MODEL BASED 3D GAZE ESTIMATION FOR PROVISION OF VIRTUAL EYE CONTACT

W. Waizenegger[1,2]*, N. Atzpadin*[1]*, O. Schreer*[1]*, I. Feldmann*[1] *and P. Eisert*[1,2]

[1]Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Berlin, Germany
[2]Humboldt University, Berlin, Germany

ABSTRACT

In recent years, video communication has received a rapidly increasing interest on the market. Still unsolved is the problem of eye contact. The conferee still needs to decide whether to look into the camera or directly to the screen. Recently, a solution to this problem was presented which is based on a real-time 3D modeling of the conferees [1]. In order to achieve direct eye contact the authors defined a virtual camera directly on the screen in the eyes of the remote conferee. This paper discusses the problem of adequately positioning this virtual camera. A new approach will be presented which performs an eye and gaze tracking directly on the real-time 3D model rather than on the 2D image. Our methods not only provides robust and highly accurate results but is also able to additionally measure the distance between the conferees eye and the display with high precision.

Index Terms— stereo gaze tracking, 3D, video communication, camera calibration

1. INTRODUCTION

In recent years, large video conferencing system providers such as CISCO, Polycom, LifeSize Communications and others have paid a lot of attention to increase the level of experience and the sense of *"Being there"* mainly due to high resolution video, large screens and convincing audio quality. As customers are now getting used to high frame rate and high resolution video, the development has to cater to other non natural aspects of the communication. One of the most disturbing drawbacks that prevents natural conversation and therefore hinder the user acceptance [2] is missing eye contact between the conferees.

Within the last two decades attempts were made to account for this problem, e.g. [3][4][5]. However, these approaches did not succeed in terms of application to commercial video conferencing systems. The main reason for this is the significant amount of processing power and the unavailability of robust and efficient 3D analysis algorithms to achieve acceptable and convincing corrected views.

Recently, algorithmic concepts based on Hybrid Recursive Matching (HRM) and Patch-Sweeping [6][1] have been proposed to account for these requirements. However, the provision of convincing eye contact not only depends on the high quality of the novel rendered view but also on the accuracy of the estimated perspective of the virtual camera. Further on, the correct perspective depends on the line of sight between the local conferee and the position on the display where the conferee is looking at. This requires a continuous and accurate estimation of the eyeballs and the line of sight while the conferee is looking at the display. So far eye gaze detection has been performed using infrared light in combination with video analysis. As at least two cameras are available anyway to create virtual views based on 3D video analysis, this information can be exploited for more accurate eye gaze detection together with the 3D position. Compared to previous approaches, the paper presents a novel algorithm for eye gaze detection using stereo information from standard cameras mounted on the display.

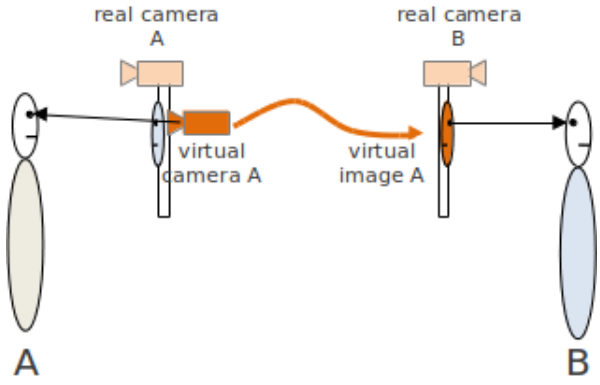

Fig. 1. Setup of real and virtual cameras for the provision of eye contact.

The paper is organized as follows. In section 2 we briefly introduce a concept for eye contact provision in video communication. Section 3 gives an overview of the developed algorithm. Section 4 and 5 give some insight on eye detection and iris segmentation as well as stereo processing of iris data. In section 6, the placement of the virtual camera is discussed. The applicability and accuracy of the approach is proven by experiments described in section 7.

2. A CONCEPT FOR EYE CONTACT PROVISION

In figure 1, the general set-up of a video communication system is depicted. Participant A and B are both captured with real cameras A and B on top of the display. The resulting image is then transmitted to the other party and shown on the display. Both participants do not perceive eye contact as they are looking onto the display and not into the camera. The eye contact problem can be solved by introducing a new camera, named as virtual camera for both sides. These virtual cameras are placed exactly in the eyes of both participants on the related display surfaces. Having this situation both participants can look into each others eyes while being captured from the top of the displays. Technically, a real-time 3D model of both conferees is computed. For a calibrated stereo camera setup on top of the displays arbitrary virtual cameras can be defined in relation to this model. Details about this approach can be found in [1].

According to the setup presented in figure 1, a major challenge is to compute the perspectively correct positions and viewing directions of both virtual cameras in real-time. Convincing eye contact will be generated only if the virtual cameras on both sides represent exactly the positions and gaze directions of their remote counterparts, i.e. the real eyes of the remote conferees. In order to solve this problem, a geometrical relationship between the real camera and a given position on the display needs to be established, which is required for virtual view rendering. For this process, knowledge about the gaze directions is required. It is the only possible way to identify where the users are looking at. The remaining part of this paper will discuss these problems in more detail. Firstly, the next section will describe the general algorithm. Afterwards, the process of eye and iris detection will be described. In combination with a high precision depth estimation as well as a virtual eyeball modeling the exact gaze directions and 3D positions of the conferees eyes can be determined.

3. ALGORITHMIC OVERVIEW

The illustration in figure 2 provides a brief overview of our algorithmic processing chain. The first step is a detection of eye regions in a stereo image pair and the segmentation of the iris in order to precisely locate the pupil centers. If the results successfully pass the filter rules described in section 4 a subsequent depth estimation in the eye regions is performed to enable the eyeball modeling in 3D and a constitutive gaze estimation. Finally, based on geometric properties of video communication setups and the computed viewing direction the synthetic camera placement is done.

4. EYE DETECTION AND IRIS SEGMENTATION

As outlined in figure 2 the first step of our model based gaze estimation algorithm is a purely image-based detection of the eyes and a segmentation of the iris region in order to precisely locate the center of the pupil. Therefore we initially perform a rough detection of the eye region with an improved version of the Viola-Jones object detection framework [7][8]. It is an *AdaBoost* based algorithm that uses cascaded Haar-like features. We use the publicly available implementation provided by [9] for our work. In order the get more robust and consistent results we apply the algorithm twice. A first cascade is used to detect the face of the conferee which leads to a valid region of interest for possible eye locations. Afterwards a second, nested cascade is used to compute potential eye locations within the previously identified facial area.

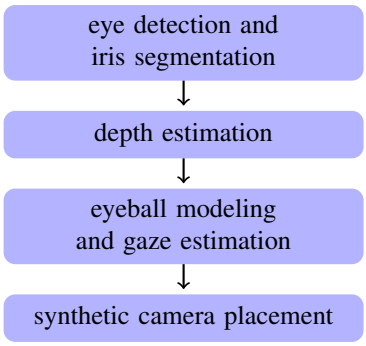

Fig. 2. Brief overview of our algorithmic processing chain.

Since the cascade classifiers output is the location of the eye area and the center of an eye region does not necessarily coincide with the center of the pupil a further refinement needs to be done. For simplification we assume that the edge of an iris is a circular contour. Regarding the binary edge image of an eye region $\boldsymbol{I}_e$, e.g. obtained with [10], we are interested in the arguments that solve the optimization problem

$$\max_{r\in[r_l,r_h]} \max_{\boldsymbol{M}\in\boldsymbol{\Omega}} \int_0^{2\pi} \boldsymbol{I}_e(r\cos(\alpha)+m_x, r\sin(\alpha)+m_y)\mathrm{d}\alpha, \quad (1)$$

where r_l and r_h are the lower and upper bounds for the iris radius in image space and $\boldsymbol{M} = (m_x, m_y)^T$ denotes the center of the iris in the image domain $\boldsymbol{\Omega}$. In order to find the maximizers for this problem we apply a circular Hough transformation on a pixel wise discretization of equation (1) and lookup the peak value in Hough space. In order to minimize the occurrence of false positives a subsequent filtering of the computed results based on plausibility constraints in image domain and in 3D is done as follows. A result is completely rejected if

- the number of eye candidates is not exactly two in both images.
- the orientation of the line connecting the center of the two iris candidates deviates more than ten degree from horizontal alignment.
- the re-projection error of the triangulated eye pair exceeds one pixel.

- the distance in 3D between the triangulated eye pair is smaller than 60 mm or greater than 85 mm.
- the computed eyeball radius deviates more than 1 mm from the anatomical mean of adults (see section 5).

5. STEREO GAZE ESTIMATION

In the following we introduce our novel stereo gaze estimation approach which is based on the iris segmentations of a stereo image pair as described in section 4. For each eye the gaze is estimated individually by modeling an initial eyeball and perform a subsequent iterative refinement. First we look up the depth values for the iris region. These values are available due to the 3D analysis done by the proposed video communication system [1].

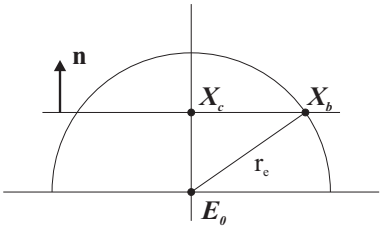

Fig. 3. Initial estimation of the eyeball center.

In order to obtain an initial position of the eyeball center we fit a spatial *eye plane* according to the depth values in the iris region by a least squares approach. As illustrated in figure 3 the eyeball is modeled by using the fact that the radius r_e of an adults eyeball is about $11.5\,\text{mm}$. Together with the 3D positions on the *eye plane* of the iris center $\boldsymbol{X}_c$ and one point on the iris border $\boldsymbol{X}_b$ a first estimate of the eyeball center $\boldsymbol{E}_0$ can be computed by $\boldsymbol{E}_0 = \boldsymbol{X}_c - \sqrt{r_e^2 - \|\boldsymbol{X}_c - \boldsymbol{X}_b\|^2} \cdot \boldsymbol{n}$, where $\boldsymbol{n}$ is the normal of the *eye plane* pointing in camera direction. For further refinement of the eye position we minimize the expression $\boldsymbol{F}(\boldsymbol{E}_0) = \int_{\boldsymbol{\Omega}} |\boldsymbol{I}(\boldsymbol{x}) - \boldsymbol{J}(\boldsymbol{x}'(\boldsymbol{E}_0))|\,\mathrm{d}\boldsymbol{x}$, where $\boldsymbol{I}, \boldsymbol{J}$ are the stereo images $\boldsymbol{x} \in \boldsymbol{\Omega}$ are the image points residing in the iris area of image $\boldsymbol{I}$ including some border pixels and $\boldsymbol{x}'(\boldsymbol{E}_0)$ are the coordinates of those pixels transfered via the surface of the eyeball onto the image plane of $\boldsymbol{J}$. For this task we employ the gradient descent approach $\boldsymbol{E}_{i+1} \leftarrow \boldsymbol{E}_i - \Delta\boldsymbol{F}(\boldsymbol{E}_i)$, where $\Delta\boldsymbol{F}(\boldsymbol{E}_i)$ denotes the gradient of our optimization target. Once the gradient descent step converged the gaze direction reads as the eyeball normal of the point that corresponds to the pupil center in image space obtained through iris segmentation.

6. SYNTHETIC CAMERA PLACEMENT

Naturally, the possibility for user interaction is limited regarding online applications like video communication. In the following we propose a novel method for the positioning of a virtual camera for eye contact provision. The required manual interaction is just a single orthogonal peek onto the display plane. The outcome provides the position of the virtual camera in space and the rendering position for the remote conferee on the screen.

For simplicity, we discuss the setup illustrated in figure 4, where the cameras are mounted on top of the display and a *reference camera* with spatial position C_r is in centered position. Other configuration with cameras below or beside the display can be handled analogously. To enable the calibration process the distance $\|C_r - D\|_2$ has to be measured physically in order to get the exact relative position of a *reference camera* and the display. This needs to be done only a single time as an offline pre-calibration step. Please note, that the point D indicates where the display is located and does not refer to the border of the monitor.

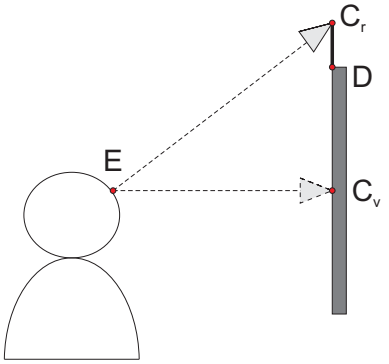

Fig. 4. Geometric illustration for virtual camera placement.

Once the orthogonal viewing direction of the conferee is available, the center of the virtual camera C_v can be computed by intersecting the viewing ray with the plane that contains C_r and is orthogonal to the gaze direction. Subsequently the virtual camera can be placed in opposite gaze direction as depicted in figure 4. As mentioned earlier we also need to know at which position the remote conferee should be rendered on the screen. Since the physical display size and its resolution are always available we use this information to compute suitable screen coordinates. Concretely, using the known space points D and C_v and the physical display properties we can transfer C_v to pixel values on the screen.

7. EXPERIMENTAL RESULTS

The systematic evaluation of our algorithm with respect to the quality of eye contact provision is hard to boil down to some easy measurable parameters. For the conference presentation we will prepare recordings of various synthesized eye contact sequences that allows the audience to evaluate the subjective quality of the feeling of *being looked at*.

In the following we only provide still images of typical results for our gaze estimation and the synthetic view based

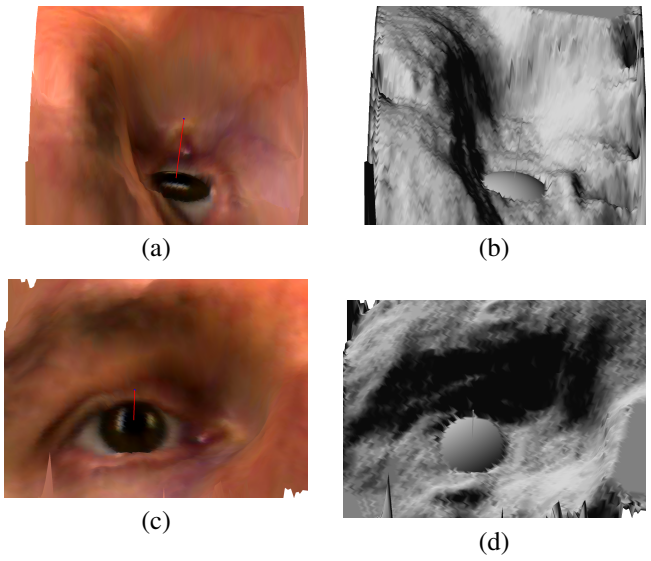

Fig. 5. 3D representation of an outcome of our stereo gaze estimation for one eye. (a) Textured 3D model at side perspective. The red arrow indicates the computed gaze direction. (b) 3D model, polygonal mesh representation, the same perspective as (a). (c) Textured 3D model from below. The red arrow indicates the computed gaze direction. (d) Polygonal mesh representation, same perspective as (c).

on the calibration of an eye contact providing video communication system where we integrated our algorithm. Our implementation is completely done in C++. On full HD input images the resulting runtime is about 35ms for the gaze estimation of one eye and 60ms for the cascade classifier. We use one thread for the classifier and one for each eye in order to minimize delay and processing time. Therefore, our method is suitable to interactively calibrate the video communication system and maintain user acceptance. In figure 5 a typical outcome for the stereo gaze estimation from two different perspectives is illustrated. The estimated gaze leads to a quite natural feeling of being looked at as depicted in figure 6. Moreover, due to the filter rules described in section 4 and the robustness of our algorithm we almost never observed wrong calibrations. At the same time the duration of the calibration process usually does not exceed two seconds.

8. CONCLUSION

In this paper we presented a novel algorithm for the calibration of a virtual camera used for eye contact provision and the correct placement of the virtual view on the display plane. It directly benefits from the target application by reusing the computed 3D information for gaze estimation. At the same time it effectively solves the calibration problem with almost no user interaction and at video frame rate. Moreover, we applied our algorithm successfully on the adjustment of an eye contact providing video communication system.

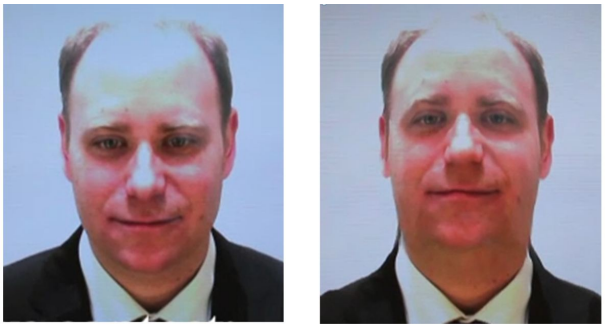

Fig. 6. Typical video conferencing setup, left) original view, no eye contact due to displacement angle between real camera and display, right) proposed method, virtual eye contact based on detection of eye position and gaze direction.

9. REFERENCES

[1] W. Waizenegger, I. Feldmann, and O. Schreer, "Real-time patch sweeping for high-quality depth estimation in 3D video-conferencing applications," in *Proc. of Real-Time Image and Video Processing*, San Francisco, California, United States, 2011.

[2] L. Muhlbach, M. Bocker, and A. Prussog, "Telepresence in videocommunications: A study on stereoscopy and individual eye contact," 1995.

[3] M. Ott, J. P. Lewis, and I. Cox, "Teleconferencing eye contact using a virtual camera," in *INTERACT '93 and CHI '93 conference companion on Human factors in computing systems - CHI '93*, Amsterdam, The Netherlands, 1993, pp. 109–110.

[4] J. Gemmell, K. Toyama, C.L. Zitnick, T. Kang, and S. Seitz, "Gaze awareness for video-conferencing: a software approach," *Multimedia, IEEE*, vol. 7, no. 4, pp. 26–35, 2000.

[5] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung, "GAZE-2: conveying eye contact in group video conferencing using eye-controlled camera direction," in *Proceedings of the conference on Human factors in computing systems - CHI '03*, Ft. Lauderdale, Florida, USA, 2003, p. 521.

[6] W. Waizenegger, N. Atzpadin, O. Schreer, and I. Feldmann, "Patch-Sweeping with robust prior for high precision depth estimation in real-time systems," in *Proc. International Conference on Image Processing (ICIP 2011)*, Brussels, Belgium, 2011.

[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001*. 2001, vol. 1, pp. I–511– I–518 vol.1, IEEE.

[8] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *2002 International Conference on Image Processing. 2002. Proceedings*. 2002, vol. 1, pp. I–900– I–903 vol.1, IEEE.

[9] "OpenCV," http://opencv.willowgarage.com/.

[10] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.