Realtime AAM based User Attention Estimation

Sebastian Hommel
Computer Science Institute
Hochschule Ruhr West
Germany, Bottrop 46240
E-Mail: Sebastian.Hommel@hs-ruhrwest.de
Telephone: +49 (0) 208 882 54 - 811
Fax: +49 (0) 208 882 54 - 834

Uwe Handmann
Computer Science Institute
Hochschule Ruhr West
Germany, Bottrop 46240
E-Mail: Uwe.Handmann@hs-ruhrwest.de
Telephone: +49 (0) 208 882 54 - 802
Fax: +49 (0) 208 882 54 - 834

***Abstract*—In this paper a method of automatic real-time capable visual user attention for a face to face human machine interaction is described. This method based on *Active Appearance Models* (AAMs) and *Multilayer Perceptrons* (MLPs) to map the *Active Appearance Parameters* (AAM-Parameters) onto the current head pose. Afterwards, the chronology of the head pose becomes classified to attention or inattention. This visual attention estimation will be used in service robotic by human-robotic interaction to get a feedback whether the user is interested in the current dialog and for correct interpretation of the current emotional condition. To allow a more natural dialog the head pose is also very efficient interpreted as head nodding or shaking by the use of adaptive statistical moments. Especially, the head movement of many demented people are restricted, so they often only use their eyes to look around. For that reason this paper examine a simple gaze estimation with the help of an ordinary webcam.**

***Index Terms*—Active Appearance Model, Head Gesture Recognition, Visual Attention, Gaze Estimation**

I. Introduction

Due to the growing occurrence of service robots more and more unexperienced and non-instructed users are getting in touch with service robots. Therefore a lot of effort has been spent on enabling a natural human-robot dialog in service robotics during the last years. Nowadays, service systems like shopping robots, ticket machines or entertainment electronics are established in our society. Furthermore, service systems are getting more and more important in home environment. This ranging from robotic animals for amusement through to service robots which help with housework, scheduling, home health care etc.. Especially for the acceptance of these systems they must be easy to use, since non-instructed users should be able to operate these systems. The most intuitive kind to interact with a technical system is the human like communication. An essential part in human like communication is to know whether the dialog partner is attention to the service system or to any other. This is necessary to adapt the dialog and for the correct interpretation of the current users condition like the emotion. There are many methods to estimate the user attention, like biometric parameters, full body movements or the used attention estimation which based on the direction of view. Trefflich [1] showed that the head movements have a strong correlation to eye movements, because the head movements are the low-pass filtered eye movements. So the 3D head pose is used for the attention estimation. To estimate the 3D head pose in a 2D image a head description is needed which model a robust 3D head. The Active Appearance Model approximation from Stricker [2] is used to realize this head model in real time.

Active Appearance Models (AAMs) have been established to characterize non-rigid objects, like human heads, and can be used to analyze the users state based on visual features. Therefore, the parameters of the AAM are adapted, so that the model fits to the current face. Due to the fact that the AAM represents the shape as well as the texture of the face, the whole appearance is captured by the model parameters. Afterwards, the parameters of the AAM can be utilized to extract information about the users state. The main advantage of using an AAM is the holistic representation of the face.

Especially in home health care for demented people it is often not possible to use the head pose for attention estimation, because many demented people are restricted in their head movements. This people mostly look around only by using eye movements. To allow a visual attention estimation in this case, a simple eye tracker is considered which operate with an ordinary webcam. This eye tracker based on the eye position determined by an AAM, too.

A further essential information for a gentler and more natural dialog, which can be estimated with the help of the head pose is a head nodding and shaking. Afterwards, the head nodding and shaking can be interpreted as Yes or No to allow a simple nonverbal gestural answering.

To classify the chronology of the head poses to attention or inattention an adaptive variance is used in this work. To classify the chronology of the head poses to head nodding, shaking or others an adaptive excess kurtosis is used.

II. Related Work

The work comprises the tasks of extraction of the visual focus, and head gesture classification. All of these tasks require information about the users head. A wide range of different methods can be found in the literature using different kinds of feature extraction and classification approaches. In the following, we give a brief overview of different methods, which have been applied for the specific tasks.

1) Visual Focus of Attention: Basically, a persons visual focus is determined by eye gaze. However, the proposed

978-1-4577-1974-5/11/$26.00 ©2011 IEEE

systems in the literature require high resolution of the eyes [3]. Nevertheless, the head pose can be regarded as a low pass filtered eye gaze and therefore the head pose can be utilized to get information about the visual focus [4]. Hidden Markov Models are a very common way to extract the focus of attention from a sequence of head poses [4], [5], [6]. However, the mentioned methods try to extract a focus of attention in terms of certain objects or persons, which lies not in the scope of this paper.

2) Head Gesture Recognition: Known approaches for head gesture recognition are quite similar to visual attention estimation. Again, the head pose is extracted and evaluated over time. A common way to enable the time based evaluation is to apply Neural Networks as shown in [7]. Furthermore, SVM Classification as utilized in [8] or Hidden Markov Models [9] can be applied for head gesture classification. Nevertheless, the proposed methods are also not able to learn head gestures online and therefore are not appropriate to learn new head gesture semantics during the human-robot dialog.

III. System Architecture

The system in this paper is based on the parameters of a fitted AAM. The system interpret the AAM-Parameters as head pose and analyzes them for attention and gesture estimation. For some applications like the emotion estimation a very detailed but only frontal face model is needed. Due to their complexity the model does not fit enough to different head poses. However, for attention estimation and detection of head shaking or nodding a model is needed that fits good to a rotated face. This is possible by using model that is not detailed in shape. In attention estimation and head gesture classification, the variance and the excess kurtosis of the extracted head poses are calculated. Whereas, the variance is used for attention estimation and the excess kurtosis is used for head gesture classification. An overview of this architecture is shown in Fig. 1.

A. Head Pose Estimation

Both, the attention and the gestures Yes and No can be determined by head poses. Trefflich shows that the visual attention correlates with the head pose [1]. The first part in the estimator is to extract the head pose. Afterwards, the attention and head gestures are estimated by using statistics. Experiments have shown that some AAM-parameters correlate with the head pose once the training dates contain head poses. For head pose estimation an own dataset is used which consist in mixed facial image sequences of male and female people who rotate their heads around. Each image of this dataset is labeled by the current head pose which is determined by the so called *Flock of Birds*. The *Flock of Birds* is a two parted system which determined the head pose by using magnetic fields. The one part is fixed and must be positioned near by the camera, the other part must be mounted on the top of the head. This system must be calibrated for each user to get correct values. Few samples of this dataset is shown in Fig. 2 .

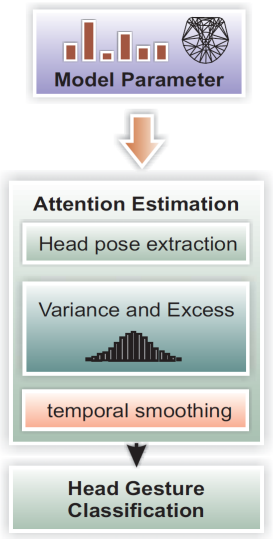


Fig. 1. **System Architecture** The attention estimation is done by extracting the head pose from AAM-parameters. The poses are aggregated over time to compute the variance and excess kurtosis. The variance is used to compute the users attention and the excess kurtosis is applied for head gesture recognition. Temporal smoothing is applied for both subsystems to reduce input parameters noise.

Fig. 2. **Head Pose Dataset** This is an example of the used dataset for the head pose estimation.

The correlation between head poses and AAM-parameters can not be generalized, thereby it is necessary to test each AAM. In the used AAM-approximation the first four shape parameters are generated by a face detector. This so called *global shape parameters* describe only the position of the face in the image. The other shape parameters which describe the individual form are so called *local shape parameters*. Fig. 3 shows the correlation between the first ten possible local shape parameters and the head rotation. By reason of this correlation it is possible to use an AAM with only the first two local shape parameters.

Through this, the used model includes 50 texture parameters but only six shape parameters, whereof two shape parameters describe the head rotation, one for horizontal and one for vertical head rotation. To generate a similar model a method

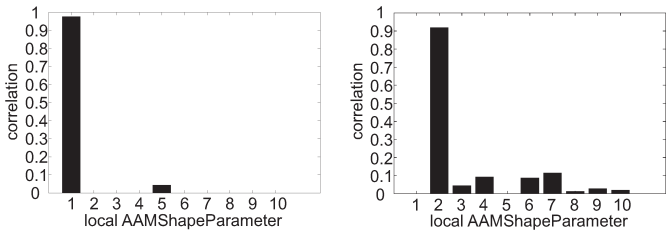


Fig. 3. **Correlation between the local AAM-Parameters and the Head Poses** The left diagram shows the correlation with the horizontal head rotation and right diagram shows the correlation with the vertical head rotation.

to generate an Active Appearance Model with two separate example sets are favorable [2]. In the used method the texture is learned first. After that, the shape could be learned by separate examples. For learning the texture the FG-Net dataset [10] is used. An own dataset of one person who move the head only horizontal and later only vertical is used for learning the shape parameters. To map the two shape parameters which describe the head pose onto the correlated real world head pose, a simple *Multilayer Perceptron* is used.

B. Attention Estimation

To analyze the head poses, statistics are used since there allow a direct attention and gesture estimation without training data. For efficient calculation of these statistics, an *adaptive recursive method* is used, which was developed by GRIESSBACH [11]. All used *adaptive recursive methods* employ a constant weight c which range from 0 to 1. The variance Z^2 (Fig. 4) of the head poses are used to get a classification onto attentive or inattentive.

$$M_0 = m_0 \tag{1}$$

$$M_{t+1} = M_t + c_M \cdot (X_{t+1} - M_t) \tag{2}$$

$$Z_0^2 = z_0 \tag{3}$$

$$Z_{t+1}^2 = Z_t^2 + c_{Z^2} \cdot \left((X_{t+1} - M_{t+1})^2 - Z_t^2 \right) \tag{4}$$

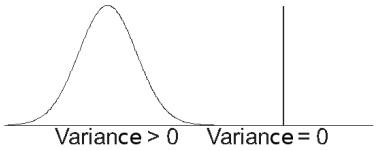


Fig. 4. **Variance** When the variance is greater than $40°$, the current part of the sequence is classified as inattention and otherwise as attention.

A sequence of head poses are classified as attentive once the variance is lower as a threshold value of 40 degree. The aim in this work is to get a continuous value for the attention A_t. For this an adaptive recursive mean is updated by 100 once the person is attentive and with 0 otherwise.

$$A_0 = 100 \tag{5}$$

$$A_t = A_{t-1} + c_A \cdot (100 - A_{t-1}) \quad \text{, if attention} \tag{6}$$

$$A_t = A_{t-1} + c_A \cdot (0 - A_{t-1}) \quad \text{, if inattention} \tag{7}$$

When a person is attentive mostly it is in interest where the people look too. This point of interest I can be estimated by the adaptive mean of the head poses p.

$$I_t = I_{t-1} + c_I \cdot (p_t - I_{t-1}) \tag{8}$$

C. Head Gesture Estimation

Furthermore, the developed estimator is able to detect a head shake and nodding from a face image sequence. Afterwards, a context-sensitive interpretation as Yes or No is possible. When a person shake their head, only few changes of the head pose in the vertical are generated, but lots of changes in the horizontal are expected. Thereby, the excess kurtosis of the horizontal head poses is platykurtic and the excess kurtosis of the vertical head poses is leptokurtic. The effect is reverse by nodding. Since this condition is unequivocal it can be used for detection. (Fig. 5) The adaptive excess kurtosis ϵ can also calculated by a method of GRIESSBACH [11].

$$Z_0^4 = z_0 \tag{9}$$

$$Z_{t+1}^4 = Z_t^4 + c \cdot \left((X_{t+1} - M_{n+1})^4 - Z_t^4 \right) \tag{10}$$

$$\epsilon = \frac{Z_n^4}{(Z_n^2)^2} - 3 \tag{11}$$

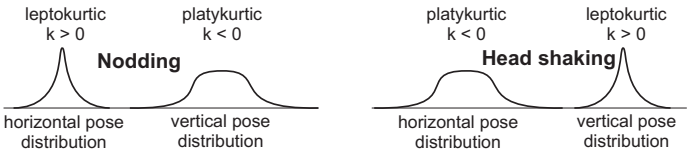


Fig. 5. **Head Gesture** Performing a nodding gesture involves changes in the vertical head pose. Therefore, the vertical head pose distribution becomes platykurtic while the horizontal distribution is leptokurtic. Head shaking involves changes of the horizontal head pose. Therefore, the distribution characteristics are inverted.

D. Eye Tracker

To allow an attention estimation for demented people the eye gaze is needed, since their are mostly unable to move the head fully. So this people looks around only with their eyes and without head movements. To estimate the eye gaze a ordinary webcam is used, too. First, the AAM affords the eye position, so it is possible to focus only to this image parts. A gray level eye consist in a white plane a darker iris and the black pupil, so the pupil becomes detect as the darkest point in the eye. To establish the eye gaze it is necessary to know the possible eye movements. Speckmann and Hescheler reported in [12] that the healthy eye is able to move $20°$ to the left and $20°$ to the right. This is an interest area of $40°$ onto the curvature of the eye. Up to this value it is possible to assume a linear

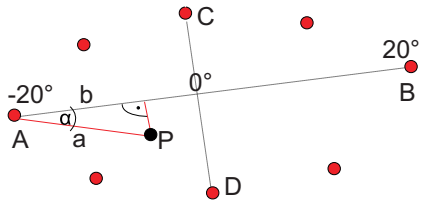


Fig. 6. **Eye Tracker** Triangulation to calculate the pupil position in horizontal and vertical for estimate the eye gaze

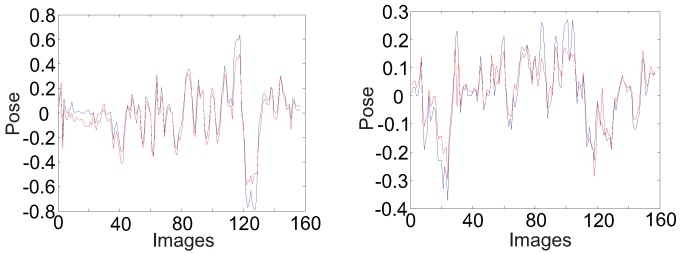


Fig. 7. **MLP-Output vs. "Flock of Birds"** The blue curve represent the head poses which are determined with the help of the "Flock of Birds" and the red curve represent the head poses which are estimated by the presented system. *left:* horizontal head poses; *right:* vertical head poses

correlation between the pupil position into the 2D image and the pupil position onto real world eye. The same assumption is used for the vertical eye movement. Thereby, it is possible to estimate the line of gaze for each pupil position by the use of a triangulation. This is exemplary shown in Fig. 6. To estimate the global eye gaze it is necessary to add the head pose, too.

IV. Experimental Results

This section presents experimental results achieved by using the described approach. Furthermore, a number of test sequences are recorded to evaluate the proposed attention measuring system, the head gesture recognition and the eye tracker.

A. Head Pose Estimation

To evaluate the presented AAM and MLP based head pose estimation 8 further sequences are recorded whereby the people can look around. By recording this sequences the head pose which is determined by the *Flock of Birds* is simultaneous recorded. Than the head poses which are estimated with the help of the presented system is compared to the head poses of the *Flock of Birds*. Thereby, the RMS for vertical pose is $0.1387°$ and the RMS for horizontal is $0.1546°$. This result is shown in Fig. 7.

B. Attention Estimation

To evaluate the proposed attention measure under real world conditions, 23 test sequences from 8 different persons became recorded. The people were asked to watch several video clips in front of a computer monitor, while they were monitored with the help of a frontal camera. During the first stage, the person were watching an exciting movie. In the second stage, the same persons were watching a boring video, while another person enters the room and tries to distract the test persons by talking to them or letting things falling on the ground (Fig. 8).

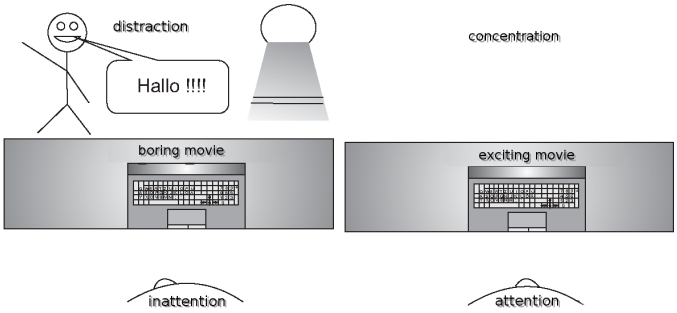


Fig. 8. **Attention Estimation Experiment** This sketch shows the used experiment structure. *left:* the test person looks a boring video; *right:* the test person looks an exiting video

TABLE I
Attention Estimation This table shows the rate of attention for each evaluation sequence which is determined by the presented system and by human interpretation.

System	Human	Difference
23.6%	0%	23.6%
16.8%	15.5%	1.3%
3.1%	0%	3.1%
1.1%	2.3%	1.2%
8.5%	12.9%	4.4%
4.9%	6.7%	1.8%
0%	0%	0%
0%	0%	0%
25.5%	7.5%	18%
0%	0.7%	0.7%
0.2%	0%	0.2%
4.3%	0%	4.3%
1.1%	0%	1.1%
0%	0%	0%
8.6%	0%	8.6%
14.8%	7.8%	7%
31.6%	26.9%	4.7%
0%	0%	0%
0%	0%	0%
27.1%	20.6%	6.5%
13.4%	15.6%	2.2%
28.4%	27.1%	1.3%
69.6%	68.3%	1.3%

Afterwards, the recorded image sequences were labeled manually by several people in terms of visual attention. For evaluation purposes the same sequences were presented to the proposed attention system. The average divergence between the presented system and the labels is only 4 percentage points which is shown in Table I.

The described system only fails massively by three image sequences. By visual analyzing this sequences it is prominent that the high divergence of these sequences is generally caused by a bad model fitting leading the head direction estimation to fail.

Before the variance for the attention estimation is used, the excess kurtosis was also tried to use, since the excess kurtosis have no scale unit. However, the use of the excess kurtosis failed because it is zero when the head pose histogram is

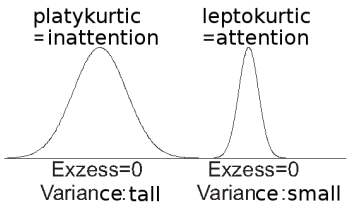
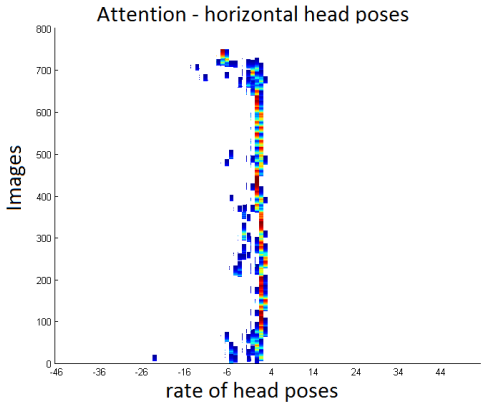
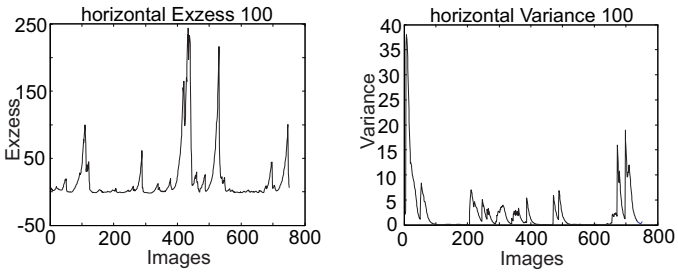


Fig. 9. **Excess Kurtosis vs. Variance** A wide and a small histogram can be Gaussian distribution, so the excess kurtosis can be 0, once the variance become tall when the user is inattentive.

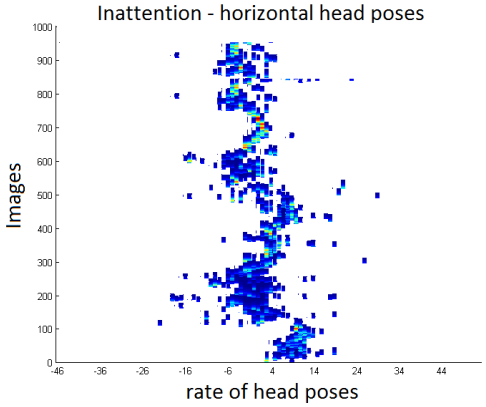


(a) **Horizontal Head Poses** increasing rate from blue to red

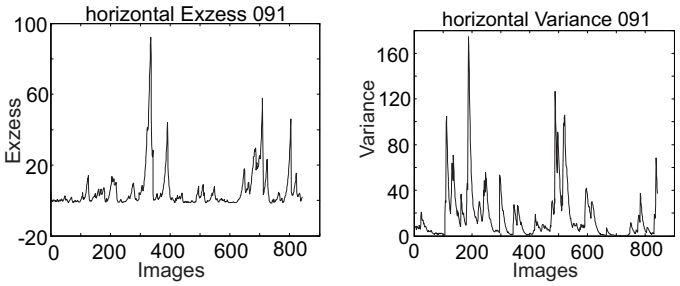


(b) **Exzess Kurtosis** (c) **Variance**

Fig. 10. **Excess Kurtosis vs. Variance of an attention sequence**

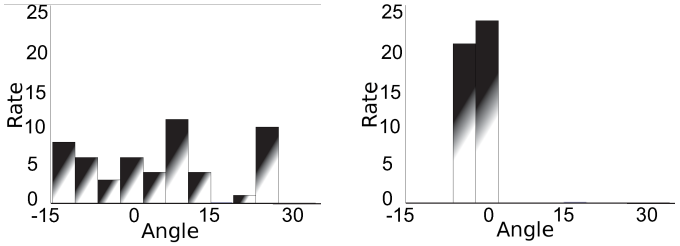


(a) **Horizontal Head Poses** increasing rate from blue to red



(b) **Exzess Kurtosis** (c) **Variance**

Fig. 11. **Excess Kurtosis vs. Variance of an inattention sequence**



(a) **horizontal head pose** (b) **horizontal eye gaze**

Fig. 12. **Histogram of horizontal head pose vs. eye gaze**

normally distributed and this is able during attention and during inattention (Fig. 9).

Fig. 10 and Fig. 11 shows exemplary the head poses, the excess kurtosis and the variance of an attention and a inattention sequence.

C. Head Gesture Estimation

To evaluate the Head Gesture Estimation, 4 sequences from several people, which nod and shake the head between several other head movements, became recorded. On this experiment each head nodding and shaking could be detected by no false positive detection. Already a single fully Yes or No head gesture could be detected, where the system reaction is shortly delayed, once the used adaptive calculation.

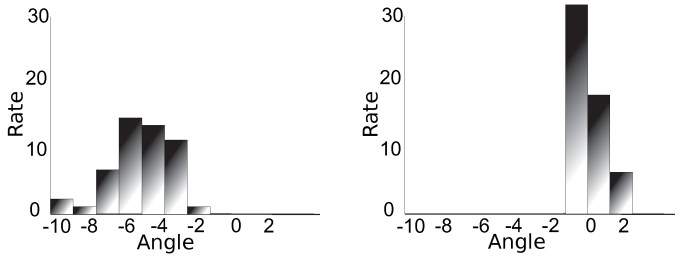
D. Eye Tracker

The simple eye tracker is only evaluated by three sequences. In this sequences the people look straight ahead onto the camera and move only their heads, so the eye gaze should roughly 0° where the head pose differ. In Fig. 12 and Fig. 13 the vertical and horizontal eye gazes and head poses is exemplary visualized for one sequence.

Furthermore, the Table II shows the variance of the head poses and the eye gazes of the three sequences. This results shows that it is principal possible to estimate the eye gaze by this simple method with a small error. This estimation should be exact enough to use it in an attention estimation.

V. Conclusion

In this paper a way is presented to extract user attention and head gestures utilizing the shape and texture parameters from a fitted Active Appearance Model. This paper focus on improving the human-robot interaction and therefore applies an attention and head gesture estimation, which use the AAM



(a) **vertical head pose** (b) **vertical eye gaze**

Fig. 13. **Histogram of vertical head pose vs. eye gaze**

TABLE II
VARIANCE HEAD POSE VS. EYE GAZE THIS TABLE SHOWS THE VARIANCE OF THE VERTICAL AND HORIZONTAL HEAD POSES AND EYE GAZES OF 3 TEST SEQUENCES FROM SEVERAL PERSON.

Person	hor. head pose	hor. eye gaze	ver. head pose	ver. eye gaze
1	71.4	4.3	15.6	1.8
2	39.5	6.7	2.0	0.8
3	176.8	3.0	2.8	0.6

shape parameters to estimate the users head pose. For the measure of attention the distribution of the head pose over time are used. By comparison the hand labeled attention values with the system output, the presented system seems to be able to estimate the attention value quite well. In addition, a head gesture recognition based on the temporal event mapping approach became proposed. Continuing this work, it is possible to integrate the proposed system into a dialog system [13]. This will be very helpful to examine how the proposed attention values can be utilized to enable a more natural human-robot interaction. Furthermore, the possibility to track the human eyes with the help of an ordinary webcam with a small estimation error is shown.

REFERENCES

[1] B. Trefflich, "Videogestützte Überwachung der Fahreraufmerksamkeit und Adaption von Faherassistenzsystemen." Ph.D. dissertation, Technische Universität Ilmenau, 2009.

[2] R. Stricker, C. Martin, and H.-M. Gross, "Increasing the robustness of 2d active appearance models for real-world applications," in *Proceedings of the 7th International Conference on Computer Vision Systems: Computer Vision Systems*, ser. ICVS '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 364–373. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04667-4_37

[3] P. Smith, S. Member, M. Shah, and N. D. V. Lobo, "Determining Driver Visual Attention with One Camera," *IEEE Trans. on Intelligent Transportation Systems*, vol. 4, p. 2003, 2003.

[4] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel, "From Gaze to Focus of Attention," 1998.

[5] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez, "Tracking the Visual Focus of Attention for a Varying Number of Wandering People," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1212–1229, 2008.

[6] S. Ba and J. Odobez, "Recognizing Visual Focus of Attention from Head Pose in Natural Meetings," 2009.

[7] H. N. L.M. King and P. Taylor, "Hands-free Head-movement Gesture Recognition using Artificial Neural Networks and the Magnified Gradient Function," *IEE Conf. ot the Engineering in Medicine and Biology Society*, pp. 2063–2066, 2005.

[8] L.-P. Morency and T. Darrell, "Head gesture recognition in intelligent interfaces: the role of context in improving recognition," in *IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces*. New York, NY, USA: ACM, 2006, pp. 32–38.

[9] P. Lu, M. Zhang, X. Zhu, and Y. Wang, "Head nod and shake recognition based on multi-view model and hidden markov model," in *Computer Graphics, Imaging and Vision: New Trends, 2005. International Conference on*, july 2005, pp. 61 – 64.

[10] F. Wallhoff, "Facial expressions and emotion database http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html," *Technische Universität München*, 2006.

[11] G. Grießbach, "Weiterentwicklung und Anwendung komplexer adaptiver Schätzalgorithmen in der Biosignalanalyse, der Bildverarbeitung und der Klassifikation zur EBG-Analyse kognitiver Prozesse," *DFG-Antrag Gr1 55511-2*, 1998.

[12] E.-J. Speckmann and R. K. Jrgen Hescheler, *Repetitorium Physiologie*, 2nd ed. Urban & Fischer Verlag, 2008, ISBN 978-3-437-42321-5.

[13] C. S. S. Müller and H.-M. Gross, "Aspects of user specific dialog adaptation for an autonomous robot," in *IWK*, 2010.