Attention and Emotion Based Adaption of Dialog Systems

Sebastian Hommel, Ahmad Rabie, and Uwe Handmann

Abstract. In this work methods are described, which are used for an individual adaption of a dialog system. Anyway, an automatic real-time capable visual user attention estimation for a face to face human machine interaction is described. Furthermore, an emotion estimation is presented, which combines a visual and an acoustic method. Both, the attention estimation and the visual emotion estimation based on *Active Appearance Models* (AAMs). Certainly, for the attention estimation *Multilayer Perceptrons* (MLPs) are used to map the *Active Appearance Parameters* (AAM-Parameters) onto the current head pose. Afterwards, the chronology of the head poses is classified as attention or inattention. In the visual emotion estimation the AAM-Parameter will be classified by a *Support-Vector-Machine* (SVM). The acoustic emotion estimation also use a SVM to classifies emotion related audio signal features into the 5 basis emotions (neutral, happy, sad, anger, surprise). Afterward, a *Bayes network* is used to combine the results of the visual and the acoustic estimation in the decision level. The visual attention estimation as well as the emotion estimation will be used in service robotic to allow a more natural and human like dialog. Furthermore, the human head pose is 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 work examine a simple gaze estimation with the help of an ordinary webcam. Moreover, a full body user re-identification method is described, which allows an individual state estimation of several people for hight dynamic situations. In this work an appearance based method is described, which allows a fast people re-identification over a short time span to allow the usage of individual parameter.

Keywords: Multilayer Perceptron, Bayes network, active appearance model, support-vector-machine.

Sebastian Hommel · Ahmad Rabie · Uwe Handmann
University of Applied Sciences Ruhr West,
Computer Science Institute, Tannenstr. 43, 46240 Bottrop, Germany
email: {sebastian.hommel,ahmad.rabie,
uwe.handmann}@hs-ruhrwest.de

E. Pap (Ed.): Intelligent Systems: Models and Applications, TIEI 3, pp. 215–235.
DOI: 10.1007/978-3-642-33959-2_12

1 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 or ticket machines are established in our society. Furthermore, service systems are getting more and more important in home environment. This ranges from robotic animals for amusement 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. Essential parts in human like communication is to know the emotional state of the dialog partner. For example, in tutoring systems or computer games, knowing about the user's feeling of boredom, frustration or happiness can increase learning success or fun in the game. Especially in human-robot interaction, affective reactions of the robot, following the recognition of the user's emotional state, can make the interaction more natural and human-like.

A further helpful information, for a user interaction is to know whether the dialog partner is attentive to the service system or to any other [9]. There are many methods to estimate the user attention, like full body movements or the used attention estimation which based on the direction of view. To estimate the emotion of the dialog partner, it is possible to interpret the facial shape and texture, the voice and some times full body movements. In this work an emotion estimation method is used, which combines the cues of facial expressions and audio signal information.

To estimate the user attention and the facial expression in a 2D image, a head description is needed which model the head of the current user. Two *Active Appearance Models* (AAMs) are used to realize this head models in real time. *Active Appearance Models* 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. Therefore, the parameters of the AAM are adapted, so that the model fits to the current face in shape and appearance.

Especially in home health care for demented people, often it is not possible to use the head pose for attention estimation, because many demented people are restricted in their head movements. These people mostly look around by moving merely their eyes. 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. Further essential informations for a gentler and more natural dialog, which can be estimated with the help of the head pose are 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.

For the estimation of users state it is often helpful to use an individual reference, like the *Individual Mean Face* which is described in [10]. Especially in home environment the number of interaction partners is limited and the most interaction partners come again. So it is helpful to use a former estimated reference. For this reason, a user identification is needed. A full body method for identification is described in this work. This method interprets the full body appearance which is used for a re-identification in a brief span.

2 Related Work

The work comprises the tasks of extracting the visual focus, head gesture classification as well as emotion estimation. All of these tasks require information about the user's head. A wide variety of methods with different kinds of feature extraction and classification approaches can be found in the literature. In the following, we give a brief overview of different methods, which have been applied for the specific tasks.

2.1 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 [23]. 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 [25]. Hidden Markov Models are a very common way to extract the focus of attention from a sequence of head poses [25, 22, 1]. 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.

2.2 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 [12]. Furthermore, SVM Classification as utilized in [14] or Hidden Markov Models [13] 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.

2.3 Audio-Visual Emotion Recognition

Joining several modalities in a single multimodal emotion recognition system could be achieved in several fusion levels, which will be detailed discussed in Sec 4.3. Paleari and Lisetti proposed a general framework for multimodal information fusion towards multimodal emotion recognition. They discussed that the fusion of the

information takes place at signal, feature and decision levels. However, the work did not report any practical implementation and experimental results [17]. De Silva and Chi exploited a rule based method for decision level fusion of speech and vision based systems. The multimodal results showed an improvement over both of the individual systems [21]. Zeng et al. used a voting method to combine output of audiobased and vision-based recognition systems for person-dependent emotion recognition [32].

However, the above listed works of audio-visual-based emotion recognition did not consider the influence of the speech-related configuration of the face on the emotion-related facial changes and consequently on the accuracy of recognizing emotions in natural and unconstrained conversational human robot interaction, which is challenged in this work.

3 Head Pose Interpretation

In this work the user's head poses are extracted and used for an attention and head gesture estimation. The attention estimation in this paper is based on the parameters of a fitted AAM. Trefflich [27] 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. The described system interpret the AAM-Parameters as head pose and analyzes them for attention and gesture estimation [11]. 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 a model that is not detailed in shape. An overview of the used architecture is shown in Fig. 1.

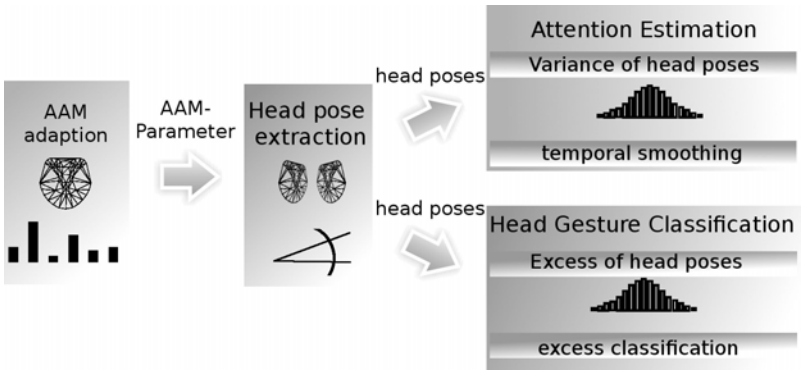


Fig. 1 Head Pose Interpretation. The attention estimation and head gesture classification is done by extracting the head pose from AAM-parameters. The poses are aggregated over time to compute the variance and excess kurtosis. The variance is used to compute the user's attention while the excess kurtosis is applied for head gesture recognition. Temporal smoothing is applied for both subsystems to reduce input parameters noise.

3.1 Head Pose Estimation

Both the attention and the gestures of Yes and No can be determined by head poses. Trefflich shows that the visual attention correlates with the head pose [27]. 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 dataset contain head poses. For head pose estimation an own dataset is used which contains 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 .

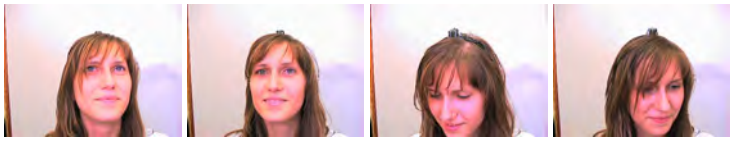


Fig. 2 Head Pose Dataset. This is an example of the used dataset for the head pose estimation.

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. These so called *global shape parameters* only describe 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 have a similar model a method to generate an Active Appearance Model with two separate example sets is favorable [26]. In the used method the texture is learned first. After that, the shape could be learned by separate examples. For texture training the FG-Net dataset [30] is used. An own dataset of one person who move the head horizontally and vertically in separated time frames 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, two simple *Multilayer Perceptrons* (MLPs) are used. One MLP maps the AAM-Parameter which describes the horizontal head pose onto the correct linear head pose. Equivalent to this, the other MLP maps the AAM-Parameter which describes

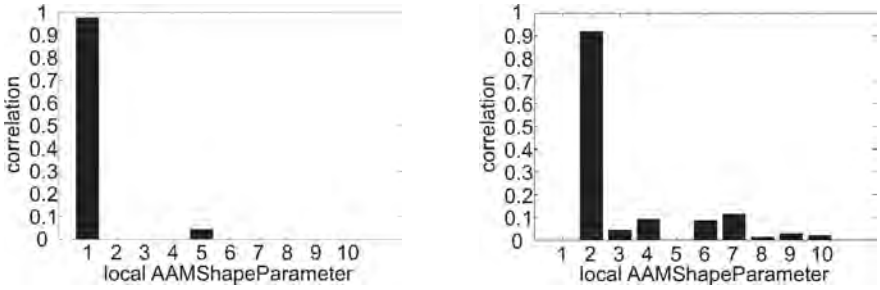


Fig. 3 Correlation between the local AAM-Parameters and the Head Poses. The left diagram shows the correlation with the horizontal head rotation and right diagram shows the correlation with the vertical head rotation.

the vertical head pose onto the correct linear head pose. In this way the horizontal and vertical head poses are described in degree. Each MLP consists in one input neuron, two hidden neurons and one output neuron.

3.2 Attention Estimation

To analyze the head poses, the statistics is used, since it allows 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 [6]. All used *adaptive recursive methods* employ a constant weight c which range from 0 to 1. It is possible to use different weights for each adaptive recursive method. For that reason we use the indexed weights c_M, c_{Z^2}, c_A and c_I. By the use of a tall c, the method is more sensitive to input chances. The variance Z^2 (Fig. 4) of the head poses are used to get a classification onto attentive or inattentive. To calculate the variance at the time $t+1$, the temporal mean M_{t+1} and the input value (current head pose) X_{t+1} is needed. In this work, the temporal mean is also adaptive recursive calculated. For the use of adaptive recursive methods, well initializations are favourable. In this work, we postulate that the user is initially attention. Therefore, we use 0 for m_0 and z_0:

$$M_0 = m_0,$$
$$M_{t+1} = M_t + c_M \cdot (X_{t+1} - M_t),$$
$$Z_0^2 = z_0,$$
$$Z_{t+1}^2 = Z_t^2 + c_{Z^2} \cdot \left((X_{t+1} - M_{t+1})^2 - Z_t^2 \right).$$

A sequence of head poses are classified as attentive once the variance is lower than a threshold with a 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:

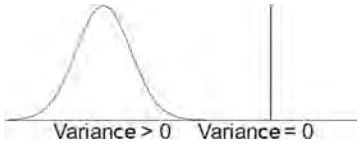


Fig. 4 Variance. When the variance is greater than 40°, the current part of the sequence is classified as inattention and otherwise as attention

$$A_0 = 100,$$
$$A_t = A_{t-1} + c_A \cdot (100 - A_{t-1}) \text{ , if attention,}$$
$$A_t = A_{t-1} + c_A \cdot (0 - A_{t-1}) \text{ , if inattention.}$$

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:

$$I_0 = 0,$$
$$I_t = I_{t-1} + c_I \cdot (p_t - I_{t-1}).$$

3.3 Head Gesture Estimation

Furthermore, the developed estimator is able to detect head shake and nodding from a face image sequence. Afterwards, a context-sensitive interpretation as Yes or No is possible. When a person shakes 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:

$$Z_0^4 = z_0,$$

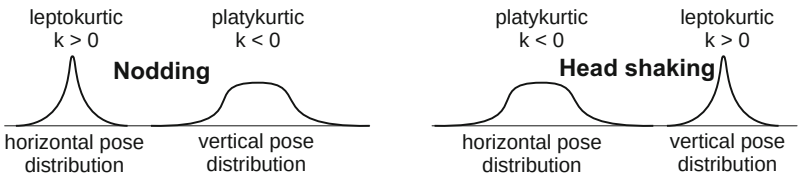


Fig. 5 Head Gesture. Performing a nodding gesture involves changes in the vertical head pose. Therefore, the vertical head pose distribution becomes platykurtic while the horizontal distribution is leptokurtic. Head shaking involves changes of the horizontal head pose. Therefore, the distribution characteristics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

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

$$\varepsilon = \frac{Z_n^4}{(Z_n^2)^2} - 3.$$

3.4 *Eye Tracker*

To estimate the attention of demented people the eye gaze is needed, since they are mostly unable to move the head full freely. So these people look around only with their eyes and without head movements. To estimate the eye gaze an ordinary webcam is used. First, the AAM affords the eye position, so it is possible to focus only to this image parts. A gray level eye consist of a white plane, a darker iris and a black pupil, so the pupil is detected 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 that the healthy eye is able to move 20° to the left and 20° to the right [24]. This is an interest area of 40° onto the curvature of the eye. Up to this value it is possible to assume a linear correlation between the pupil position into the 2D image and the pupil position onto real world eye.

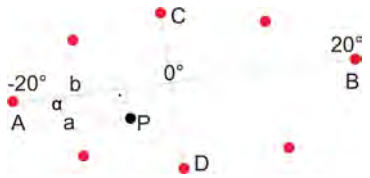


Fig. 6 Eye Tracker. Triangulation to calculate the horizontal and vertical pupil position for the eye gaze estimation

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 triangulation. This is exemplary shown in Fig. 6. To estimate the global eye gaze it is necessary to add the head pose, too.

4 Emotion Recognition

Possible modalities to exploit for automatic recognition are language (acoustic-tic and linguistic information), facial expressions, body gestures, bio signals (e. g. heart rate, skin conductance), or behavioral patterns (such as mouse clicks). Though one modality alone can already give information on the affective state of a user, humans always exploit all available modalities, and if an automatic systems attempts to reach human performance, the need for multi-modality is obvious. Thereby not only consent results of different modalities lead to more confident decisions, but also conflicting results can be helpful [17], e. g. to detect pretended or masked emotions, or

to find out more reliable modalities for certain emotions. The most obvious modalities in human-human conversation, and also in human-robot conversation which we aim to enhance, are speech and facial expressions.

In our work we challenge this approach by analyzing the auditory and visual stimuli with respect to their general discriminative power in recognizing emotions. Note that in our work we focus on interactive scenarios and are thus targeting at systems that are able to work online. The approaches we present in this paper are, therefore, not only being tested offline on existing databases but have proven their applicability in robotic applications in real world settings [8, 19, 18]. This is in contrast to other work (e.g. [3]), which has focused on offline emotion recognition only. The following three sections will provide a brief introduction on the respective unimodal analysis techniques as well on the proposed probabilistic decision level fusion.

4.1 Visual Facial Expression Recognition

In order to recognize basic emotion visually, we take a closer look into the interlocutor's face. The basic technique applied here are Active Appearance models (AAMs) first introduced by Cootes et al [5]. The generative AAM approach uses statistical models of shape and texture to describe and synthesize face images. An AAM, that is built from training set, can describe and generate both shape and texture using a single appearance parameter vector, which is used as feature vector for the classification. The active component of an AAM is a search algorithm that computes the appearance parameter vector for a yet unseen face iteratively, starting from an initial estimation of its shape. The AAM fitting algorithm is part of the integrated vision system [19] that consists of three basic components. Face pose and basic facial features (BFFs), such as nose, mouth and eyes, are recognized by the face detection module [4]. The coordinates representing these features are conveyed to the facial feature extraction module. Here, the BFFs are used to initialize the iterative AAM fitting algorithm. After the features are extracted the resulting parameter vector for every image frame is passed to a classifier which categorizes it in one of the six basic emotions in addition to the neutral one.

Besides the feature vector, AAM fitting also returns a reconstruction error that is applied as a confidence measure to reason about the quality of the fitting and

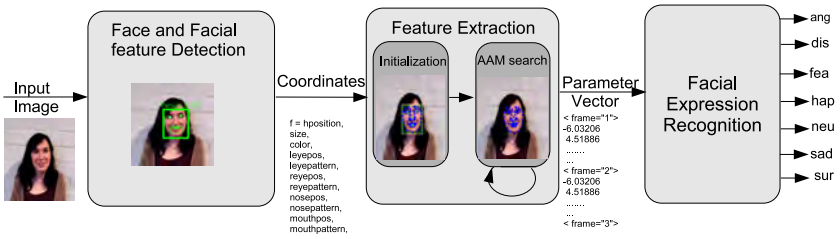


Fig. 7 Emotion Recognition. Architecture of the facial analysis sub-system

also to reject prior false positives resulting from face detection. As classifier a one-against-all Support Vector Machine is applied. The whole system is applicable in soft real-time, running at a rate of approximate 5 Hz on recent PC hardware.

4.2 Emotion Recognition from Speech

For the recognition of emotions from speech, EmoVoice, a framework that features offline analysis of available emotional speech databases, as well as online analysis of emotional speech for applications, is used [29]. This approach is purely based on acoustic features, that is no word information is used. The first step is the extraction of emotion-related features from audio signal. A large vector of statistical features based on prosodic and acoustic properties of the speech signal was calculated for each utterance in the DaFEx database. To reduce the size of feature vectors (over 1400 attributes) correlation-based feature subset selection is used. This selection is necessary to increase performance as well as speed of classification. By this way, 71 features related to pitch, energy, MFCCs, linear regression, range of the frequency spectrum of short-term signal segments, speech proportion, length of voiced and unvoiced parts in an utterance, and the number of glottal pulses remained. For classification support vector machines are used, but with a linear kernel. The feature selection is typically done offline, but the feature extraction and classification can be done either offline or in real-time.

4.3 Probabilistic Decision Level Fusion

Multimodal information fusion is the task of combining some interrelated information from multiple modalities. In an emotion analysis system, while a unimodal system incorporates features of a single modality (visual, audio, tactile, or body information) the multimodal systems use information from multiple different modalities simultaneously.

As affective states in interaction are usually conveyed on different cues at the same time, we agree with other works summarized in [31] that a fusion of visual and acoustic recognition yields significant performance gains. Hence, we followed the idea of an online integration scheme based on the prior offline analysis of recognition results on a database.

In current fusion research, three types of multi-modal fusion strategies are usually applied, namely data-/signal-level fusion, feature-level fusion, and decision-level fusion [17]. Signal-level fusion is applicable solely to sources of the same nature and tightly synchronous. Generally it is achieved by mixing two or more physical signals of the same nature (two auditive signals, two visual signals of two cams, etc). This type of mixing is not feasible for multimodal fusion due to the fact that different modalities always have different captors and different signal characteristics (auditive and visual).

Feature-level fusion means concatenation of the features outputted from different signal processors together to construct a joint feature vector, which is then

conveyed to the affect analyzer. It is used when there is evidence of class-dependent correlation between the features of multiple sources. For example, features can be extracted from a video processor (facial expression) and speech signal (emotion-related prosodic features). Feature-level fusion benefits of interdependence and correlation of the affective features in both modalities but is criticized for ignoring the differences in temporal structure, scale and metrics. Although, feature-level fusion demands synchronization of some extent between modalities. Yet another drawback of such a fusion strategy is that it is more difficult and computationally more intense than combining at the decision level.

The third fusion strategy combines the semantic information captured from the individual unimodal systems, rather than mixing together features or signals. Due to the advantages of (I) being free of synchronization issues between modalities, (II) using relative simple fusion algorithms, and (III) their low computational requirement in contrast to the feature-based methods, decision-level fusion methods are adopted from the vast majority of researchers in the field of multimodality emotion recognition. Following this conclusion we decided a probabilistic-based decision level fusion method to join the facial expression-based, and the acoustic information-based emotion recognizers into bimodal one.

The proposed decision-level fusion method is probabilistic approach based on a top-down-reasoning Bayesian network with a rather simple structure depicted in Fig 8. Based on the classification results of the individual visual and acoustic classifiers, we feed these into the Bayesian network as evidence of the observable nodes (Acoustic and Visual, respectively). By Bayesian inference the posteriori probabilities of the unobservable affective fusion (Fusion) node are computed as:

$$\mathbf{P}(\text{Fusion} = e_f | \text{Visual} = e_v, \text{Acoustic} = e_a), \tag{1}$$

where, e_f, e_v, e_a can belong to any one of seven emotion classes mentioned above, and taken as a final result.

The required probability tables of the Bayesian network are obtained from a performance evaluation of each individual classifiers in an offline training phase based

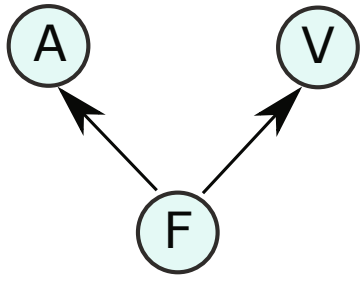


Fig. 8 The Bayesian network. The structure of the Baysian network used to fuse cues of both uni-modals. Evidences of observable nodes acoustic (A) and visual (V) are fed as input into the corresponding nodes. The posteriori probabilities of the unobservable node are computed, wich give fusion (F) as the final result.

on ground-truth-annotated databases [12]. Therefore, confusion matrices of each classifier are turned into conditional probability tables modeling the dependent observation probabilities of the model according to the arrows in Fig. 8.

5 User Re-identification

To use individual settings of a dialog adaption system for different users, a user recognition is essential. To allow a contactless interaction, the system could use visual or acoustic information. Whereas acoustic informations of the user are not evaluable all the time, visual informations are used in this work. Especially for the described facial methods a facial user re-identification could be used [7]. Certainly, in the case of searching or monitoring the dialog partner, facial information are often not available. In this case and to boost the facial re-identification, a whole-body user recognition is used. This work shows an appearance based user re-identification. Before an re-identification is possible, the person must be detected. In this work Histograms of Oriented Gradients are classified by a Support Vector Machine. To be real time capable, a GPU (Graphics Processing Unit) based algorithm is used [16].

The appearance based whole-body user re-identification is useful for monitoring multiple dialog partner in one scene. This kind of re-identification is not helpful for an application over a long period of time since humans whole-body appearance varies by changing clothes. Certainly, the appearance based features are divided into color and texture features. Whereas the texture is naturally independent of the illumination. Typically, color image pixels are described with the help of three color channels, which represents the red, green and blue parts of the current color (RGB). This results in different pixel representations given different illuminations. For this reason the RGB color representation is transformed into a color representation which allows to separate the illumination influence (Formula 2). The used color space represents the pixel color also in three channels, which describes separately the hue, saturation and the value (HSV). The hue range from 0° to 360°, the saturation range from 0% to 100% as well as the value. In this work only the illumination independent hue and saturation of the color are used:

$$
\begin{aligned}
&R,G,B \in [0,1]; MAX = max(R,G,B); MIN = min(R,G,B)\\
&H = \begin{cases} 0^\circ, & \text{if } MAX = MIN \Leftrightarrow R = G = B,\\ 60^\circ \cdot (0 + \frac{G-B}{MAX-MIN}), & \text{if } MAX = R,\\ 60^\circ \cdot (2 + \frac{B-R}{MAX-MIN}), & \text{if } MAX = G,\\ 60^\circ \cdot (4 + \frac{R-G}{MAX-MIN}), & \text{if } MAX = B, \end{cases}\\
&H = H + 360^\circ, \text{ if } H < 0^\circ,\\
&S = \begin{cases} 0, & \text{if } MAX = 0 \Leftrightarrow R = G = B = 0,\\ \frac{MAX-MIN}{MAX}, & \text{other}, \end{cases}\\
&V = MAX.
\end{aligned}
\tag{2}
$$

Three fixed parts of the detected people are selected to eliminate color and texture informations Fig. 9. One rectangle part of the lower body is separated to determine the mean hue and saturation. Furthermore, the mean horizontal and vertical texture rates as well as the mean hue and saturation is calculated from a rectangle part of the upper body. One histogram of the hues and one histogram of the saturations are calculated from an oval area of the upper body.

Fig. 9 Feature extraction. The used features for the full body people re-identification will be extracted from three areas. The location of this areas is relative to the people detection.

Whereas, the upper body appearance is typically more complex the texture rates and the histograms are calculated only from this part. The mean horizontal and the mean vertical texture rates describe the strength of the texture at the selected area. To calculate the horizontal and vertical texture, the Scharr filter [20] is used. The mean hue and saturation describe the ground color of the user's lower and upper body, while the histograms describe more detailed the upper body colors. By using the hue and saturation histograms, even detailed color prints etc. at the clothes are represented in a very compact form. In this work normalized histograms are used due to their scale independency. To handle minor changes of the color, the hue and saturation are both divided into only 16 parts for the histograms. To allow more robust identification, all the features will be tracked over the time. This tracks are used to build a user feature space. This kind of people description is very compact and easy to save. To compare the saved feature spaces with the current features, only the normalized differences of all the features are calculated and summarized. Certainly, during the calculation of the difference D between the hues of the searched person feature space (H_i) and the hue of the current hypotheses (H_c), it must heeded that the hue is represented as a circle. To calculate the difference of the hue (Formula 3) is used:

$$\begin{aligned} D &= |H_i - H_c|, \\ D &= 360^\circ - D, \text{ if } 360^\circ - D < D. \end{aligned} \tag{3}$$

In this way, the summarized differences are used as a score for user identification. A small score means a hight probability to find the correct person.

6 Experimental Results

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

6.1 Head Pose Estimation

To evaluate the presented AAM and MLP based head pose estimation, 8 test 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. Then 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. 10.

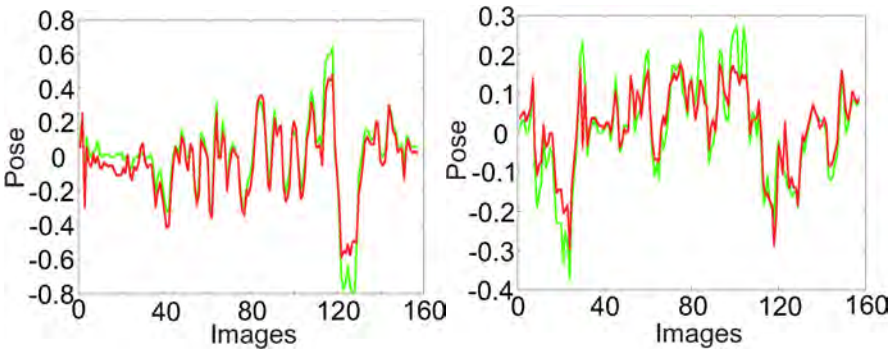


Fig. 10 MLP-Output vs. ”Flock of Birds”. The green curve represent the head poses which are determined with the help of the ”Flock of Birds” and the red curve represent the head poses which are estimated by the presented system. *left:* horizontal head poses; *right:* vertical head poses.

6.2 Attention Estimation

To evaluate the proposed attention measure under real world conditions, 23 test sequences from 8 different persons are 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. 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 estimation system. The average divergence between the presented system and the labels is only 4 percentage points which is shown in Table 1.

Table 1 Attention Estimation. This table shows the rate of attention for each evaluation sequence which is determined by the presented system and by human interpretation

Sequence	1	2	3	4	5	6	7	8	9	10	11	...
System	23.6%	16.8%	3.1%	1.1%	8.5%	4.9%	0%	0%	25.5%	0%	0.2%	...
Human	0%	15.5%	0%	2.3%	12.9%	6.7%	0%	0%	7.5%	0.7%	0%	...
Difference	23.6%	1.3%	3.1%	1.2%	4.4%	1.8%	0%	0%	18%	0.7%	0.2%	...

Sequence	12	13	14	15	16	17	18	19	20	21	22	23
System	4.3%	1.1%	0%	8.6%	14.8%	31.6%	0%	0%	27.1%	13.4%	28.4%	69.6%
Human	0%	0%	0%	0%	7.8%	26.9%	0%	0%	20.6%	15.6%	27.1%	68.3%
Difference	4.3%	1.1%	0%	8.6%	7%	4.7%	0%	0%	6.5%	2.2%	1.3%	1.3%

The described system only fails massively at three image sequences. By visual analyzing of these 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 normally distributed and this is able during attention and during inattention (Fig. 11).

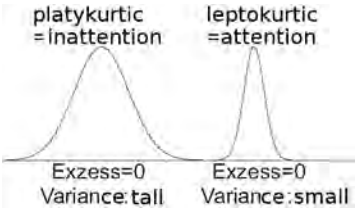
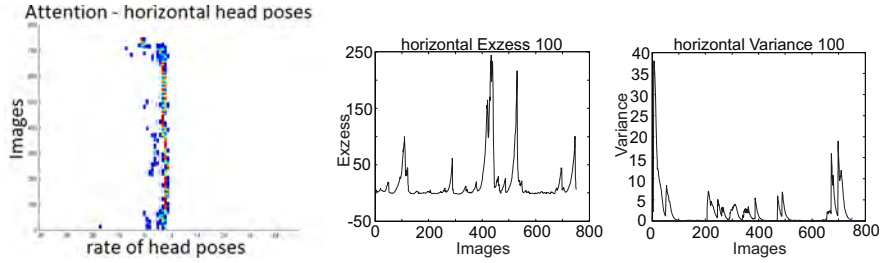


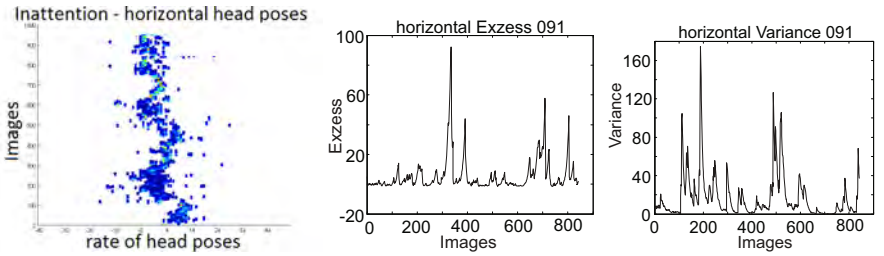
Fig. 11 Excess Kurtosis vs. Variance. A wide and a small histogram can be Gaussian distribution, so the excess kurtosis can be 0, once the variance become tall when the user is inattentive

Fig. 12 and Fig. 13 shows exemplary the head poses, the excess kurtosis and the variance of an attention and an inattention sequence.



(a) **Horizontal Head Poses** increasing rate from blue to red *(b)* **Excess Kurtosis** *(c)* **Variance**

Fig. 12 Excess Kurtosis vs. Variance of an attention sequence. (a) shows the temporal histogram for each time. A small histogram with many red areas leads to a small variance (c) but to a fluctuating excess kurtosis (b).



(a) **Horizontal Head Poses** increasing rate from blue to red *(b)* **Excess Kurtosis** *(c)* **Variance**

Fig. 13 Excess Kurtosis vs. Variance of an inattention sequence. (a) shows the temporal histogram for each time. A wide histogram with less red areas leads to a large variance (c) but to a fluctuating excess kurtosis (b).

6.3 Head Gesture Estimation

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

6.4 Eye Tracker

The simple eye tracker is only evaluated by three sequences. In these 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. 14 and Fig. 15 the vertical and horizontal eye gazes and head poses is exemplary visualized for one sequence.

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

Fig. 14 Histogram of horizontal head pose vs. eye gaze

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

Fig. 15 Histogram of vertical head pose vs. eye gaze

Furthermore, the Table 2 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.

Table 2 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

6.5 Bimodal Emotion Recognition

As we are striving in this work to give the robot a bimodal emotion recognition ability that is based on analyzing facial expressions and speech information, the systems are afresh evaluated on data set with subjects in a real-life conditions. Four subjects have participated in this test (one female and three males). The whole procedure is divided into training and test phases. For one subject both phases were conducted in the same day; for two others the test was is conducted in the following day, while for the fourth subject the time interval was two days.

In the training phase the subjects are asked to display facial expressions of live emotion classes: anger, happiness, neutral, sadness, and surprise with and without speaking. The average amount of data captured from each subject for each facial expression class was 246 images. To create conditions of real-life human-robot interaction as much as possible, the subjects are allowed to move arbitrarily in front of the camera. During this phase a person-independent AAM, which is built from a subset of the DaFEx database of talking and non-talking subjects, is used to extract the emotion-related facial features. These features are then conveyed to train a person-dependent SVM.

In the test phase the subjects are asked to display facial expressions and utter a few sentences (in general five) expressing as much an emotions as possible [1]. The above-mentioned AAM is used to extract facial features, which are labeled with the proper emotional class by the above-trained SVM. In this session a person-independent speech-based emotion recognizer is utilized to categorize each utterance into the proper emotional class. An average of 145.25 images from each subject for each emotion are used as test data. The validation matrix for the fusion scheme of each subject was an averaged confusion matrix (CPT), which is obtained from the performance of both individual systems on the three remaining subjects.

Table 3 Emotion Recognition. The performance of each stand-alone unimodal systems against the one of the bimodal system. All results are obtained from a test in a real-life condition.

	Anger	Happiness	Neutral	Sadness	Surprise	Total
Vis	75.00	43.75	50.00	60.28	47.92	55.39
Aco	33.04	15.42	36.25	23.06	10.42	23.63
Audio-Visual	75.00	50.00	68.75	49.03	47.92	58.14

Table 3 illustrates the results obtained from both the stand-alone and bimodal systems. The low rates delivered by the speech-based emotion analysis system - the first raw - could be because a person-independent classifier is used, which is trained on a speech-based emotion database that does not include the subjects participating in the evaluation procedure. Nevertheless, it can be seen that the whole performance of the bimodal system has an advantage over both facial-expression- and speech-information-based systems, which satisfy the goal of the fusion scheme proposed

[1] The sentences were emotional words free.

previously. However, when the performance of each channel on each emotion is considered it is notable that the recognition rate of happiness and neutral is enhanced when the bimodal system is employed, which indicates that the cues of both modalities comprise complementary information for these two emotions. In contrast, from the first and fifth rows, it is noticeable that both unimodal cues comprise only redundant information so that combining both modalities yields no improvement with regard to discrimination ability for the recognition of anger and surprise. Furthermore, the fourth column indicates that both modalities deliver conflicting information, which causes sadness to be recognized even less than the stand-alone facial-expression-based modality.

7 Conclusion

In this paper, approaches for extracting the user attention, the head gestures and the emotions are presented. These approaches utilize the shape and texture parameters from a fitted Active Appearance Model. This paper focus on improving the human-robot interaction. An attention and head gesture estimation, which use the AAM shape parameters to estimate the user's head pose is applied. For the measurement 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 is proposed.

In order to enhance the naturality of the interaction an approach, which provide the ability for the dialog system to analyze the emotional state of the interaction partner, is implemented. The proposed approach fused a facial-expression-based emotion analysis system with a speech-based analysis system in a bimodal emotion analysis system. A probabilistic decision level fusion approach is used. That makes benefits of its simplicity, no mandatory of synchronization, and the general discriminative power of each unimodal compared to the other fusion methods. Five of the seven basic emotions of Ekman are considered in a life-like scenario of human robot interaction. The bimodal system outperforms both uni-modal system in a natural and life-like conversational human-robot interaction.

Continuing this work, could be possible by integrating the proposed systems into a dialog system [15]. This will be very helpful to examine how the proposed attention values as well as the emotion 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. To enable the use of individual parameter, a re-identification method is presented.

Non-basic emotions such as frustration, sleepiness and satisfaction could be considered in future works by adopting the dimensional approach of emotion categorization. A further work on the topic of attention and emotion estimation could be the usage of full body motion. Furthermore, it is possible to use full body motion as a further feature for the full body re-identification.

Acknowledgements. This work was partly funded by the German Federal Ministry of Education and Research (BMBF) in the framework of the APFel project.

References

1. Ba, S., Odobez, J.: Recognizing Visual Focus of Attention from Head Pose in Natural Meetings (2009)
2. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. In: Proc. Int. Conf. Multimodal Interfaces (2004)
3. Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A., Karpouzis, K.: Modeling naturalistic affective states via facial and vocal expressions recognition. In: Proc. Int. Conf. Multimodal Interfaces, pp. 146–154. ACM, New York (2006)
4. Castrillón, M., Déniz, O., Guerra, C., Hernández, M.: ENCARA2: Real-time detection of multiple faces at different resolutions in video streams. Journal of Visual Communication and Image Representation 18(2), 130–140 (2007)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. PAMI 23(6), 681–685 (2001)
6. Grießbach, G.: 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)
7. Handmann, U., Hommel, S., Brauckmann, M., Dose, M.: Face Detection and Person Identification on Mobile Platforms. Springer Tracts in Advanced Robotics (STAR). Springer, Germany (2012)
8. Hegel, F., Spexard, T., Vogt, T., Horstmann, G., Wrede, B.: Playing a different imitation game: Interaction with an empathicandroid robot. In: Proc. Int. Conf. Humanoid Robots, pp. 56–61 (2006)
9. Hommel, S.: Zeitliche Analyse von Emotionen auf Basis von Active Appearance Modellen. GRIN Verlag GmbH (2010)
10. Hommel, S., Handmann, U.: AAM based Continuous Facial Expression Recognition for Face Image Sequences. In: 2011 12th IEEE International Symposium on Computational Intelligence and Informatics (CINTI), Budapest, Hungary, pp. 189–194 (2011)
11. Hommel, S., Handmann, U.: Realtime AAM based User Attention Estimation. In: 2011 IEEE 9th International Symposium Intelligent Systems and Informatics (SISY), Subotica, Serbia, pp. 201–206 (2011)
12. King, L.M., Taylor, P.B.: Hands-free Head-movement Gesture Recognition using Artificial Neural Networks and the Magnified Gradient Function. In: IEEE Conf. of the Engineering in Medicine and Biology Society, pp. 2063–2066 (2005)
13. Lu, P., Zhang, M., Zhu, X., Wang, Y.: Head nod and shake recognition based on multiview model and hidden Markov model. In: Computer Graphics, Imaging and Vision: New Trends, pp. 61–64 (2005)
14. Morency, L.-P., Trevor, D.: Head gesture recognition in intelligent interfaces: the role of context in improving recognition. In: Proceedings of the 11th International Conference on Intelligent User Interfaces, IUI 2006, pp. 32–38. ACM, New York (2006)
15. Müller, S., Schröter, C., Gross, H.-M.: Aspects of user specific dialog adaptation for an autonomous robot. In: IWK (2010)
16. Prisacariu, V., Reid, I.: fastHOG - a real-time GPU implementation of HOG. Department of Engineering Science, Oxford University, Tech. Rep. 2310/09 (2009)

17. Paleari, M., Lisetti, C.L.: Toward multimodal fusion of affective cues. In: Proc. ACM int. Workshop on Human-Centered Multimedia, pp. 99–10. ACM, New York (2006)
18. Rabie, A., Handmann, U.: Fusion of Audio- and Visual Cues for Real-Life Emotional Human Robot Interaction. In: DAGM 2011 (2011)
19. Rabie, A., Lang, C., Hanheide, M., Castrillon-Santana, M., Sagerer, G.: Lang, Ch., Hanheide, M., Castrillon-Santana, M., Sagerer, G.: Automatic Initialization for Facial Analysis in Interactive Robotics. In: Proc. Int. Conf. Computer Vision Systems, Santorini, Greece (2008)
20. Scharr, H.: Optimal operators in digital image processing. Ph.D. thesis, Interdisciplinary Center for Scientific Computer, Ruprecht-Karls-Universität, Heidelberg (2000)
21. Silva, L.D., Chi, P.: Bimodal Emotion Recognition. In: Fourth IEEE Int. Conf. on Automatic Face and Gesture Recognition (2000)
22. Smith, K., Ba, S.O., Odobez, J.-M., Gatica-Perez, D.: Tracking the Visual Focus of Attention for a Varying Number of Wandering People. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 1212–1229 (2008)
23. Smith, P., Member, S., Shah, M., Lobo, N.D.V.: Determining Driver Visual Attention with One Camera. IEEE Trans. on Intelligent Transportation Systems 4 (2003)
24. Speckmann, E.J., Hescheler, J., Köhling, R.: Repetitorium Physiologie, 2nd edn. Urban & Fischer Verlag (2008)
25. Stiefelhagen, R., Finke, M., Yang, J., Waibel, A.: From Gaze to Focus of Attention (1998)
26. Stricker, R., Martin, C., Gross, H.-M.: Increasing the Robustness of 2D Active Appearance Models for Real-World Applications. In: Fritz, M., Schiele, B., Piater, J.H. (eds.) ICVS 2009. LNCS, vol. 5815, pp. 364–373. Springer, Heidelberg (2009)
27. Trefflich, B.: Videogestützte Überwachung der Fahreraufmerksamkeit und Adaption von Faherassistenzsystemen. Technische Universität Ilmenau (2009)
28. Viola, P., Jones, M.: Robust Real-time Object Detection. In: Second Int. Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sympling (2001)
29. Vogt, T., André, E., Bee, N.: EmoVoice — A Framework for Online Recognition of Emotions from Voice. In: André, E., Dybkjær, L., Minker, W., Neumann, H., Pieraccini, R., Weber, M. (eds.) PIT 2008. LNCS (LNAI), vol. 5078, pp. 188–199. Springer, Heidelberg (2008)
30. Wallhoff, F.: Facial Expressions and Emotion Database. Technische Universität München (2006), `http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html`
31. Zeng, Z., Pantic, M., Roisman, M.I., Huang, T.S.: A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Transaction on Pattern Analysis and Macine Intellegence 31, 39–58 (2009)
32. Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T.S., Roth, D., Levinson, S.: Bimodal HCI-related Affect Recognition. In: ICMI (2004)