Emotion based human-robot interaction

Karsten Berns * and *Zuhair* Zafar

Robotics Research Lab (RRLAB), Department of Computer Science, TU Kaiserslautern, Germany

Abstract. Human-machine interaction is a major challenge in the development of complex humanoid robots. In addition to verbal communication the use of non-verbal cues such as hand, arm and body gestures or mimics can improve the understanding of the intention of the robot. On the other hand, by perceiving such mechanisms of a human in a typical interaction scenario the humanoid robot can adapt its interaction skills in a better way. In this work, the perception system of two social robots, ROMAN and ROBIN of the RRLAB of the TU Kaiserslautern, is presented in the range of human-robot interaction.

1 Introduction

An increasing interest in emotion can be seen in the behavioural, biological and social sciences but also more and more in advanced robotics. Research over the last two decades suggests that many phenomenon, ranging from individual cognitive processing to social and collective behaviour, cannot be understood without taking into account affective determinants (i.e., motives, attitudes, moods, and emotions) [1]. The emerging field of affective science seeks to bring together the disciplines which study the biological, psychological and social dimensions of affect [2]. The fact that emotions are considered to be essential to human reasoning suggests that they might play an important role in autonomous robots as well.

Providing a robot with social skills has been studied for many years and still an active research area. Several challenges have been recorded in this field due to the complexity of interpersonal communication. A social robot should work in the human's daily life environment and perceive human's feelings, intentions, and demands. Furthermore, a social robot needs to recognize and understand human's emotions to interact naturally. However, the perception of human emotions is not an easy task even for humans themselves which is a challenge in humanoid robotics.

How social robots express emotions is equally significant along with perceiving emotional states of a person. Like humans, social robots should be able to express different emotional states using their facial expressions or body movements in order to express their feelings on any event. Currently, there exists few robots that are able to generate different emotions, however these systems are quite primitive and perform poorly in real-time.

The work in this paper discusses the importance of perception system that enables a social robot to analyse complex human behaviours and interactive mode in order to interact naturally for intelligent human-robot interaction. The rest of the paper is organized as follows: previous research has been discussed in Section 2, in Section 3 we describe the low-level features and high-level perception system of social robot, ROBIN in detail. In Section 4 we discuss experiments conducted and conclude the paper in Section 5.

2 Previous research

In the last decade, numerous researchers developed task specific robots. However, few social service robots have been developed with the aim to develop robust human-robot interaction [19-23]. One such robot, ROMAN, has been developed by RRLAB as a test platform for human-robot interaction [3]. It consists of an upper body, two arms, neck, and an expressive head. ROMAN can generate facial expressions, gestures, and expressive body postures. It can generate nonverbal expressions using 47 DOF. Furthermore, it is also equipped with an expressive speech synthesizer. Figure 1 shows the ROMAN robot.

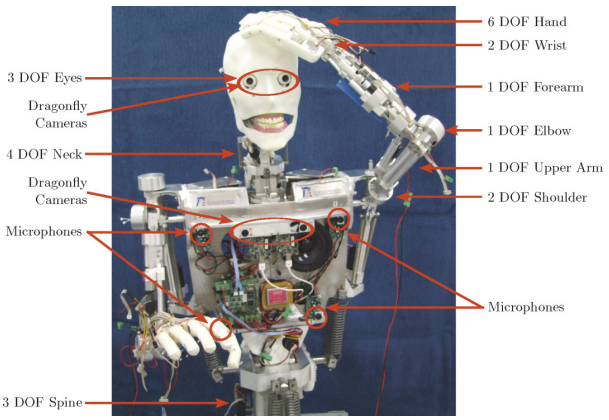

Fig. 1. Interactive Humanoid Robot, ROMAN.

* Corresponding author: berns@cs.uni-kl.de

The control architecture of the robot, emotion-based architecture, is developed by Hirth [4]. The goal of the implementation is to realize the functions of emotions: regulative (emotion), selective (percepts), expressive (habits), motivational (motives), and rating (emotion), as well as the secondary functions.

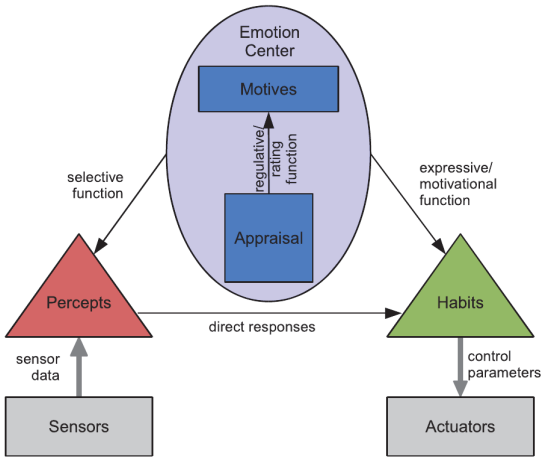

Fig. 2. Emotion-based control architecture, consisting of four main groups, motives, emotional state, habits of interaction, and percepts of interaction.

Compared to previous publications [3] where the rating and the regulative function were not included, all five functions of emotion mentioned in [5] are realized, since the rating and the regulative function are very important with a view to more cognitive or learning applications. The perception system perceives and interprets information of the environment. Depending on this information, direct responses, performed by the habits, are activated and the motives calculate their satisfaction. This satisfaction changes the current emotional state of the robot. Besides this, the motives activate several habits to change the robot's behaviour in order to reach a satisfied state. They also determine which information of the percepts is needed in the current situation. That way, the selective function is realized. This architecture has been examined using Tangram game [6] playing scenario as shown in the Figure 2.

Fig. 3. ROMAN playing Tangram game with human.

Beside the control architecture, ROMAN has been equipped with expressive function which enables it to produce different emotional states. Figure 4 depicts the facial expression of six basic emotions. A silicon skin has been glued to ROMAN's head. By moving small metal plates the different expressions have been generated [7].

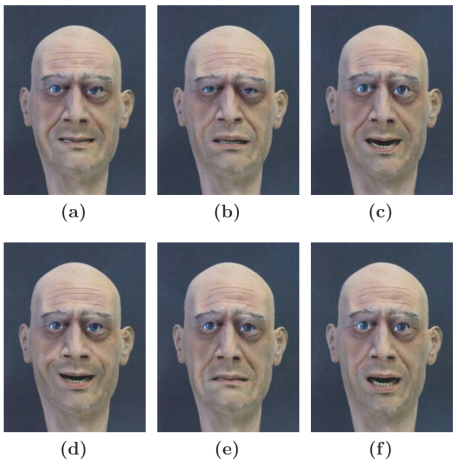

Fig. 4. Facial expressions of ROMAN realized with a silicone skin (anger, disgust, fear, happiness, sadness and surprise).

However, generating expressions accurately using silicon skin is quite challenging and cannot be generalized to different facial expressions. Moreover, ROMAN lacks in robustness when performing different gestures in real-time. Apart from emotion-based control architecture, important interactive modules like human feedback, context-aware perception and personality traits inference are missing in ROMAN.

In order to address these shortcomings, Technical University of Kaiserslautern has developed a social humanoid robot, ROBot-human-INteraction (ROBIN). It is equipped with a backlit projected face, arms, hands and torso. It can speaks via its built-in speech synthesis module in English and German language. The face makes use of projective technology to express almost any facial expression using action units. The head is able to move sideways for about ±45 degrees. RGB-D sensor is installed on the chest of robot. Additionally, a high definition camera is also installed on the head. The whole arm has 14 degrees of freedom, where hands are able to perform nearly all gestures. ROBIN has its own processor that can handle all the movements of joints. Figure 5 shows the ROBIN robot. In the following section, we describe the perception system of ROBIN.

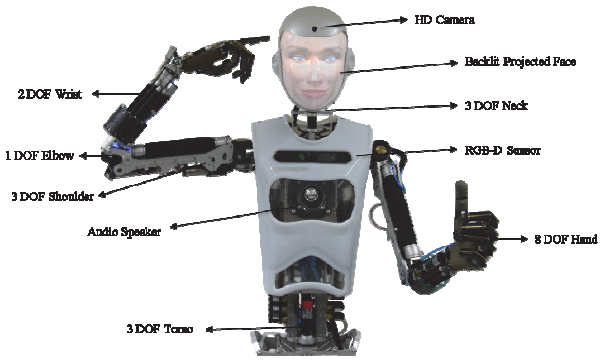

Fig. 5. Interactive humanoid robot, ROBIN, developed by TU Kaiserslautern, Germany.

3 Perception system of ROBIN

The human perception system is a part of the central nervous system, which is responsible for detecting and interpreting the surrounding environment. It consists of two eyes and the perception part of the brain. Eyes play the role of the camera to perceive the environment whereas the brain does all the complex image processing. Human, in general, depends on the visual system more than all other sensory systems. This is not only due to the quantity of information it provides, but also the robustness and efficiency in processing the information being a key factor.

Perception of the environment includes analysing and interpreting human behaviours. Most of the human behaviours are realized using verbal and nonverbal cues. Psychologically, nonverbal cues convey a lot of information and sometimes more than what verbal cues can express. Analysing and interpreting the nonverbal cues is a challenging task even for humans. ROBIN perception system can be divided into 2 subcategories: low-level perception and high-level perception system. Former deals with low-level features, e.g., facial expressions, posture, head pose, etc. while latter deals with complex human behaviours, e.g., feedback perception, context-aware perception and personality traits perception. We discuss each of these in the following subsections.

3.1 Low-level perception system

Before recognizing complex human behaviours and gearing a robotic system with emotional intelligence, recognition of low-level features accurately in real-time is highly critical. In this level, information from sensors are processed and analysed on the frame basis. The sensor used is an RGB-D sensor that provides depth information as well as RGB images. The function of low-level perception is to extract *percepts*, which are low-level information such as face location, head pose estimation, posture recognition, hand recognition, facial expressions recognition etc. In the section, we discuss major low-level features implemented in ROBIN.

3.1.1 Face detection

Face detection is an essential skill that a social robot should possess. It is the core process for many other skills like face tracking, facial expression recognition, face recognition, gender/age recognition, and head pose estimation. The present work utilizes a Haar cascade classifier to find candidate faces in 2D RGB images. The faces detected by the Haar cascade classifier is not always representing a real face. This is due to the used 2D features that can be found in any variations of intensities (colors) analogous to a human face. To reduce the false positives that Haar cascade classifier produces, the corresponding depth information acquired by an RGB-D sensor is used [8] as shown in Figure 6.

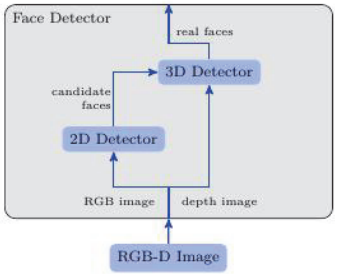

Fig. 6. Working of face detection module.

3.1.2 Facial expression recognition

Facial expressions of humans play a significant role in the interpersonal interaction. They are the primary source of expressing emotions and feelings. The work at hand uses action units to recognize the facial expression of the interaction partner [9]. It uses a deep convolutional neural network (CNN), or deep learning to estimate the activation of each of the related action units and combinations. The architecture comprises six layers (excluding input layer) as shown in Figure 7. For each action unit, CNN receives a human face as a 32×32 gray image and outputs the confidences of two classes active and not active of the corresponding action unit. It uses two convolution layers each with its own subsampling layer (max-pooling). 6 basic and neutral facial expressions have been recognized using action units results with an overall accuracy of 90.2%.

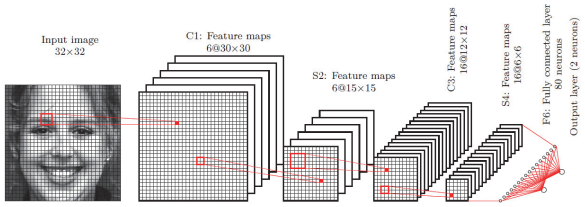

Fig. 7. Deep architecture of facial expression recognition module.

3.1.3 Posture recognition

Human pose also shows a whole lot of information about human emotional state. Human pose can convey information about human internal state whether he/she is nervous, or confident, or thinking, or not interested etc. In order to realize this on the robot, we use an approach proposed by Zafar et al. [10] that uses depth data along with NiTE library to detect human joint positions and then convert them into meaningful angles for feature vector generation task. The resultant feature vector is quite unique for each posture and is invariant to height, body shape, illumination, proximity and appearance of human. The working schematics of the proposed approach is shown in figure 8. The system is able to recognize overall 21 gestures real-time when classified by using multi-class SVM with an accuracy of 98%.

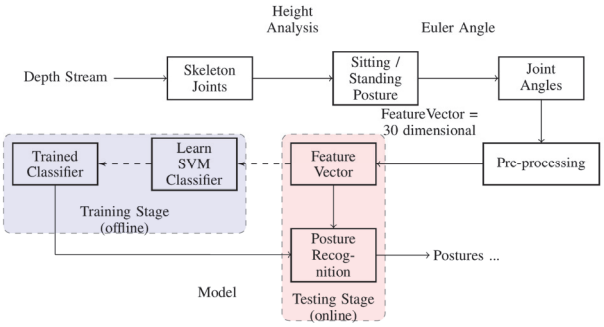

Fig. 8. Working of Posture Recognition Module.

3.1.4 Static hand gesture recognition

Like other nonverbal cues, hand gestures are also used in interpersonal face-to-face communications. They also reveal some of the human's emotions and feelings. In this work, the hand gesture recognition approach proposed by Zafar and Berns [11] is used as shown in Figure 9. This approach has been implemented to recognize 18 different hand gestures and shown high efficiency and robustness. It uses bag-of-features (BoF) of scale invariant feature transform (SIFT) key points and support vector machines (SVMs) to recognize hand gestures. Interest point approaches extracts useful information from an image and bag of feature approach is used to represent them in terms of a feature vector. A total of 15 different hand gestures are recognized with more than 94% accuracy in real-time.

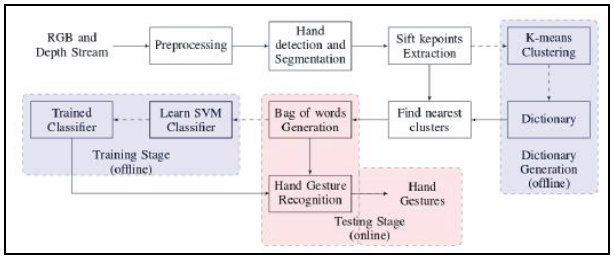

Fig. 9. Schematic flow of hand gesture recognition module.

3.1.5 Dynamic hand gesture recognition

Similar to static, dynamic hand gesture also shows complex human behaviour when interpret accurately along with the context. The system uses the approach presented by Zafar et al. [12]. The recognizer uses only depth image information, and the hand position provided by a hand tracker library, in order to construct its feature vectors. The recognizer builds two types of feature vectors to increase accuracy; the frame feature vectors that describe a static hand, and the sequence feature vectors that describe a contiguous segment of frames. The recognizer also uses two statistical classifiers. The frame feature vectors are utilized by the frame classifier. The results of the classifier, then become part of the sequence feature vector, which in turn are utilized by the sequence classifier as shown in Figure 10. A total of 16 different interactive dynamic hand gestures has been recognized with an approximate accuracy of 87%.

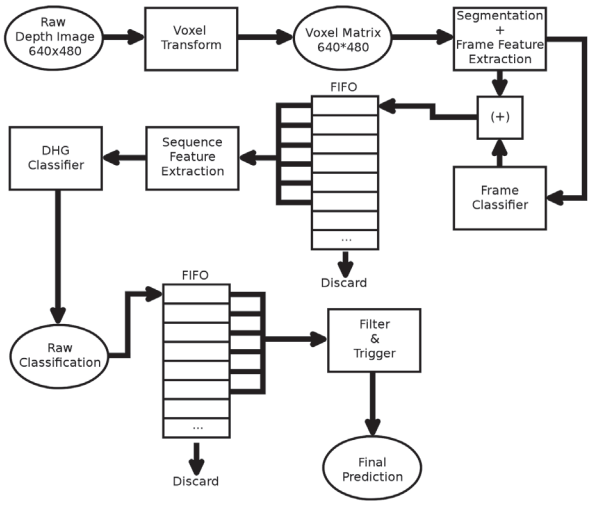

Fig. 10. Working flow of dynamic hand gesture recognition module.

3.1.6 Head pose estimation

Head movements convey a lot of information in inter-human communication. Humans have the ability to interpret these movements quickly and effortlessly, while it is regarded as a difficult challenge in computer systems and robotics. Detecting the human head movement requires estimating the head pose (position and orientation) over time. This work uses depth information for head pose estimation. This approach is called Direction-Magnitude Pattern (DMP). These features are suitable for depth information because they result in a surface map of the face. Three linear Support Vector Machines for Regression (ε -SVRs) are trained to detect the pose angles roll, pitch and yaw as presented in [8] and extended in [13]. The input of these SVRs is the feature descriptor of the face under test which is DMP features. Dynamics of the approach is shown in Figure 11. The reports a mean error of ±3 degrees when tested in real-time. Head gestures include nodding, shaking, tilting, looking up, down, right, left, forward are realized in this work.

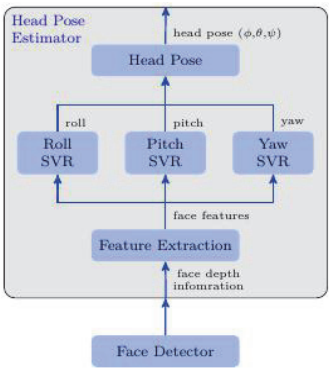

Fig. 11. Working flow of dynamic hand gesture recognition module.

3.1.7 Human identity recognition

Recognizing human identity is an important aspect in intelligent human-robot interaction. In human-human interaction, humans recognizes other human's identity and according to his past experiences all together changes his whole behaviour. According to some psychologists, recognition of identity is one of the basic and equally important step on which we build our interaction. In order to recognize human identity, we use Local Binary Pattern Histogram approach to extract local feature from human face and classify them [14]. More than 20 different subjects have been stored in the system and robot with an accuracy of 97.3% recognizes human identity correctly.

3.2 High-level perception system

High level perception deals with more complex human behaviour which uses combination of low level features stated above. Human behaviour can be recognized by recognizing some basic low level features and realize it over longer period of time to extract some useful cues.

In order to realize robust and natural human-robot interaction, we have developed higher level human perception system that extracts human feedback features, context aware features and lastly the human personality features. These features are all equally important to realize a natural and efficient multimodal human-robot interaction. The following subsections discuss the details of high-level features.

3.2.1 Human feedback perception

Humans use feedback in their conversation to exchange information about four communicative functions. These functions are continuation, perception, understanding and attitudinal reactions. Any expression, verbal or non-verbal, can be regarded as a feedback if it serves one or more of these functions.

Human feedback perception by a robot requires many low-level perception processes. These processes include a frame-based perception of human articulators such as the face, head, hands, body, and so on. In this work, the outputs of these processes are called percepts. Head pose, facial expression, hand static gestures, and posture are examples of the percepts. These low-level percepts then can be accumulated over time to form higher-level information. High-level percepts include dynamic head gestures, dynamic hand gestures, body movements, and so on.

In order to communicate successfully, according to Allwood [15], two participants should establish a contact with each other. It is necessary that participants show their ability and willingness to continue in the interaction. Continuation signal tells the interlocutor about the desire of continuing in the interaction regardless of the contents of the message. Two types of continuation signal can be distinguished: You go on and I go on [16].

Once the contact is established, one participant possibly produces a message. The receiver should be able and willing to perceive the message and he may or may not understand the message. Understanding the perceived message means that the contents of the message is interpreted in the same way that the speaker intends to. Once the receiver understands the message, he/she may give attitudinal and behavioural reaction according to the acceptance of the message. Attitudinal reaction can be positive, negative or neutral. The implementation of feedback perception is described in [16] in detail.

3.2.2 Context-aware perception

One of the key aspect in human-human interaction is the knowledge of context. Humans use this feature intelligently in order to reduce the number of possible interactive topics whenever they interact with each other. As soon as person A meets person B, person A automatically recalls every information regarding person B starting from some basic things how person B looks like, his interests, his views and opinions about different things to some more subjective things like his personality and his personal relation with him etc. In short all that information acquired over past interactions becomes active. This information plays a huge role during interaction. Human after an interaction with other person know which things are more interested to him and which are not and hence ask questions or talks on those topics.

Similarly, in human-robot interaction, this feature can make the whole interaction intelligent. In order to implement this context aware interaction feature, robot has to store all the information of interaction with each person in a separate file. This file is accessed only when that person is recognized using face recognizer module. Robot can access the information acquired in the past interactions and use it to bring topics that are more relevant to that person for interaction.

How this file is being generated? The process is quite similar to how human acquires knowledge. According to psychological studies, humans examine each other and ask questions starting from some basic questions like "What is your name?", "What do you do?", "Where do you come from?" etc. and with time they ask about their interests, their hobby's, their goals and so on. Similarly, the robot recognizes the person in the start and checks whether there is any information regarding that person is present in its database. If not, then robot activates the introductory question script where, it started asking about person whereabouts and his/her occupation and then moved more towards asking his/her goals, interests, favourite things, etc. Currently, the system has more than 60 questions which robot can ask and saves the responses in a separate file. Based on the information stored in this file, robot finds relevant things, topics or ask questions.

This type of interaction is intelligent as it is not scripted or fixed. Robot can talk on topics which are relevant to person interacting with him. We have conducted experiments where robot first asks questions in the start to know about the person. The person responses are saved in a file. This file is stored with a name of the

person in the system. Next time when that person again interacts with robot, robot recognizes him and brings all the information regarding that person for context aware human-robot interaction. Experiments have shown the effectiveness of the system and reports promising results.

3.2.3 Personality traits perception

Knowledge of human personality is a bonus in human-robot interaction. Robot can behave differently according to human personality type. However, recognizing human personality is a very challenging task. Even for humans, sometimes it is difficult to know the personality of our counterparts. Numerous psychologists have presented their theories which reports some features that represent human personality type. Most of these features are not the defining features and hence do not points directly towards a personality type. However there exists some studies which maps different non-verbal cues to different personality types. According to psychologists, it is almost impossible to derive personality type from nonverbal cues which makes the problem difficult. Since tone of human speech conveys a lot about human personality, recognizing this feature is highly critical in precise recognition of human personality. However, recognizing tone of human speech is not realized in this project.

Other methods include recognizing some cues about human personality type from human body gestures and appearance, and asking questions regarding different situations and analyse how the person responds. Though these methods may not be precise alone but combining both of these methods provide us a good hint about the personality type of human. According to psychological survey conducted by Jensen [17], different personality types can be recognized using non-verbal patterns namely: (a) Extraversion, Introversion (b) Open to new experiences, Traditionalist, (c) Conscientiousness, Careless, (d) Agreeableness, Self-centred and (e) Neuroticism, Emotionally stable.

Some of the features the exploits personality type (E) including proximity feature, gesticulation features, facial expression features etc. Individuals high on extraversion prefer to stand close to conversation partner and the also like to sit close to co-communicator. Individuals low on extraversion, prefers to stand and sit at a distance when in a conversation. Similarly extraversion is more related to frequent use of rapid body movements while neuroticism in involved in more self-touching behaviour. Extraversion is also associated with more intense facial expressions, e.g., smiling a lot during conversation. Agreeableness personality type is positively correlated to laughter and sympathetic facial display. These and more features represent (roughly) some of these personality type.

The five-factor model dealing personality traits plays a critical role in the field of communication studies, psychology and philosophy. Based on the verbal and nonverbal cues, the personality trait of a human can be guessed and to some extent the specific trait can be detected. We require a human to judge in which dimension the personality of a human lies. The judgmental process is extremely fuzzy as there are so many facets ingrained in every human. Interestingly, there is no quantitative standard to judge the severity of each and every facet. Intuitive and perceptive skills do the trick for the person judging personality of another person. However, as mentioned earlier, researchers have come up with some theories and methods which can help to recognize some of these personality traits.

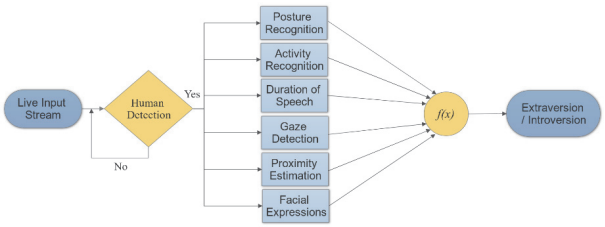

Fig. 12. Working flow of extroversion-introversion trait detection.

Figure 12 shows the work flow of our approach which uses human nonverbal cues for recognition of extroversion-introversion trait. Robot uses ASUS Xtion sensor along with OpenNI and NiTE Library to detect humans in the scenario and recognize major facets relevant to extroversion-introversion dimension. A score is estimated based on weighted function f(x) which determines extroversion-introversion personality trait.

Our approach uses nonverbal features, i.e., human posture, rapid body movements (activity), duration of speech, proximity, gaze and facial expressions as presented by Zafar et al. [18]. Integration of these nonverbal cues in a meaningful way is highly critical for overall outcome. Simplest approach is to add all the binary outcome of each facets to generate a score. However, the problem lies in the situation when one or more nonverbal feature(s) would either be detected wrongly or not detected at all. This would cause misleading outcome. In order to address this, a weighted outcome strategy has been employed. This ensures that the outcome is not sensitive to any particular nonverbal facet, rather it considers the appropriate impact of each facet in general. According to empirical studies and literature survey, it has been found out that the activity facet plays the most vital role in the recognition of extroversion-introversion. Similarly, duration of speech and posture are also considered significant in the recognition process. Equation (1) shows the weighted function f(x), which can have a value ranging from 0 - 1.

$$f(x) = 0.3\,(A) + 0.2\,(S) + 0.15\,(P) + 0.15\,(G) + \\ + 0.1\,(D) + 0.1\,(E), \quad (1)$$

where A is the activity detection, S is the duration of speech, P is the posture, G is the gaze detectors, D is the distance detector and E is the facial expressions detector. Since facial expressions are quite dynamic and vary from person to person especially when a person is speaking, they are prone to false recognition. Hence, this facet has low weight. Similarly, proximity of a person happens only for a short time during the interaction. Hence it also has low weight. The score

generated by the function f(x) is used to recognize extroversion/introversion. Higher value indicates strongly extrovert person and lower value shows extremely introvert person.

4 Experimentation

For experimentation of feedback system, 5 different subjects play 20 questions game with the robot. As few subjects don't know how to play the game, robot explains the game and game-play at the beginning. To answer multiple choice question, basic hand gestures are used. An information card is placed beside the robot for the user to consult which hand gesture represents which answer option. For evaluation of the experiments, subject plays the game with the robot. System writes the activation of all the behaviours in a separate file along with head, hand, face gestures and the question for every frame. A Human expert analyses the subject behaviours in the recorded videos. He observes the feedback behaviours whenever a question has been asked by the robot and fills in the questionnaire. He also takes notice of subtle changes in the behaviour, for example lack of interest, understanding or not understanding behaviours etc. At the end of the experiments, the recorded file by the system is compared with expert questionnaire.

Figure 13 shows the plot of extroversion-introversion score of a job interview experiment. Initially, candidate acts as an introvert which can also be visualize from the plot. Candidate has crossed his arms and looking downwards with minimal body movements. This behaviour results in low score for extroversion. After period of time, candidate behaves like an extrovert as seen in the figure. Different features like open arms, body movements, facial expressions, mutual eye gaze, etc. are active and lead to high score for extroversion. Images at each time-stamp are used to validate the weighted score.

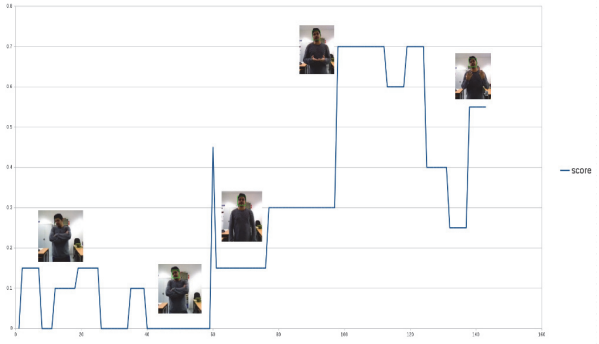

Fig. 13. Weighted score plot of extroversion-introversion trait.

5 Conclusion

In this work, we have discussed the perception system of humanoid robot ROBIN. In order to achieve emotion-based human-robot interaction, there is a need of robust perception system that is able to perceive and understand some basic human nonverbal behaviour as well as some high-level perception tasks which include understanding human feedback communication channel, context-aware interaction and personality traits inference. The system is tested in real-time with active low-level and high-level perception features.

References

1. Affective Science. http://www.ncbi.nlm.nih.gov
2. A.G. Francis, M.J. Mehta, A. Ram, *Handbook of Research on Synthetic Emotion and Sociable Robotics* (Hershey, NewYork, 2009)
3. J. Hirth, K. Berns, Humanoid robots, 97–116 (2009)
4. J. Hirth, N. Schmitz, K. Berns, International Journal of Social Robotics, **3** (3), 273–290 (2011)
5. H.R. Lueckert, I. Lueckert, *Einfuehrung in die Kognitive Verhaltenstherapie: Allgemeine Grundlagen* (Ernst Reinhardt Verlag, Munich 1994)
6. J. Elffers, M. Schuyt, *Tangram: The Ancient Chinese Puzzle* (Evergreen, 1999)
7. K. Berns, J. Hirth, Proc. Of IROS, 3119–3124 (2006)
8. S. Saleh, A. Kickton, J. Hirth, K. Berns, Proc. of ACHI (2013)
9. S. Al-Darraji, K. Berns, A. Rodic, Proc. of RAAD, 413–420 (2016)
10. Z. Zafar, R. Venugopal, K. Berns, Real-Time Recognition of Human Postures for Human-Robot Interaction, Proc. of ACHI, Rome, Italy, (2018)
11. Z. Zafar, K. Berns, Proc. of ACHI, 333–338 (2016)
12. Z. Zafar et al., International Conference on Robotics in Alpe-Adria Danube Region, 649–656 (2017)
13. S. Saleh, K. Berns, Proc. of PETRA, 5 (2015)
14. T. Ahonen, A. Hadid, M. Pietikainen, Proc. of ECCV, 469–481 (2004)
15. J. Allwood, International Conference on Cooperative Multimodal Communication, **2155**, 113–124 (2001)
16. S. Al-Darraji, Z. Zafar, K. Berns, Proceedings of the 30th International BCS Human Computer Interaction Conference: Fusion!, 27 (2016)
17. M. Jensen, Personality, J. of Social Science, **4,** 57–70 (2016)
18. Z. Zafar, S.H. Paplu, K. Berns, Proc. of RAAD, (2018)
19. T. Hashimoto, S. Hiramatsu, T. Tsuji, H. Kobayashi, Proc. of RO-MAN, 326–331 (2007)
20. M. Noma, N. Saiwaki, S. Itakura, H. Ishiguro, Proc. of Humanoids, 163–168, (2006)
21. M. Bennewitz, F. Faber, D. Joho, S. Behnke, Proc. of RO-MAN, 1072–1077 (2007)
22. K. Papoutsakis et al., Proc. of PETRA, 56 (2013)
23. T. Asfour, K. Welke, P. Azad, A. Ude, R. Dillman, Proc. of Humanoids, 447–453 (2008)