort, G. Frank, M. Lainscsek, C. Fasel, I. and Movellan, J. Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition* (CVPR '05), pp. 568-573, 2005.

[8] Bartlett, M.S. Littlewort, G. Frank, M.G. Lainscsek, C. Fasel, I. and Movellan, J. Fully Automatic Facial Action Recognition in Spontaneous Behavior, *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition* (AFGR '06), pp. 223-230, 2006.

[9] Whitehill, J. and Omlin, C.W. Haar Features for FACS AU Recognition, *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition* (AFGR '06), pp. 217-222, 2006.

[10] Valstar, M. Pantic, M. and Patras, I. Motion History for Facial Action Detection from Face Video, *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics* (SMC '04), vol. 1, pp. 635-640, 2004.

[11] Lucey, S. Ashraf, A.B. and Cohn, J.F. Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face, *Face Recognition*, K. Delac, and M. Grgic, eds., pp. 275-286, I-Tech Education and Publishing, 2007.

[12] Cid, F. Prado, J.A. Bustos, P. Nunez, P. A real time and robust facial expression recognition and imitation approach for affective human-robot interaction using Gabor filtering, *Intelligent Robots and Systems (IROS)*, 2013 IEEE/RSJ International Conference on , vol., no., pp.2188,2193, 3-7 Nov. 2013

[13] Yi Li and Hashimoto, M., Effect of emotional synchronization using facial expression recognition in human-robot communication, *Robotics and Biomimetics (ROBIO)*, 2011 IEEE International Conference on , vol., no., pp.2872,2877, 7-11 Dec. 2011

[14] Schramm W. *The Process and Effects of Mass Communication*. University of Illinois Press, 1974.

[15] Shannon, C. E. and Weaver, W. *The mathematical theory of communication*. Illini books edition. Univ. of Illinois Pr., Urbana, 1972.

[16] Hjelmas, E. and Low, B.K. Face detection: A survey. *In Computer Vision and Image Understanding*, volume 83, pages 236–274, Sept. 2001.

[17] Yang, M.H. Kriegman, D.J. and Ahuja, N. Detecting faces in images: a survey. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 24, pages 34–58, Jan 2002.

[18] Saleh, S. Kickton, A. Hirth, J. and Berns, K. Robust perception of an interaction partner using depth information. *In Proceeding of the International Conference on Advances in Computer-Human Interactions (ACHI)*, Nice, France, February 24-March 1 2013.

[19] Breitenstein, M. Kuettel, D. Weise, T. van Gool, L. and Pfister, H. Real-time face pose estimation from single range images. *In Computer Vision and Pattern Recognition*, 2008. CVPR 2008. IEEE Conference on, pages 1 –8, june 2008.

[20] Saleh, S. and Berns, K. Nonverbal Communication With a Humanoid Robot Via Head Gestures. *In PETRA'15: Proceedings of the 8th International Conference on PErvasive Technologies Related to Assistive Environments*, ISBN 978-1-4503-3452-5. Curfo, Greece, 1-3 July 2015.

[21] Paggio, P. and Navarretta, C. Feedback and gestural behaviour in a conversational corpus of Danish. In *NEALT (Northen European Association of Language Technology) Proceedings Series. 2011*, pp. 33-39.

[22] Savran, B. Sankur, M.T. Bilge, Comparative Evaluation of 3D versus 2D Modality for Automatic Detection of Facial Action Units , *Pattern Recognition*, Vol. 45, Issue 2, p767-782, Feb. 2012.

[23] Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, Bosphorus Database for 3D Face Analysis, *The First COST 2101 Workshop on Biometrics and Identity Management (BIOID 2008)*, Roskilde University, Denmark, 7 - 9 May 2008.