in: Proc. 55th Int. Scientific Colloquium, Ilmenau, Germany, 2010, pp. 629-634, Verlag ISLE 2010

REALTIME USER ATTENTION AND EMOTION ESTIMATION ON A MOBILE ROBOT

Ronny Stricker[*], *Sebastian Hommel*[*], *Christian Martin*[†] *and Horst-Michael Gross*[*]

[*]Neuroinformatics and Cognitive Robotics Lab
Ilmenau University of Technology
98684 Ilmenau, Germany

[†]MetraLabs GmbH
98693 Ilmenau, Germany

ABSTRACT

Within the scope of service robotics a natural and adaptive dialog is getting more and more important to offer intuitive interaction for users not familiar with the system. Therefore, service robots have to be aware of the person's mood and visual attention. Furthermore, robotic systems should be able to recognize simple head gestures like head shaking or nodding as these gestures are a natural way of human communication. This paper presents a method to extract these information from image sequences containing the head of the interaction partner with the help of Active Appearance Models (AAMs). Therefore, a variant of AAM robust to illumination is fitted to a sequence of face images to get a parametric description of the user's face. The paper shows how to apply these parameters for human state estimation in terms of attention and emotion positivity. Furthermore, we show how to utilize *Temporal Event Mapping* with the AAM parameters to learn head gestures and head gesture semantics online during the human-robot dialog.

Index Terms— Active Appearance Model, Gesture Recognition, Visual Attention, Human Emotion Recognition

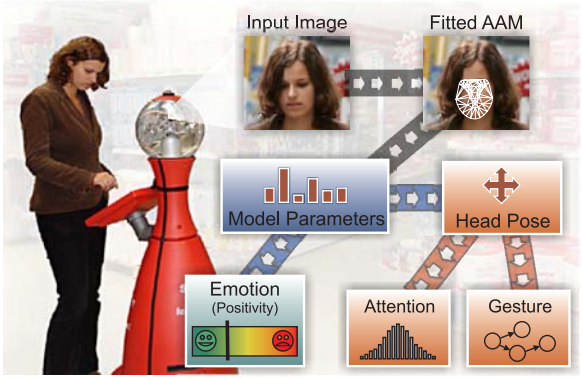

Fig. 1. *System Overview. First, an AAM is fitted to the input image containing the user's face. Based upon the obtained model parameters the emotion estimation module tries to estimate a linear user positivity value. The attention and gesture recognition model extracts the head pose from the model parameters and estimates the user attention and head gesture.*

1. INTRODUCTION

Due to the growing occurrence of assistive robots more and more inexperienced and non-instructed users are getting in touch with such robots. Hence, a lot of effort has been spent on enabling a natural human-robot dialog in service and assistive robotics during the last years.

Using visual information about the user can make a contribution in this context by means of emotion classification, head gesture interpretation, and extraction of the user's visual focus of attention.

One exemplary application, where the human-robot interaction can be significantly improved with the help of visual features is the touch screen based dialog for companion robots. In such situations it is very important for the system to know if the user is attentive and can follow the instructions or information displayed on the screen. This becomes even more important if the robot should encourage the user to do some specific task he might not really want to do. For example, home companion robots for people suffering from cognitive impairment should activate the user to do some cognitive training to delay the impact of their disease [1]. In that case, it is very important to have information about the user's visual focus to derive a measure of attention. Furthermore, the system should be able to recognize whether the user is getting bored or happy by doing the exercises. Therefore, some kind of emotion scale is necessary to measure the positivity of the user's emotion over time. The interactivity of the dialog could be further improved by supporting some basic kinds of head gestures like head nodding or head shaking.

Active Appearance Models (AAMs) [2] have been established to characterize non-rigid objects, like human heads, and can be used to analyze the user's state based on visual features. The AAM is fitted to an input image with the users face to give an exact match in terms of texture and shape by adjusting the model parameters. Afterwards, the parameters of the AAM can be utilized to extract information about the user's state (Fig. 1). The main advantage of using an AAM is the holistic representation of the face.

This work suggests a holistic framework to extract the user's attention as well as the user's emotion and to learn head gestures and head gesture semantics online based on an AAM. The paper is organized as follows: After an overview of the related work, Sect. 3 gives a brief description of the basics of AAMs. Sect. 4 introduces the holistic evaluation system. Sect. 5 shows the results which have been achieved with the proposed methods. The paper concludes with a summary and an outlook to ongoing work in Sect. 6.

This work is partially funded from the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216487. (CompanionAble project)

2. RELATED WORK

The work comprises the tasks of emotion estimation, extraction of the visual focus, and head gesture classification. All of these tasks require information about the user's 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.

2.0.1. Emotion Recognition

Different authors focused on examining various methods to classify emotions using AAMs. The different classifiers used for that task range from Support Vector Machines (SVMs) [3] and Neural Networks [4] to Hidden Markov Models (HMM) [5]. The similarity of these approaches is that they try to map the user's emotion into discrete basis emotions. Using the discrete emotion classes is not the only way to explain human emotions. Breazeal and Scassellati show in [6] that emotions can also be arranged in a tree dimensional continuous emotion space. Whereas one dimension of the continuous emotion space is the positivity of the emotion. However, to the best of our knowledge there is no approach which tries to map the emotion space onto a linear positivity scale with the help of AAMs.

2.0.2. Visual Focus of Attention

Basically, a person's visual focus is determined by eye gaze. However, the proposed systems in the literature require high resolution of the eyes [7]. However, the head pose can be regarded as a low pass filtered eye gaze and therefore, can be utilized to get information about the visual focus [8]. Hidden Markov Models are a very common way to extract the focus of attention from a sequence of head poses [8, 9, 10]. 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.

2.0.3. Head Gesture Recognition

Known approaches for head gesture recognition are quite similar to that of 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 [11]. Furthermore, SVM Classification as utilized in [12] or Hidden Markov Models [13] can be applied for head gesture classification. Nevertheless, the proposed methods are not able to learn head gestures online and therefore are not appropriate to learn new head gesture semantics during the human-robot dialog.

3. BASICS OF ACTIVE APPEARANCE MODELS

Active Appearance Models, first introduced in [2], provide a good way to model non-rigid objects within the scope of image processing and are, therefore, very popular to model human faces or viscera. The AAM itself is a combination of two statistical models. First, the shape model represents the geometry of the object. Secondly, the appearance model allows the modeling of the object texture within the normalized mean shape of the model. The models are built by training images, which are labeled with landmark points on certain positions of the object. These n landmark locations build up the shape $\mathbf{s} = (x_1, y_1, ..., x_n, y_n)^T$ of an AAM instance. Using a Principle Component Analysis (PCA) for all training shapes, the resulting shape model can be represented by a set of shape parameters $\mathbf{p}$ combined with the basis shapes $\mathbf{s_i}$:

$$\mathbf{s}(\mathbf{p}) = \mathbf{s_0} + \sum_{i=1}^{n} p_i \mathbf{s_i}. \quad (1)$$

Afterwards, a triangulation of the mean shape $\mathbf{s_0}$ is used to establish a relation between the labeled points and the surface of the object. With the help of surface triangles, every single point on arbitrary shape $\mathbf{s_i}$ can be warped to a destination shape $\mathbf{s_j}$ using an affine transformation. As shown in [14] we can describe this transformation as $W(\mathbf{x}; \mathbf{p})$, which maps a point $\mathbf{x} = (x, y)^T$ within the model shape to the shape defined by the parameters $\mathbf{p}$. This transformation is used afterwards to build the appearance model, which is very similar to the shape model. The important difference is that each texture sample A_i, defined by the training images, is warped to the mean shape $\mathbf{s_0}$, using the described affine transformation. The texture parameters resulting from the subsequent PCA are denoted as λ. Therefore the texture object is very similar to the *Eigenface* approach:

$$\mathbf{A}(\lambda) = \mathbf{A_0} + \sum_{i=1}^{m} \lambda_i \mathbf{A_i} \ \ , \forall \mathbf{x} \in \mathbf{s_0}. \quad (2)$$

The resulting AAM can represent any object instance M covered by the training data using the shape parameter vector $\mathbf{p}$ and the appearance parameter vector λ using (3).

$$M(W(\mathbf{x}, \mathbf{p})) = \mathbf{A_0}(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \mathbf{A_i}(\mathbf{x}) \ \ , \forall \mathbf{x} \in \mathbf{s_0}. \quad (3)$$

The goal of fitting an AAM to an unknown image, as defined by [2], is to minimize the squared difference between the synthesized model and the given image. Using gradient descent to solve this problem leads to a very efficient fitting algorithm. To overcome the problem of simultaneous optimization of shape- and appearance parameters, Baker and Matthews introduced the *Project-Out* gradient descent image alignment algorithm [15]. As the exact formulation of the fitting algorithm lies beyond the scope of this paper, the reader is referred to [14, 15] for more detailed information.

Increasing AAM Robustness Due to the principle of minimizing the difference between the input image and the synthesized model, the fitting process is very sensitive to differences between the training images and the images used during model fitting [16]. Such changes are caused by varying illumination, induced by head movement or changing lighting conditions. Furthermore, as a result of the local optimization characteristics of AAMs the fitting process is quite sensitive to getting stuck in local minima and, therefore, may cause a bad match. As we have already shown in [17] the fitting performance of AAMs under real world conditions can be significantly improved with the help of the *adaptive retinex filter* and the *adaptive parameter fitting*. The detailed description of both approaches lies beyond the scope of this paper, so the user is referred to [17] for detailed information.

4. SYSTEM ARCHITECTURE

The fitted AAM provides a parameterized version of the user's face in terms of texture and shape, it is a solid base to build

the user state estimation system. Hence, the system can be divided into three basic subsystems induced by the different user state features (Fig. 3).

4.1. Emotion Positivity Estimation

The emotion classification tries to map the user's visual emotional state into a 1D emotion space coding the emotion positivity. Unfortunately, the common facial emotion databases are labeled in terms of the seven common emotion classes (neutral, happy, sad, disgust, surprise, fear, anger). Nevertheless, according to [6], these basis emotions can be mapped onto a linear positivity scale to form a one dimensional emotion space (Fig. 2). The task of the positivity estimation module is to learn a generalized relationship between the AAM parameters $\mathbf{f} = (\mathbf{p}, \lambda)$ and the corresponding positivity value v_p of the one dimensional emotion space. As the desired mapping is only defined for a sparsely distributed set of samples points the function approximator should provide good generalization abilities. Therefore we use the method of epsilon Support Vector Regression provided by the LibSVM [18] software library for function approximation.

Although the proposed system should be able to provide the desired function approximation there are still two problems which have to be considered in advance. The first problem addresses the dimension of the feature vector $\mathbf{f}$ used for regression. Due to the high number of model parameters, needed for synthesis of a wide range of different people, the feature vector has a typical dimension between 50 and 100. The training samples on the other side depend upon the used database and are typically limited to a few hundred samples. To reduce the problems arising from that high dimensional feature vector we apply a weighting of the single feature dimensions. Thus, we have computed the correlation values cv of every dimension i with the positivity values to obtain the weighted parameter set $\hat{\mathbf{f}}_\mathbf{i}$:

$$\hat{\mathbf{f}}_\mathbf{i} = \mathbf{f}_\mathbf{i} \cdot cv(\mathbf{f}_\mathbf{i}) \tag{4}$$

The second problem, the system has to face, are the different characteristics and strength of emotions of different people. Several experiments have shown that a single frame based evaluation is likely to fail for certain people if the neutral face of a specific person is quite different from the mean neutral face. This, for example, is the case if a person is always looking a bit sad or bored by pulling down the corners of the mouth even when expressing a neutral emotion. In this case, the classification fails for almost all different emotions. One way to cope with this problem is to integrate temporal information into the classification process. In that case, the model can be adapted to that specific person by adjusting the base line for the extracted model parameters. As a result, the parameters used for regression $\hat{\mathbf{f}}_\mathbf{r}$ are coded as a difference between the current weighted model parameters $\hat{\mathbf{f}}_\mathbf{i}$ of the fitted model and the determined neutral parameters $\hat{\mathbf{f}}_\mathbf{n}$ of the specific person. Obviously, the proposed method relies

Fig. 2. *Mapping from the emotion space with 7 emotions to a linear positivity space.*

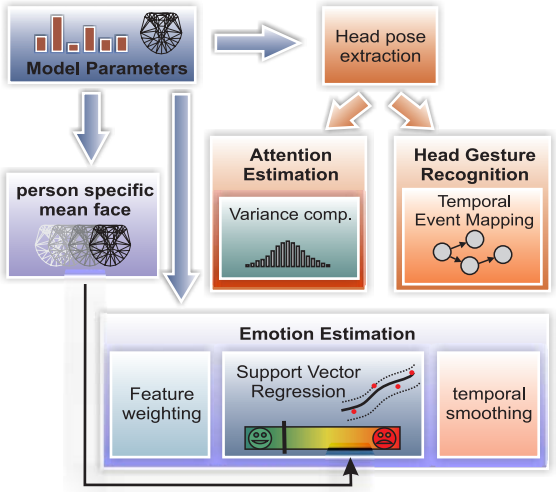

Fig. 3. *System Architecture. The 1D positivity measure is realized by applying SVR. Therefore, the AAM parameters are scaled and the person specific mean is removed. The attention estimation and head gesture recognition rely on the extracted head pose. For the attention estimation the head poses are aggregated over time to compute the variance and derive an attention measure. The head gestures in turn are recognized by applying the Temporal Event Mapping [19] with clustered head poses. Temporal smoothing is applied for all subsystems to reduce input parameter noise.*

on the known neutral reference parameters $\hat{\mathbf{f}}_\mathbf{n}$ for every user. Unfortunately, gathering these reference can be quite difficult under real world conditions, as the users initial emotion is not necessary a neutral one. To overcome this limitation, we try to estimate the user's neutral face over time. Therefore we assume, that a neutral mean face of every person can be estimated as the median of the person's model parameters over different time scales (short term or long term depending on the expected dialog time).

Furthermore, a person identification by means of a nearest neighbor classification of the model parameters is applied to enable the system to deal with several persons. The mean face parameters are acquired and stored in a database separately for every single person for later usage (Fig. 3).

4.2. Attention and Gesture Estimation

As already mentioned in Sect. 2, the head pose can be regarded as a good indicator for the visual focus of attention of the user. As the AAM parameters contain information about the shape of the user's head these parameters can also be applied to extract information about the user's head pose.

4.2.1. Estimation of Head Pose from AAM Parameters

The shape parameters of an AAM depend on the images used during the training stage. As a result, the shape parameters will also contain head movements like panning and tilting, if these movements are part of the training set. Unfortunately, due to the nature of PCA, which is applied during the model creation stage, these specific head movements are not naturally split into separate shape parameters. However, the separation of the panning and tilting movements of the head can be forced in two different ways. First, the movements can be created synthetically as already shown in [14] for global shape parameters like scaling and rotation. The drawback of this method is that some information about the 3D shape of

the head is required in advance. The second approach consists of a careful design of the training data set, which can be achieved by adding rotated faces of up to 45 degree to the training database. In this way, the variance of the panning and tilting movements will exceed the variance of the other shape parameters and therefore are split into different shape parameters. In the reverse direction the values of these two parameters can be used to obtain information about the head panning and tiling when the AAM is fitted to the user's face. Own experiments have shown that using this kind of model construction can lead to correlation values of 0.93 to 0.98 between the mentioned shape parameter and ground truth panning and tilting of the user's face.

4.2.2. Measuring Visual Attention

Extracting the user's head pose over time leads to a good measure of the user's visual focus. The visual focus in turn can be regarded as a good measure of the visual attention of the user. Hence, the visual attention can be derived from the statistics of the head movement over time. If a user focuses on one specific region, the head movement over time is minimal. Consequently, the visual attention focus can be extracted using the mean values for the horizontal and vertical head angles of the last t head poses. Furthermore, computing the variance of the last t head poses leads to a measure of the straightness of the user's visual focus. By combining the mean head pose and the head pose variance a measure of visual focus and visual attention can be extracted. If the head pose variance is low, then the user is focused on one specific point which can be obtained from the mean head pose values. If the variance, in terms, is high, then the user is not focused on one specific point and therefore can be regarded as inattentive for several fields of applications. The specific boundary value which has to be defined for the variance to distinguish attentive from inattentive head pose sequences can be experimentally obtained by evaluating the minimum and maximum variance values of attentive head pose sequences.

4.2.3. Head Gesture Recognition

The recognition of arbitrary head gestures involves the detection of a defined sequence of head poses. Common head gesture recognition approaches apply HMM [13], SVM [12] or Neural Networks [11] for sequence classification part. However, all of these classifiers have to be trained offline. Furthermore, the representation of time, which seems to be crucial for gesture detection, cannot be naturally modeled.

In [19] Schill and Zimmer introduced the *Temporal Event Mapping* approach, which is quite similar to the well known HMMs but has an explicit representation of time and no separation into training and run-time phase due to lifelong learning. Hence, the approach models system observations and transitions between these observations as a connected observations graph whereas the transitions are annotated with transitions durations. Whenever a new system state is observed, energies are generated for corresponding observation receptors R. These energies are passed to the connected receptors as the system observation changes. If the expected observation and transition duration (coded by the connected receptors) matches the new system observation the corresponding energy is boosted or damped otherwise. Therefore the energy value is increasing as long as the observation sequence matches the expected observations defined by the temporal event map. An observation sequence is recognized if the energy exceeds a defined recognition value (Fig. 4). The lifelong and online learning is guaranteed as new transitions and systems states are generated if the current observations cannot be recognized by the temporal event map.

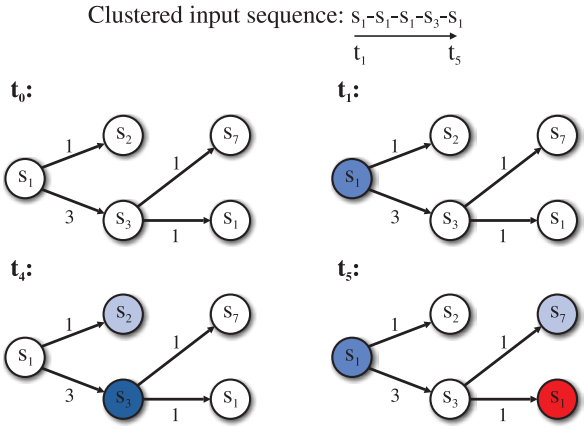

Fig. 4. *Energy levels in the Temporal Event Map after different time steps of an observation sequence (disregarding the creation of new receptors). Energies are coded using different colors (white - no energy, dark blue - high energy, red - energy above recognition boundary). The edges are labeled with the duration of the state changes. New energies are generated or passed to the following receptors whenever the observation cluster changes. Energies are boosted or damped according to the expected observations (following receptor state and edge duration). The observed gesture is recognized in timestep 5.*

Since the temporal event map is learned without a teacher, the semantics between a recognized observation sequence and the corresponding gesture meaning is missing. As the gestures should be learned and recognized online no previous knowledge about the observation sequences or even the meaning of the gestures are known in advance. As a result, the semantics have to be learned online by the dialog system. We added a variety of hebbian learning to enable the learning of gesture semantics. Whenever the dialog system detects a interaction from the user - for example detecting the Yes Keyword - it creates a new gesture class for this specific action. Each active receptor in turn has the ability to establish and reinforce hebbian weights to the new gesture class, whenever the class is generated by the dialog system. Thus, receptors, which can be associated with a specific gesture class will have a strong weight to this particularly gesture class. In reverse, if a receptor is active and has a strong weight to a specific gesture class the class will be activated and the gesture is recognized.

For the realization of the head gesture recognition system the system observations are coded according to the relative head panning and tilting movement in two consequtive image frames. The panning and tilting is converted to a 360 degree direction vector. Afterwards the direction vector is associated with 6 equally sized head direction classes. As an extra direction class is introduced for very small head movements the system state can be described with 7 different observation classes.

5. EXPERIMENTAL RESULTS

This section presents experimental results we achieved by using the described approaches. Therefore, the system is evaluated on the publicly available FG-Net database [20], which consists of images sequences of 18 people showing the ba-

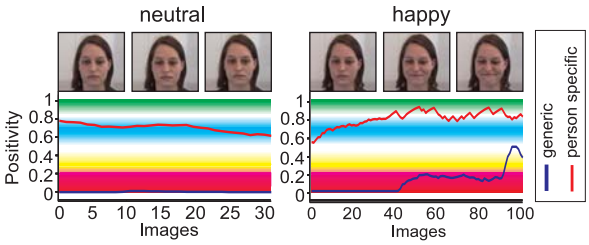

Fig. 5. *Comparison between the generic and the person-specific evaluation. For the neutral sequence the generic evaluation (blue line) fails to detect the persons emotion while the modified version (red line) is able to adapt to the persons neutral face (left). The same results are presented for a sequence starting with a neutral emotion and ending with a happy emotion. Again the generic evaluation fails, because it is not able to cope with the emotions of the specific person.*

sic emotions. The database is quite challenging and realistic as authors of the database tried to capture realistic emotions. Furthermore, we recorded a number of test sequences at our lab, to evaluate the proposed attention measuring system and the head gesture recognition.

5.1. Positivity Measure

Since all of the sequences of the FG-Net database starting with a neutral emotion, images with clearly expressed emotions have to be identified and labeled with the corresponding positivity values. After this, the Support Vector Regression can be trained using cross validation applying the leave one out strategy for every single person. As the neutral face for every person in the database is known, the average mean face can be easily computed. Unfortunately, due to the used regression technique the system cannot be directly compared to well known classification systems. Nevertheless, the RMS error between the ground truth emotion value and the obtained value can be computed as an indicator of system performance. To show the usefulness of the proposed extensions, namely the person specific neutral face (NF) and the feature weighting (FW) Table 1 shows the different RMS errors obtained for the FG-Net database. Since the positivity difference between certain emotion can go down to 0.1, the RMS error of the proposed system seems to be quite high. Nevertheless, the system can still give a good hint on person positivity for the challenging FG-Net database. The proposed extensions lead to a more robust estimation. Particularly if the neutral face of a person is quite different from the mean neutral face, the proposed integration of the person specific mean face can lead to a great improvement (Fig. 5).

Table 1. *Results of the Support Vector Regression*

used extension	train RMSE	test RMSE
SVR using **NF + FW**	0.04	**0.16**
SVR using **FW**	0.04	**0.21**
SVR using **NF**	0.04	**0.19**
SVR	0.05	**0.24**

5.2. Attention Estimation

To evaluate the proposed attention measure under real world conditions, we have recorded 23 test sequences from 8 different persons at our lab. The people where asked to watch several video clips in front of a computer monitor, while they were monitored. During the first stage of the experiment, the test person where watching an exciting movie, to get an attentive video sequence. In the second stage, the test persons where watching a boring video, while another person enters the room and tries to distract the test person by talking to her or letting things fall on the ground. Afterwards, the recorded image sequences were labeled manually in terms of visual attention and compared to the results of the proposed attention system.

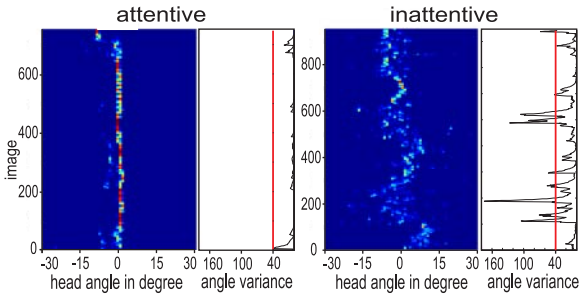

Fig. 6. *Comparison between the head poses of an attentive and an inattentive person. The aggregated head pose angle histogram (5 frames) given along the x-axis, where red is indicating a high count and blue is indicating a low count. The corresponding variance is displayed on the right side of each angle histogram. The variance values for the attentive sequence are always below the variance threshold (red line) indicating an attentive person, whereas the values for the inattentive sequence crosses the threshold several times.*

Exemplary plots of the horizontal head pose and the corresponding variance values of two sequences are given in Fig. 6. Obviously, the variance of the head pose is a good measure for the straightness of the head pose. As a consequence, the variance is a feasible measure to derive the user's attention. If the variance is above a certain threshold the attention value will decrease. We have tested the described method against hand labeled attention values of all sequences comparing the attention values. The resulting ROC curve generated by varying the variance threshold is given in Fig. 7. A major problem of the current attention evaluation arises from the missing consideration of the user's eye gaze. This leads to false attention values in situations in which the eye gaze is more important than the head direction. However, since the head pose can be regarded as a low pass filtered eye gaze, these effects are not important on a larger time scale of several seconds. Therefore, the proposed system is able to detect user's inattentiveness which last for at least 1 second quite well and can be used as a good indicator for user's visual attention.

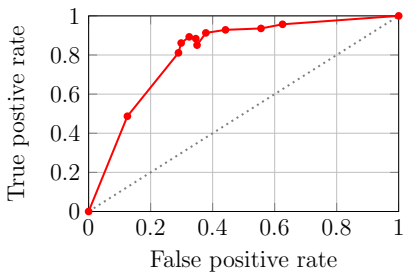

Fig. 7. *ROC curve of the presented attention estimation system. The curve is obtained by varying the attentive-inattentive threshold.*

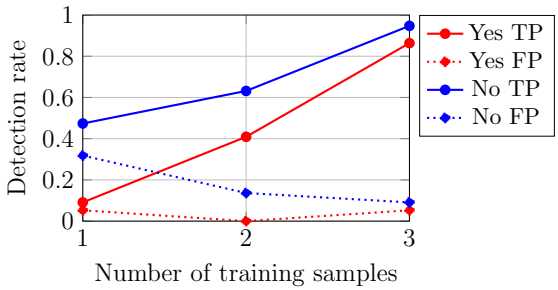

Fig. 8. *True positive (TP) and false positive (FP) rates for the detection of both head gestures depending on the number of training samples.*

5.3. Head Gesture Recognition

For evaluating the head gesture recognition, test sequences of different people from our lab were recorded. Each person was asked to perform a series of natural head shaking and head nodding gestures. They sequences start with three repetitions of head shaking and head nodding which are used from the dialog system to create the gesture classes for the introduced hebbian weights of the Temporal Event Mapping. Afterwards the sequences contains a series of five head shaking and five head nodding gestures in random order.

Afterwards the random sequence of head gestures was evaluated with a changing number of gesture class training samples. Accordingly, the number of triggered gesture classes which result in hebbian learning steps is varying between one and three. The results of the true positive and false positive rates for both head gestures are given in Fig. 8. Obviously, the recognition rate is quite low if only one gesture learning step is performed. However, the true positive rate for both gestures increase considerably as the number of training samples increases. This effect can be explained with the varying strenght of the different gestures even for one single person. Performing the same gesture with varying strength might lead to other active receptors in the temporal event map and therefore the detection fails. However, three repetitions for head shaking and head nodding gestures seems to be sufficient for robust detections. Futhermore, the false positive rates for both gestures are low even for a changing number of training samples. For that reason, the proposed gesture classification can be used as a valuable input channel for the dialog system to support natural human-robot interaction.

6. CONCLUSION AND FUTURE WORK

In this paper we present a way to extract user attention and emotion utilizing the shape and texture parameter from a fitted Active Appearance Model. We focus on improving the human-robot interaction and, therefore, apply a linear emotion measure, which determines the emotion positivity. Furthermore, to improve the emotion estimation, we suggest to use a correlation based weighting of the AAM parameter and a person specific neutral reference. In combination with Support Vector Regression (SVR) the proposed system is tested on the FG-Net database showing promising results for determining user emotion positivity over time. We also show how the AAM shape parameters can be utilized to estimate the user's head pose and derive a measure of attention using the distribution of the head pose over time. Compared to hand labeled attention values, the system is able to estimate the attention value quite well. In addition, we propose a head gesture recognition based on the temporal event mapping approach. Continuing our work, we will integrate the proposed system into a dialog system [21]. This will be very helpful to examine how the proposed emotion and attention values can be utilized to enable a more natural human-robot interaction.

7. REFERENCES

[1] "Companionable project," Website: `http://www.companionable.net`.

[2] T.F. Cootes, G. Edwards, and C.J Taylor, "Active appearance models," in *Proc. of the ECCV*, 1998.

[3] Y. Saatci and C. Town, "Cascaded classification of gender and facial expression using active appearance models," *FGR*, vol. 1, pp. 393–398, 10-12 April 2006.

[4] H. van Kuilenburg, M. Wiering, and M. den Uyl, "A model based method for automatic facial expression recognition," *Machine Learning: ECML 2005*, pp. 194–205., 2005.

[5] L. Shang and K.-P. Chan, "Nonparametric discriminant hmm and application to facial expression recognition," *CVPR*, pp. 2090–2096, 2009.

[6] C. Breazeal and B. Scassellati, "How to build robots that make friends and influence people," in *Proc. of the IROS*, 1999, vol. 2, pp. 858–863.

[7] P. Smith, Student Member, M. Shah, and N. Da Vitoria Lobo, "Determining driver visual attention with one camera," *IEEE Trans. on Intelligent Transportation Systems*, vol. 4, 2003.

[8] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel, "From gaze to focus of attention," in *Proc. of Workshop on Perceptual User Interfaces: PUI 98*, 1998, pp. 25–30.

[9] 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 Trans. on PAMI*, pp. 1212–1229, 2007.

[10] S. O. Ba and J.-M. Odobez, "Recognizing visual focus of attention from head pose in natural meetings," *Trans. Sys. Man Cyber. Part B*, pp. 16–33, 2009.

[11] L.M. King, H.T. Nguyen, and P.B. Taylor, "Hands-free head-movement gesture recognition using artificial neural networks and the magnified gradient function," *IEEE Conf. of the Engineering in Medicine and Biology Society*, pp. 2063–2066, 2005.

[12] L.-P. Morency and T. Darrell, "Head gesture recognition in intelligent interfaces: The role of context in improving recognition," in *Proc. 11th Int Conf. on Intelligent User Interfaces*, 2006, pp. 32–38.

[13] P. Lu, M. Zhang, X. Zhu, and Y. Wang, "Head nod and shake recognition based on multi-view model and hidden markov model," *Proc. Int. Conf. on Computer Graphics, Imaging and Visualization*, pp. 61–64, 2005.

[14] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *IJCV*, pp. 221–255, 2004.

[15] S. Baker and I. Matthews, "Equivalence and efficiency of image alignment algorithms," in *Proc. of IEEE Conf. on CVPR*, 2001, pp. 1090–1097.

[16] F. De la Torre, A. Collet, M. Quero, J. Cohn, and T. Kanade, "Filtered component analysis to increase robustness to local minima in appearance models," *IEEE Computer Society Conf. on CVPR*, pp. 1–8, 2007.

[17] R. Stricker, Ch. Martin, and H.-M. Gross, "Increasing the robustness of 2d active appearance models for real-world applications," *Proc. 7th ICVS LNCS 5815*, pp. 364–373, 2009.

[18] C.-C. Chang and C.-J. Lin, "Libsvm - a library for support vector machines. 2001.," `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[19] F. Schill and U.R. Zimmer, "Robust asynchronous temporal event mapping," in *Intelligent Robots and Systems. IEEE/RSJ International Conference on*, 2002, vol. 1, pp. 190 – 195.

[20] F. Wallhoff, "Facial expressions and emotion database," `http://www.\\mmk.ei.tum.de/~waf/fgnet/feedtum.html`, Munich Technical University 2006.

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