Nonverbal Communication With a Humanoid Robot
Via Head Gestures

Salah Saleh*
Robotics Research Lab.
Dept. of Computer Science
University of Kaiserslautern
Germany
saleh@cs.uni-kl.de

Karsten Berns
Robotics Research Lab.
Dept. of Computer Science
University of Kaiserslautern
Germany
berns@cs.uni-kl.de

ABSTRACT
Social interactive robots require sophisticated perception and cognition abilities to behave and interact in a natural humanlike way. The proper perception of behavior of interaction partner plays a crucial role in social robotics. The interpretation of these behaviors and mapping them to their exact meanings is also an important aspect that interactive robots should have. This paper proposes an interaction model for communicating verbally and nonverbally with human. Human behavior, during the interaction with the robot, is perceived and then interpreted depending on the situation in which the behavior has been detected. In this model, head gestures are used as a back channel (feedback) for the robot to adapt the interaction scenario. The back channel signals can be consciously or unconsciously generated by human. Simultaneously, the eye gazes are also detected to ensure right interpretation of head gestures. In order to recognize the human head gestures, head poses have been tracked over time. A stream of images with their corresponding depth information, acquired from a Kinect sensor, are used to find, track, and estimate the head poses of human. The proposed model has been tested in various experiments with different scenarios in interaction with human.

Categories and Subject Descriptors
I.2.10 [Vision and Scene Understanding]: 3D/stereo scene analysis, Motion; I.4.8 [Scene Analysis]: Depth cues, Motion, Time-varying imagery, Tracking

General Terms
Design, Experimentation, Human Factors

Keywords
Human-Robot Interaction, Nonverbal Communication, Humanoid Robot

1. INTRODUCTION

The demand for a social and interactive robot with humanlike skills is increasing rapidly due to the wide areas of application in which this robot can be used, e.g., entertainment, housekeeping or elderly care. Nevertheless the development of such a robot still covers a small aspect of the human-robot interaction. The interactive capabilities of current robots are limited to specific scenarios. These limitations are because of the complex inter-human interaction system.

Figure 1: The humanoid robot Roman of the University of Kaiserslautern interacting with human.

Compared to traditional applications where robots cooperated with specially trained experts, interactive robots need a completely different set of abilities. The operating environment of these robots is “every day life” of humans. Therefore, they need special interaction capabilities that enable socially acceptable behavior. Moreover, they have to be able to use the tools that human use in their communications. Humans use a complex combination of speech, gestures, mimics, and body poses in their communication [19]. This combination is usually referred as verbal and nonverbal communications. Social psychologists reported that, nonverbal communications represent two-third of all the communications [8].

During face-to-face communication, people use nonverbal cues either consciously or unconsciously. Head movement is a good example of nonverbal feedback. Many studies have
shown the fundamental role of head movement in face-to-face communication. It can be used for turn-taking, answering yes/no questions, and tuning with the interaction partner. Head movements, in addition to facial expression, are the most frequent feedback-related cues in human communication [13]. The head gesture functions can be dramatically changed according to the situation and some physical properties such as amplitude, frequency, and cyclicity of the head movement. Ordinary cyclic head movements are used to signal yes and no, while wide linear movements mean turn taking. Narrow linear movements can also regard as a synchrony signal if it happens during other’s speech [5]. In this work, the above facts have been considered to design a communication system for the humanoid robot ROMAN, fig. 1. This communication system perceives the nonverbal cues of the interaction partner as a feedback for the robot and interprets them in order for a natural interaction.

The contribution of this paper is twofold. First, a biologically inspired model that realizes verbal and nonverbal communication between humans and robots. This model is based on the inter-human interactive model of communication proposed by Schramm [16]. This model consists of two parts: perception and cognition. The perception part detects human face, eye gaze, and estimates head poses. The cognition part fuses the detected information from the perception part in order to recognize and interpret the partner’s head gestures. The Head gestures are used to answer questions and to maintain the communication between the robot and the interaction partner. Second, a novel feature extraction method has been implemented for head pose estimation of the interaction partner. This method has been used with depth information of the perceived human face. It describes each pixel according to the relation between the depth of the pixel and the depth of the neighboring pixels.

The remainder of this paper is organized as follows: In Section 2, a brief related work will be presented. Section 3 present the inter-human social communication. Section 4 presents the proposed system. Section 5 presents the experiments. In Section 6, conclusion and future work will be discussed.

2. RELATED WORK

Several researchers have studied the head gestures recognition and interpretation. Jian-Zheng and Zheng [9] have presented a method for head movement recognition by locating and tracking nostril. They selected the nostril as a feature because it is stable and easily recognizable feature point in human face as they argued. Their method is based on Lucas-Kanade (LK) algorithm and GentleBoost. Sakaia and Das [14] have described a real time head gesture recognition system using optical flow based classification. They have used the adaptive Gaussian Mixture Model (GMM) in background subtraction. They only considered the gesture from videos. Terreven et al. [20] presented a method for recognizing head gestures using Hidden Markov Model (HMM). Their work considers all possibilities of the head pose in any order to detect six head gestures. Most of these studies have used only color images which are very sensitive to illumination.

Other studies have focused on using nonverbal cues to control machines. For example, Morency and Darrell [12] presented a context-based head gesture recognition technique for human-computer interaction. Their technique uses stereo-vision to perceive human and SVM-based classifier to recognize head gestures. They used head gestures for responding to computer dialog boxes and for document browsing. Gast et al. [4] presented a head gesture recognition method to be used in human-machine interaction. They used their head gesture recognition to control a coffee machine and to control a service robot. A 3D face model fitting has been used with a Hidden Markov Model to recognize the head gestures. These researches have shown good results, but they still restricted to very narrow scenarios and there is no real interaction between human and machine.

Some other researchers focused on the perception of human nonverbal cues to naturally interact with robots. Gaschler et al. [3] have shown that humans use body posture and head poses as social signals to initiate and terminate the interaction with another human. They collected and analyzed a data corpus of a lot of interactions in bars between human customer and human bartender. Based on these data, they used body posture and head pose to model and train Hidden Markov Model (HMM). Calisgan et al. [2] conducted human-human interaction experiments to understand the series of nonverbal communication cues that organize the turn-taking process. After that, they utilized these nonverbal cues to govern the human-robot collaborative communication. They have used Hidden Markov Model (HMM) to detect the nonverbal cues of the person during collaboration. The work is limited that human should behave in a restricted way to be in tune with the robot. Han et al. [6] presented a method for using nonverbal information in communications with Nao robot. They used the face position over time to detect nodding and shaking. Face detection process, speech direction detection process, and sonar sensor data are used as the conversation trigger. It is obvious that using face position for detecting head gestures is not precise enough and the movement of the human can be interpreted as a head gesture which may affect the interaction.

In spite of the diversity of the research related to the nonverbal communication, they are still limited to specific scenarios. The current work is an attempt to make the nonverbal cues as a back channel for communication that can be used by 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 [16].

3. SOCIAL COMMUNICATION

Communication is the process in which we, the human beings, exchange information and meanings using various technical or natural means. Many models have been proposed by scientists to describe interpersonal communication. Most of these models are based on the mathematical linear model of Shannon and Weaver [17]. One of the important models was presented by Schramm [16]. 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. 2 shows Schramm’s interactive model.
As stated by Schramm, the communication between two individuals is highly dependent on the overlap between the field 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 [16] 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. This system consists of two main modules; human perception (decoder) and cognition of nonverbal cues (interpreter). In the following, these two modules will be described in details. Fig. 3 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 the 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 [7]. 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 [22]. The difficulty 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 proper human-robot interaction.

The present paper detects human face in two stages. The first stage uses Haar cascade classifier [21] to detect candidate faces in 2D scene. The second stage takes the candidate faces and uses depth information to verify if they are real human or not. In this stage, some depth criteria are used to verify the candidate faces. These criteria are detailed in our previous work [15]. Depth information reduces the false positives tremendously.

The location of the detected faces then passed to the next step to estimate the poses of these faces.

4.1.2 Head Pose Estimation

Another nonverbal aspect, which this paper considers, 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 systems and robotics. Detecting human head movement requires estimating head pose (position and orientation) 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 [1]. 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 [18] are trained to detect the pose angles roll, pitch and yaw as in our previous work [15]. In this work, we propose new features to be used in head pose estimation. These features are called DMP (Direction-Magnitude Pattern). The DMP feature vector is created in the following manner:

- The examined window, human face in this case, is divided into cells.
- For each pixel in a cell, its depth value is subtracted from the depth values of its eight neighbors to get the effect of the neighbors on the current pixel.
- By assuming that these resulting neighbor values are forces, we can calculate the direction and magnitude of the resultant force of the neighbors for each pixel as in equations 1 through 4.

\[
x = \sum_{i=1}^{8} v_i \cos(\theta_i)
\]

\[
y = \sum_{i=1}^{8} v_i \sin(\theta_i)
\]

\[
magnitude = \sqrt{x^2 + y^2}
\]

\[
direction = \tan^{-1}\left(\frac{y}{x}\right)
\]
Where $v_i$ are the values of the eight neighbors and $\theta_i$ are the angles of these values (forces). The angles $\theta_i$ are $0, 45, 90, 135, 180, 225, 270, \text{ and } 315$ degrees.

- The magnitude values are then categorized into four groups depending on three thresholds $\delta_1, \delta_2, \text{ and } \delta_3$ as in formula 5. These thresholds have been selected upon a lot of experiments.

- For each cell, a histogram of frequencies of each combination of direction (8 possibilities) and magnitude (4 possibilities) is calculated.

- The feature vector is the concatenation of the histograms of all cells.

\[
\text{category} = \begin{cases} 
0 & \text{if magnitude } \leq \delta_1 \\
1 & \text{if magnitude } \leq \delta_2 \\
2 & \text{if magnitude } \leq \delta_3 \\
3 & \text{otherwise} 
\end{cases}
\]

Fig. 4 shows the DMP feature calculation method. These features are suitable for depth information because they result in surface map of the face. Fig. 5c and 5d visualize the direction and magnitude values respectively. The histogram calculation is shown in fig. 6. These features have given better results in head pose estimation compared to the previous work [15].

This information is passed to the cognition module for further processing.
4.1.3 Gaze Detection

Another essential skill of an interactive robot is gaze detection. Gaze detection is the process of measuring the point to which a human is looking [10]. It plays an important role in the interaction between human and robot [23]. An interactive robot needs to know about the attention and interest of the interaction partner through the eye gaze.

In the present work, the human gaze is detected using a color thresholding and contour finding. The human eyes are detected using the Haar cascade classifier after finding the face. Then, each eye is filtered using color thresholds to discriminate the white area (sclera) from the dark area (iris) in order to get two small images, one for iris and another for the sclera. Using OpenCV contour finding, the location and the size of the iris can be determined. The sclera around the iris is also detected using the contour finding. The rate of the size of right and left parts of the sclera determines the x-direction of the gaze. In the same way, the rate of the size of top and bottom parts of the sclera determines the y-direction of the gaze.

4.2 Cognition of Head Gestures

During inter-human conversations, people look at each other to understand the feelings, emotions, and intentions. The interactive robots need also to look at the interaction partner to interpret his/her movements. Depending on these movements, the 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, eye gaze, and head pose over time. 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. Psychologically, the head gestures can convey a lot of information to the interaction partners. They mainly tell two things. The first is how someone is agreeable and in rapport with another. The second is how someone engaged and enthusiastic to the current event, and if he/she is not - where his/her true interest lies. The most important gestures considered in the present work are nodding, shaking, head up, head down, and head tilt. Head gestures have been divided, in the present work, into moving gestures and still gestures.

4.2.1 Moving Head Gestures

The moving head gesture is a sequence of head poses over time. The moving gestures considered in the present work are nodding and shaking.

Head Nodding.

Head nod is a signal for “yes” or agreement in most cultures. In some cases, it can also be used to give a positive signal to the speaker. The speed and rhythm of the nod also affect the meaning of this gesture. Slow nodding sends a signal that the listener is understanding and interested in what the speaker is saying. Fast nodding tells the speaker to finish the talk or turn-taking [11]. If the fast nodding is accompanied by touching ear or rubbing face, then it means impatience and the speaker should move on.

In this work, if the robot receives this signal after a question from the robot, then it will be regarded as “Yes” answer. If it is received during the talk and it was slow, then it will be regarded as “I am interested in your talk” or “I understand you”. Fast nodding is interpreted as “finish your talk, it is my turn”.

This gesture can be detected, by the robot, as a sequence of poses where the pitch angle of the head varies within a specific range. If the pitch angle exceeds the thresholds \([-\theta_n, \theta_n]\), then it is regarded as a nod. The number of nods can be counted as the number of times the pitch angle exceeds the thresholds. The speed of the nods depends on the number of nods and the time in which the nods occurred.

\[
r_n = \frac{N}{t}\]

Where
- \(N\) is the number of nods
- \(t\) is the time in which the nods are occurred
- \(r_n\) is the nodding rate

\[
\text{Nodding Speed} = \begin{cases} 
\text{slow} & \text{if } r_n < \rho_n \\
\text{fast} & \text{if } r_n \geq \rho_n 
\end{cases}
\]
Head shaking is the rotation of the head from side to side. It usually means “no” or disagreement [11]. Just like with the head nod, the speed and rhythm of the shake also affect its meaning. Slow and irregular shaking, usually, signals misunderstanding while slow and regular shaking can signal disbelief. Fast shakes during the other’s speech means “this is not true, I disagree”.

In this work, if the robot receives this signal after a question from the robot, then it will be regarded as “no” answer. If it is received during the robot’s talk, then it will be regarded as “I disagree with you” or “I do not believe in what you are saying”. Fast shaking can also be interpreted as “I think that your information is not true”. It can be detected, by the robot, as a sequence of poses where the yaw angle of the head varies within a specific range. If the yaw angle exceeds the thresholds $[-\theta_s, \theta_s]$, then it is regarded as a shake. The number of shakes can be counted as the number of times the yaw angle exceeds the thresholds. The speed of the shakes depends on the number of shakes and the time in which the shakes occurred.

$$r_s = \frac{S}{t}$$

Where

- $S$ is the number of shakes
- $t$ is the time in which the shakes are occurred
- $r_s$ is the shaking rate

**Shaking Speed** = \begin{cases} 
    \text{slow} & \text{if } r_s < \rho_s \\
    \text{fast} & \text{if } r_s \geq \rho_s 
\end{cases}

Where $\rho_s$ is the shaking speed threshold.

4.2.2 Still Head Gestures

In addition to the direction of interest of the human, the still head gestures can also express the emotional state or the intention of the human. There are three basic head directions that are considered in this work. They are *Head Up*, *Head Tilt*, and *Head Down*.

*Head up* gesture or head position can tell about a neutral attitude about what is being said. It can also refer to the superiority, fearlessness, or arrogance when the head is lifted high. *Head Down* gesture signals disapproval or rejection. It shows that a negative, judgmental, or aggressive attitude exists. *Head Tilt* or *Neutral* gesture means that the person is quite comfortable and interested in what is being said. It is a positive sign because it means that the listener is in tune with the speaker. It is an important gesture for a presenter to watch and to evaluate the situation of the presentation.

The present work considers the above head gestures to select a suitable scenario for the robot communicating with human. The *Head Tilt* or *Neutral* gesture during the robot talk can be interpreted as a comforting signal, and this motivates the robot to continue in his talk. If the interaction partner was looking at somewhere else than the robot during the robot talk, then this can be interpreted as “not interested” and this will decrease the motivation of the robot to continue in his talk.

These gestures can be detected by calculating the rate of a specific head pose in an interval of time. If the rate of a specific head pose exceeds the 80% of the time, then the corresponding gesture is regarded as *active*. Otherwise, the gesture is regarded as *inactive*.

4.2.3 Pose-Gaze Conflict Processing

In head direction still gestures, the eye gaze also plays a crucial role. It can refer to the place where real interest of human lies. Even if the human head is directed towards the robot, the eye gaze tells the robot if the human is really interested in his talk or not. In this work, the robot considers the eye gaze instead of the head direction gestures when there is a conflict between the eye gaze and the still head gestures. In case of moving head gestures, like nodding or shaking, the eye gaze will be omitted.

5. EXPERIMENTS

The proposed interaction model has been implemented on the humanoid robot ROMAN of the Robotics Research Lab of the University of Kaiserslautern, Fig. 7. ROMAN is a 47 degrees of freedom expressive humanoid robot designed as a test platform for human-robot interaction topic. In addition to its four dragonfly cameras, a Kinect sensor has been attached to get a precise depth information of the environment. A real time stream from this sensor with a resolution of 640x480 is examined. The Kinect sensor in this context is used without any additional software for human or gesture detection.

Figure 7: The humanoid robot ROMAN of the University of Kaiserslautern.

As a scenario for the current experiment, the human wants to make conversation should stand facing the robot. The robot, after detecting the human standing for a while, greets the interaction partner. After that it offers, through speech, some information about different topics asking the partner to select one of these topics nonverbally using nodding or shaking (“yes” for selecting a topic and “no” for rejecting a topic). Depending on the partner’s reply, yes or no, the robot either presents the offered information or offers a different topic. After selecting a specific topic and along with the presented information, the robot watches the partner’s behavior. If the partner is interested in this information, the robot continues in its presentation; otherwise it asks the
partner if he/she wants to continue with this topic or to select another one. If the robot receives a synchrony signal from the interaction partner during the robot’s talk, like slow nod, then this will motivate it to speak more.

To evaluate the proposed model, different experiments have been conducted. Interaction sessions were recorded for 16 participants. They have been asked to interact with the robot and to answer its questions nonverbally.

5.1 Head Gestures Experiment

In order to assess the interpretation of head gestures, the human should use them in the interaction with the robot. These gestures include the moving head gestures (nodding and shaking) and still head gestures (head up, head down, head tilt). Answering Yes or “nodding” to the robot’s offer lets the robot present the offered topic. Nodding the head during the robot’s talk means that the partner agrees with this talk. Answering No or “shaking” lets the robot to offer a new topic to present. Shaking the head during the robot’s talk means that the partner doesn’t agree with this talk. The participants have been already told about these simple rules to be used in the interaction.

During the robot’s talk, the robot watches the interaction partner and evaluates the situation. If the partner is interested in the presentation (looking at the robot or head tilted at a small angle), the robot will continue in its talk. Otherwise, the robot tells the partner that he looks like not interested in this topic and offers new topics to present.

More than 90% of the head gestures are perceived and interpreted correctly. Table 1 shows the results of the head gestures recognition rate. In this table, positive means that the gesture has been recognized correctly, whereas negative means that the gesture has been recognized as different gesture.

<table>
<thead>
<tr>
<th>Gestures</th>
<th>Positive</th>
<th>Negative</th>
<th>Not Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodding</td>
<td>96%</td>
<td>4%</td>
<td>0%</td>
</tr>
<tr>
<td>Shaking</td>
<td>90%</td>
<td>2%</td>
<td>8%</td>
</tr>
<tr>
<td>Interested</td>
<td>99%</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Not Interested</td>
<td>92%</td>
<td>4%</td>
<td>4%</td>
</tr>
</tbody>
</table>

5.2 Interaction Motivation Experiment

During the interaction, the robot maintains its internal motivation. The robot’s motivation to interact increases when the interaction partner is interested, and it decreases when the interaction partner is not interested. This experiment is to demonstrate the behaviors that increase or decrease the robot’s motivation. Two scenarios were used in this experiment. In the first scenario, the interaction partner behaves as interested in the robot’s talk. In the second scenario, the interaction partner behaves as not interested in the talk. The experiment has shown that if the motivation value approaches zero, then the robot will ask the partner to be alone because the partner does not like to interact. Otherwise, it continues in his talk until there is no information to present. Figure 8 shows the relation between the motivation and the scenario selected by the interaction partner.

Figure 8: Motivation chart during interested and not interested interaction. It shows the internal motivation of the robot to continue talking. The motivation is increased when the interaction partner is interested in the talk of the robot.

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 eye behavior.

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 gaze detection. The face detection, from previous work, has been used to detect the presence or absence of the interaction partner. The head pose estimation process used three SVMs for regression, with only depth information. A new feature extraction method has been implemented to be suitable with the face depth information.

After the perception of the interaction partner, the cognition module should be able to interpret the human emotions and intentions. The cognition module detects the head gestures by finding the head poses over time. The considered gestures in the present work are nodding, shaking, head up, head down, and head tilt. The head nodding and head shaking were interpreted by the robot as “yes” and “no” respectively. The head direction gestures were interpreted by the robot as “interested” or “not interested” signal.

The developed work has been evaluated in an interactive scenario between the human and the robot ROMAN. The experiments have shown good cognition of the nonverbal cues of the interaction partner.

As a future work, the system should be extended to include facial expressions, 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.

7. ACKNOWLEDGMENTS

The authors would like to thank DAAD, Germany and the Ministry of Higher Education and Scientific Research of Iraq for funding of Salah Saleh.
8. REFERENCES


