Modelling State of Interaction from Head Poses for Social Human-Robot Interaction

Andre Gaschler
fortiss GmbH
Guerickstr. 25
80805 München, Germany
gaschler@fortiss.org

Kerstin Huth
Universität Bielefeld
Universitätsstr. 25
33615 Bielefeld, Germany
khuth@uni-bielefeld.de

Manuel Giuliani
fortiss GmbH
Guerickstr. 25
80805 München, Germany
giuliani@fortiss.org

Ingmar Kessler
fortiss GmbH
Guerickstr. 25
80805 München, Germany
kesslein@in.tum.de

Jan de Ruiter
Universität Bielefeld
Universitätsstr. 25
33615 Bielefeld, Germany
jan.deruiter@uni-bielefeld.de

Alois Knoll
Technische Universität München
Boltzmannstr. 3
85748 München, Germany
knoll@in.tum.de

ABSTRACT

In this publication, we analyse how humans use head pose in various states of an interaction, in both human-human and human-robot observations. Our scenario is the short-term, every-day interaction of a customer ordering a drink from a bartender. To empirically study the use of head pose in this scenario, we recorded 108 such interactions in real bars. The analysis of these recordings shows, (i) customers follow a defined script to order their drink—*attention request*, *ordering*, *closing of interaction*—and (ii) customers use head pose to nonverbally request the attention of the bartender, to signal the ongoing process, and to close the interaction.

Based on these findings, we design a hidden Markov model that reflects the typical interaction states in the bar scenario and implement it on the human-robot interaction system of the European JAMES project. We train the model with data from an automatic head pose estimation algorithm and additional body pose information. Our evaluation shows that the model correctly recognises the state of interaction of a customer in 78.3% of all states. More specifically, the model recognises the interaction state "attention to bartender" with 83.8% and "attention to another guest" with 73.0% correctness, providing the robot sufficient knowledge to begin, perform, and end interactions in a socially appropriate way.

Keywords

Head Pose Estimation, State of Interaction, Social Human-Robot Interaction

1. INTRODUCTION

Gaze in HRI: From Modelling to Communication Workshop 2012 Boston, Massachusetts USA

Figure 1: The robot bartender of the JAMES project uses head pose estimation to infer the state of interaction of human customers in a multi-party bar scenario.

Humans use head pose and gaze direction as nonverbal cues to express communicative acts and their visual focus of attention. This has been decribed in literature many times, for example by Kendon [8], who recorded conversations of two humans. He found that humans use gaze direction differently depending on the length of their verbal utterance. When humans speak long utterances, for example to discuss a personal view, they look away from their interaction partner at the beginning of the utterance and look back to their partner at the end of the utterance to signal the opening of a possibility to enter the discussion. When using short utterances, for example to ask questions to gather information, humans mostly look directly at their interaction partner.

Clark [3] lists head pose as one of the signals that humans use for *directing-to* communicative acts, i.e. humans use head movements to direct the attention of a person they are speaking with to an entity in the environment of the interaction partners or to themselves. Langton et al. show in [10], that gaze direction is an important cue to express visual attention, but head direction and pointing gestures also

contribute significantly to the computation of the attention of an interaction partner. Furthermore, Langton et al. [11] found that a combination of head pose estimation and recognition of gaze direction is needed to exactly determine the visual attention of an interaction partner. Stiefelhagen [20] found that for recorded meetings, head pose alone is enough to correctly recognise the visual focus of attention of a meeting participant in 89% of the time.

In this publication, we recorded and analysed ordering sequences in bars, in which customers order and receive drinks from a human bartender. The quantitative analysis of these recordings shows that customers follow a fixed sequence to order a drink. The data further suggests that head pose is an important cue that is used by customers and bartenders in all steps of the ordering sequence. Based on these findings, we design and train a hidden Markov model using head pose estimation, allowing a robot bartender to automatically recognise the interaction state of humans in a bar scenario. We then implement and evaluate this model on the human-robot interaction system of the European project JAMES[1]—Joint Action for Multimodal Embodied Social Systems. **Figure 1** shows the robot during an interaction with a customer. The robot setup consists of two industrial robot arms (Mitsubishi Melfa RV-6SL) with compliant hands (Meka H2), and an animatronic head (Philips iCat) which is capable of producing emotions and lip-synchronised speech. The robot is equipped with two stereo cameras (PointGrey Bumblebee) and a depth sensor (Microsoft Kinect).

2. RELATED WORK

The challenge of *visual head pose estimation* is a well-studied field that offers various approaches and techniques. Murphy et al. [13] give an overview of head pose estimation approaches. The authors argue that head pose estimation is the first step that enables recognition of other nonverbal cues, including detection of gaze direction and emotion recognition. Humans can easily estimate the head pose of an interaction partner, but for a computer vision system this ability is rather difficult to achieve. Technically, head pose estimation is the recovery of the 3D pose of a human head from digital images or continuous image sequences. The pose usually covers pitch, yaw and roll angles, as well as the approximate position of the head in space.

The diversity of head pose estimation techniques can roughly be classified into appearance-based methods and model-based methods. Appearance-based methods, or image-based methods, consider the image region of the face as a whole. One way to recover head poses is to apply coarse-to-fine classifiers, which are previously trained on the pose space [12]. This classification approach naturally delivers only discrete output, especially when a small set of combined detection-pose estimation classifiers is used [7]. Another common appearance-based method is image-based face tracking, which allows rather accurate relative motion recovery, but usually faces the problem of proper initialisation of a perfectly frontal face.

Model-based methods, or feature-based methods, abstract from the actual image region and recover low-dimensional features, such as the position of facial features. These features can be processed by pure geometry [5], by a trained artificial neural network (ANN) [21], or by other non-linear regression methods [13]. Of course, appearance-based and feature-based techniques can be combined, for example by tracking facial features [14] or using dense stereo images and neural network processing [18].

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

3. APPROACH

In this section, we present our findings on the usage of head pose in human-human bar scenarios (Section 3.1), our implementation for automatic head pose estimation (Section 3.2), and our model of the interaction state of the human customers (Section 3.3).

3.1 Human-Human Observation

In everyday interactions between two or more humans, mutual attention is established automatically. As Kendon [9] proposed, it is functional and necessary to establish mutual attention to begin a conversation. For an everyday conversation, Kendon's domain was defined by turning to each other and exchanging mutual gazes. On the basis of the importance of mutual gaze, Bavelas et al. [1] showed that gaze windows are produced by speakers to request a listener's responses. Furthermore the authors found that listeners tend to look at speakers more often so they are able to respond quickly and appropriately, if their response is requested. As communicational behaviour and gaze exchange is done with routine by humans in everyday life it is important to train social robots to be able to interpret the head pose of humans, too.

To empirically study how humans interact in bar scenarios, we recorded bartenders and customers in several bars in Germany. For the recordings, we installed two cameras and two microphones to record bartenders and customers in a horizontal viewing angle of approximately 45 °, covering an area of 3 to 4 meters in front of the bartender's working area. For this study, we put our focus on situations in which the customers ordered drinks from the bartender. After annotating the video data with ELAN [19][2], our data corpus contains 108 successful ordering interactions. The annotation includes the interaction state of the customer (waiting-at-bar, bidding-for-attention, talking-to-other-customers, etc.), the focus of attention of bartender and customers, gestures, and speech.

We found that from the viewpoint of the customer a typical ordering interaction consists of three states: (1) *attention request* towards bartender, (2) *ordering* of one or more beverages, and (3) *closing of interaction* by payment and exchange of polite phrases. The bartender reacts to these states by (1) acknowledging the attention request, (2) serving the ordered drink, and (3) asking for payment. The sequence of interaction states can also include substates, for example some customers ask for information about the available choice of drinks. However, we found that the usage of head pose does not differ from the three main interaction states that we are focusing on in this work. Please refer to [6] for a complete overview of all states of interaction that we found. The data analysis shows that mutual attention is crucial to initiate an ordering interaction as well as for

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

Customer I think I'd like to have a "Cape Cide", if it is in the fridge over there...
[customer orders drink and mentions uncertainty about the availability of the drink]

Bartender Yes
[bartender verbally acknowledges request but continues working and does not look up to customer]

Customer ... because it's not in the one right here ...
[customer did not hear the verbal acknowledgement of bartender and continues with ordering]

Customer ... if I saw it right, but maybe I was blind-folded
[since the customer gets no visual acknowledgement by the bartender, she gets uncertain about her order]

Figure 2: Sample dialogue in which a bartender does not follow the established protocols of using head pose to acknowledge a customer's request. In the consequence, the customer gets uncertain about the order and the interaction slows down.

a successful and fluent interaction in the remaining states of interaction. Thus, the bartender has to understand the verbal and nonverbal signals by the customer. For an example, it would be inappropriate for the bartender to ask a customer for another order while he or she is engaged in a conversation.

In the data analysis of our data corpus, we found that in interaction state (1) *attention request*, customers mainly request the bartender's attention nonverbally. They use their physical presence at the bar, and in particular the head and body direction as cues to display their request for attention. In our corpus, we found 103 requests for attention. Out of these 103 attention requests, 92 persons directed their head towards the bartender to request the attention of the bartender. From these 92 persons, 63 directed their body to the bartender as well. Only 11 participants directed their body to the bartender but not their head, because they were engaged in other activities, including talking to other customers or reading the bar menu.

During interaction state (2) *ordering*, 101 out of 108 persons directed their head to the bartender, the remaining seven persons did not direct their head straight to the bartender because they preferred talking sideways to increase the loudness of their voices, which was due to loud music that was played in some of the locations in which the recordings took place. Again, head pose served as the main cue for the attention of customers who spoke to the bartender at the same time. Due to environmental factors at the locations, such as bad lighting conditions and glasses hiding customers' eyes, we were not able to analyse gaze of all of the customers. However, 65 customers clearly looked at the bartender during the interaction, and in the remaining cases, the bartender was still able to acknowledge their requests. Thus, we assume that the typical gaze window, as introduced by Bavelas et al., can also be observed in our data. To strengthen this point, in **Figure 2** we present a dialogue between a customer and a bartender, which was taken from one of our recordings. In this example, the bartender mixes a drink while asking for the customer's order.

In analogy to Bavelas' observation, which was that story tellers become uncertain when the listener's appropriate responses are lacking, in the example dialogue the customer becomes uncertain because she does not hear the bartenders acknowledgement. She looks at the bartender who does not look up but finishes mixing a drink first. So, the customer begins to justify her order, still waiting for a verbal or nonverbal response by the bartender.

In interaction state (3) *closing of interaction*, 48 customers did not look at the bartender while performing the last part of the interaction (paying, getting change or thanking the bartender for the service). From these 48 customers, 27 customers directed their head downwards and the remaining 21 customers directed the head to other directions, but away from the bartender. 40 customers directed their head away from the bartender right after the interaction was fully completed. From these 40 customers, 22 customers directed their head downwards. Out of the remaining 19 customers, 14 customers continued to talk to the bartender and 5 customers did not direct their head to the bartender while finishing their interaction, but looked up to the bartender as they left the bar. The reason for many customers to direct their head downwards was mainly functional, because they needed to handle money or their drink. However, the bartender can infer from this cue that the interaction is finished.

To summarise this section: we found that head direction is an important cue to recognise the intention of a human in all interaction states of an ordering sequence during the interaction with a human bartender. We infer from these findings that head pose estimation is an important cue for a robot bartender to infer the interaction state of a human user in an ordering sequence and that head pose estimation should be an adequate source to automatically recognise these states.

3.2 Visual Head Pose Estimation

In order to recognise the status of interaction of human customers, the robot needs to estimate their head poses from visual input. From the above results of our human-human study, we can expect that head pose estimation will supply the information necessary to follow the state of interaction.

Our visual head pose estimation algorithm is mostly based on the works by Vatahska, Bennewitz, and Behnke [21]. Their approach for head pose estimation follows two steps: first, the algorithm detects faces and facial features. Second, it processes the positions of the facial features with a trained neural network to yield the three head pose angles pitch, yaw, and roll. We chose this approach because it is simple, yet computationally efficient, robust and sufficiently accurate for our human-robot interaction scenario.

The face and facial feature detection is based on the well-known Haar-feature classifier by Viola and Jones [22]. Frontal and profile faces are handled by separate classifier cascades. Then, we detect eyes, nose and mouth by appropriate classifier cascades in the expected regions of interest, assuming an upright head pose with a roll angle within $\pm 25^\circ$. For the detection, we used classifiers that are available from the OpenCV library [2].

Provided that at least one of the four facial features—eyes, nose and mouth–is detected, we extract their position and mutual distances, as shown in **Figure 3**. This information vector is normalised and fed into one of 22 three-layer perceptrons, one for each possible combination of features and view, frontal or profile. Each artificial neural network was trained by the annotated Head Pose Image Database [4] and synthesised images based on the Basel Face Model [15],

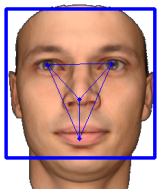

Figure 3: Position and distance features used for head pose estimation. 3D model for rendering taken from [15].

totalling a number of 6000 training images.

The neural networks were designed with one intermediate layer of six neurons and the symmetric sigmoid $f(x) = (1 - e^{-x})/(1 + e^{-x})$ as an activation function. All input and output variables were normalised to the range $[-0.5; 0.5]$; for input positions and distances, the outer box in **Figure 3** is taken as a reference, with the coordinate origin in the centre. For training, the RPROP learning algorithm [17] was used with the adaptation parameters $\eta^+ = 1.2$ and $\eta^- = 0.5$.

Table 1: Head pose angular accuracy on test data sets. All columns show the mean deviation from the known angle.

Test data set	Pitch	Yaw	Roll
Synthesised images from [15]			
Frontal view, all features	8.5°	5.4°	4.3°
Frontal view	12.2°	14.3°	10.5°
Profile view	11.2°	10.9°	8.9°
Labelled images from [4]			
Frontal view, all features	12.3°	9.1°	1.3°
Frontal view	14.9°	16.8°	3.8°
Profile view	14.5°	16.7°	3.2°

With the described head pose estimation technique, we are able to recover the face orientation angles within an uncertainty of about 15 degrees, as shown in **Table 1**. When all features are detected, the accuracy is improved to 10 degrees. This accuracy is perfectly within our requirements, as it allows us to recognise the attention and the state of interaction of the human guests in our HRI scenario. Please note that for comparison with the results in [21], our results in **Table 1** also include cases of incompletely detected facial features and non-synthetic images.

As a result, we are able to estimate head poses of multiple interaction partners of the JAMES robot. The computer vision system was implemented on commodity hardware with GPU acceleration and delivers real-time results at 15–30 fps, depending on the number of people in the field of view of the camera. Now that we can robustly estimate head poses, the next processing step covers the modelling of the interaction states in our bar scenario. As we have seen in Section 3.1, understanding the meaning of the head pose as a focus of attention is vital to successful interaction. One way to automatically recognise the state of interaction of the participants is to model the sequence of interaction, which we describe in the following section.

3.3 Modelling State of Interaction

A common approach to model sequential, statistical processes are hidden Markov models (HMM). Sequential processes are very common in nature, and sequential data arise in distinctive fields, such as sound and speech processing, DNA sequencing or economics. A HMM consists of a set of hidden states, whose transitions are modelled, and a set of observable output signals. For our application in socially appropriate human-robot interaction, the hidden states correspond to the states of interaction with a particular person, and the output signals correspond to the measured head poses of that person.

Following the notation by Rabiner [16], an HMM λ consists of n hidden states X, m observation variables Z, an $n \times n$ transition matrix A, and an $n \times m$ emission matrix B. The state transition matrix 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 B maps the hidden states to the probability of observed variables. To complete the definition of an HMM, an initial state probability vector π may be added to list the initial probability of each hidden state.

Combining all of the above, 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

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

From **Equation 1** one can deduce the probabilities of observation sequences and hidden state sequences. The set of model parameters $\lambda = \{\pi, A, B\}$ is therefore sufficient to describe the HMM.

In addition to this model definition with only discrete observations, our head pose measurements are both multidimensional and (spatially) continuous. To reflect this, the simple emission matrix of the above HMM is replaced by a d-dimensional continuous emission distribution of k multivariate Gaussian distributions, forming a continuous multidimensional HMM. For that, we define Z_t as a d-dimensional vector with the emission distribution $b_j(Z_t)$ for each hidden state j:

$$b_j(Z_t) = \sum_{k=1}^{K} c_{jk} \cdot \mathcal{N}(Z_t|\mu_{jk}, \Sigma_{jk}) \tag{2}$$

The emission distribution is parameterised by a mixture of k Gaussian distributions with D-dimensional mean vectors μ_{jk} and $d \times d$ covariance matrices Σ_{jk}, which are weighted by the mixing coefficients c_{jk}. Please note that we constrain Σ to a diagonal matrix to reduce the number of parameters and avoid over-fitting. Finally, the set of parameters $\lambda = \{\pi, A, c, \mu, \Sigma\}$ completely governs our continuous multidimensional HMM.

The two fundamental problems for our Hidden Markov Model of the states of interaction are *model training* and *state decoding*. For model training, a sufficiently large observation sequence Z needs to be labelled with the hidden state sequence X. From this, the optimal HMM λ is solved, that generates the observation sequence with maximum likelihood, which can be achieved by expectation maximization (EM) [16]. The second problem, state decoding, determines

the most likely sequence X for a given observation sequence Z and a HMM λ. This can be achieved by the Viterbi algorithm [16], and is part of the on-line recognition component of our robot setup.

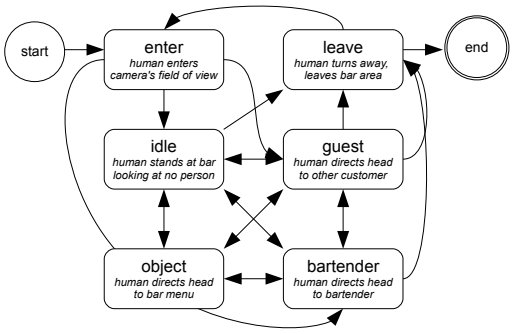

Figure 4: Hidden Markov model of the states of interaction.

Figure 4 shows the states of interaction and possible transitions in our hidden Markov model. The model reflects the state of a single human guest—in the case of multiple guests, each of them corresponds to a separate model. It was designed with two thoughts in mind: on the one hand, it has to distinguish between states that are necessary for the robot to recognise attention requests, on-going interactions and the closing of interactions. On the other hand, the number of states should be minimal to ensure an acceptable recognition performance.

4. EVALUATION

For evaluating our model of interaction states, we collected video recordings of more than 60 typical interactions with the robot bartender. The interactions included one, two or three human guests.

As head poses can by nature not be recognised at all times, we additionally collected torso poses given by a Kinect depth sensing device. With this, positions of people are even available when they are not facing the robot and we avoid the otherwise likely case of confusing people. To complete our set of observed variables, we added two fuzzy values f_1 and f_2 that respond to other guests being of the field-of-view of a person. The first fuzzy value f_1 ranges from 1, when another guest is located directly on the line-of-sight, down to 0, when no other guests are in the field-of-view. The second fuzzy value f_2 likewise responds to other people being in the field-of-view, but furthermore decreases with distance. Finally, our set of observed variables is a vector of $d = 14$ dimensions for each person, 6 for head pose, 6 for torso pose, and 2 fuzzy values for other people being in the field-of-view.

For model training and testing, all recorded data were manually labelled following the state definitions in **Figure 4**. Two thirds of the data were used for training, one third for testing.

After systematic exploration of possible model parameters, we chose an inner state count of $s = 5$ inner states for each state in **Figure 4** and an emission distribution of $k = 6$ Gaussians with diagonal covariance matrices for each inner state. With this model, we achieved 89.90% correctly recognised states of interaction with 8.21% false insertions, yielding an accuracy of 81.59% on the training set.

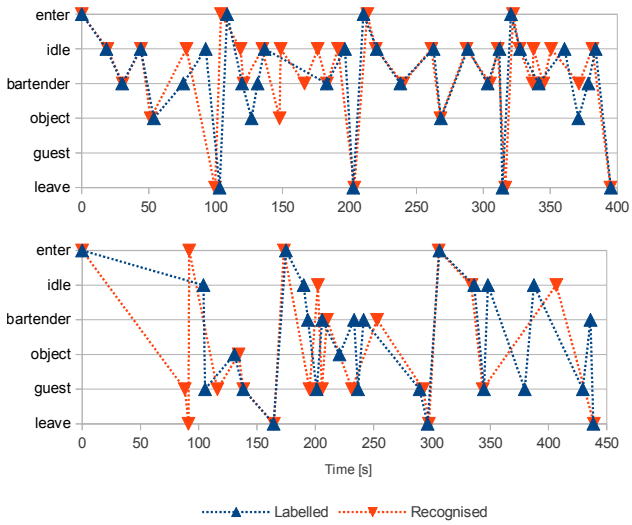

Figure 5: Sample recognition results in comparison to labelled states over time. Top view shows a single participant scenario, bottom view shows a participant within a group.

Figure 5 shows an exemplary comparison of recognised and labelled states of interaction in the testing set over time. The observation from this comparison is that, even though states are not always correctly recognised, labelled and recognised state change usually coincide in time, a systematic delay is not observed.

A complete overview on the testing results are shown in **Table 2**: All in all, 78.3% of the states of interaction could correctly be recognised. Even though we still observe a significant rate of 21.1% false insertions, the robot is able to identify the most important state changes in the interaction with human guests: 83.8% of all interactions with the bartender are detected, and together with 94.7% recognised idle states and 73.0% detected interactions among guests, the robot gains sufficient knowledge when to handle an interaction request or when to close an interaction.

5. CONCLUSIONS

This work yielded two main contributions: first, we took video recordings of human-human interactions in several bars to empirically research the use of head pose in an everyday situation. In the recorded sequences, human customers order drinks from a human bartender. The analysis of this data showed that humans follow a certain predefined sequence of interaction states when they order drinks: first, they request the bartender's attention; second, they place their order; third, they close the interaction. More importantly, we found that head pose is an important part of nonverbal communication that humans use in all of these states of interaction to express their intentions, first to ensure to have the attention of the bartender and to keep the interaction alive during the ordering process, then to signal that the interaction has come to an end.

The second contribution of this publication is that we used the findings from the human-human interaction recordings to design and implement a model of interaction states for

Table 2: Results and confusion matrix of the test data set.

Correctness	78.3%	H/N
Accuracy	57.2%	(H–I)/N
Correctly recognised states	141	H
Deletions	16	D
Substitutions	23	S
Insertions	38	I
Number of states	180	N

Recognised states	**Labelled states**							
	e	i	b	o	g	l	D	%Corr.
enter	20	0	0	0	0	0	0	100.0
idle	1	36	1	0	0	0	2	94.7
bartender	0	4	31	1	0	1	8	83.8
object	0	4	1	6	0	0	1	54.5
guest	3	7	0	0	27	0	5	73.0
leave	0	0	0	0	0	21	0	100.0
I	4	6	12	0	10	6		

a robot bartender. For the modelling, we used a hidden Markov model (HMM), which was trained with information from an automatic head pose estimation algorithm and additional body pose information. We proved in an evaluation, that the robot can recognise interaction states correctly in 78.3% of all test cases for all interaction states. More specifically, we achieved recognition rates of 83.8% for the crucial interaction state "attention to bartender" and 73.0% for "attention to other guest", allowing the robot to perform a socially appropriate interaction with multiple human guests.

6. ACKNOWLEDGEMENTS

This research was supported by the European Commission through the project JAMES[3] (FP7-270435-STREP).

7. REFERENCES

[1] J. B. Bavelas, L. Coates, and T. Johnson. Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, 52(3):566–580, 2002.

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[3] H. H. Clark. Pointing and placing. *Pointing: where language, culture, and cognition meet*, page 243, 2003.

[4] N. Gourier, D. Hall, and J. Crowley. Estimating face orientation from robust detection of salient facial structures. In *FG Net Workshop on Visual Observation of Deictic Gestures*, pages 1–9, 2004.

[5] T. Horprasert, Y. Yacoob, and L. Davis. Computing 3-d head orientation from a monocular image sequence. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 242–247. IEEE, 1996.

[6] K. Huth. Wie man ein Bier bestellt. Master's thesis, Universität Bielefeld, 2011.

[7] M. Jones and P. Viola. Fast multi-view face detection. *Mitsubishi Electric Research Lab TR-20003-96*, 2003.

[8] A. Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26(1):22–63, 1967.

[9] A. Kendon. Features of the structural analysis of human communicational behavior. *Aspects of Nonverbal Communication*, pages 29–43, 1980.

[10] S. R. Langton, R. J. Watt, and V. Bruce. Do the eyes have it? cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50–59, 2000.

[11] S. R. H. Langton, H. Honeyman, and E. Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Attention, Perception, & Psychophysics*, 66(5):752–771, 2004.

[12] J. Meynet, T. Arsan, J. Mota, and J. Thiran. Fast multiview face tracking with pose estimation. 2007.

[13] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626, 2009.

[14] M. Pateraki, H. Baltzakis, P. Kondaxakis, and P. Trahanias. Tracking of facial features to support human-robot interaction. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3755–3760. IEEE, 2009.

[15] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments*, 2009.

[16] 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.

[17] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *Neural Networks, 1993., IEEE International Conference on*, pages 586–591, 1993.

[18] E. Seemann, K. Nickel, and R. Stiefelhagen. Head pose estimation using stereo vision for human-robot interaction. In *Proceedings of the sixth IEEE international conference on automatic face and gesture recognition 2004*, pages 626–631. IEEE, 2004.

[19] H. Sloetjes and P. Wittenburg. Annotation by category: Elan and iso dcr. *Proceedings of the Sixth International Language Resources and Evaluation (LRECŠ08), Marrakech, Morocco, may*, 2008.

[20] R. Stiefelhagen. Tracking focus of attention in meetings. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces, 2002*, pages 273–280. IEEE, 2002.

[21] T. Vatahska, M. Bennewitz, and S. Behnke. Feature-based head pose estimation from images. In *Humanoid Robots, 2007 7th IEEE-RAS International Conference on*, pages 330–335. IEEE, 2007.

[22] P. Viola and M. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[3] `http://www.james-project.eu`