Proceeding of the 2nd
RSI/ISM International Conference on Robotics and Mechatronics
October 15-17, 2014, Tehran, Iran

An RGB-D Based Social Behavior Interpretation System for a Humanoid Social Robot

Aolfazl Zaraki*, Manuel Giuliani†, Maryam Banitalebi Dehkordi‡
Daniele Mazzei*, Annamaria D'ursi*, and Danilo De Rossi*
*Research Center "E. Piaggio"
University of Pisa
Pisa, Italy
†Dept. of Cyberphysical Systems
Fortiss GmbH
Munich, Germany
‡Perceptual Robotics Laboratory
Scuola Superiore Sant'Anna
Pisa, Italy

***Abstract*—Humanoid social robots that interact with people need to be capable of interpreting the social behavior of their interaction partners in order to respond in a socially appropriate way. In this paper, we present a social behavior interpretation system that enables a humanoid robot to recognize human social behavior by analyzing communicative signals. The system receives the constructed RGB-D scene from a Kinect sensor, extracts information about body gesture and head pose from the scene using Microsoft Kinect SDK, and recognizes eight human social behaviors using a Hidden Markov Model (HMM). We trained the eight-state HMM with a corpus of 35 recorded human-human interaction scenes. The evaluation of the system shows a weighted average recognition rate of 81% for all states.**

***Index Terms*—Human-robot interaction, hidden Markov model, social behavior recognition, humanlike robot**

Fig. 1. The humanoid social robot FACE interacting with a group of people.

I. Introduction

With the rapid advancement of mechanical design, computational methods and related works in the field of robotics, humanoid social robots that interact with people are gradually becoming more integrated into human daily life [1]. This class of robots is being designed in order to assist humans in different human-centered scenarios, e.g., for domestic and public applications, collaborative tasks, or education purposes [2]. Thus, in addition to task-performing capabilities, humanoid social robots need to be able to interpret the social behaviors of their interaction partners in real-world situations in order to respond in an appropriate way. For example, Fig. 1 shows the humanoid robot FACE [3]–[7] that was built to appropriately respond to the social behavior of multiple humans. In order to increase the social capabilities of FACE, in this paper we are following this question: how can we equip a humanoid robot with the ability to interpret the social behavior of humans?

Several studies [8]–[11] have shown that nonverbal cues are an important source of information through which humans express social behaviors and intentions in everyday interactions. Among the wide spectrum of nonverbal communicative cues, *head pose* and *body gesture* play a fundamental role in delivering meaningful information and thus, recognizing human social behaviors should also rely on these communication modalities. Therefore, detecting nonverbal communicative cues is essential to simulate human perception and to implement an interpretation system for a humanoid robot.

In this paper, we present a social behavior interpretation system (SBIS) that recognizes the behaviors of humans by analyzing their social signals, based on head pose and body gesture. In our target scenario [12], the robot should be able to interact with multiple humans. For that, the robot needs to be able to greet entering persons, to offer a drink to those who attract the robot's attention, and to say goodbye to leaving persons. In addition, the robot should not disturb two

978-1-4799-6743-8/14/$31.00 ©2014 IEEE

humans who are talking to each other. Therefore, the SBIS needs to recognize scenario-related social behaviors including bidding for the robot's attention, drinking, picking an object from the desk, socializing with other people, entering the room, and leaving the room. For this reason, the system consists of a perception and an interpretation layer. The perception layer receives input from a Microsoft Kinect, and provides perceptual information extracting several human nonverbal communicative cues from the data stream. It then creates a feature vector that contains all extracted features, and passes it to the interpretation layer. This layer uses a trained Hidden Markov Model (HMM) [13] to recognize social behaviors from the feature vectors.

The remainder of the paper is organized as follows: Section II reviews related works. Section III gives details about nonverbal communicative cues in human-human interactions and technical information about HMMs. Section IV presents the general structure of our SBIS. Section V contains the description, results, and discussion of an evaluation of the SBIS. Finally, Section VI concludes this publication.

II. Related Works

In the last decades, social behavior recognition and intention estimation, have attracted the attention of many researchers. The common goal of the previous works in these fields was to design a system that observes a complex scene by a vision sensor, detects humans, and recognizes human activity through analyzing their body geometry. To achieve this goal, the following procedure is widely followed: first, sequences of human actions are recorded on video and annotated. Then, a feature vector is extracted for each single frame of every recorded video. Each feature vector contains the most important information of an image, and thus, an action can be represented by a feature vector sequence. Finally, a classifier is trained using these sequential feature vectors in order to classify different actions. In an application using the trained classifier, the system receives as input the visual scene and returns as output the name of activities of people who are presented in the scene [14].

In spite of similarity in the behavior recognition process, previous works differ in the feature vector creation process methods and the type of employed classifier. Some researchers proposed frameworks that recognize human activity analyzing 2D scene with a pattern matching techniques (e.g., [15], [16]). Scheutz et al. [17] implemented a cognitive behavior recognition system for receptionist and waiter robots that recognizes and generates affective behavior. Brand et al. [18] utilized coupled HMMs to recognize two-handed activities. The feature vectors of the above-mentioned works were created from 2D images selected from video frames. Thanks to the advancement of 3D vision devices, RGB-D sensors can benefit recognition systems by constructing more detailed information of scenes. For this reason, some other works implemented behavior recognition systems with depth-based sensors. For example, Sung et al. [19] proposed an activity recognition system for unstructured environments using a hierarchical maximum entropy Markov model. The feature vector of [19] was created from human 3D body pose, motion and point cloud information. Sung et al. evaluated their system for 12 activities in different environments.

Similar to our work, human activity recognition and social behavior recognitions have also be implemented specifically for HRI applications. Yang et al. [20] presented a full-body automatic activity recognition system using HMMs for a HRI application that robustly recognizes several meaningful body movements such as jumping, walking, and hand waving. In this work, the feature vector was created from the angular relationship between human body parts. Yang et al. integrated the proposed system in a mobile robot. Through an empirical study, Gaschler et al. [21] found that people mostly express their intentions non-verbally through head pose and body gestures. Thus, Gaschler et al. proposed an intention recognition system using HMM for a bartender robot that recognizes the social signals of customers who want to initiate an interaction with the robot. The feature vector of this work was created of head pose and 3D positions of body joints.

In our work, we propose a social behavior interpretation system using an HMM that recognizes eight social behaviors (activities) of a human in social contexts. We create the feature vector from head and body 3D information, which are reconstructed by the Kinect as the input sensor.

III. Background

This section discusses the importance of nonverbal communicative cues in human social behavior, particularly those cues that we used in the feature vector extraction process (i.e., body gesture and head pose). Furthermore, we briefly describe the theoretical aspects of HMMs, the classifier that we are using for recognizing social behaviors.

A. Nonverbal communicative cues in a human-human social interaction

Nonverbal communicative cues consist of a wide range of wordless cues that people use mostly to regulate a social interaction with other people. Several studies in human behavior science [8]–[11] have shown that nonverbal cues (e.g., body gesture, head pose, body orientation of others, physical distance between people, facial expressions, etc.) are the most important sources of information through which we express social behaviors and intentions in daily human-human interactions. The variety of nonverbal cues and lack of sensing technologies results in the challenge on how to find the optimal set of cues that people mostly use to perform social behavior in a social interaction. Knowing this issue is fundamental when designing a social behavior recognition system.

As literature indicated, among the wide spectrum of nonverbal communicative cues, *head pose* and *body gesture* play a fundamental role in delivering meaningful information to other people and thus, recognizing human social behaviors should rely on these communication modalities. Gaschler et al. [21] demonstrated that people use *body gesture* and *head*

pose as social signals to initiate a social interaction. Due to the importance of these cues, other work e.g., [8] and [10] use the term *bodily communication* for nonverbal communication. Following this notation, we design our social behavior interpretation system to estimate human behaviors analyzing body gesture and head pose as reliable sources of information.

B. Hidden Markov Model

Hidden Markov Models (HMM) are a machine learning method that is used to classify a sequence of states based on a given sequence of observations over time. As Ramage reported in [22], HMMs can be used when we cannot have any observation of the real series of states, and instead we observe only some outputs that are generated by each state. Assume $x = [x_1, x_2, .., x_V]$ be a sequence of observed output (i.e., $x_t \in V,\ t = 1...T$), the sequence of states $z = [z_1, z_2, ..z_N]$ exist that are taken from a state S, in which $z_t \in S,\ t = 1..N$. However, it should be noted that the value of states are unobservable. The definition of a HMM is completed defining two matrices: transition and emission.

Following this notation, the transition matrix A_{ij} is defined as transition from a hidden state i to j within a time step. Given the hidden state s_j, the emission matrix B_{jk} is defined to encode the probability of s_j generating the observed output. In addition, to use HMM, we should initialize the states and observations.

The hidden Markov model then can be defined as

$$\lambda = (A, B, \pi) \tag{1}$$

where, vector π is the initial probability of all hidden states at time zero and A_{ij} and B_{jk} are transition and emission matrices. In the case of continuous observations, instead of discrete probabilities, we should employ a continuous probability density function. Assume the emission distribution $b_j(x_t)$ is estimated by M Gaussian distributions with mean vectors μ_{jm}, $b_j(x_t)$ can be written as

$$b_j(x_t) = \Sigma_{m=1}^{M} C_{jm} \mu(\mu_{jm}, \Sigma_{jm}, x_t) \tag{2}$$

where, Σ_{jm}, C_{jm} and μ_{jm} are covariance matrix, weighted coefficient, and mean vectors, respectively.

In this work, our observations are human body joints position, orientations, and head pose (yaw, pitch, and roll angles of head) that are captured by the vision sensor (see section IV-A), while the states are different human social behaviors (see section IV-B). Furthermore, due to our scenario we deal with a continuous multidimensional HMM, which can be defined as

$$\lambda = (A, C, \mu, \Sigma, \pi) \tag{3}$$

After implementing the HMM for human behavior recognition, we should pre-train the model performing an offline process. At this point, we train the HMM model with the observations recorded in several videos of human-human interactions.

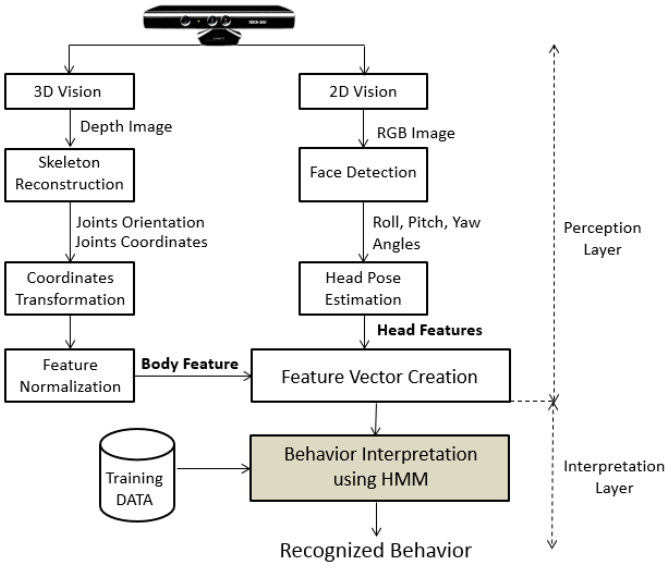

Fig. 2. General structure of the proposed Social Behavior Interpretation System (SBIS). The system receives the RGB-D scenes from a Kinect, analyzing the data in the perception layer, and recognizes human social behaviors in the interpretation layer.

IV. Social Behavior Interpretation System (SBIS)

This section gives details for the implementation of our human social behavior interpretation system (SBIS). To recognize human behaviors, the system should be able to actively observe a human body gesture and head pose, and compare the corresponding motions to the predefined motions over time. Therefore, we propose an SBIS, which monitors a visual scene using Kinect, detects humans in the scene, analyzes corresponding social signals, and uses an HMM model to estimate the behavioral intention of the detected humans. Fig. 2 shows an overview of the SBIS. It consists of two distinct layers, the perception layer and the interpretation layer. Our current implementation aims to recognize behaviors, analyzing those human-relevant features that have a communicative role in a social HRI scenario (see III-A). The following section describes the two SBIS layers.

A. Perception Layer

The perception layers consists of two sub-components: data acquisition and feature vector creation. They are deputed to prune data and create in real-time the feature vector that consists of human body and head information, from RGB-D scenes. In the perception layer, a skeleton composed of 20 joints represents a human body gesture. Each joint is represented by its 3D position as well as a rotation matrix that shows the joint's orientation angles (roll, pitch, and yaw) in real-world coordinates. For the implementation of the SBIS, we use the Microsoft Kinect SDK to detect and track seven joints of the upper body and the head pose of tracked humans (see Fig. 3). Additionally, the system

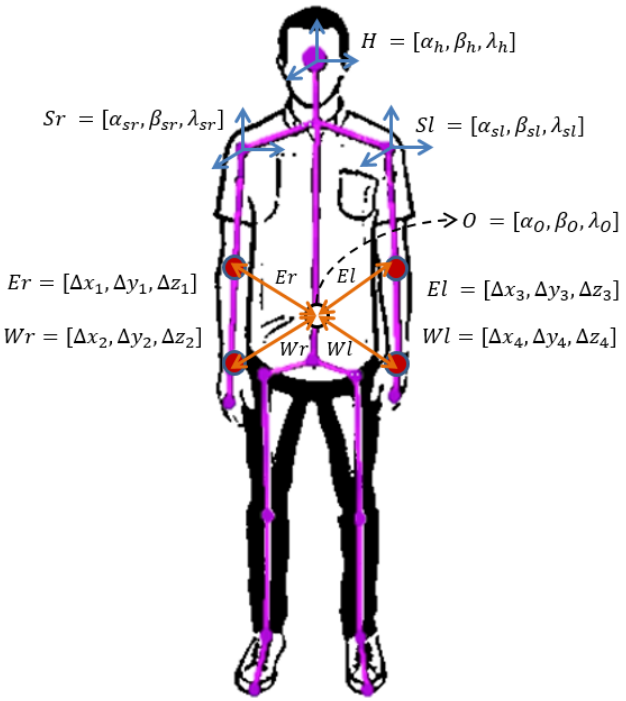

Fig. 3. Body gesture and head pose consist of rotational movements and displacements of 7 among 20 joints recognized by the perception system. The hip center joint in the center of the body is the origin of the local coordinates.

analyzes the extracted features and creates a feature vector that represents the current behavior of the humans. The set of features (displacements and rotations) of each RGB-D frame represents a human static body gesture. Furthermore, the human behaviors are represented as a time sequence of static gestures obtained from corresponding RGB-D frames, in real time. As can be seen in Fig. 3, the total number of features are 24, as follows:

- **Arms** (elbow and wrist, displacement only), $2 \times 2 \times 3 = 12$ features, $[El, Er, Wl, Wr]$.
- **Shoulders** (orientation only), $2 \times 1 \times 3 = 6$ features, $Sl = [\alpha_{sl}, \beta_{sl}, \lambda_{sl}]$ and $Sr = [\alpha_{sr}, \beta_{sr}, \lambda_{sr}]$.
- **Hip Center** (orientation only) $1 \times 3 = 3$ features, $O = [\alpha_o, \beta_o, \lambda_o]$.
- **Head** (orientation only), $1 \times 3 = 3$ features, $H = [\alpha_h, \beta_h, \lambda_h]$.

The feature extraction and feature vector creation processes are described as follows:

1) Body gesture features: As shown in Fig. 3, the extracted features of human body are rotational movements and displacements of body joints. It should be noted that in order to increase the performance of the system, the optimal set of features are defined. The set of features are obtained in a way to satisfy two important issues: (i) the features allow the system to recognize activities independent of the human or sensor positions; (ii) the features perfectly support all human rotational and transitional motions.

In order to allow the system to be independent of Kinect and user position, we define local coordinates on the hip center joint of the human body (as shown in Fig. 3) and measure the body parameters in the local coordinates. The joints' displacements $(\Delta x, \Delta y, \Delta z)$ of elbows and wrists of left and right hands are obtained in the local coordinates. In addition, the orientations (roll, pitch, and yaw angles) of left and right shoulders and hip center joint are obtained with respect to the torso.

2) Head pose features: As shown in Fig. 3, the human head pose is represented using Euler angles as $[\alpha_h, \beta_h, \lambda_h]$, which are roll, pitch and yaw angles of the head. In order to detect and track head pose in real-time, we integrated an additional component called "*Kinect toolkit for face tracking*" [23] into the perception layer. It detects and tracks a human face in the RGB-D frames and allows the perception layer to create in real-time the head pose feature vector.

3) Feature vector creation: The feature vector is a 1×24 vector, which consists of the obtained body and head features:

$$Feature\ Vector = [body\ features,\ head\ features]_{1\times 24} \quad (4)$$

The system creates the feature vector every $1/30$ second. A time sequence of vectors represents human behavior over time. The perception layer streams out the created feature vector to the interpretation layer that allows the SBIS to estimate the human behaviors.

B. Interpretation Layer

Having a feature vector that contains body gesture and head pose features, the interpretation layer is able to interpret a set of pre-defined human social behaviors. Inspired by human social behavior in a human-human interaction, we define the eight social behaviors Entering, Idle, Bidding for attention, Waiting, Taking object from the table, Drinking, Socializing, and Leaving. According to these states, we design, parametrize, and train our HMM that should be able to interpret behaviors from a time sequence of feature vectors. Fig. 4 shows the states of our HMM and their transitions. As can be seen, we selected eight states for the human behavior, which empower the robot with sufficient information to properly respond in a social interaction with humans. These states are defined in the way to recognize the most important activities and to minimize the complexity of the model by restricting the possible states.

The human activities in a social interaction are started by *Entering* and ended by *Leaving* states. Considering the human behavior during staying in a social interaction, we defined the default state as *Idle* in which the observed human neither performs any activities nor interacts with the robot. The *Bidding for attention* state describes the human when he/she tries to initiate a social interaction with the robot. The *Waiting* state describes the human when he/she waits to receive something (e.g., bottle of water) from the robot. In the *Taking object* state, the human picks up an object from the table. In the *Drinking* state, the human drinks water with her/his left or right hand. The *Socializing* state, describes the human interacting with another person and not the robot.

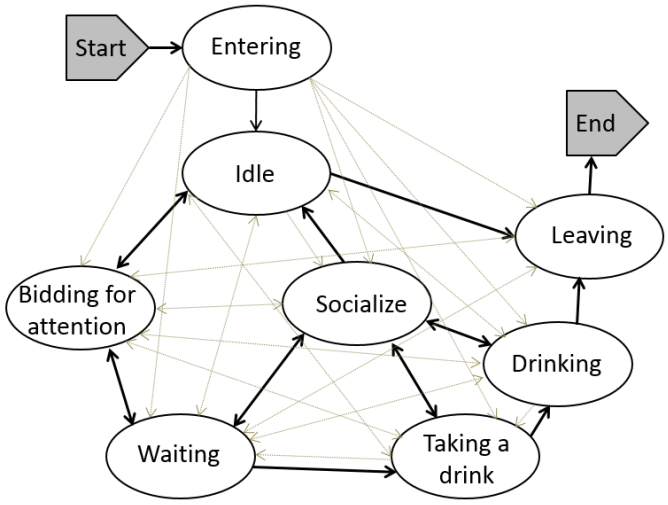

Fig. 4. The Hidden Markov Model consists of eight states. The model is fully connected, as shown by the dashed lines. Bold lines show the most often observed behavior transitions in human-human interactions.

V. Evaluation

In this section, we evaluate our SBIS. This evaluation includes details about the data collection and data analysis process, in order to train and test the HMM-based interpretation layer of the SBIS. Furthermore, we present the evaluation results and a discussion of the results.

A. Data Collection and Analysis

Through a Kinect device and using Kinect Studio software [24], we recorded the RGB-D of 35 interaction scenes, in which several humans of both genders interacted with each other or with our robot. The participants were instructed to reenact the social behaviors of our target scenario: entering and leaving the room, bidding for the robot's attention, staying idle in the room, waiting for the robot to offer a drink, drinking, and socializing with other humans. The recordings yielded 35 video files with corresponding feature vector files. We annotated the video files by hand using Anvil [25], in order to obtain the sequence of social behaviors of the recorded humans. The annotation process resulted in a total number of 524 annotated states with separated log files. Table I gives an overview of the number of labeled files per social behavior. The associated feature vector of each state was stored in sub-log file with the same name of the state. Each log file represents the body gestures and head poses of the recorded humans while performing the target social behaviors. This data set was then used to train the HMM, as described in the following section.

B. Training HMM and System Evaluation

In order to implement an HMM that contains the model of eight social activities of our scenario, we integrated the Accord .Net Framework [26] to the interpretation layer. Using 80% of the collected feature vectors, we trained a single continuous multidimensional HMM model that estimates the eight social behaviors of our target scenario. We used the remaining 20% of the collected feature vectors to evaluate the performance and accuracy of the system. Table I contains the exact number of samples used for training and testing.

TABLE I
Number of labeled sequences for each of the target social behaviors.

States	Enter	Idle	BidAt.	Wait	TakeDrink	Drink	Social	Leave
Total No.	35	64	45	37	69	150	89	35
Training (80%)	28	51	36	30	55	119	71	28
Test (20%)	7	13	9	7	14	31	18	7

TABLE II
Confusion matrix and recognition accuracy of SBIS for used test data.

	Labeled states							
Recognized states	E	I	B	W	T	D	S	L
Enter	7	3	0	0	0	0	0	0
Idle	0	8	0	0	0	0	0	0
Bid attention	0	0	4	0	0	0	0	0
Wait	0	0	0	3	0	0	0	0
Take a drink	0	1	5	4	10	1	0	0
Drink	0	1	0	0	3	29	0	0
Socialize	0	0	0	0	0	1	18	0
Leave	0	0	0	0	1	0	0	6
Accuracy (%)	100	62	44	43	71	94	100	100

C. Results

To evaluate the performance of the proposed system, we used the test set (20% of data) as the input of the system and compared the output of the system (recognized social behavior) to the labels that we indicated manually through annotation process. The confusion matrix in Table II shows the recognized states and the percentage of states that were recognized correctly.

As shown in the table, the system is capable of recognizing the social behaviors entering, socializing and leaving with 100% accuracy. The social behaviors idle at room, bidding for attention, and waiting for drink, were recognized with lower accuracies of 62%, 44%, and 43%, respectively. The social behaviors taking a drink and drinking were recognized with 71% and 94% accuracy. The overall weighted average accuracy over all states is 81%.

D. Discussion

The Experimental results show that the system was able to correctly recognize different social behaviors, using only social signals based on body gestures and head poses. Therefore, using this system the humanoid robot is able to interpret the social behaviors of the interaction partners and respond in an appropriate way that results an acceptable human-robot interaction.

The accuracy of the model is limited due to the uncertainties and inaccuracies exist on the data of human body gestures (e.g., the ambiguous states prevent the precise labeling) and the inevitable inaccuracies in the Kinect data

flow (e.g., the light condition affects the Kinect acquired data). Both of these issues are inherent and unavoidable, however, to avoid an inappropriate human-robot interaction, we set a threshold on the recognition probability. In this way, system interprets the activity with the probability lower than threshold as *No Recognized Action*. It should be noted that, this interpretation is fundamental for the control of the robot's behavior.

In addition to the above mentioned sources of unavoidable uncertainties on the model, the low recognition accuracy is most likely due to the similarity of social signals in different activities as well as the diversity in number of training sequences. However, the system is able to discriminate activities in which the social signals are much different from other activities (e.g., entering, and leaving).

VI. Conclusion and Future Development

This paper presented the design and implementation of a social behavior interpretation system for a humanoid social robot. The system is based on features that represent body posture and head pose of a human, which we extracted with a Kinect RGB-D sensor. For behavior recognition, we trained an HMM that recognizes eight different social behaviors. For training and testing the social behavior interpretation system, we recorded and annotated 35 human-human interactions, which resulted in a data set of 524 labeled human social behaviors. The results of an evaluation of the system showed that the implemented model was able to recognize human social behaviors with a weighted average accuracy rate of 81%.

As a future work, we will develop a model considering more states, in order to allow the system to recognize a wide range of human activities in a human-robot interaction scenario. Furthermore, we plan to integrate the interpretation system with the other input modalities (e.g., speech recognition) of our humanoid robot.

References

[1] C. L. Breazeal, *Designing sociable robots*. MIT press, 2004.

[2] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: a survey," *Foundations and Trends in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007.

[3] A. Zaraki, D. Mazzei, M. Giuliani, and D. D. Rossi, "Designing and evaluating a social gaze-control system for a humanoid robot," *IEEE Transactions on Human-Machine Systems*, vol. PP, no. 99, pp. 1–12, February 2014.

[4] D. Mazzei, N. Lazzeri, D. Hanson, and D. De Rossi, "Hefes: An hybrid engine for facial expressions synthesis to control human-like androids and avatars," in *Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on*. IEEE, 2012, pp. 195–200.

[5] D. Mazzei, L. Billeci, A. Armato, N. Lazzeri, A. Cisternino, G. Pioggia, R. Igliozzi, F. Muratori, A. Ahluwalia, and D. De Rossi, "The face of autism," in *RO-MAN, 2010 IEEE*. IEEE, 2010, pp. 791–796.

[6] A. Zaraki, D. Mazzei, N. Lazzeri, M. Pieroni, and D. De Rossi, "Preliminary implementation of context-aware attention system for humanoid robots," in *Proceedings of the Second international conference on Biomimetic and Biohybrid Systems*. Springer-Verlag, 2013, pp. 457–459.

[7] A. Zaraki, M. B. Dehkordi, D. Mazzei, and D. De Rossi, "An experimental eye-tracking study for the design of a context-dependent social robot blinking model," in *Biomimetic and Biohybrid Systems*. Springer, 2014, pp. 356–366.

[8] M. Argyle, *Bodily communication*. Routledge, 2013.

[9] V. P. Richmond, J. C. McCroskey, and S. K. Payne, *Nonverbal behavior in interpersonal relations*. Prentice Hall Englewood Cliffs, NJ, 1991.

[10] M. Argyle, "Non-verbal communication in human social interaction." 1972.

[11] E. T. Hall and E. T. Hall, *The hidden dimension*. Anchor Books New York, 1969, vol. 1990.

[12] D. Mazzei, L. Cominelli, N. Lazzeri, A. Zaraki, and D. De Rossi, "I-clips brain: A hybrid cognitive system for social robots," in *Biomimetic and Biohybrid Systems*. Springer, 2014, pp. 213–224.

[13] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[14] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[15] H. Ning, T. X. Han, D. B. Walther, M. Liu, and T. S. Huang, "Hierarchical space-time model enabling efficient search for human actions," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 6, pp. 808–820, 2009.

[16] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis, "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2012–2019.

[17] M. Scheutz, J. Kramer, C. Middendorff, P. Schermerhorn, M. Heilman, D. Anderson, and P. Bui, "Toward affective cognitive robots for human-robot interaction," in *Proceedings of the National Conference on Artificial Intelligence*, vol. 20, no. 4. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005, p. 1737.

[18] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. IEEE, 1997, pp. 994–999.

[19] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 842–849.

[20] H.-D. Yang, A.-Y. Park, and S.-W. Lee, "Gesture spotting and recognition for human–robot interaction," *Robotics, IEEE Transactions on*, vol. 23, no. 2, pp. 256–270, 2007.

[21] A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter, and A. Knoll, "Social behavior recognition using body posture and head pose for human-robot interaction," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 2128–2133.

[22] D. Ramage, "Hidden markov models fundamentals," *Lecture Notes. http://cs229. stanford. edu/section/cs229-hmm. pdf*, 2007.

[23] C. R. Souza. (2014) Toolkit for face tracking @ONLINE. [Online]. Available: http://msdn.microsoft.com/en-us/library/jj131022.aspx

[24] MSDN. (2014) The kinect studio @ONLINE. [Online]. Available: http://msdn.microsoft.com/en-us/library/hh855389.aspx

[25] M. Kipp, "Multimedia annotation, querying and analysis in anvil," *Multimedia information extraction*, vol. 19, 2010.

[26] C. R. Souza. (2012) The accord.net framework @ONLINE. [Online]. Available: http://accord.googlecode.com