Social Signal Recognition Using Body Posture and Head Pose for Human-Robot Interaction

Andre Gaschler, Sören Jentzsch, Manuel Giuliani, Kerstin Huth, Jan de Ruiter and Alois Knoll

***Abstract*— Robots that interact with humans in everyday situations, need to be able to interpret the nonverbal social signals of their human interaction partners. We show that humans use body posture and head pose as social signals to initiate and terminate interaction when ordering drinks at a bar. For that, we record and analyze 108 interactions of humans interacting with a human bartender. Based on these findings, we train a Hidden Markov Model (HMM) using automatic body posture and head pose estimation. With this model, the bartender robot of the project JAMES can recognize typical social signals of human customers. Evaluation shows a recognition rate of 82.9 % for all implemented social signals and in particular a recognition rate of 91.2 % for bartender attention requests, which allows the robot to interact with multiple humans in a robust and socially appropriate way.**

I. Introduction and Related Work

The European project JAMES[1] develops a robot, which is shown in Figure 1, that works as a bartender in a scenario in which it takes drink orders from human customers and hands out beverages. We believe that the interaction between humans and a robot in this everyday interaction scenario—in contrast to for instance industrial contexts—needs to be social to ensure that the robot successfully completes its task. Thus, the robot needs to be able to correctly interpret the intentions of its human interaction partners.

In everyday situations, such as the JAMES bar scenario, humans often express their intentions nonverbally. In this publication, we show that humans predominently use body posture and head pose to initiate various steps of an interaction to order drinks in bars. These findings are based on the analysis of empirical data: actual human-human interactions that were recorded in several bars. Analogous to these results, we train a Hidden Markov Model (HMM) with data from automatic body posture and head pose estimation. This model is an effective classification system that allows the robot to robustly recognize social signals of human customers, for example whether they want to order a drink or are just staying at the bar to chat with friends.

In recent years, a few projects presented robot baristas, for example FusionBot [1] or Care-O-Bot [2]. The focus of these projects was research on robot motions and object manipulation capabilities, unlike in the JAMES project that researches socially appropriate robotics in everyday situations. Thus, the work of other groups, which focused on recognition of social signals in similar interactions as in JAMES, is better comparable to our work: Scheutz et al. [3] presented a framework for a robot that recognizes and generates affective behavior and evaluated parts of it on a mobile robot receptionist and waiter. Castellano et al. [4] described a system to recognize affective states and attentiveness. They found that sensing the intention and affective state of the human participants is crucial for HRI.

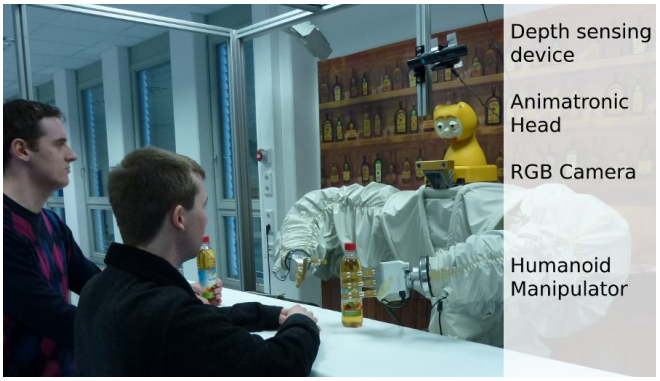

Fig. 1. The JAMES robot is able to recognize the intentions of humans in a multi-party bar scenario. It uses a camera and a depth sensor to analyze body posture and head pose of its human customers to infer their intentions.

The application of HMMs to model and recognize human actions from image sequences dates back to the work of Yamato, Ohya and Ishii in 1992 [5]. Considering human motion, Yang et al. [6] presented a full body gesture recognition system for HRI. They trained a Hidden Markov Model with data from visual human pose reconstruction to implement an automatic human gait recognition approach. With this recognition method, they were able to robustly detect human motions, including walking, running, jumping, and sitting. Lenz et al. [7] trained an HMM with data from a hand tracking device to automatically recognize the states of an interaction in an industrial human-robot interaction scenario. Finally, Vinciarelli et al. [8] give a general overview on the field of social signal recognition. They introduced a taxonomy of social signals in which body posture is listed as a social signal that is used to express emotion, personality, status, dominance, persuasion, regulation, and rapport.

II. Approach

Humans use social signals to communicate their intentions, for example in order to request the attention of

A. Gaschler, S. Jentzsch and M. Giuliani are with the fortiss GmbH affiliated to Technische Universität München, Munich, Germany. Correspondence should be addressed to gaschler@fortiss.org

K. Huth and J. de Ruiter are with the Department of Linguistics, Universität Bielefeld, Bielefeld, Germany.

A. Knoll is with the Department of Informatics, Technische Universität München, Munich, Germany.

[1] http://www.james-project.eu

a bartender, or to end the interaction when an ordering procedure is finished. The usage of these signals underlies certain regularities, thus, we collected a data corpus of over one hundred interactions in bars, in which human customers ordered drinks from a human bartender, and analyzed these data to find how humans use body posture and head pose as social signals (as described in Section II-A). Based on these data, we use body posture and head pose recognition (as described in Section II-B) to model and train an HMM with which we can recognize the most important social signals in a bar scenario. We have already applied head pose estimation for attention recognition in a work-in-progress paper [9], which we briefly summarize in Section II-C; in this work, emphasis is placed on the additional visual features, body posture and spatial group arrangements. Finally, we present the results of an evaluation of the trained HMM in Section II-D.

A. Body Posture and Head Pose in Everyday Human-Human Interaction

In everyday interaction between two or more humans mutual attention is a crucial feature for a successful interaction, which is established by the interactants without conscious effort. Many authors have pointed out that mutual attention and mutual gaze cannot be established without a foregoing body posture by interlocutors to each other ([10], [11], [12]). Therefore, body posture is the initial starting point for gaining mutual attention in an interaction. In addition, Clark [13] mentioned placement of persons as crucial for initiating an interaction in a study involving interactions at counters, which further motivated our analysis.

Ciolek and Kendon [14] researched how humans position themselves when they are interacting with each other. During conversation, humans build small spatial-orientational arrangements by facing each other around a smaller space. These arrangements are perfectly aligned to exchange speech, gazes, and gestures for an effective communication. Ciolek and Kendon formalized these communication arrangements and showed that their shape is influenced by internal factors (relationship between interactants, attitudes towards each other) as well as external factors (physical space can be crowded, noisy, or physically constrained). Kendon introduced the term f-formations for these spatial arrangements, which are also known as facing formations in literature. F-formations are typically built by humans during interactions in everyday situations to engage in joint action or in a conversation. Ciolek categorized six different f-formations that are used by two conversation partners, from which we only consider the four arrangements that we found in our recordings of human-human interactions: H-formation, conversation partners directly face each other and have their body panes in parallel; N-formation, conversation partners face each other and have their body planes in parallel while they are staying slightly displaced from each other; V-formation, the conversation partners' body planes are not parallel but form an angle of approximately 45°; L-formation, conversation partners stand at a right angle to each other.

To empirically research the frequency and importance of body posture and f-formations in everyday interactions, we recorded 108 interactions between bartenders and customers at bars in several German clubs. The recordings were made with two HD video cameras, whose viewing angles were roughly pointed in 45° horizontally at bartender and customer respectively, to ensure that the whole interaction can be seen. Following the recordings, we annotated the videos with ELAN [15][2] in order to count the usage of body posture to form f-formations between bartender and customer.

Our findings confirm that body posture is a crucial signal to initiate an interaction. Out of 108 interactions only two customers were hindered from reaching the bar and stood more than one meter away from the bar while ordering a drink. The other customers placed themselves physically directly in front of the bar to indicate their intention for an ordering interaction. 99 customers stood nearly parallel to the bar worktop and faced the bartender with their body, their head or both (94 customers directed their body in the direction of the bartender independent of the bartender's position). 94 customers initiated an H-, N- or V-shaped f-formation for interacting with the bartender. In 40 interactions it was due to the bartender's body movements, that the f-formation became an open L-shape or turned from an intended N-shape to a V-shape. After the interaction was finished, 26 persons remained at the bar. Out of these, 25 had their drink visibly in front of them. Furthermore, 15 persons turned their body away so it neither faced the bartender nor the front of the worktop. In groups, 18 cases out of 26, at least one group member turned to the other members and therefore away from the bartender. If the customers did not turn away, they wanted to continue talking to the bartender. Furthermore, we observe that head pose serves to emphasize body posture and helps to infer if a customer is in company. Thus, 94 of 106 attention requesting customers used their head posture as well to indicate that they requested the attention of the bartender.

In analogy, head pose serves as an indicator for the beginning of the end of an interaction as well. Due to functional reasons the head posture can precede the body posture. Thus if customers turn their head away from the bartender, for example downwards or sidewards, it can be inferred that they are about to end the interaction and might leave the bar area. In our analysis, we observed 88 customers who turned their head away while initiating the ending of the interaction. Out of the remaining 19 persons, 14 wanted to continue talking to the bartender after ordering a drink or asked questions related to the ordering interaction. 5 persons did not look at the bartender while finishing their interaction but looked up to the bartender as they left the bar[3]. Only

[2]ELAN is an annotation program by the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. It can be downloaded at `http://www.lat-mpi.eu/tools/elan/`.

[3]One person was not visible at the end of their interaction. Thus, this recording was excluded from the head pose analysis.

one person did not use head or body posture to signal an attention request; this single customer was a member of a group and chatting with a friend, who had his body directed to the bartender.

In a nutshell, *body orientation, body posture, head pose* and, to a lesser extent, spatial arrangement in a group are *relevant social signals* in this kind of interaction. Most notably, interactions are almost always initiated and ended using these signals. It is therefore our belief that for social human-robot interaction, the robot needs to be able to recognize these signals in order to perform a socially appropriate interaction. Please note that additional findings of our human-human interaction study are published in [16].

B. Body Posture and Head Pose in Human-Robot Interaction

The understanding of the relevant communication signals among humans is necessary to model the analogous domain for HRI. In our scenario, which is shown in Fig. 1, the robot takes the role of a bartender. Multiple human customers may approach the bar area, order drinks, ask for the menu or a specific drink, chat with each other or simply spend time at the bar. We chose this setting as it focuses on short-horizon, multi-party social interaction.

For the robot to behave in a socially appropriate way, it needs to recognize a customer's request for attention when initiating and ending interaction with the bartender. This knowledge allows the robot to recognize and serve customers, respect ongoing conversations, and even glance to newly arrived customers while still proceeding with an ongoing interaction. Corresponding to the results of the human-human study, we designed the JAMES robot to recognize and process three types of features, as shown in Fig. 2: first, we recognize faces and estimate gaze by a head pose estimation routine. This software component was already presented in [9] and only slightly adapted for this work. Second, we capture human motion from a depth camera (Microsoft Kinect), which reconstructs human joints and motion data from infra-red structured light illuminated images in real-time. These human motion data corresponds to body posture, as referred to in the human-human interaction study. Third, we include spatial arrangement among the group of human participants. This feature is computed from the set of recognized participants and their body orientations by applying two fuzzy-valued functions for each person.

C. Image Recognition Process

Fig. 2 shows an overview of the image processing steps that are necessary to obtain the three social signals head pose, body posture, and spatial group arrangement. This section explains these steps in more detail.

1) Head Pose Estimation: The head pose estimation component was implemented in the course of our earlier work in [9], where it is also explained and evaluated in detail. Following the algorithm described by Vatahska, Bennewitz, and Behnke [17], the head pose estimator is divided into a face and facial feature detection step and a neural network-based head pose estimation. First, faces are detected based

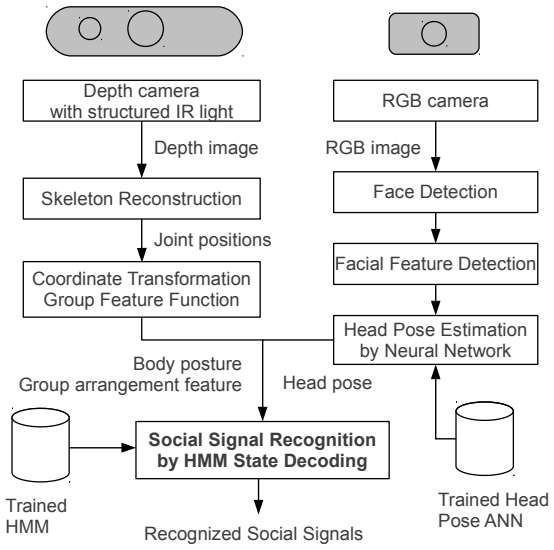

Fig. 2. Overview of image processing steps to obtain body posture, head pose, and group arrangement.

on the well-known Haar-feature classifier by Viola and Jones [18]. Following that, we are using facial feature detection to estimate the location of eyes, nose, and mouth of the recognized person. Second, all visually detected feature positions and their mutual distances are processed by a 3-layer artifical neural network (ANN), depending on which set of features is detected. Finally, the head pose estimation component outputs the set of visually detected face positions and their respective pitch, yaw, and roll angles in space. As evaluated in [9], this head pose estimation technique allows us to recognize head pose angles at an accuracy of approximately 15°. Even though the algorithm is rather simple, its robustness and accuracy meet the requirements of the JAMES bar scenario.

2) Body Posture Reconstruction: The advent of affordable, infra-red-based depth cameras, originally designed for console games, has made it easy and straightforward to capture human motion without using intrusive sensors or markers. The scene is illuminated by static, structured infra-red light, captured by a monochrome camera, and processed on-board to generate real-time depth images. In a second step, human skeletons are fitted and tracked by the PrimeSense NITE software [19]. Both accuracy and robustness of the resulting joint positions are sufficient for our scenario. The skeleton reconstruction step is followed by a coordinate transformation step to output joint positions with respect to the coordinate frame formed by the position and orientation of the respective person. This transformation serves two purposes: first, the relation between gestures and body posture vectors becomes simpler, as body posture is formulated independent from absolute position and orientation. Second, feature space can be reduced more easily by simply omitting less relevant coordinates, which allows us to simplify the design of our HMM.

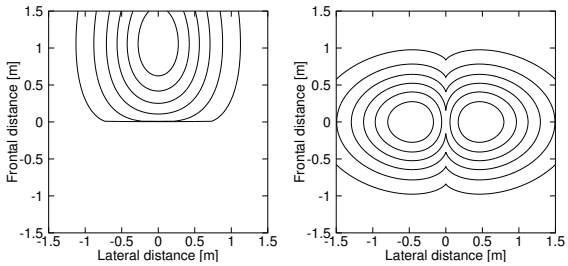

Fig. 3. Fuzzy features used to detect spatial group arrangements (f-formations) between two interacting humans.

3) Spatial Group Arrangement: As we showed in Section II-B, not only personal body posture and head pose are informative social signals, but also the spatial arrangement in a group is relevant to express engagement in an interaction. Following the observations of Ciolek and Kendon [14], we apply two fuzzy functions to map the spatial arrangement of a person in a group to a simple 2-dimensional feature vector, describing the group arrangement, i.e. the f-formation, with respect to that person. These two very simple functions, which we show in Fig. 3, are dependent on the positions of other humans; the left function responds to other humans being in front of a person, the right function to other humans standing beside a person. The first function corresponds to Ciolek's H-, N-, and V-formations, whereas the second corresponds to L-formations. We found this 2-valued feature vector to be sufficient to automatically recognize engagement and disengagement of interactions between participants.

D. Social Signal Recognition with Hidden Markov Models

Having outlined the image processing steps to obtain a feature vector from body posture, head pose, and spatial group arrangement, in this section we explain our social signal recognition approach that is based on these features. For modeling the temporal sequence of behavioral states and social signals, we make use of HMMs [20]. HMMs are a powerful approach to model sequential and statistical processes that allow indirect observation. In our scenario, it is obvious to model the sequence of social signals as the hidden states of a Markov process. Accordingly, measured body posture and head poses correspond to the observable emissions of the hidden states of that model.

As defined by Rabiner [20], a hidden Markov model $\boldsymbol{\lambda}$ consists of a set of n hidden states $\mathbf{X}$, a set of m observation variables $\mathbf{Z}$, an $n \times n$ transition matrix $\mathbf{A}$, and an $n \times m$ emission matrix $\mathbf{B}$. The state transition matrix $\mathbf{A}$ contains the probability for a transition from one hidden state to another within a time step and therefore models the stationary stochastic process. The emission matrix $\mathbf{B}$ maps the hidden states to the probability of observed variables. To complete the definition of an HMM, an initial state probability vector $\boldsymbol{\pi}$ may be added to list the initial probability of each hidden state.

Following this notation, the joint distribution of a specific sequence with hidden states $X_1, X_2, \ldots, X_T$ and observable variables $Z_1, Z_2, \ldots, Z_T$ over T time steps is then

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\lambda}) = \underbrace{p(X_1|\boldsymbol{\pi}) \left[\prod_{t=2}^{T} p(X_t|X_{t-1}, \mathbf{A})\right]}_{\text{transition}} \underbrace{\prod_{t=1}^{T} p(Z_t|X_t, \mathbf{B})}_{\text{emission}} \tag{1}$$

Eq. 1 completely governs the model, as all further probablities of observation sequences and hidden state sequences can be deduced from it. The above defined set of model parameters $\boldsymbol{\lambda} = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}\}$ is therefore sufficient to describe a hidden Markov model. In order to reflect our real-valued and multi-dimensional observations—body posture and head pose—the emission matrix $\mathbf{B}$ is augmented to a d-dimensional emission distribution $b_j(\boldsymbol{Z_t})$, and the observation $\boldsymbol{Z_t}$ to a d-dimensional vector, for each hidden state j:

$$b_j(\boldsymbol{Z_t}) = \sum_{k=1}^{K} c_{jk} \cdot \mathcal{N}(\boldsymbol{Z_t}|\boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \tag{2}$$

The emission distribution is parameterized by k Gaussian distributions with d-dimensional mean vectors $\boldsymbol{\mu}_{jk}$ and $d \times d$ covariance matrices $\boldsymbol{\Sigma}_{jk}$, which are then weighted by the mixing coefficients c_{jk}. All in all, the set of parameters $\boldsymbol{\lambda} = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ completely governs our Continuous Multidimensional HMM.

The two fundamental problems for HMM-based recognition components are *model training* and *state decoding*. The model training step is used to find the parameter set that generates a given training set of observations with maximum likelihood. This procedure may be performed offline and serves the trained social signal recognition model as an output. Mathematically, the objective of the training is to maximize the likelihood of the given observation sequences $\mathbf{Z}$ finding the optimal model $\boldsymbol{\lambda}$ maximizing $p(\mathbf{Z}|\boldsymbol{\lambda})$.

Even though this problem cannot be directly maximized, typical practical instances can be solved by expectation maximization (EM), which yields an iterative local maximization of the likelihood function [20]. However, this local solution only allows optimization of our continuous learning parameters $\boldsymbol{\lambda}$, assuming we define a reasonable $\boldsymbol{\pi}$ in advance. The free model parameters—the number of hidden states per action and the number of Gaussian mixtures—can only be found through grid search, performing the actual EM learning procedure for all reasonable discrete model parameter settings.

E. Social Signal Model Definition

In this section, we describe the design of our HMM and its parameterization. Fig. 4 shows the states and transitions of our model for social signal recognition in the JAMES scenario. Each human participant is modeled by a separate state model. The states were chosen both to allow detection of crucial social signals in our scenario, but also to limit complexity and thus allowing reasonable accuracy of the recognized states. Through experimental observation, we chose a model with eight states, which sufficiently describes

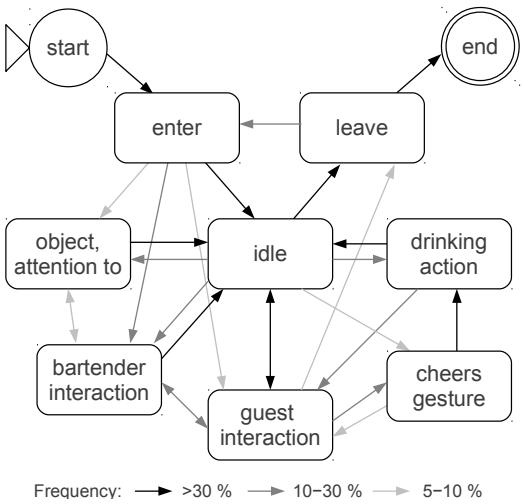

Fig. 4. Hidden Markov Model and observed frequency of transitions of most common interaction states in the JAMES bar scenario.

the behavior of the human customers and allows the robot to perform a socially appropriate interaction: the state *idle* is the default state of a person located at the bar; state *attention to object* is used to model humans, who look at a menu located behind the bar; state *cheers gesture* describes the action of two persons clinking glasses; state *bartender interaction* describes humans, who order drinks from the robot; state *drinking action* describes a human, who drinks from an already received beverage; the states *enter* and *leave* describe the humans entering or leaving the bar, respectively. The shades of the arrows in Figure 4 reflect the observed frequency of the transitions in the training data set.

F. Model Training and Parameterization

In order to train the HMM, we used the robot's camera and depth sensor to collect recordings of more than 200 scenes in which humans interacted with the robot and with each other. The scenes included one, two, or three participants. All data were labeled by hand, resulting in a total of 1720 states. The scenarios were divided into a training set of 1010 states, a cross-validation set suitable for model parameterization of 319 states and a testing set of 391 states. From the labeled training data set, the transition frequencies were counted and further used as additive transition probabilities. We also pruned all transitions with a frequency less than 5%, allowing a reasonably simplified model and slightly improved performance. All model parameterization steps were evaluated on the cross-validation data set.

First, we evaluated several feature vector definitions and found that not all body posture data are equally relevant for recognition performance. We observed that the best set of features contains torso and hand positions (excluding arm positions), body alignment, head pose given both as a normal vector and as pitch and yaw angles, as well as the two spatial group arrangement feature values. In total, our chosen feature vector contains 19 normalized, real-valued components.

Second, we evaluated different hyperparameters on the cross-validation data set. We observed a slight advantage of full covariance matrices as emission models compared to diagonal variance matrices. This advantage could possibly be explained by the geometric nature of the feature vector components and their inherent cross-correlation. Ultimately, an HMM with just one inner state and an emission mixture of three Gaussians with full covariance matrices for each social signal showed best performance.

TABLE I

STATE CHANGE PENALTY ON CROSS-VALIDATION DATA SET.

Transition Penalty	**%Corr**	**%Acc**	**H**	**D**	**S**	**I**
0%	95.92	3.45	306	5	8	295
10%	90.91	45.14	290	16	13	146
20%	86.21	61.76	275	27	17	78
25%	84.33	65.20	269	32	18	61
30%	83.07	67.08	265	38	16	51
35%	82.13	68.03	262	40	17	45
40%	79.31	67.08	253	50	16	39
50%	78.37	67.71	250	61	8	34

Third, we optimized the state change penalty parameter. This parameter is necessary to define a compromise between false deletions (D) and false insertions (I), as shown in Table I. Even though HMMs are mostly independent of time scale, some adjustment is necessary to take the different orders of magnitude between camera frame rate and typical transition rate into account. Table I shows that false insertions are substantially reduced when applying a state change penalty, resulting in a slight decrease of correctly recognized states (H) by increasing the number of deletions (D); substitutions (S) are hardly affected. In our scenario, false deletions—ignoring a customer's request or interrupting a conversation—are less acceptable than false insertions—asking an idle participant to order. Therefore, we chose a moderate state change penalty of 30%, slightly below that of the optimal accuracy.

III. EVALUATION

The results of our social signal recognition model are shown in Table II. The confusion matrix summarizes substitutions of recognized states compared to manually labeled states. Furthermore, false insertions and deletions, and the percentage of correctly recognized states are shown. All in all, 82.9% of the states were recognized correctly. More specifically, the crucial social signal "bartender interaction" is recognized at a rate of 91.2% and with few false insertions. Similarly, "interaction with other guests" is recognized at a rate of 94.7%. With these recognized signals, the robot is able to initiate, perform and close interactions in a socially appropriate way.

It should be noted that there are a number of factors limiting the accuracy of the HMM: input data of body posture and head pose may be inaccurate or even wrong, such as in the case of occlusions. Also, the states are sometimes ambiguous and do not allow exact and precise

TABLE II

CONFUSION MATRIX AND RESULTS ON THE TEST DATA SET

Recognized states	**Labeled states** b	o	g	c	d	e	i	l	D	%Corr.
bartender	52	1	0	1	1	2	0	0	6	91.2
object	0	24	0	0	0	0	1	0	2	96.0
guest	0	0	36	0	0	0	1	1	7	94.7
cheers	2	0	0	12	1	1	1	1	6	66.7
drink	0	0	1	0	35	1	1	0	6	92.1
enter	0	0	0	0	0	30	0	0	1	100.0
idle	0	1	1	1	0	1	105	0	17	96.3
leave	0	0	0	0	0	0	0	30	1	100.0
I	8	10	5	7	4	5	18	8		

Correctness	82.9%	H/N
Accuracy	66.2%	(H–I)/N
Correctly recognized states	324	H
Deletions	46	D
Substitutions	21	S
Insertions	65	I
Number of states	391	N

labeling. These factors are inherent and cannot be overcome by collecting more training data or changing properties of the model.

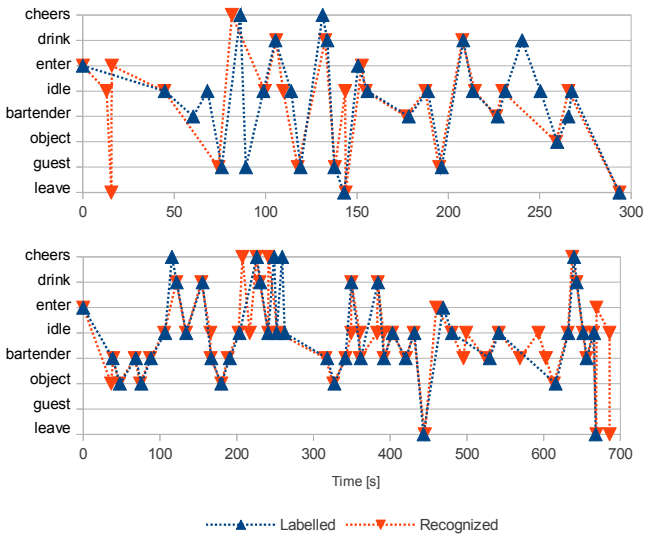

Fig. 5. Sample social signal recognition results compared to labeled states over time. Upper view shows a participant within a group, lower view shows a single participant.

Besides the results of the confusion matrix, we also reviewed the accuracy of the social signal recognition over time. Fig. 5 shows two scenes with labeled and recognized states in comparison; one of a participant in a group, one of a single participant interacting with the robot. The important observation from this diagram is that there is no visible systematic delay between labeled and recognized social signals.

IV. CONCLUSION

The contribution of this work is twofold: first, we empirically researched the frequency and importance of body posture and f-formation in an everyday situation, namely the interaction with a bartender. Quantitative analysis showed that three signals are crucial to non-verbally initiate, answer and end such kind of interaction: body posture, f-formation, and head pose. Second, we applied these findings to a human-robot interaction scenario and implemented image processing components to recognize the signals. Finally, we trained a hidden Markov model with recorded human-robot interactions. Evaluation showed a correct social signal recognition in 82.9% of all test cases. Most notably, we allowed the robot to recognize the vital social signal "bartender interaction" at a rate of 91.2%, which enables the robot to interact with humans in a socially appropriate way.

V. ACKNOWLEDGMENTS

This research was supported by the European Commision through the project JAMES (FP7-270435-STREP).

REFERENCES

[1] D. Limbu, Y. Tan, C. Wong, R. Jiang, H. Wu, L. Li, E. Kah, X. Yu, D. Li, and H. Li, "Experiences with a barista robot, FusionBot," *Progress in Robotics*, pp. 140–151, 2009.

[2] B. Graf, U. Reiser, M. Hagele, K. Mauz, and P. Klein, "Robotic home assistant Care-O-bot® 3-product vision and innovation platform," in *Advanced Robotics and its Social Impacts (ARSO), 2009 IEEE Workshop on*. IEEE, 2009, pp. 139–144.

[3] 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, 2005, p. 1737.

[4] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. McOwan, "Affect recognition for interactive companions: challenges and design in real world scenarios," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 89–98, 2010.

[5] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Computer Vision and Pattern Recognition, Conf on*, 1992, pp. 379–385.

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

[7] C. Lenz, A. Sotzek, T. Röder, H. Radrich, A. Knoll, M. Huber, and S. Glasauer, "Human workflow analysis using 3d occupancy grid hand tracking in a human-robot collaboration scenario," in *Intelligent Robots and Systems (IROS), Intl Conf on*, 2011, pp. 3375–3380.

[8] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[9] A. Gaschler, K. Huth, M. Giuliani, I. Kessler, J. de Ruiter, and A. Knoll, "Modelling state of interaction from head poses for social Human-Robot Interaction," in *Proc. of the Gaze in Human-Robot Interaction Workshop held at the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012)*, Boston, MA, 2012.

[10] C. Heath, "Talk and recipiency: Sequential organization in speech and body movement," *Structures of social action: Studies in conversation analysis*, pp. 247–265, 1984.

[11] A. Kendon, "Features of the structural analysis of human communicational behavior," *Aspects of Nonverbal Communication, Swets and Zeitlinger, Lisse, Netherlands*, pp. 29–43, 1980.

[12] E. A. Schegloff, "Opening sequencing," *Perpetual contact: Mobile communication, private talk, public performance*, p. 326, 2002.

[13] H. H. Clark, "Pointing and placing." *Pointing: where language, culture, and cognition meet*, p. 243, 2003.

[14] T. M. Ciolek and A. Kendon, "Environment and the spatial arrangement of conversational encounters," *Sociological Inquiry*, vol. 50, no. 3-4, pp. 237–271, 1980.

[15] H. Sloetjes and P. Wittenburg, "Annotation by category: Elan and iso dcr," *Proceedings of the Sixth International Language Resources and Evaluation (LREC08), Marrakech, Morocco, may*, 2008.
[16] K. Huth, S. Loth, and J. de Ruiter, "How to order a beer; distilling computational models from natural data for social robotics," in preparation.
[17] T. Vatahska, M. Bennewitz, and S. Behnke, "Feature-based head pose estimation from images," in *Humanoid Robots, 2007 7th IEEE-RAS International Conference on*. IEEE, 2007, pp. 330–335.
[18] P. Viola and M. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.
[19] "PrimeSense NITE," http://www.openni.org/.
[20] L. R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. Proceedings of the IEEE, Vol. 77, No. 2, February, 1989.