Multimodal Face Detection, Head Orientation and Eye Gaze Tracking

Frank Wallhoff and Markus Ablaßmeier and Gerhard Rigoll
Institute for Human-Machine Communication
Technische Universität München
80290 München, Germany
{wallhoff,ablassmeier,rigoll}@tum.de

***Abstract*— For several applications within the Human-Machine-Interface domain a person's face plays a key role as an information source, such as the identification, the computation of the affective state or to predict the awareness of an user. Therefore, this paper presents a multi-modal approach for finding and tracking a face and estimating the head's gaze as well as the eyes' view direction. Throughout the paper several measurements relying on two different camera inputs are introduced which can be used to form a robust computation of the head orientation and the viewing direction of a person.**

I. INTRODUCTION

The detection of faces even in cluttered scenes is often a fundamental and basic step for Human-Machine-Interfaces (HMI). Typical examples with a strong demand for highly reliable face detection results are numerous, such as an user identification over biometric face features, or the recognition of the affective state using the facial expression. In order to construct such a robust detection system the entire decision process, i.e. to say where a face is located, should better not be based on a single measurement but on several ones. However, integrating too many measures is not affordable with respect to computational reasons.

Another, more specific task after locating the faces' positions is to recognize what a person is looking at. Obviously there is a difference between the head gaze and the view direction of the eyes. A head can face a specific object, a monitor for instance, and follow several activities on the screen, such as the mouse pointer, by moving the eyes without moving the head. In many application domains the knowledge about the view direction of the eyes is more important than the orientation of the head, respectively the face. But the measurements relying on the eyes and the head are usually related to each other. A typical example in the automotive context using information about the view angle is to measure the duration of distraction while driving and operating the infotainment system at once. Another demanding application is the proper perspective projection of a 3-dimensional scene in a virtual reality environment or to highlight the area with the field of view.

In order to detect the position of a face while simultaneously recognizing the head and eye view direction we propose a dual camera setup as follows: one camera captures a scene using the regular visual wavelengths without additional illumination. The measurements based on the visual queue are mostly for face detection:

- face detection by skin color models
- face likelihood using neural networks
- eye likelihood computation
- estimation of the head orientation

The second camera records the infrared spectrum of the scene. It makes use of self-emitted infrared illumination which will be introduced in more detail. The measures to detect and follow the eyes with the infrared camera are:

- appearance based face detection
- pupil detection
- eye view angle detection

All these listed measures can be interpreted as weak classifiers. In order to combine the single measures several fusion strategies can be considered. However, with the goal of integrating the static decisions of single frames over the time and fusing the queues an extended random sampling technique is applied here for face detection and gaze estimation. The infrared eye view angle detection is currently independent.

The main part of the paper is structured as follows: in the next section the isolated queues for face and eye detection are presented separately. Then a framework to incorporate these isolated results by fusing and integrating temporal information is introduced. After the eye gaze recognition results within an automotive application the paper closes with a discussion and an outlook.

II. FACE DETECTION

Although several efforts and a lot of work has been carried out to robustly and precisely detect faces in still images with different lighting conditions and in front of arbitrary backgrounds there is no final solution which can be applied to unconditioned scenarios.

A detailed review over work in the face detection domain can be found in [1] and [2]. Depending on the obeyed technique the search for faces is often restricted to frontal upright shots or color photos. On the other hand approaches with a high degree of freedom are usually computationally rather expensive. To bridge this gap we apply the following detection queues which may be fused later.

A. FACE DETECTION USING SKIN COLOR

Color is a key feature for the detection of hands and heads in images. It is probably one of the most used methods for the detection of human body parts which may be rested on its low computational cost. The disadvantage is

Proc. MFI, Heidelberg
IEEE

the low reliability caused by the change of skin-tone color appearance under different lightning conditions.

We use an approach to recognize skin color under varying illumination and brightness conditions by transforming the *RGB*-color intensities into the normalized *rg*-chroma space. The basic assumption is that a skin colored pixel lies within a certain area in this *rg*-chroma plane, the so-called skin locus [3].

In this approach a skin color candidate has to be between two circles g_{up} and g_{down} in the *rg*-plane, where $g_{up} = a_{up}r^2 + b_{up}r + c_{up}$ and $g_{down} = a_{down}r^2 + b_{down}r + c_{down}$ ($a_{up} = -1.8423$, $b_{up} = 1.5294$, $c_{up} = 0.0422$, $a_{down} = -0.7279$, $b_{down} = 0.6066$ and $c_{down} = 0.1766$,). Furthermore whitish and grayish pixels which lie in a circle with the radius 0.02 around the color white ($r = g = 0.333$) do not represent skin. Together with the neighboring pixels, a skin color probability $p(c|\Omega_k)$ can be introduced by computing the normalized number of skin colored pixels in a rectangular box.

The skin color pixels inside the skin locus are depicted in Fig. 1, a typical skin mask after applying this technique in Fig 1.

Fig. 1. Skin locus and example image with typical skin color segments

B. EXAMPLE-BASED FACE DETECTION

In addition to the search of faces using skin color a second technique is considered to calculate face-likelihoods and to reject non-face regions. Although several classification paradigms for appearance-based face detectors exist, they are all using the same fundamental processing which is depicted in Fig. 2. A binary classifier with a fixed input size is trained with positive and negative examples of monochromatic intensity distributions. Thus the real properties of the object to be found are learned over their appearance. Later in the application phase an unknown input image has to be scanned sequentially at all positions and sizes. Depending on the variance against changes in translation and scaling the subsampling can be increased with appropriate step sizes. To allow a higher reliability against illumination changes all scanned images are commonly preprocessed as suggested by Sung [4].

For our purposes we use an elaborated implementation of an artificial neural network (NN) based classifier similar to that one which has already been introduced by Rowley [5]. This technique has established itself as being highly robust but computationally expensive. It can be applied to regular gray-scale and even infrared images. However, a combination of this queue together with skin color will already lead to

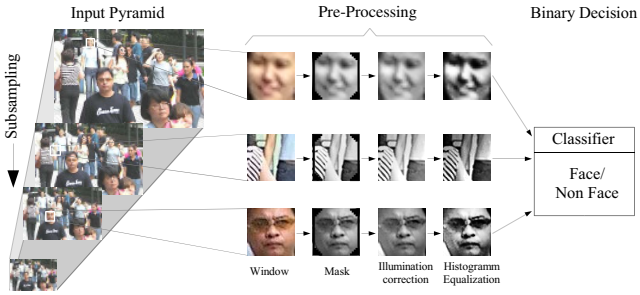

Fig. 2. Image sub-sampling pyramid and preprocessing[4]

a fast and robust system by leaving out those blocks with a low number of skin pixels.

With the help of a feed-forward multi layer perceptron (MLP) structure being trained with frontal upright and moderately tilted in depth rotated faces (approx 10°) we can compute the likelihood of a given image to be a face by applying a threshold of 0.5. The input layer has a size of 20×20 pixels.

Since there are typically several detections in the region of the true face position the resulting hypotheses have to be merged. This can be done by discarding isolated detections and merging those positives which have an overlap to others.

III. HEAD GAZE ESTIMATION

With the previously presented example-based approach it is possible so far to detect frontal upright faces. The following extension aims at also detecting faces rotated in depth which means that they are rotated in the horizontal direction while estimating the rotation angle of the head $\varphi_{\text{disc.}}$. A head rotation within the projection plane, i.e. tilting is not within the scope of the estimation but could be covered by expanding the pyramidal sampling presented earlier. The appearance of vertical rotations or elevations of more than $\pm 15°$ are rather seldom within the aimed applications.

Due to the nature of a head being a complex three-dimensional object, the addressed rotation influences the appearance of the face with respect to the observing camera plane. Furthermore the resulting appearance is less compact than the relation between frontal captured faces. However, because of the good performance and functionality of the example-based NN face detector this approach was extended allowing rotations in depth by expanding the hidden layer and the output vector with probabilities for discrete view directions. Therefore the previously deployed hidden layer of the NN was expanded by additional parallel instances. At the same time the number of output neurons was increased from 1 to 8.

After an back-propagation training with bootstrapping the first neuron of the NN still represents the probability $p(\lambda^{\varphi}_{\text{face}}|x)$ of seeing a face at the input of the network. The other seven values can be interpreted as probabilities $p(\bar{\lambda}^{\varphi}_{\text{face}}|x)$ for the angle φ between the optical axis and the head gaze.

The values of the discrete angles are $-90°, -45°, -22,5°, 0°, 22,5°, 45°$ and $90°$ resulting

from the available material in the FERET database [6]. The expanded output vector together with typical training examples is summarized in Fig. 3.

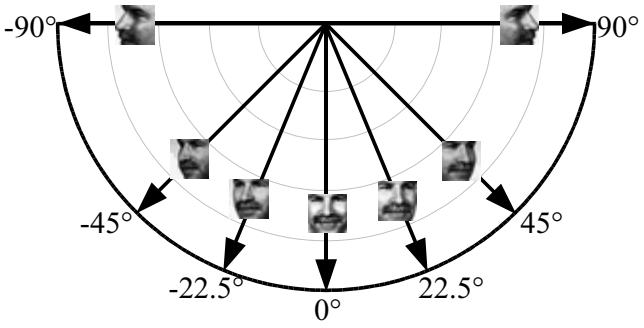

Fig. 3. Output vector for the discrete azimuthal gaze estimation

Since the real gaze of a face typically lies between the prior learned prototype views, the discrete output probabilities have to be transfered to continuous ones by computing the weighted vectors according to Eq. 1.

$$\varphi_{\text{cont.}} = \angle \left[\sum_{\varphi_{\text{disc.}} \in \{0,\pm 22{,}5,\pm 45,\pm 90\}} p(\lambda_{\text{face}}^{\varphi_{\text{disc}}} | x) \cdot e^{j\varphi_{\text{disc}}} \right] \quad (1)$$

IV. EYE AND MOUTH FINDING

A. APPEARANCE BASED EYE AND MOUTH DETECTION

As the detection of faces even in complex scenarios has successfully been integrated with the presented method using a NN the eyes and the mouth have to be localized in a subsequent step.

After the detection of a face region seeking the eyes and the mouth can be restricted to empirically pre-defined areas [7]: eyes are assumed to be found in the upper half of the region, the mouth in the lower half as depicted in Fig. 4a.

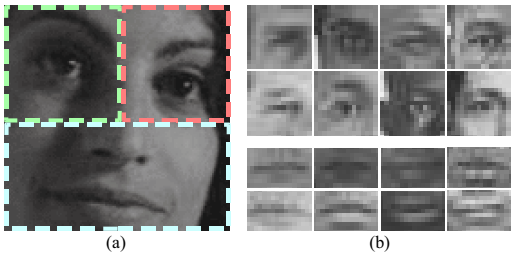

Fig. 4. Evident search regions for eyes and mouth

For the actual task, i.e. to detect the *eyes* and the *mouth*, modified NN approaches similar to the face detector are applied. Besides the modification of the network architecture the training material has to be adapted to the new tasks. Typical examples of positive training material are depicted in Fig 4b.

The positive examples for the eye region are taken from quadratic boxes concentric to the iris. The size of the box is derived from the eye-to-eye distance and normalized to 20×20 pixels. The organization of the empirically found network is as follows (20×20): the input layer is retinally linked to 28 hidden layers, where four of them (20×5 neurons) are reserved for larger regions such as eye brews and eyelashes, another four (10×10 neurons) are for the detection of the pupil. For a finer resolution 16 additional hidden layers (each 5×5 neurons) are foreseen. All layers are fully connected with the output neuron.

The eye-to-eye distance also specifies the width for the mouth region and the double of its height. The box is concentric to center of the mouth. The architecture for the mouth NN consists of an input layer with 25×15 neurons which is linked to receptive hidden layers: three for the upper and lower lip (25×5 neurons) and another five (5×15 neurons) to recognize dimples and the corners of the mouth.

B. RED-EYE-EFFECT BASED DETECTION

The red-eye-effect is well-known especially in conjunction of compact photo cameras with a flash light. If the photographed person is looking into the camera lens, the pupils often appear in red instead of black color. The red-eye-effect is based on the reflection of the flash light by the blood-rich retina of the eye. If a person is looking directly into the lens of the camera the observed phenomena has the highest impact. In this case the camera and eye axis are parallel. This effect will be used for the detection of the pupils' positions.

By using this effect the localization of pupils becomes more precise than for example a limbus based one since it is less affected by the lids.

With a special technical setup alternately bright and dark pupil reflections can be produced by lighting the eye from different located light sources [8] and [9]. For this reason two infrared LEDs together with a gray-scale camera producing interlaced images are used. The first half-image is illuminated with an infrared LED very close to the axis, the second with the objective-far LED (indicated by T and T' in Fig. 8). The arising reflections are depicted in Fig. 5 and Fig. 6 [10].

Fig. 5. First field with on-axis LED

By differencing both fields the eyes can be found easily since their intensity distribution only differs in the reflectance of the eyes and a constant difference in the gain which can be compensated by applying histogram normalization. As reflections from any other constant lighting source are present in both fields they are not observed in the difference image and do not interfere with the current measure.

Consequently, only two blobs remain that represent the two pupils as shown in Fig. 7. Noise can be dynamically

Fig. 6. Second field with far axis LED

suppressed by flooring the minimal intensity relative to the actual histogram. The position of the eyes can be computed by the centers of blobs with the highest peaks in the difference image.

Fig. 7. Difference image of even and odd field

C. RED-EYE BASED GAZE RECOGNITION

To allow calculating the gaze direction the reflection of the retina and additionally the first reflection on the cornea are needed. The first one P is the center of the red-eye from Fig. 5 caused by the on-axis LED. The second point P' is the bright spot, also-called first Purkinje reflection in Fig. 6 caused by the off-axis LED.

From the horizontal and vertical distance $h_{\Theta_{h,v}} = P_{x,y} - P'_{x,y}$ between these two points $P(x,y)$ and $P'(x,y)$ the angles $\Theta_{h,v}$ between the camera lens axis and the eye axis can be calculated in accordance with the following simplified formulas:

$$h_{\Theta_{h,v}} = (r_{\text{Eye}} - r_{\text{Cornea}}) \cdot \sin\Theta_{h,v} \tag{2}$$

$r_{\text{Eye}} - r_{\text{Cornea}}$ is the difference between the radii of eye and cornea, $h_{\Theta_{h,v}}$ are the distances between the pupil center P and the cornea reflection P' in the image plane in pixels. Without committing a large error for small angles $\Theta_{h,v} < 4^\circ$ the term $\sin\Theta_{h,v}$ can be replaced by $\Theta_{h,v}$. Thus the formula can be simplified to $h_{\Theta_{h,v}} = (r_{\text{Eye}} - r_{\text{Cornea}}) \cdot \Theta_{h,v}$.

Due to the physical nature of the light reflectance on the cornea, a certain upper limit arises, which is dependent on the individual properties of the observed eyes. This limit is at approx. $\Theta_h < \pm 30^\circ$ in the horizontal direction and vertically at approx. $\Theta_v < \pm 15^\circ$. For larger angles $\Theta_{h,v} > 30^\circ$ the proposed model and the formula are not valid anymore.

The remaining parameters, i.e. $r_{\text{Eye}} - r_{\text{Cornea}}$ can be approximated by averaged or empirically gathered values. However, to provide precise estimations of the direction, the geometrical values of the eyes should be measured individually in a calibration phase for each person. An example with distances in the vertical and horizontal direction between the reflections is shown in Fig. 9.

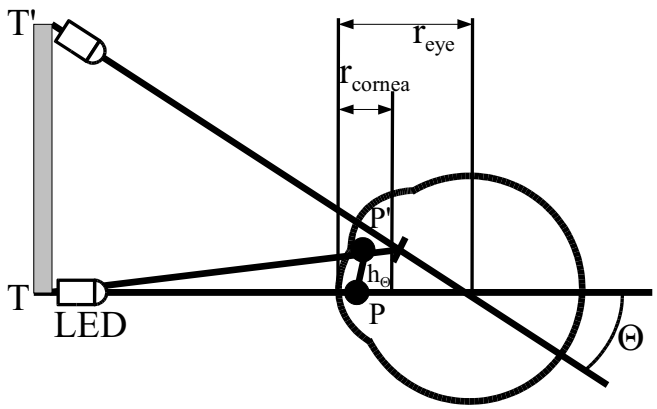

Fig. 8. Red-eye based gaze detection principle

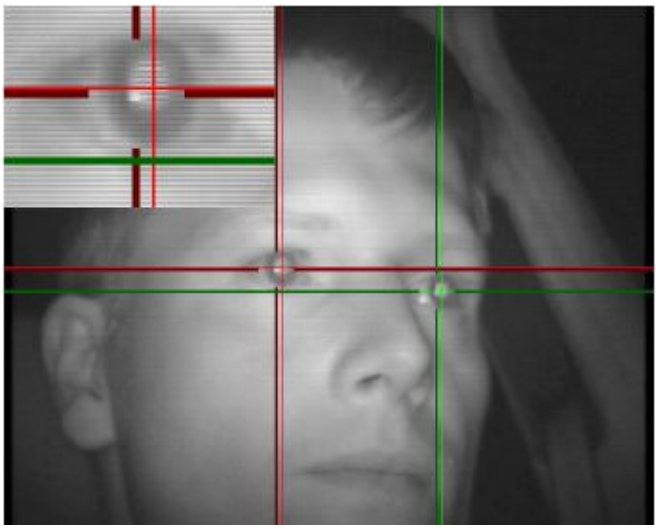

Fig. 9. Cornea reflection and pupil position

V. CONDENSATION ALGORITHM

Assuming a Markov-State-Space model with hidden states $\{\mathbf{x}_t\}$ describing position, size and dynamics of a face, the prior described observations, such as skin color and face-likelihood $\{\mathbf{z}_t\}$ are used to estimate the state of the system through the filtering distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t})$. In most cases, this probability cannot be derived directly but is calculated recursively by

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|\mathbf{x}_t) \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}) d\mathbf{x}_{t-1}.$$

The prior distribution $p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})$ describing the system state in the last time step is predicted with dynamics $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. Then the observation $p(\mathbf{z}_t|\mathbf{x}_t)$ updates the predicted distribution according to the measurement of the image to generate the current distribution.

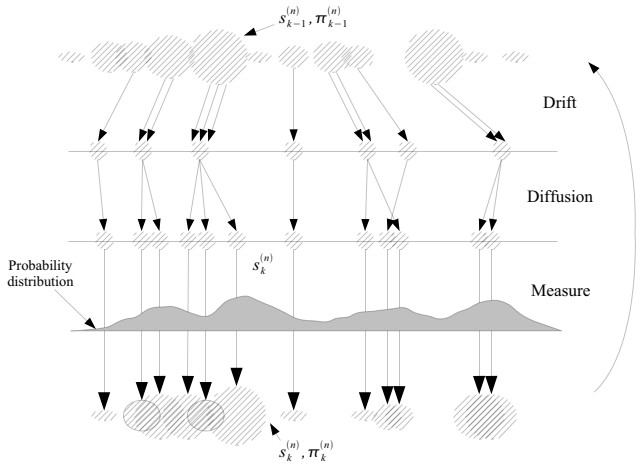

Fig. 10. Overview of the Condensation algorithm [11]

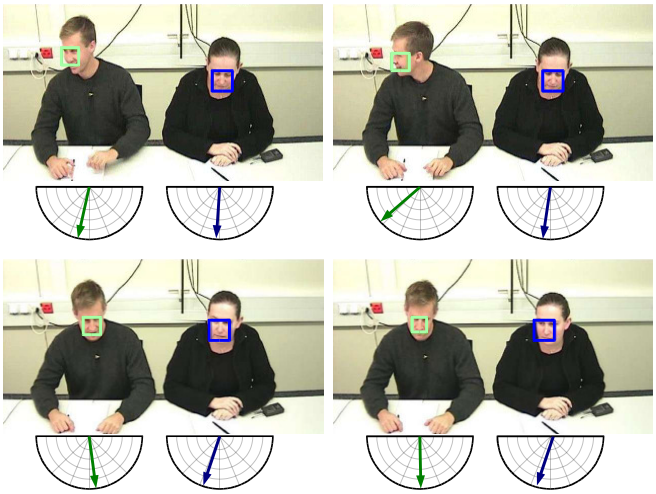

Fig. 11. Detection results with estimated head gazes

The filtering distribution is approximated with a set of weighted samples, called particles. These are containing information about the system state, such as position, size, and dynamics. This way the distribution becomes $\widehat{p}_N(\mathbf{x}_t|\mathbf{z}_{1:t}) = \sum_{i=1}^{N} \pi_t^{(i)} \delta(\mathbf{x}_t - \mathbf{x}_t^{(i)})$. This method is known as condensation algorithm (particle filter, sequential monte carlo) [11].

In the first step $N = 200$ particles are initialized with the output of the face detector. Then each particle is predicted by a linear regressive dynamical motion model with constant velocity plus random noise. The parameters of this dynamical model are determined by training an adaptive linear network (ADALINE). For each particle the probability for containing a skin colored region out of the skin color mask is derived, and a face likelihood using the pre-described neural network is measured. These observations are linked together by multiplication $p(\mathbf{z}_t|\mathbf{x}_t^{(i)}) = p(\mathbf{z}_t^{skin}|\mathbf{x}_t^{(i)}) \cdot p(\mathbf{z}_t^{face}|\mathbf{x}_t^{(i)})$ and deliver the weight for each particle $\pi_t^{(i)} \propto \pi_{t-1}^{(i)} \cdot p(\mathbf{z}_t|\mathbf{x}_t^{(i)})$. A resampling step for the particle set, using the new weights keeps the particles in regions with high "face likeliness".

To allow tracking of faces from people entering the scene, 10% of the particles are initialized by the face detection algorithm at each time step. For determining the number and locations of faces, connected regions of particles with a specific size, minimum amount of particles and minimum probability are searched. For each so found location of a face a minimum number of particles is kept.

VI. EXPERIMENTS AND RESULTS

In the following sections the performance of the proposed detection and gaze recognition approaches are summarized.

A. FACE DETECTION AND HEAD GAZE ESTIMATION

The resulting head gaze detection approach has been tested on a sequence of a multi-modal meeting room scenarios with two participants [12]. Fig. 11 depicts typical detection results of faces together with their estimated azimuthal view angles represented by arrows.

A qualitative evaluation of the detected face areas in this sequence results in superior localization capabilities even with heads rotated in depth. However, the estimation of the head gaze angle can just be interpreted as a rough guess. In peaks the difference between real head direction and the measured one was up to 20°, the mean difference was $\approx 5°$.

A more detailed resolution of the output vector will dramatically increase the reliability of the estimation. On the other hand such an extension would also mean higher computational efforts.

B. EXAMPLE-BASED EYE DETECTION

To test the previously introduced nets for detecting eyes and mouth the face regions of 50 persons are pre-segmented with the face detector module in a first step. In the second step the nets are separately applied to the earlier mentioned search regions within the face regions. Multiple detections are merged by averaging. The coordinates of the eyes are given by the center of the detected boxes. Typical results are depicted in Fig. 12.

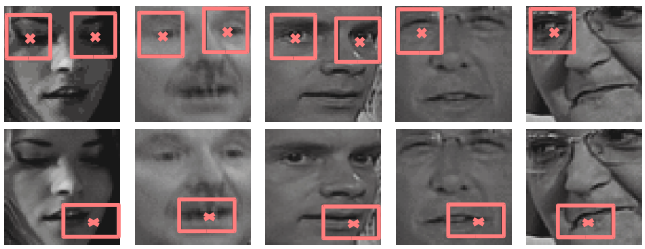

Fig. 12. Detection results for eyes (upper block) and mouths (lower block)

A quantitative analysis of the obtained results can be achieved by defining an empirical maximum distance to the manually marked ideal location. For subsequent applications a marker is found to be acceptable if the relative Euclidean distance to the ground truth position is less than 25% of the box height. The results gained by this rule are summarized in table I.

The detection rates for the eyes with this approach are already very promising. The main reason for missing eyes are

too small bounding boxes from the previous face detection stage which can be easily corrected by enlarging the search region. Indicated by its low precision the detection of the mouth seems to be a more serious problem arising from its larger elastic properties.

Type	Detection Rate [%]	Acceptable [%]
Eyes	72	72
Mouth	36	18

TABLE I

LOCALIZATION RESULTS FOR EYES AND MOUTH

C. RED-EYE-EFFECT BASED MEASURES

For the calculation of the gaze direction the test setup consists of a camera with an infrared filter and an IR LED control. Aiming at the use for in-vehicle applications the image sensor is not facing the head frontally but is mounted on the center console. It has to be emphasized that the eye with a higher distance to the camera yields inferior values. The system is calibrated to the eye parameters of each subject with several fixed points on a map. Then, to measure the error between the estimated and the real angle several projected points from a beamer have to be fixated by the test person [13].

Subjective results confirm the function of our algorithm. However, the horizontal eye tracking of the line-of-sight seems to work better than the vertical, probably because of the larger horizontal expansion of the pupil.

In order to gather some quantitative results for the accuracy and reliability eight subjects accomplished the task described above. The test resulted in an average deviation between actual and calculated angle of 3.6° horizontal and 4.7° vertically. This resolution was mainly limited by the PAL resolution of the obeyed camera.

The successful gaze detection, where both the pupil and cornea reflection are found, is 88% for the nearer eye and 77% for the more distanced eye. If the duration is considered where the eyes are closed (in practice between 5% and 10%) the average rate of both eyes almost reaches 90%.

VII. CONCLUSIONS AND FUTURE WORKS

The present treatise introduced several measures to detect and track faces as well as estimating the gazes of the head and the eyes.

The integrated face detection and tracking algorithm already works with high precision. By applying a random sampling technique the algorithm runs on-line with 25 frames per second. The estimation of the head gaze is currently working quite well. In the future the accuracy might be improved by reducing the intervals of the learned views.

Experiments with the example-based detection of the eyes turned also out to be promising. Due to the low computational costs the system is able to run in real-time. The system might further be improved by adjusting the network architecture. Unfortunately, the precision of the mouth detection approach is rather weak. Due to its flexible nature example-based models do not seem to be useful here.

The detection capabilities based on the red-eye-effect are also very forward-looking and outperform the example-based system. A stronger coupling of both systems is mandatory in the future. View direction recognition based on the red-eye-effect shows reliable results and can be honed by employing a high resolution camera.

Besides improving all modules, we are working on a stronger fusion of the currently independent measures. It is planned to incorporate a stronger early fusion and the use of Graphical Models. Therefore all measures have to refer to same world coordinates which have to be derived from the currently independent camera coordinates. The use of additional image devices that can acquire depth information of the scene are desirable, for example a sensor measuring the run length of an emitted light impulse.

VIII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of Ronnie Mittermair and Florian Dibiasi. The red-eye based gaze detection was mainly based on their preliminary experiments and evaluations.

REFERENCES

[1] E. Hjelmas and B. K. Low, "Face Detection: A Survey," *Computer Vision and Image Understanding (CVIU)*, vol. 83, no. 3, pp. 236–274, Sept. 2001.

[2] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 24, no. 1, pp. 34–58, Jan. 2002.

[3] M. Soriano, S. Huovinen, B. Martinkauppi, and M. Laaksonen, "Skin detection in video under changing illumination conditions," in *Proc. 15th International Conference on Pattern Recognition, Barcelona, Spain*, 2000, pp. 839–842.

[4] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 20, no. 1, pp. 39–51, Jan. 1998.

[5] H. A. Rowley, *Neural Network-Based Face Detection*, Ph.D. thesis, School of Computer Science, Carnegie Mellon University, May 1999.

[6] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 22, no. 10, pp. 1090–1034, Oct. 2000.

[7] J. Blaesing, "Integration of a system for finding faces and face-parts for person identification.," Diploma Thesis, Institute for Computer Science, Gerhard-Mercator-Universität Duisburg (in German), Supervisor: F. Wallhoff, 2002.

[8] C. Morimoto, D. Koons, A. Amir, and M. Flickner, "Pupil detection and tracking using multiple light sources," in *Fifth European Conference on Computer Vision (ECCV 98)*, 1998.

[9] A. Haro, M. Flickner, and I. Essa, "Detecting and tracking eyes by using their physiological properties, dynamics, and appearance.," in *In IEEE Computer Vision and Pattern Recognition*, 1999, vol. 1, pp. 163–168.

[10] R. Mittermair, "Design of a system for the gaze direction in the automotive context using the red-eye effect.," Diploma Thesis, Human-Machine Communication, Technische Universität München (in German), Supervisor: M. Ablaßmaier, 2005.

[11] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *International Journal of Computer Vision (IJCV)*, vol. 29, no. 1, pp. 5–28, 1998.

[12] S. Renals, "The multimodal meeting manager (m4)," Project Homepage `http://www.dcs.shef.ac.uk/spandh/projects/m4/`, September 2005.

[13] F. Dibiasi, "Software implementation for the gaze direction in cars using the red-eye effect.," Bachelor Thesis, Human-Machine Communication, Technische Universität München (in German), Supervisor: M. Ablaßmaier, 2005.