3D FACE CAPTURE FOR RAPID PERSON AUTHENTICATION

Markus Kettern[1], David Blumenthal-Barby[2], Wolfgang Paier[3], Frank Fritze[4]
and Peter Eisert[5]

{markus.kettern, david.blumenthal, wolfgang.paier, peter.eisert}@hhi.fraunhofer.de
[1, 2, 3, 5] Fraunhofer Heinrich Hertz Institut, Einsteinufer 37, 10587 Berlin
[2, 5] Humboldt University of Berlin, Unter den Linden 6, 10099 Berlin

[4] *frank.fritze@bdr.de*
Bundesdruckerei GmbH, Oranienstrasse 91, 10969 Berlin

Abstract

We propose a novel face capture system for security gateways which allows for inexpensive rapid automated or computer-aided face-based person authentication employing 3D head and face data. By face-based person authentication we refer to the process of comparing the appearance of a person to a visual representation of that person stored on a security document. This comparison can be done either manually or automatically and the data to be compared may be a standard facial image or a 3D representation of the person's head, depending on the capabilities of the security document. We propose algorithms for 3D reconstruction of persons passing the security gate as well as for head pose estimation and compensation to enable precise alignment of the 3D representation to be compared to the document. Furthermore, we show how eyeglasses affect reconstruction results and propose methods to compensate for these effects as on-going research.

Keywords: 3D Modelling, Gateway, Authentication, Border Security.

INTRODUCTION

Automatic and Computer Aided Rapid Person Authentication

Most person authentication is still done manually with an official comparing the appearance of a person to a standardized photograph contained on a security document. The reliability of this process is very high provided the security document is hard to forge or manipulate and the official is trustworthy. On the other hand, the speed of this process is limited and it requires human time – a resource becoming more and more valuable.

Automatic authentication systems on the other hand may provide scalable throughput at low cost if the capturing stage and the algorithms carrying out the authentication are robust enough to enable automatic authentication of most individuals. As of today, automatic systems struggle to be more effective than human officers.

To ease the work of the officials performing the authentication, computer-aided systems may provide them with images of the person to be authenticated that are adapted to the standard image on the document.

We propose a security gate system capable of rapidly capturing persons passing through it and creating novel views of these persons that are adapted to be easily compared to a standard representation like an image stored or depicted on the security document. More precisely, we propose to compensate the pose and lighting of the person's head and face in order to simulate what the person captured would look like if captured under the circumstances used to acquire the image stored on the document, e.g. frontal pose, even lighting, etc.

System overview

The proposed gateway is mounted with two video depth sensors and several high-resolution static cameras. A person stepping through the gate is detected and tracked by the video depth sensors. These sensors trigger the static cameras once the person has reached their volume of convergence. Also, the depth output of these sensors is integrated over time to generate a coarse 3D model of the person's head. This model is used as initialization for a 3D reconstruction algorithm which uses the high-resolution still images as input. Once the detailed 3D model of the head is completed, it can be used to render novel views of the head matching the pose in the authentication image stored / printed on the security document.

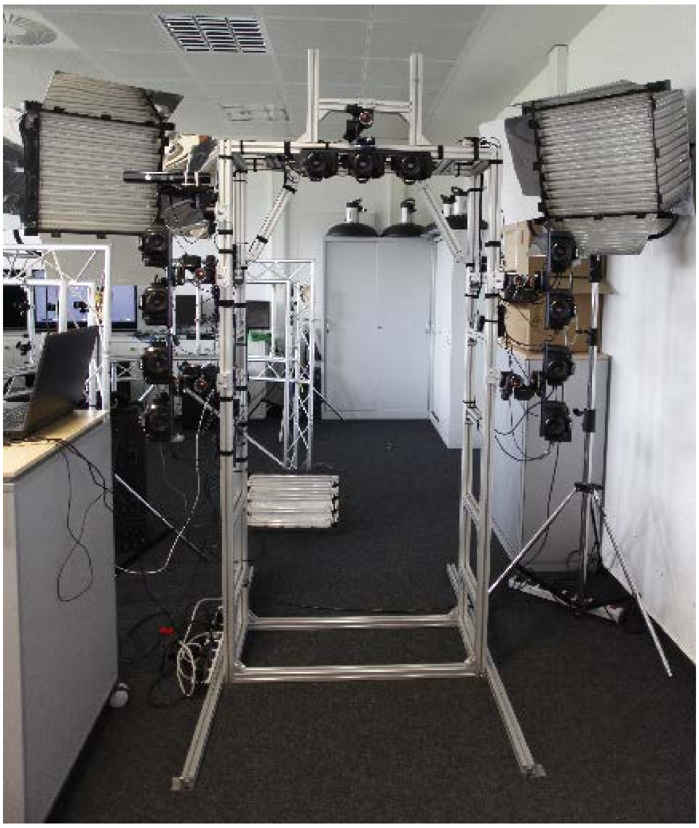

Fig. 1: Prototype of the proposed security gate

Related and previous work

Numerous approaches to automatic and semi-automatic person authentication and security gates are either published or being used in practice [1, 2, 3]. To enhance the robustness of automatic procedures, multi-modal approaches are commonplace [4].

Fast and accurate 3D reconstruction methods for authentication as described here can be used to enhance any authentication approach involving the comparison of a facial image to an image stored on a document or in a database.

We build our system on experience and work we have acquired in past projects covering 3D reconstruction [5], image segmentation [6] and camera calibration [7].

DEPTH SEONSORS FOR TRACKING, TRIGGERING AND INITIAL 3D MODELING

A person passing the security gate is first detected and subsequently being tracked in real-time by specific algorithms constantly evaluating the output of the video depth sensors. For the prototype, we used the easily available Microsoft Kinect which provides robust low-resolution depth maps without disturbing visible projections onto the person's face. When the person reaches the convergence volume of the static cameras, the tracking program triggers these cameras via a specially made triggering module to obtain a high-quality multi-view image set of the person. The output of the video depth sensors is also used to create an initial guess of the 3D shape of the person.

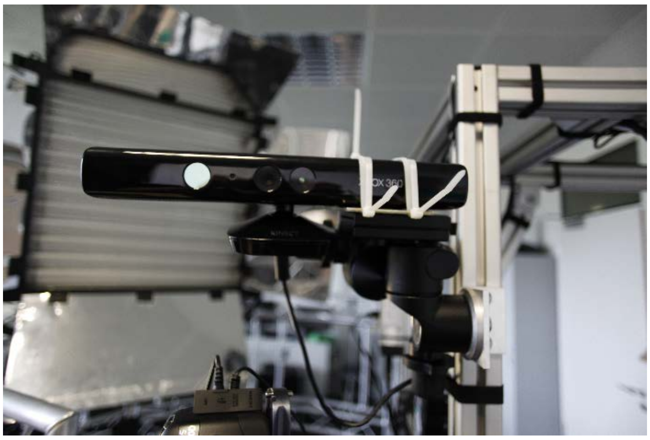

Fig. 2: Kinect video depth sensor mounted on the prototype

Person tracking

The open-source SDK [8] used to communicate with the Kinect video depth sensors allows us to exploit Kinect's highly robust body tracking capabilities at 30 fps. Also, the SDK provides methods to monitor the distance of the person being tracked. This enables the system to automatically trigger the static cameras once the person has reached the centre of the convergence volume.

Camera and depth sensor calibration

Each Kinect 3D video sensor consists of a colour video camera, an infra-red projector for projecting a structured pattern onto the surfaces viewed by the sensor and an infra-red camera for acquiring the distortions the surfaces introduce into the projected pattern, which are used to create the depth map. If we cover up the infra-red projector, this infra-red depth camera reproduces low-intensity grayscale images which can be used to calibrate it against other cameras, specifically against the static D-SLRs using standard correspondence-based calibration routines (e.g. [9], using a checkerboard with known geometric properties). The same method is also used to obtain the calibration of the two video depth sensors towards each other.

Initial 3D reconstruction from 3D video sensors

A coarse localization of the head in space can be obtained from the tracking results. The video depth sensors provide real-time depth maps which can be used together with the calibration information obtained above to create point clouds in 3D space. However, the outputs are of low resolution, both spatially and in the number of points. Furthermore, they exhibit several types of noise: Holes appear and disappear in objects from frame to frame; depth measurements are noisy; object boundaries tend to fray and depth discontinuities (like the person's chin in fig. 3) may produce shadow artefacts with no measurement at all. In order to overcome these limitations, we have developed a method of creating smooth, consistent 3D models by combining the synchronized outputs of the two depth video sensors and integrating them over time.

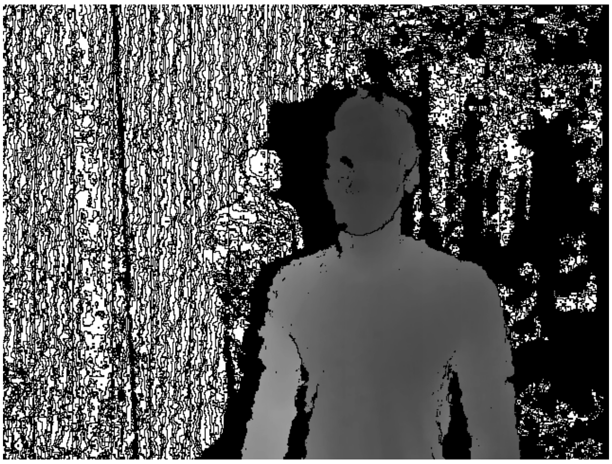

Fig. 3: Depth output of the Kinect Sensor

Synchronization of the two depth video streams is achieved by inferring the stream delay which minimizes the point-cloud difference in the overlap region of both sensors. Next, the merged point clouds from each frame within the timeframe to be integrated over are registered towards each other. This is done by using the *iterative closest points* method (ICP) [10] to find a rigid motion of one point cloud to minimize the distances between pairwise closest points. The registered point clouds are integrated and smoothed to obtain a smooth and clean 3D model of the captured face (fig. 4).

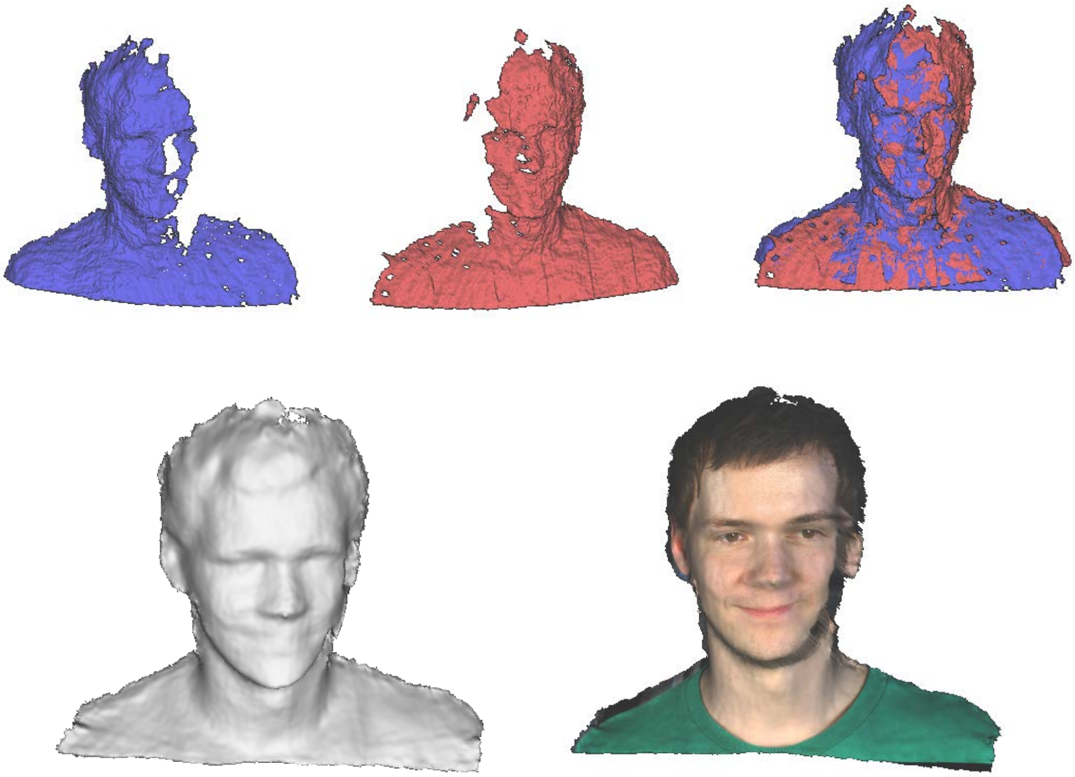

Fig. 4: Top: Fusion of two depth frames, bottom: Integrated depth video output without and with texture captured by the D-SLR cameras

Since the depth sensors are equipped with low-resolution RGB cameras as well, we can use the captured texture together with the calibration information to register these integrated 3D models to the images captured by the D-SLR cameras. This is done by estimating a rigid transformation of the 3D model to match one of the D-SLR still images. After an overall adaption of brightness and contrast between the Kinect RGB texture and the still image, pixel intensity differences can be used as matching error function and are optimized in the least squares sense using the Gauss-Newton algorithm.

Camera triggering module

Synchronous triggering of several D-SLR cameras is known to be a difficult task. Even in the case that all trigger signals are dispatched in the same time instant, each camera may exhibit an individual offset between receiving the signal and the actually triggering the shot.

To address this problem, a camera triggering module has been designed and constructed which allows defining an individual time offset for the trigger signal of each camera. The desired offsets can be measured using the same triggering module and capturing a high resolution timer with several simultaneously triggered cameras.

The triggering module provides one individual triggering channel for each camera and a triggering delay can be chosen for each channel. Thus it allows having all cameras record a shot at precisely the same time instant.

HIGH-RESOLUTION 3D MODELS USING STATIC CAMERAS

For the capture of detailed shots, the prototype comprises 11 static high-resolution D-SLRs. Three of these cameras are mounted in upside down landscape position above the gateway to capture a trifocal top-frontal view of the person. Four cameras are mounted one above the other in portrait position on either side of the gateway to provide lateral and bottom-up views of the face and flexibility towards differently tall persons.

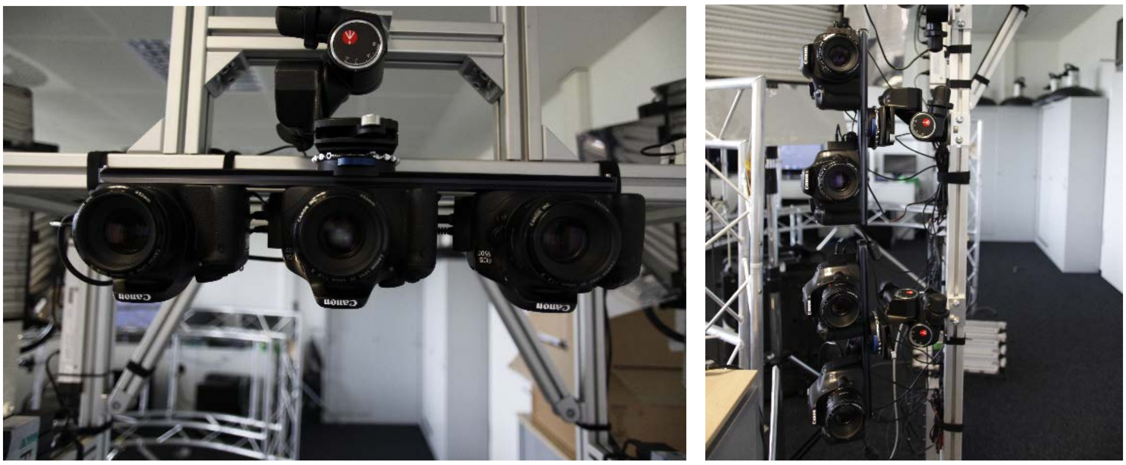

Fig. 5: D-SLR cameras mounted on the prototype

Since the person is moving while being captured at high resolution we have to aim at low shutter times. On the other hand, to keep the length of the capturing volume (depth of field) large enough, using high aperture values is advisable. The resulting requirements on balancing lighting and camera sensibility are relaxed however by the fact that today even of entry-level D-SLRs provide very good imaging characteristics in high ISO ranges.

Background and lighting

Several types of background can be chosen to enable clean segmentation of the images captured by the D-SLR cameras.

- White plates can be brightened up with extra lights to eliminate shadows and provide saturated white backgrounds, however, they interfere with bright hair
- Retro-reflective curtain can be used to dynamically adapt the background colour without visible changes in the illumination of the person. However, the LED ring-lights needed for each camera to provide lighting from the required angles represent a significant cost factor
- Neutral dark, static backgrounds may be easily extracted using background subtraction methods but introduce similar problems as white plates

We tested these background types and found that a segmentation algorithm is needed for all types (see section *Image Segmentation*).

To avoid problems for people sensitive to sudden lighting changes, we refrain from using flashlights and use floodlights instead. The significant brightness of these floodlights can be reduced when cameras capable of capturing with high ISO sensitivity are used.

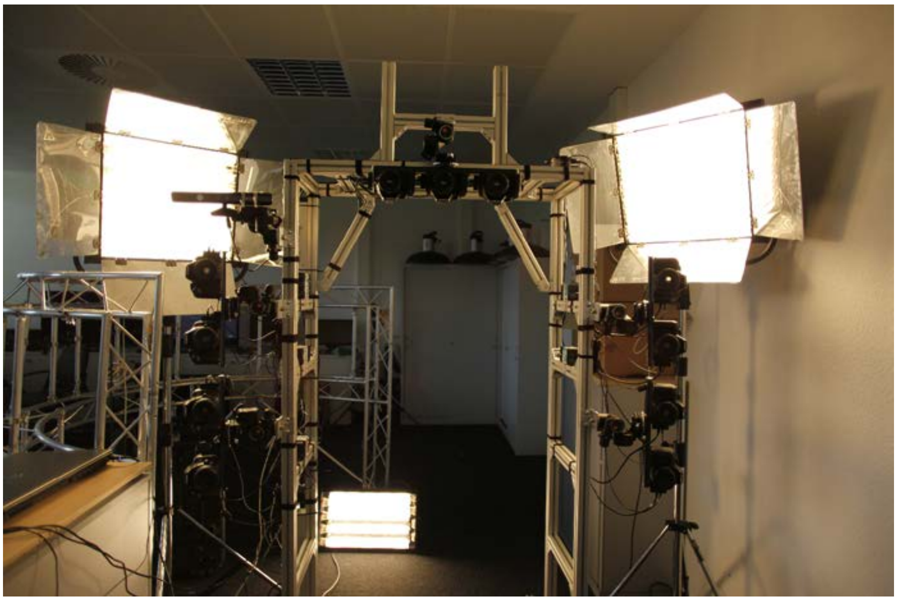

Fig. 6: Lighting setup used to create test shots

Image segmentation

As an input additional to the images, the algorithm creating the high-detail 3D reconstruction needs foreground-background segmentation masks of these images. These masks can be created for static environments by using background subtraction. We have devised a fast and reliable method for background subtraction relying on thresholding a difference likelihood image (calculated from intensity and hue differences) between a recording of the background and the actual image. The method identifies stationary points in the cumulative histogram such that the size of segmented foreground region lies within a predefined range [6].

If semi-transparent textures shall be used for rendering, they can be inferred from the input images using methods for digital alpha-matting [11]. The trimaps used as input to these methods can be computed consistently over multiple views from the binary segmentation masks and the camera calibration using approximate visual hulls [6]. A state-of-the-art matting approach has been extended to multi-view inputs by the authors in [12].

High resolution 3D reconstruction from static cameras

The 3D shape of the person’s head is reconstructed by densely matching a pair of input images (stereo) and later fusing the results of several stereo reconstructions. For stereo reconstruction, we establish a warp function that maps each pixel in one image to a pixel in the other image via an assumed depth for each pixel. This warp function is spanned by a triangle mesh and inferred by minimization of the intensity error between mapped pixels. High vertex counts enable detailed surfaces and a smoothness term relating adjacent depth values controls the smoothness of the reconstruction. The minimization is carried out by a high-performance sparse implementation of the iterative Gauss-Newton algorithm [5].

The stereo reconstruction results can be fused into one integrated model by rigidly attaching them to a morphable head model [13] and then deforming the morphable model to match the attached geometry slices. This yields a complete, closed 3D model with the mesh topology of the morphable model.

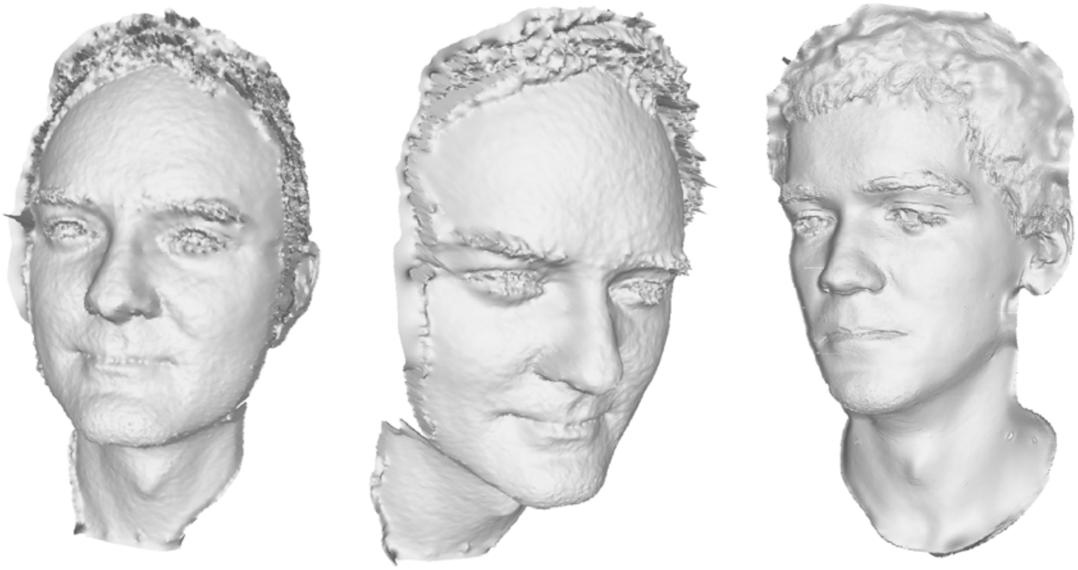

Fig. 7: Results of the high-resolution stereo 3D reconstruction

RENDERING OF STANDARDIZED VIEWS

Once a 3D model of the person's head is acquired, it can be used to render the recorded head images from any point of view. Several approaches can be taken to match a standard pose as desired. If the lighting situation of the capturing stage has been measured, the lighting of the images can also be estimated and compensated using the 3D model.

Pose compensation

The most precise way of compensating the pose for comparison between rendered view and document image is a rigid, image-based matching of the 3D model to the document image. We have implemented a method of minimizing the mutual information [14] between rendered and original image over the 6 parameters of a rigid motion of the 3D model. Mutual information is insensitive to lighting differences between the images and therefore advisable if the images are taken under different circumstances. On the other hand, since fusing the stereo reconstructions requires attaching them to a morphable head model, automatic localization of facial features such as the eyes is trivial and may also be used to find the best pose for a comparison.

Once found, the textured 3D model is rendered in this pose to yield the image to be compared to the image stored on the document.

RESEARCH IN PROGRESS: GLASSES

Glasses are worn constantly by more than 50% of Germany's population [15] so any method that aims at robustly processing people's faces under real-world conditions must be able to cope with people wearing glasses. However, their thin, often shiny frames and their biggest parts (the actual eyeglasses) being transparent and only visible from reflection and refraction make glasses extremely hard to reconstruct in 3D (see fig. 8 for an example).

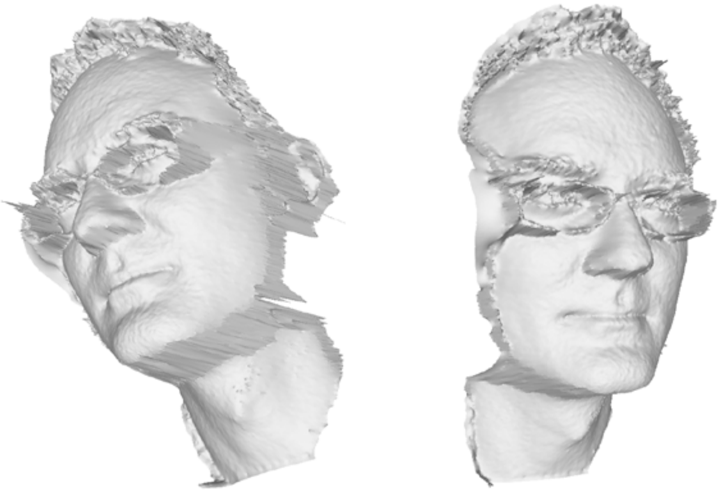

Fig. 8: Obvious difficulties in reconstructing the fine structure of the spectacle frame and subtle errors introduced by the refraction of the lenses (distorted eye regions)

We are currently researching methods of generating a geometric representation of glasses allowing for compensation of their refractive effects as well as rendering them into novel views, but can only give an outline of this research here since it is still very much in progress.

Detection and localization

We are developing an algorithm that detects glasses in a single image using active contours [16] and a suitable low-dimensional functional model representing the outline of each lens. Initialized with the output of a standard eye-tracker, this method will provide us with the precise form and location of the lens boundaries as well as an initial guess of the shape and colours of the frame. Once the outlines are localized in several images, we can compute the outline curve in 3D space.

Geometric representation and rendering

Spectacle Frames

If we want to create a 3D representation of the spectacle frame, the 3D curve of the lens outlines will provide a good initialization to volumetric approaches like [17] which yield the highest chance of success for reconstructing these thin and shiny objects. Furthermore, we can use the colour information gained from the detection to create the segmentation mask images typically used by these approaches. However, since all images are recorded from similar viewing directions, the representation created will be incomplete where view coverage is low. We will apply model knowledge in order to reduce these errors and robustly create a realistic representation of the spectacle frame.

Lenses

The reconstruction of transparent surfaces in front of an object with unknown shape and texture is pioneering work which has to our knowledge not yet been achieved by the computer vision community. We aim at exploiting visual correspondences between at least three calibrated views in an inverted ray-tracing framework [18]. Reflections will be detected and removed first with a correspondence map between the reflections in different views hinting at a coarse approximation of the frontal surface of the lens. Equipped with this initialization we will aim at jointly estimating the parameters of the lens model and the depth of the trifocal correspondences.

CONCLUSION

We have proposed a 3D face capture system enabling rapid automatic acquisition of 3D representations of people passing through a security gate and shown how they can be used to ease or automate the task of face-based visual person authentication. We have outlined several directions of on-going and future research we will conduct to achieve highly robust reconstruction and authentication results in real-life situations.

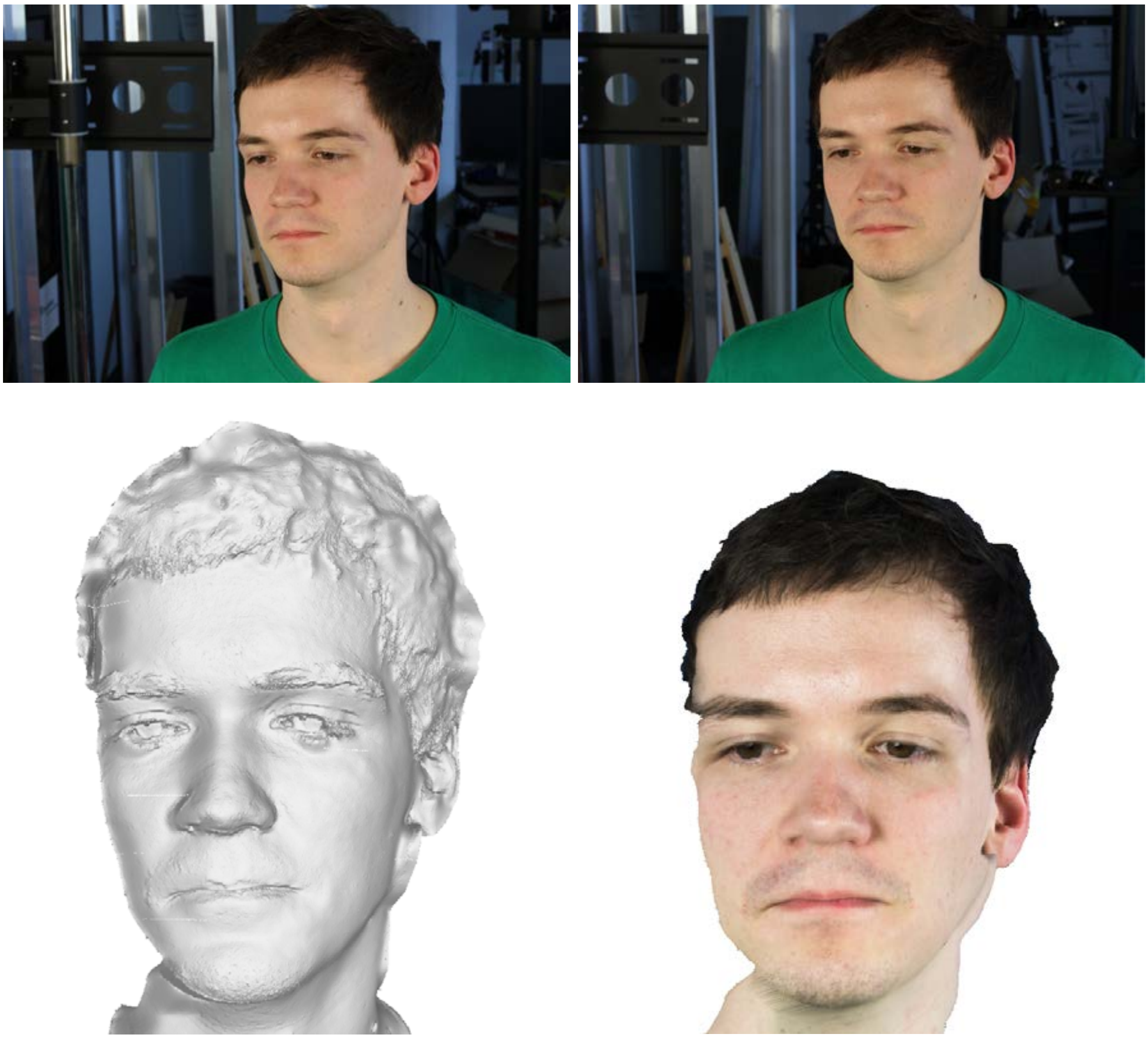

Fig. 9: Sample input, pose compensated stereo 3D mesh and textured model

ACKNOWLEDGEMENT

The work presented in this paper has been supported by the German ProFIT research project *FreeFace* (grant no. 10149078).

REFERENCES [Arial, 12-point, bold, left alignment]

Reference [Arial, 11-point, left alignment, upper and lower case]

[1] Tistarelli, M. and Grosso, E (2000). *Active vision-based face authentication*, Image and Vision Computing 18, pp 299-314

[2] Crawford, M. (2011) *Facial recognition progress report*, Defense and Security, DOI: 10.1117/2.2201109.01

[3] Beumier, C. and Acheroy, M. (1998). *Automatic Face Authentication from 3D Surface*, Proc. BMVC, pp. 449-458

[4] Arandjelović, O. et al (2007). *Towards Person Authentication by Fusing Visual and Thermal Face Biometrics*, Signals and Communication Technology: Face Biometrics for Personal Identification, pp. 75-90

[5] Schneider, D.C. et al (2011). *Deformable Image Alignment as a Source of Stereo Correspondences on Portraits*, NORDIA workshop, Proc. CVPR 2011

[6] Kettern, M., Schneider, D.C., and Eisert, P. (2011). *Multiple View Segmentation and Matting*, Proc. CVMP 2011

[7] Kettern, M., Prestele, B. and Eisert, P. (2010). *Recording Consistent Multiview Image Data*, Proc. 3DNordOst 2010

[8] http://www.openni.org/

[9] http://opencv.org/

[10] Zhang, Z. (1992). *Iterative Point Matching for Registration of Free-form Curves*

[11] He, K. et al, *A global sampling method for alpha matting*, Proc. CVPR 2011, pp. 2049-2056

[12] Kettern, M. and Eisert, P. (2011). *Multiview-consistent Color Sampling for Alpha Matting*, Eurographics poster, 2012

[13] Blanz, V. and Vetter, T. (2003). *Face Recognition Based on Fitting a 3D Morphable Model*, IEEE PAMI, 25, 9, pp. 1063-1074

[14] Pluim, J. et al (2000). *Image registration by maximization of combined mutual information and gradient information*, IEEE Trans. on Medical Imaging, pp. 809-814

[15] http://newsroom.sehen.de/allensbach-brillenstudie/
http://www.zva.de/studien/

[16] Chan, T.F. (2001). *Active Contours without Edges*, IEEE Trans. on Image processing 10, 2, pp. 266-277

[17] Laurentini, A. (1994). *The Visual Hull Concept for Silhouette-Based Image Understanding*, IEEE PAMI, 16, 2, pp. 150-162

[18] Ben-Ezra, M. and Nayar, S.K. (2003). *What Does Motion Reveal About Transparency?*, Proc. ICCV 2003, pp 1025-1033