Fast 3D Reconstruction of Faces with Glasses

Fabio Maninchedda[1] Martin R. Oswald[1] Marc Pollefeys[1,2]
[1]Department of Computer Science, ETH Zurich [2]Microsoft

Abstract

We present a method for the fast 3D face reconstruction of people wearing glasses. Our method explicitly and robustly models the case in which a face to be reconstructed is partially occluded by glasses. We propose a simple and generic model for glasses that copes with a wide variety of different shapes, colors and styles, without the need for any database or learning. Our algorithm is simple, fast and requires only small amounts of both memory and runtime resources, allowing for a fast interactive 3D reconstruction on commodity mobile phones. The thorough evaluation of our approach on synthetic and real data demonstrates superior reconstruction results due to the explicit modeling of glasses.

1. Introduction

In this paper, we target the mobile 3D reconstruction of human faces that are possibly covered with glasses. Fig. 1 provides an overview of our method. The instant creation of 3D face reconstructions has a number of applications enjoying growing demands, such as the creation of 3D selfies that can be immediately viewed or printed in 3D, the creation of 3D avatars for augmented or virtual reality applications, content creation for games or movies, or 3D face authentication.

Challenges of Mobile Face Reconstruction. In contrast to many other 3D reconstruction methods, which often rely on a controlled image acquisition setup, in mobile face reconstruction we are confronted with a large number of additional difficulties like significant variations in lighting conditions, higher amounts of noise due to low cost sensors, motion blur, rolling shutter distortions, image artifacts due to dirty lenses, non-rigid deformations of the face during scanning, and finally also fewer computing resources.

Importance of Modeling Eyeglasses. Surprisingly, only few works in the face reconstruction literature address the topic of face reconstruction in the presence of glasses even though a large number of people wear them because they require a visual aid. Modeling glasses explicitly is beneficial for many applications. An important example is face authentication. In [46] Sharif et al. have shown that an attacker wearing eyeglasses can easily fool a state-of-the-art 2D authentication system into believing that he is another individual. Such a simple attack would not work for 3D face authentication systems because they heavily rely on the 3D shape, but this study highlights two important facts. First, the expressiveness and generality of pure 2D approaches is limited. Second, glasses that cover significant portions of the face can have a big impact on authentication systems and hence deserve to be modeled separately. This is certainly also true for 3D face authentication. In this paper we try to fill this gap and propose a robust multi-view 3D reconstruction approach for faces and eyeglasses which has interactive processing times on current commodity mobile devices.

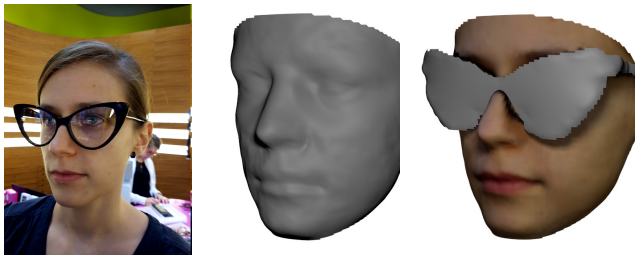

Figure 1: Result obtained using our approach. From left to right: Input image, result mesh, textured result with reconstruction of the glasses. Our approach first segments the eyeglasses to allow for a plausible reconstruction of the face while ignoring the depth values measured on the glasses.

1.1. Related Work

Our method takes advantage of several computer vision methods like image segmentation, generic 3D reconstruction, and statistical shape models. In a 3D setting there are only few works that address the segmentation and modeling of eyeglasses, but there are a number 2D approaches for which we also give a brief overview.

Eyeglass Detection and Segmentation. In [54], eyeglasses can be detected, localized as well as removed from a single input image. In [6], Fourier descriptors are used describe the boundary of the lenses and a genetic algorithm is used to

extract their contours. [17] proposes an algorithm for the detection of glasses via roughly aligned bounding boxes rather than pixel-accurate segmentations. Both works [6, 17] also give a very good overview on state-of-the-art glasses detection methods from single images. It is important to emphasize that these works target a different problem as they operate on a single color image and have no depth information which makes the accurate segmentation of the glasses more challenging.

Generic Multi-view 3D Reconstruction Algorithms. Both faces and glasses can also be reconstructed with general 3D reconstruction algorithms. The vast majority of multi-view approaches first estimate the camera parameters in a separate pre-processing step via feature point-based structure-from-motion techniques [44]. Consecutively, depth maps are usually estimated from two or multiple images for which many methods exist. A good overview of stereo estimation methods can be found in established stereo benchmarks [32, 43].

For the fusion of multiple depth maps into a joint 3D model, various approaches have been proposed, with the most prominent ones being volumetric approaches fusing occupancies [23, 25, 27, 28, 55], signed-distance functions [14, 33, 52, 56], or mesh-based techniques [15, 18, 20]. Some of these methods have also been tuned for interactive 3D reconstruction on mobile devices [26, 34, 45, 50]. In [29], a generic semantic 3D reconstruction approach was adopted for face reconstruction by using implicit face-specific shape priors to reconstruct skin, hair, eyebrows and beard as separate semantically labeled geometries.

3D Morphable Models (3DMMs). Since the amount of noise and outliers for generic 3D reconstruction techniques is generally high, a variety of approaches tailored specially to face reconstruction have been proposed. A popular approach are blend shape models or 3D morphable shape models [1, 5, 21, 36, 37] which represent human faces as a linear combination of eigenvectors coming from an eigendecomposition of a shape database. Additionally, these approaches have been extended to deal with variations in facial expressions [2, 8, 9, 22, 42, 51, 53].

3DMMs with Occluders. Both methods [16, 47] jointly estimate an occlusion map and align a statistical face model to a single input image within an EM-like probabilistic estimation process. Although they can handle arbitrary types of occlusion the segmentation is either sensitive to illumination or color changes [47], or they require substantial color differences between occluders and the face [16]. This is problematic because real occluders can be mislabeled as skin in case of similar colors which can lead to distortions in the model adaptation. The method in [9] aims at reconstructing a 3D face with unknown expression that is occluded by a head-mounted device. The method works in real-time, but requires a pre-scanned 3D model of the target person and also takes advantage of the device's inertial measurement unit for estimating the head pose.

Hybrid Approaches. While explicit face models greatly deal with missing data and large amounts of noise, they fail to recover instance-specific details like wrinkles, moles, dimples, or scars. Several approaches add these details back in a separate processing step [10, 49]. On the other hand, generic 3D reconstruction methods are able to recover such details but work much worse for higher amounts of noise, outliers or missing data. Recently, in [30] a hybrid approach was proposed in which a probabilistic face model is used to denoise a generic height map-based 3D reconstruction. Especially due to the method's low computation time and competitive output quality, we adopt their idea, but incorporate the explicit modeling of eyeglasses.

Generally, a separate modeling of occluders is not a new idea and has already been shown to be beneficial [16, 47]. However, to the best of our knowledge this is the first paper addressing explicit occlusion modeling for face reconstruction in the presence of depth information.

1.2. Contributions

The contributions of the system presented in this paper can be summarized as follows:

- We present a system that fully automatically reconstructs a human face and a rough 3D geometry of eyeglasses on a mobile phone using only on-device processing. This is achieved by detecting and segmenting the glasses prior to the reconstruction.
- We propose a general variational segmentation model that can represent a large variety of glasses and which does not require a database for learning or model retrieval.
- We show that the solution of the segmentation problem can be efficiently minimized or approximated by solving a series of 2D shortest path problems.

1.3. Problem Setting and Notation

We assume that a face is scanned by n colored input images $\{I_i\}_{i=1}^n$ for which corresponding depth maps $\{D_i\}_{i=1}^n$ are computed with a block-matching approach and corresponding camera calibration parameters $\{P_i\}_{i=1}^n$ are obtained with a feature point-based structure-from-motion approach [26, 50]. To obtain a first alignment between the input depth and the statistical shape model we perform landmark detections on each frame using the method presented in [41]. The 3D position of the face is estimated using the 2D landmarks and the camera calibration. We follow the approach presented in [30] and represent the face with a 2.5D height map. Instead of using the unified projection model [4, 19, 31] we use a cylindrical mapping which naturally fits the shape of a human face very well. A point

on the face $\mathbf{X} = (X, Y, Z)^\top$ is projected to the image point $\mathbf{x}$ using a cylindrical projection. The distance to the point that is stored in the height map H is simply given by $H(\mathbf{x}) = \sqrt{X^2 + Z^2}$, where we assume that the cylinder is at the origin and that the Y-axis is the height axis of the cylinder. The authors of [30] propose a simple way of integrating the depth into the height map representation by simply storing the mean distance of the observed distance along the ray. However, in the presence of glasses, certain rays will intersect the skin as well as the frame of the glasses and the mean will not represent a good estimate of the distance. To overcome this problem we use a cylindrical cost volume where we discretize each ray into a fixed number of bins. The depth samples are then stored in the corresponding bin along the viewing ray and the distance of the bin with the largest number of aggregated depth votes is stored in the height map. The corresponding texture image I is computed by averaging the color information from the images that have been captured from the most frontal viewpoint with respect to the projection direction. The advantage of using the projection direction instead of the surface normal to select the cameras is that it leads to a texture with less seams as the color information of neighboring pixels will come mostly from the same images.

2. Segmentation of Eyeglasses

After computing the face height map H and the corresponding texture information I our goal is to detect and segment potentially existing eyeglasses.

2.1. Eyeglass Detection

The first step of our approach is to detect whether glasses are present in a given input scan. As previously stated in the related work, eyeglass detection is a difficult problem on a single color image, but the additional height map values provide useful information that simplifies the detection significantly. For all points $\mathbf{x}$ inside a small rectangular region Ω_{ROI} around the eyes, we sum up all gradient magnitudes of the height map H, and found a single threshold θ that separates subjects wearing glasses from those not wearing glasses in our dataset. That is, we detect the existence of glasses in the height map if

$$\int_{\Omega_{\text{ROI}}} \|\nabla H(\mathbf{x})\|_2^2 \, d\mathbf{x} > \theta \,. \tag{1}$$

A classifier which further includes location and color information would be more powerful, but was not necessary in our case.

2.2. Eyeglass Segmentation

We propose a generic eyeglass model which does not need any learning, but only relies on a few simple assumptions upon existence of glasses in the input data:

1. **Appearance / Depth:** We assume that the glasses differ from the face either in color appearance, or in reconstructed depth values. In most cases both modalities indicate the shape and color of the glasses' frame.
2. **Connectivity:** To deal with large amounts of noise and outliers, we assume that the frame is a connected surface (of arbitrary shape) from the left to the right ear.
3. **Location:** We assume that the inner eye landmark points are covered by the glasses.
4. **Symmetry:** The vast majority of glasses are symmetric with respect to left-right reflection along the center.

In the following we will phrase these assumptions in mathematical terms and propose to find the segmentation as a minimizer of a variational energy term.

Variational Segmentation Model. Considering the information from the height map H and corresponding texture image I, we obtain a segmentation of the eyeglasses by computing an unknown indicator function $u : \Omega \to \{0, 1\}$ which is defined on the same domain $\Omega \subset \mathbb{R}^2$ as the inputs H and I. For better readability we introduce shorthand notions for the foreground set $\Omega_{u=1} := \{\mathbf{x} \in \Omega \mid u(\mathbf{x}) = 1\}$ and the horizontal domain boundary $\partial_x\Omega := \{\mathbf{x} \in \partial\Omega \mid \forall y : \mathbf{x} = (0, y) \vee \mathbf{x} = (x_{max}, y)\}$.

We enforce the foreground set to reach from the left ear $\mathbf{x}_l$ to the right ear $\mathbf{x}_r$ by using connectivity constraints that can be efficiently imposed as single-source tree shape priors [48] or as single pair connected path [35] by additionally enforcing both starting points to be in the foreground set. These constraints can be efficiently enforced by linear constraints defined on a precomputed tree of shortest geodesic paths. In particular, we require that there exists a connected path $C(\mathbf{x}_l, \mathbf{x}_r)$ from pixels $\mathbf{x}_l$ to $\mathbf{x}_r$ to be entirely within the foreground set. The segmented image can then be computed as the minimizer of the following optimization problem:

$$\begin{aligned} \underset{u}{\text{minimize}} \quad & \int_\Omega \Big(\lambda f u + \phi(\nabla u)\Big) \, d\mathbf{x} \\ \text{subject to} \quad & \exists C(\mathbf{x}_l, \mathbf{x}_r) \subset \Omega_{u=1} \cup \partial_x\Omega \\ & u(\mathbf{x}_l) = u(\mathbf{x}_r) = u(\mathbf{x}_p) = u(\mathbf{x}_q) = 1 \ , \end{aligned} \tag{2}$$

where the constraints on $\mathbf{x}_l, \mathbf{x}_r$ ensure the connected two-point path (Assumption 2) and the constraint on the landmarks points $\mathbf{x}_p$ and $\mathbf{x}_q$ enforce their occurrence in the foreground set (Assumption 3) - see Fig. 2 for examplary landmark locations. The appearance properties (Assumption 1) can be expressed as combination of the regional term f, e.g. via a log-likelihood ratio of appearance probabilities $f = -\log \frac{P_{\text{fg}}}{P_{\text{bg}}}$, or within the regularizer $\phi(\cdot)$.

A typical choice for the regularizer is a weighted total variation term $\phi(\nabla u) = g\,|\nabla u|_2$ in which function $g : \Omega \to \mathbb{R}_{\geq 0}$ controls the local smoothness [7]. However, we use a

more powerful anisotropic regularization function [40]:

$$\phi(\mathbf{p}) = \sqrt{\mathbf{p}^T \Sigma_{\mathbf{p}} \mathbf{p}} \qquad \Sigma_{\mathbf{p}} = \Sigma^I_{\mathbf{p}}(\mathbf{n}_I) \Sigma^{Hu}_{\mathbf{p}}(\mathbf{n}_{Hu}) \Sigma^{Hl}_{\mathbf{p}}(\mathbf{n}_{Hl}). \tag{3}$$

The matrix $\Sigma_{\mathbf{p}}$ consists of three parts, one for the texture image and two for the height map, the former one, being defined as

$$\Sigma^I_{\mathbf{p}}(\mathbf{n}) = g_I(\mathbf{x}, \mathbf{p}, \mathbf{n})^2 \mathbf{n}^T \mathbf{n} + \mathbf{n}^T_{\perp} \mathbf{n}_{\perp} \ , \tag{4}$$

being a symmetric non-singular square matrix which favors gradients to align with a given normal direction $\mathbf{n}_I = \nabla I / |\nabla I|_2$ that we extract from the image and $\mathbf{n}_{\perp}$ is the vector perpendicular to $\mathbf{n}$. We choose the image-based weighting function as

$$g_I(\mathbf{x}, \mathbf{p}, \mathbf{n}) = \exp[-\lambda_I |\nabla I(\mathbf{x})|_2] \ . \tag{5}$$

In combination with Eq. (3), this cost function has an ellipsoidal shape of magnitude one in the tangential direction and $g_I(\cdot)$ in the normal direction. Note, that the costs function is symmetric with respect to the sign of the image gradient, which is a desirable property since we want to align with the image gradient direction regardless whether the color of the glasses is brighter or darker than the skin color - see, e.g. Fig. 2 in which the frame has both darker and brighter color than the skin.

In contrast, for the alignment of the labeling function with the gradients in the height map we want an asymmetric cost function, because we know the dominant direction of the height maps gradients: Since the glasses are always in front of the face, we know that the depth gradient for the upper glass boundary is positive and, respectively, negative for the lower boundary.

We therefore use more general cost functions, called Wulff shapes [57], which can be both anisotropic and non-symmetric. We use a weighted anisotropic ellipsoidal shape for the positive normal direction and a circular shape for the negative one:

$$g_H(\mathbf{x}, \mathbf{p}, \mathbf{n}) = \mathbf{1}_{[\mathbf{p} \cdot \mathbf{n} \leq 0]} + \mathbf{1}_{[\mathbf{p} \cdot \mathbf{n} > 0]} \exp[-\lambda_H |\partial_y H(\mathbf{x})|] \, . \tag{6}$$

This term evaluates only vertical gradients ($\partial_y H$), because they are dominant features along the glass outlines in the depth maps for which we know the sign of the vertical direction. The height map in Fig. 2 shows that the vertical component gradient is always pointing upwards for the upper segmentation boundary and downwards for the lower one. Without this signed directional cost term, the segmentation boundary would often follow the depth gradients along the frame interior (which are directed in the opposite direction). Therefore, we define the matrices for height map cost functions $\Sigma^{Hu}_{\mathbf{p}}(\mathbf{n}_{Hu})$, $\Sigma^{Hl}_{\mathbf{p}}(\mathbf{n}_{Hl})$ exactly as in Eq. (4), but with weight function (6) using upward and downward pointing

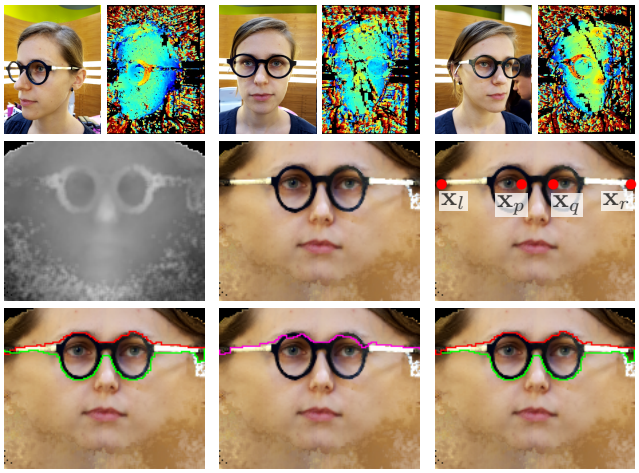

Figure 2: **First row**: example of multiple input images and corresponding depth maps computed on a mobile phone. **Second row** (left to right): height map distances, height map texture, inner eye landmark points. **Third row** (left to right): upper and lower shortest paths computed without boundary path constraint, boundary path, upper and lower shortest paths with boundary path constraint. Here, all results are computed without the symmetry constraint.

Wulff shapes with fixed normals for the upper and lower segmentation boundary, respectively:

$$\mathbf{n}_{Hu} = (0, 1)^T \qquad \mathbf{n}_{Hl} = (0, -1)^T \ . \tag{7}$$

We found experimentally, that such a regularizer provides strong features for the segmentation and makes a regional color-based term like in [16] unnecessary. The color of the human face among different people and under different lighting conditions spans a large region in the color space and a color-based segmentation is therefore either very sensitive to such changes or not very discriminative. We hence assume equal labeling likelihoods $P_{\text{fg}} = P_{\text{bg}}$ in every pixel, for which the regional term then vanishes ($f = 0$). Unfortunately, this segmentation model has two disadvantages:

1. **Speed:** Although the connectivity constraints do not have a large influence on the numerical optimization of (2), state-of-the-art algorithms like [38] still require several hundred iterations to converge.
2. **Solvability:** The efficient computation with connectivity constraints [48] can only be applied to isotropic regularizers, because the optimal shortest path tree can only be precomputed for costs that do not depend on the labeling.

We circumvent these drawbacks by leveraging the special structure of our segmentation problem.

Efficient Optimization via Shortest Paths. Since the connectivity constraints always connect the left domain bound-

ary with the right one, the foreground set boundary always consists of an upper connected path C_u and a lower connected path C_l which separate the upper and lower background region from the foreground, respectively. Without data fidelity term ($f = 0$) only the regularizer defines the pixel-wise cost $c(\mathbf{x}) = \phi(\nabla u(\mathbf{x}))$ and then both, the upper and lower boundary of the foreground set is defined by the shortest path through this cost volume. For an isotropic regularizer $\phi_u(\cdot)$ the costs for upper and lower foreground set boundary are the same and the foreground region collapses to a single connected path through the image - exactly the geodesic path which is precomputed in [48] - which is not useful for our setting. With the directional-dependent cost function in Eq. (6) we obtain two different cost functions which encode that the upper boundary contains positive depth gradients and the lower boundary contains negative ones. We simply compute the two boundary paths with Dijkstra's algorithm on each of the two different cost volumes. The upper and lower boundary paths are shown in Fig. 2 in red and green respectively.

Boundary Path Dependency. Unfortunately, the computation of upper and lower boundary paths is in general not independent. If they do not cross each other, we have found the optimal solution and we are done. This happened in the majority of our experiments. If they cross each other, a simple strategy to prevent the crossing is an iterative algorithm that takes out one edge at a detected crossing and then recomputes both shortest paths. This can then be iterated until no crossings are detected anymore or the graph is disconnected and there is no solution. There are several algorithms available which make shortest path re-computations after changing a single edge in the graph more efficient, e.g. DynamicSWSF-FP [39], or the so-called Lifelong Planning A* or Incremental A* [24].

However, since we are seeking high efficiency and we again leverage the special graph structure and propose a simple and effective heuristic. We assume that there exists a boundary path between upper and lower which bounds the domain which each path can traverse, that is, the upper path lies on or above the boundary path and the lower path lies on or below it, respectively. The problem is that we do not know a priori where this boundary is located. If we detected a path crossing we assume that the boundary path traverses the crossing points. We calculate the boundary path on the point-wise minimum of upper and lower path cost only on the domain between the previously computed upper and lower paths.

After we have obtained the boundary path, we recompute upper and lower paths on the respective domains restricted by the boundary path. The boundary path is shown in magenta in Fig. 2. This way, we get a guaranteed cross-free solution with either two or five iterations of Dijkstra's algorithm, for which two of the latter optional three iterations

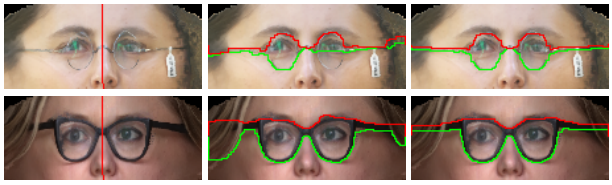

Figure 3: Segmentation results for two datasets with and without symmetry constraints. From left to right: texture image with detected symmetry axis, result without symmetry constraint, result with symmetry constraint.

are computed on only half the domain. Unfortunately, in this case we cannot guarantee to find the global solution for problem (2), but we found experimentally that the proposed heuristic does exactly what we want and more importantly this case occurred only rarely in our experiments.

Symmetry Constraints. Due to the symmetric nature of glasses it is possible to improve the segmentation by enforcing symmetry constraints. The idea is to flip and average the per-pixel costs along the symmetry axis of the glasses. Unfortunately, flipping the per-pixel costs along the vertical image axis is not accurate enough because of potential errors in the face alignment and due to the fact that the glasses are not always worn perfectly horizontally. Therefore, we optimize for an in-plane rotation $\mathbf{R}$ and translation $\mathbf{t}$ via:

$$\underset{\mathbf{R},\mathbf{t}}{\text{minimize}} \sum_{\mathbf{x}} w(\mathbf{x}) \left[\bar{H}(\mathbf{R}\mathbf{x}+\mathbf{t}) - H(\mathbf{x})\right] \quad , \qquad (8)$$

where $\bar{H}$ denotes the height map flipped around the vertical image axis and $w(\mathbf{x}) = (1 - \exp[-\lambda_I |\nabla I(\mathbf{x})|_2])(1 - \exp[-\lambda_H |\nabla H(\mathbf{x})|_2])$. This problem can be efficiently and robustly optimized using gradient descent based image alignment algorithms [3]. The final per-pixel cost for the boundary path computation is given by $c'(\mathbf{x}) = 0.5\,[c(\mathbf{x}) + \bar{c}(\mathbf{R}\mathbf{x}+\mathbf{t})]$. In Fig. 3 we show two example segmentations where the symmetry constraint is beneficial.

3. 3D Face Reconstruction

The 3D face reconstruction approach presented in [30] neglects eyeglasses, but forms the basis of our work and is therefore briefly repeated within this section. The method consists of 3 steps, namely an alignment optimization, a model fitting step and a height map optimization that regularizes the difference of the height map to the model fit.

Alignment Optimization. We found that the alignment optimization that uses a truncated cost function is very robust and is able to align faces well even in presence of glasses. The idea is to look for a similarity transform that minimizes the truncated sum of absolute differences between the height map that needs to be aligned and a reference height

Figure 4: Segmentation results for a variety eyeglass models and subjects. In the majority of cases, our segmentation model delineates the shape of the glasses very well. The segmentation is depicted by the upper (red) and lower (green) boundary path. Only in very challenging cases like transparent or frameless glasses the segmentation fails (bottom right). However, in this case the consecutive 3D reconstruction will be barely affected because of the missing depth values along the glass frame.

map of the mean face of the statistical model. More details can be found in [30].

Model Fitting. Let $\mathbf{f} \in \mathbb{R}^{WH}$ be a vector built by stacking all the pixels of a height map H on top of each other. W and H denote the width and height of the height map, respectively. By applying a covariance based PCA to a mean normalized data matrix $\mathbf{D} = [\mathbf{f}_1, \ldots, \mathbf{f}_p]$ we obtain a statistical face model $\mathcal{F} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{U})$, where $\boldsymbol{\mu}$ denotes the mean, $\boldsymbol{\sigma}$ denotes the standard deviation and $\mathbf{U}$ is an orthonormal basis of principal components [5, 37]. Fitting such a model to a vectorized height map $\mathbf{f}$ amounts to finding coefficients $\boldsymbol{\beta}$ such that $\mathbf{f} = \boldsymbol{\mu} + \mathbf{U}\,\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{\beta}$. In this work we estimate the coefficients by minimizing

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \;\; \|\mathbf{U}\,\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{\beta} - (\mathbf{f} - \boldsymbol{\mu})\|_2^2 + \|\alpha\boldsymbol{\beta}\|_2^2 \;\;, \quad (9)$$

where α is a parameter that pulls the fit towards the mean. This is a commonly used strategy that prevents overfitting [5]. While this fitting method differs slightly from [30] the big difference is that we fit the model only on regions that are not part of the eyeglasses. If the eyeglasses are ignored the model fitting is prone to cause degenerations in the facial geometry.

Height Map Regularization. Low dimensional parametric models cannot represent instance specific variations such as moles, dimples, scars, or wrinkles. We would like to bring back details that are present in the depth maps but have not been captured by the model. Furthermore, since part of the face is covered by glasses, we would like to use the model to get a plausible reconstruction of the occluded area. As in [30] we regularize a residual R obtained by subtracting the fitted model $H^{\mathcal{F}}$ from the height map H. For all the pixels belonging to the glasses we set the residual to zero. This is equivalent to a fill-in guided completely by the fitted model. The regularized residual $\mathbf{u}$ is obtained by optimizing

$$\underset{\mathbf{u}}{\text{minimize}} \;\; \sum_{i,j} \|\nabla \mathbf{u}_{i,j}\|_\epsilon + \lambda \,\|(\mathbf{u}_{i,j} - R_{i,j})\|_2^2 \;\;, \quad (10)$$

where $\lambda \in \mathbb{R}_{\geq 0}$ is a weight that trades smoothness against data fidelity and $\|\cdot\|_\epsilon$ denotes the Huber norm [11]. The final reconstruction H is obtained by adding the regularized residual to the fitted model $H = H^{\mathcal{F}} + R$.

Eyeglass reconstruction. While we use Eq. (10) for optimizing the face geometry and ignore the depth measurements within the segmented glasses, we use the same energy (10) for computing a geometry of the glass frame. We optimize the height map only within the segmented glasses and only use depth values close to the eyeglass segmentation boundary which mostly contain depth measures from the frame and set all other data fidelity values $R_{i,j}$ to zero, i.e. we effectively solve a depth inpainting problem [12]. For better consistency, we leverage the previously detected symmetry by averaging the geometry of the glasses.

4. Experimental Evaluation

We conducted a series of experiments both on synthetic and real data in order to evaluate our segmentation method, and the subsequent reconstruction with respect to 1) variations of the human face, 2) variation of the eyeglass shape, and 3) variations of the noise level on the depth maps. Furthermore, there are variations of the lightning in the dataset, because the scans have been acquired in various locations. All the test data has been either captured with a Samsung Galaxy S7 or with a Motorola Nexus 6 phone.

Segmentation Evaluation. In Fig. 4 we evaluate the segmentation of the height maps on a variety of eyeglass shapes and human faces. The red and green paths depict the upper

Figure 5: Examplary hand-labeled ground-truth segmentations for some of the datasets of Fig. 4.

and lower segmentation boundary of the glass segmentation. Our segmentation approach robustly segments the majority of eyeglass shapes. The major difficulty are frameless glasses, as depicted at the bottom of the last column, because there is very little evidence in both the height map values as well as the color image. We also performed a quantitative evaluation of the segmentation accuracy for the datasets of Fig. 4 using hand labeled ground-truth segmentations (see also Fig. 5). The average intersection-over-union score is 0.84 (0.90 when evaluated on the central region $\{x : 30 \leq x \leq 120\}$ for a height map width of 150). Without symmetry constraint the scores are 0.81 (0.88).

Robustness and Accuracy Evaluation on Synthetic Data. Due to the difficulty of acquiring ground truth face models we have performed a synthetic evaluation in which we have augmented 3D models of faces with glasses. For each model we have rendered 15 depth maps and texture images with and without glasses from varying viewpoints that cover the face area well. The depth maps have been corrupted by removing 50% of the data and by adding increasing levels of zero mean Gaussian noise. To evaluate the accuracy of the reconstruction we use the distance measure proposed in [13]. As we can see in Fig. 6 and Tab. 1 our approach copes well with noise and yields plausible reconstructions that are visually very similar to the reconstruction results on the model without glasses. The error magnitude in the occluded areas is bigger but that is fine as long as the reconstruction is visually pleasing.

$\sigma =$	0.0	2.0	4.0	6.0	8.0
no glasses (avg)	0.1	0.2	0.3	0.5	0.8
no glasses (max)	3.0	3.4	3.5	3.6	4.1
glasses (avg)	0.3	0.4	0.5	0.7	0.8
glasses (max)	4.0	4.3	4.0	3.7	4.1

Table 1: Average and maximal error in $[mm]$ for model reconstruction with and without glasses averaged over 3 different models.

Robustness and Accuracy Evaluation on Real Data. In Fig. 7 we show a variety of 3D results, again for different faces as well as different glass shapes. Our algorithm yields very plausible results also within the area that has been occluded by the glasses. Fig. 8 compares the results of our algorithm with eyeglass segmentation against the result without a preceding segmentation. Without the segmentation the model fitting tries to adapt to depth values of both the face and the eyeglasses, which leads to strongly distorted models (Fig. 8, bottom right).

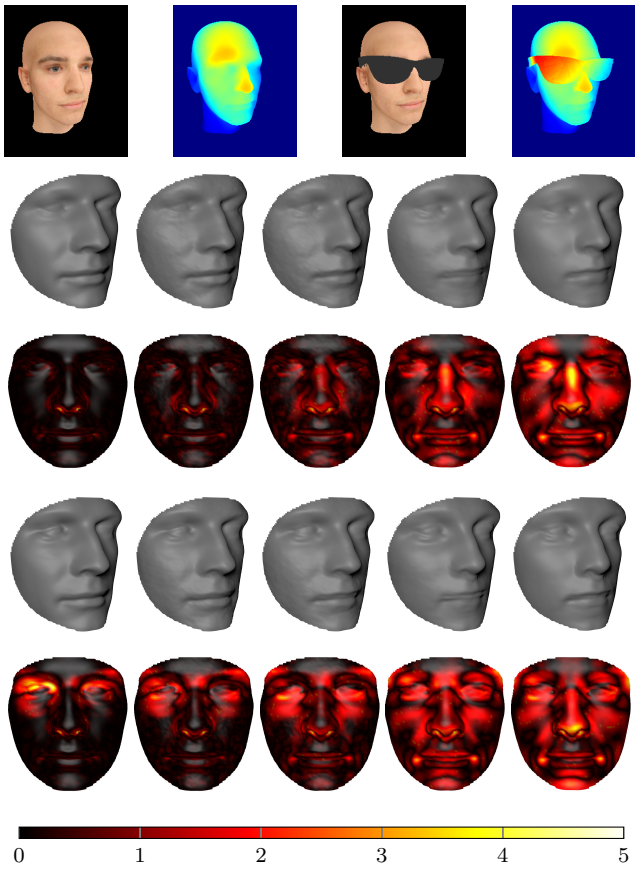

Figure 6: Synthetic evaluation of reconstruction accuracy. **First row:** example of 3D head model with and without glasses and corresponding ground truth depth. The reconstruction results are computed using 15 depth maps with 50% missing data and increasing Gaussian noise with zero mean and $\sigma = 0, 2, 4, 6, 8\ [mm]$ (from left to right). **Second and third row:** reconstruction results for a sample model without glasses. **Fourth and fifth row:** results for the same model with glasses. Color map units are in millimeters.

Runtimes. Timings on a Samsung Galaxy S7 for a height map resolution of 150 × 120 pixels and depth maps at a resolution of 320 × 240 pixels are reported in Tab. 2.

5. Conclusion

We presented a novel method for the 3D reconstruction of faces and eyeglasses which is suitable to run on mobile devices. Our method uses a cylindrical height map for the aggregation of the input depth maps with correspond-

Figure 7: Results for different subjects wearing glasses of various shapes. From left to right: sample input image, result obtained using the method presented in [30], result obtained using the proposed approach, reconstruction of glasses obtained using our approach.

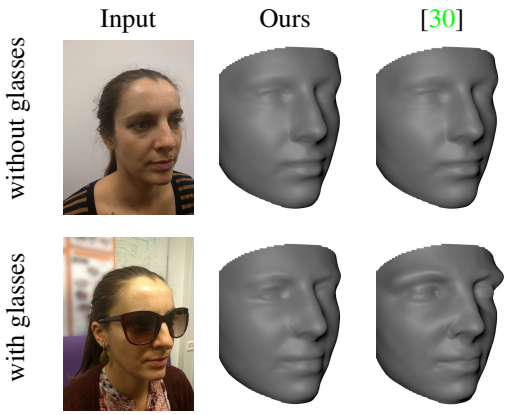

Figure 8: Model fitting results for a subject with and without glasses in comparison to [30] which does not remove the glasses before the model fitting. Ignoring the glasses can cause significant distortions in the model (bottom right).

Depth computation & integration (per depth map)	105 ms
Alignment (once)	2000 ms
Texture computation (using 30 images, once)	1800 ms
Segmentation (once, $\approx 40ms$ per shortest path)	**100 ms**
Model fitting (once)	200 ms
Regularization (once, 700 iterations)	2200 ms
Total (30 depth maps)	9450 ms

Table 2: Average runtimes (unoptimized code) on a Samsung Galaxy S7. Apart from the segmentation (bold) all steps and runtimes are similar to [30], leading to only 1% runtime overhead.

ing color information. We introduced a variational model for the segmentation of eyeglasses inside the height map and showed that the segmentation problem can be more efficiently solved or approximated by computing a series of either two or five Dijkstra shortest path computations. Our method then subsequently reconstructs the 3D face by fitting a statistical face model to the non-glass geometry and then regularizes the difference between the shape model and the aggregated height map, which allows to reconstruct instance-specific details. Similarly, we are able to reconstruct the geometry of the glasses by regularizing the aggregated height values inside the segmented glass region. Multiple experiments on synthetic and real data demonstrate that our method is robust to changes in noise, lighting conditions, various face and glass shapes. Although our model does not need a database of eyeglass to reliably segment a large variety of glasses, in future work we want to explore the performance of learned models for glasses.

Acknowledgements. This work was supported by Grant 16703.1 PFES-ES of CTI Switzerland. We are grateful to Müller Optik Zürich AG (http://www.mueller-optik.ch) for providing various eyeglass models for the experiments.

References

[1] O. Aldrian and W. A. P. Smith. Inverse rendering of faces with a 3d morphable model. *TPAMI*, 35(5):1080–1093, 2013. 2

[2] B. Amberg, R. Knothe, and T. Vetter. Expression invariant 3d face recognition with a morphable model. In *IEEE International Conference on Automatic Face & Gesture Recognition*, 2008. 2

[3] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004. 5

[4] J. P. Barreto and H. Araujo. Issues on the geometry of central catadioptric image formation. In *CVPR*, volume 2, pages II–422. IEEE, 2001. 2

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 2, 6

[6] D. Borza, A. S. Darabant, and R. Danescu. Eyeglasses lens contour extraction from facial images using an efficient shape description. *Sensors*, 13(10):13638–13658, 2013. 1, 2

[7] X. Bresson, S. Esedoḡlu, P. Vandergheynst, J.-P. Thiran, and S. Osher. Fast global minimization of the active contour/snake model. 28(2):151–167, 2007. 3

[8] A. Brunton, T. Bolkart, and S. Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *ECCV*, pages 297–312. Springer, 2014. 2

[9] X. P. Burgos-Artizzu, J. Fleureau, O. Dumas, T. Tapie, F. L. Clerc, and N. Mollet. Real-time expression-sensitive hmd face reconstruction. In *SIGGRAPH Asia Technical Briefs*, pages 9:1–9:4. ACM, 2015. 2

[10] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. *TOG*, 34(4):46, 2015. 2

[11] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *JMIV*, 40(1):120–145, 2011. 6

[12] T. F. Chan and J. Shen. Mathematical models for local nontexture inpaintings. *SIAM J. Appl. Math*, 62:1019–1043, 2002. 6

[13] P. Cignoni, C. Rocchini, and R. Scopigno. Metro: measuring error on simplified surfaces. In *Computer Graphics Forum*, volume 17, pages 167–174. Wiley Online Library, 1998. 7

[14] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Conference on Computer graphics and interactive techniques*, 1996. 2

[15] A. Delaunoy, E. Prados, P. G. I. Piracés, J.-P. Pons, and P. Sturm. Minimizing the multi-view stereo reprojection error for triangular surface meshes. In *BMVC*, 2008. 2

[16] B. Egger, A. Schneider, C. Blumer, A. Morel-Forster, S. Schnborn, and T. Vetter. Occlusion-aware 3d morphable face models. In *BMVC*, 2016. 2, 4

[17] A. Fernández, R. García, R. Usamentiaga, and R. Casado. Glasses detection on real images based on robust alignment. *Machine Vision and Applications*, 26(4):519–531, 2015. 2

[18] S. Fuhrmann and M. Goesele. Floating scale surface reconstruction. *TOG*, 33(4):46, 2014. 2

[19] C. Geyer and K. Daniilidis. A unifying theory for central panoramic systems and practical implications. In *ECCV*, pages 445–461. Springer, 2000. 2

[20] V. H. Hiep, R. Keriven, P. Labatut, and J.-P. Pons. Towards high-resolution large-scale multi-view stereo. In *CVPR*, 2009. 2

[21] P. Huber, Z. Feng, W. J. Christmas, J. Kittler, and M. Rätsch. Fitting 3d morphable face models using local features. In *ICIP*, pages 1195–1199, 2015. 2

[22] V. Kazemi, C. Keskin, J. Taylor, P. Kohli, and S. Izadi. Real-time face reconstruction from a single depth image. In *3DV*, 2014. 2

[23] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *SGP*, 2006. 2

[24] S. Koenig and M. Likhachev. Incremental a*. In *NIPS*, pages 1539–1546, 2001. 5

[25] K. Kolev, M. Klodt, T. Brox, and D. Cremers. Continuous global optimization in multiview 3d reconstruction. *IJCV*, 84(1):80–96, 2009. 2

[26] K. Kolev, P. Tanskanen, P. Speciale, and M. Pollefeys. Turning mobile phones into 3d scanners. In *CVPR*, 2014. 2

[27] P. Labatut, J.-P. Pons, and R. Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *ICCV*, 2007. 2

[28] V. Lempitsky and Y. Boykov. Global optimization for shape fitting. In *CVPR*, 2007. 2

[29] F. Maninchedda, C. Häne, B. Jacquet, A. Delaunoy, and M. Pollefeys. Semantic 3d reconstruction of heads. In *ECCV*, 2016. 2

[30] F. Maninchedda, C. Häne, M. R. Oswald, and M. Pollefeys. Face reconstruction on mobile devices using a height map shape model and fast regularization. In *3DV*, 2016. 2, 3, 5, 6, 8

[31] C. Mei and P. Rives. Single view point omnidirectional camera calibration from planar grids. In *ICRA*, pages 3945–3950. IEEE, 2007. 2

[32] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 2

[33] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, pages 127–136. IEEE, 2011. 2

[34] P. Ondrúška, P. Kohli, and S. Izadi. Mobilefusion: Real-time volumetric surface reconstruction and dense tracking on mobile phones. *IEEE transactions on visualization and computer graphics*, 2015. 2

[35] M. R. Oswald, J. Stühmer, and D. Cremers. Generalized connectivity constraints for spatio-temporal 3d reconstruction. In *ECCV*, pages 32–46, 2014. 3

[36] A. Patel and W. A. Smith. 3d morphable face models revisited. In *CVPR*, 2009. 2

[37] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. 2009. 2, 6

[38] T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *ICCV*, pages 1762–1769, 2011. 4

[39] G. Ramalingam and T. W. Reps. An incremental algorithm for a generalization of the shortest-path problem. *J. Algorithms*, 21(2):267–305, 1996. 5

[40] C. Reinbacher, T. Pock, C. Bauer, and H. Bischof. Variational segmentation of elongated volumetric structures. In *CVPR*, 2010. 4

[41] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, pages 1685–1692, 2014. 2

[42] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *CVPR*, 2015. 2

[43] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002. 2

[44] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *CVPR*, June 2016. 2

[45] T. Schöps, T. Sattler, C. Häne, and M. Pollefeys. 3d modeling on the go: Interactive 3d reconstruction of large-scale scenes on mobile devices. In *3DV*, 2015. 2

[46] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *CCS*. ACM, 2016. 1

[47] M. D. Smet, R. Fransens, and L. J. V. Gool. A generalized EM approach for 3d model based face recognition under occlusions. In *CVPR*, pages 1423–1430, 2006. 2

[48] J. Stühmer, P. Schröder, and D. Cremers. Tree shape priors with connectivity constraints using convex relaxation on general graphs. In *ICCV*, pages 2336–2343, 2013. 3, 4, 5

[49] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *ECCV*, 2014. 2

[50] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys. Live metric 3d reconstruction on mobile phones. In *ICCV*, 2013. 2

[51] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Demo of face2face: real-time face capture and reenactment of RGB videos. In *SIGGRAPH*, pages 5:1–5:2, 2016. 2

[52] B. Ummenhofer and T. Brox. Global, dense multiscale reconstruction for a billion points. In *ICCV*, pages 1341–1349, 2015. 2

[53] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. 2005. 2

[54] C. Wu, C. Liu, H. Shum, Y. Xu, and Z. Zhang. Automatic eyeglasses removal from face images. *TPAMI*, 26(3):322–336, 2003. 1

[55] C. Zach. Fast and high quality fusion of depth maps. In *3DPVT*, 2008. 2

[56] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust tv-l 1 range image integration. In *ICCV*, 2007. 2

[57] C. Zach, L. Shan, and M. Niethammer. Globally optimal finsler active contours. In *DAGM*, pages 552–561, 2009. 4