2016 Fourth International Conference on 3D Vision

Face Reconstruction on Mobile Devices Using a Height Map Shape Model and Fast Regularization

Fabio Maninchedda[1], Christian Häne[2], Martin R. Oswald[1], and Marc Pollefeys[1,3]

[1]ETH Zürich, Switzerland [2]University of California, Berkeley [3]Microsoft, USA

Abstract

We present a system which is able to reconstruct human faces on mobile devices with only on-device processing using the sensors which are typically built into a current commodity smart phone. Such technology can for example be used for facial authentication purposes or as a fast preview for further post-processing. Our method uses recently proposed techniques which compute depth maps by passive multi-view stereo directly on the device. We propose an efficient method which recovers the geometry of the face from the typically noisy point cloud. First, we show that we can safely restrict the reconstruction to a 2.5D height map representation. Therefore we then propose a novel low dimensional height map shape model for faces which can be fitted to the input data efficiently even on a mobile phone. In order to be able to represent instance specific shape details, such as moles, we augment the reconstruction from the shape model with a distance map which can be regularized efficiently. We thoroughly evaluate our approach on synthetic and real data, thereby we use both high resolution depth data acquired using high quality multi-view stereo and depth data directly computed on mobile phones.

1. Introduction

Digital 3D reconstruction of human faces has been studied extensively in the past. Reconstruction algorithms are often aimed at specific applications or a group of applications. These range from digital avatars, 3D printing to tracking of facial expressions in videos or even authentication. Recently, mobile devices have become powerful enough to generate 3D models with on-device computing and using the live imagery of built-in cameras. This opens the technology for new applications where the ability to run the 3D reconstruction on the device is crucial. One example is security critical applications where the input data should not leave the device, such as authentication through a face scan. Another example where on-device processing is desirable is for a live preview of reconstructions to ensure that the captured data is of sufficient quality for post processing. We propose a system which fully automatically reconstructs a human face in a few seconds on commodity mobile phones using only on-device processing and built-in sensors.

Impressive 3D models of faces computed with passive stereo matching were presented in [15, 5], the key requirements for high quality reconstructions are 1) high resolution data taken in excellent lighting conditions and 2) very accurate camera calibrations using bundle adjustment or a fixed multi-camera rig. None of that is given when using mobile devices in uncontrolled environments. The user takes images with the built-in camera in potentially bad lighting conditions leading to motion blur, rolling-shutter artifacts, and non-rigid deformation of the face during capturing. Moreover, currently computational resources on mobile devices do not facilitate the usage of high resolution images and bundle adjustment. All these shortcomings lead to a high level of noise and inaccuracies in the captured depth maps. Therefore, one of the main difficulties when acquiring high quality reconstructions of faces on a mobile device is tackling this high level of noise.

One of the most popular tools when dealing with noise or incomplete data of faces, are low dimensional statistical models of human faces [7, 33]. Typically the model is directly fitted into the input data. Compared to generic reconstruction algorithms, this leads to a better constrained formulation as only the parameters of a low dimensional model and its alignment to the input data are estimated. Due to the dependency between the size of faces and their shape, e.g. female faces tend to be smaller, an expensive procedure which alternates between finding the correspondences between model and data, and estimating its parameters is typically utilized [2]. Another shortcoming of such models is that they are unable to capture instance specific details such as moles, wrinkles or scars. We propose to overcome these shortcomings with the following contributions:

- A pipeline which fuses a set of noisy depth maps acquired using passive stereo into a 3D face model by using a processing pipeline which works on a 2.5D height map representation. (Sec. 2)

978-1-5090-5407-7/16 $31.00 © 2016 IEEE
DOI 10.1109/3DV.2016.59

- A statistical 2.5D height map shape model of faces, in which the scale is removed from the model through a prior alignment to the mean shape for efficient alignment and fitting. (Secs. 4 and 5)

- We propose to add instance specific details to the model with a difference map which can be efficiently regularized using convex optimization. (Sec. 6)

1.1. Related Work

Acquiring 3D reconstructions of faces from images is a broad topic. Many 3D reconstruction algorithms for generic shapes can also be used for faces.

One of the first steps that is traditionally executed for 3D reconstruction form images, is the recovery of the camera poses of the input images, i.e. solving the structure-from-motion (SfM) problem [18]. From the input images and the recovered camera poses a collection of depth maps can be computed by dense stereo matching [42, 15]. Alternatively, also active sensors such as structured light, time-of-flight or laser scanners are used to measure depth data. A variety of methods for computing a final 3D model from depth data have been proposed: A set of a few high quality depth maps [23, 28], volumetric binary labeling into free and occupied space [21, 26, 25, 43], volumetric truncated signed distance fields [13, 44, 29], and mesh based optimization [14, 19].

Using the on-device sensors of commodity mobile phones [38, 24, 31] compute 3D models interactively with only on-device processing. With specialized computer vision enabled mobile devices [22] and [36] achieve 3D reconstructions using an active structured light sensor or passive motion stereo, respectively.

Human faces have a strong shape similarity between individual faces. Statistical shape models which capture the variations of human faces in a low dimensional space are therefore a popular tool. Several models have been proposed which either only capture the shape of the neutral expression [7, 32, 33] or also add facial expressions [2, 39, 8]. One drawback of statistical face models is that they are unable to capture instance specific shape variations. Therefore, they are either discarded or added afterwards using for example shading based techniques [37] or local regressors [9]. In this paper, our objective is to reconstruct a human face in neutral expression, e.g. for authentication purposes. The main objective becomes fitting the face shape model into a potentially noisy input point cloud. Fitting the model of [33] requires an iterative process which alternates between finding correspondences and fitting the model [2, 3] leading to a running time of up to 90 seconds to fit the model to an input scan. In [8] an iterative coarse-to-fine optimization is utilized, leading to a running time for the model fitting of several seconds on a desktop computer. [20] proposes to speed up the model fitting by using a discriminatively trained random forest to estimate the correspondences between a single input depth frame, captured with an active depth sensor, and the shape model. In our work, we aim for accurate and efficient reconstruction of faces on mobile phones, which typically do not have active depth sensing available and have restricted computing resources.

1.2. Overview

The inputs to our height map face reconstruction algorithm is a set of images, $\mathcal{I} = \{I_1, \ldots, I_n\}$, depth maps $\mathcal{D} = \{D_1, \ldots, D_n\}$ and the corresponding camera parameters $\mathcal{P} = \{P_1, \ldots, P_n\}$. Each camera parameter $P_i = \{K_i, R_i, C_i\}$ consists of the camera intrinsics K_i and pose $[R_i, C_i]$. An initial alignment is established by computing a similarity transform between a few selected points of the mean face of the Basel Face Model (BFM) [33] and triangulated landmarks computed on the input images using [34]. The depth maps are then integrated into a height map representation that we introduce in Sec. 2. Details of the depth map integration procedure are explained in Sec. 3. The alignment of the height map is then further refined by an iterative optimization that is detailed in Sec. 4. The depth information is then re-integrated using the refined alignment. A face model computed directly in the height map representation is fitted to the data using a simple weighted least squares fit presented in Sec. 5. The residual obtained by subtracting the fitted model from the height map is regularized using an efficient convex optimization that we describe in Sec. 6. The optimized residual contains individual specific details that cannot be captured by the low dimensional face model. Finally, the optimized residual is added back to the fitted model to obtain the final result. Fig. 1 summarizes all the steps of the proposed algorithm as a flow diagram.

2. Height Map Representation

In order to keep the demands on computing and memory resources of our approach low, we model the 3D shape of a human face with a 2.5D height map. That is, we assume that the manifold of the human face is homeomorphic to a square. To obtain such a parametrization one needs to find a mapping $\mathbf{X} \to \mathbf{p}$ that maps each point $\mathbf{X} \in \mathbb{R}^3$ on the face to a point $\mathbf{p} \in [1, N] \times [1, M]$ in a rectangular region. In order to map all 3D points of the human face onto a height map, we assume that all these points are visible from a single point. We model the height map by a projection with a virtual omni-directional camera that is located inside the head looking toward the face and store the distance between the camera center C_0 and the face point $\mathbf{X}$ at the corresponding position. The resulting height map and camera parameters will be denoted as $\mathcal{H} \in \mathbb{R}^{N \times M}$ and $P_0 = \{K_0, R_0, C_0\}$ respectively. To be flexible in terms of field of view we use the unified projection model [16, 4, 27]. First, a face point $\mathbf{X} = (X_x, X_y, X_z)^\top$ is projected onto the unit sphere $\mathbf{X}_s = \frac{\mathbf{X}}{\|\mathbf{X}\|}$. Then, the function $\mathbf{m} = \hbar(\mathbf{X}_s, \xi)$

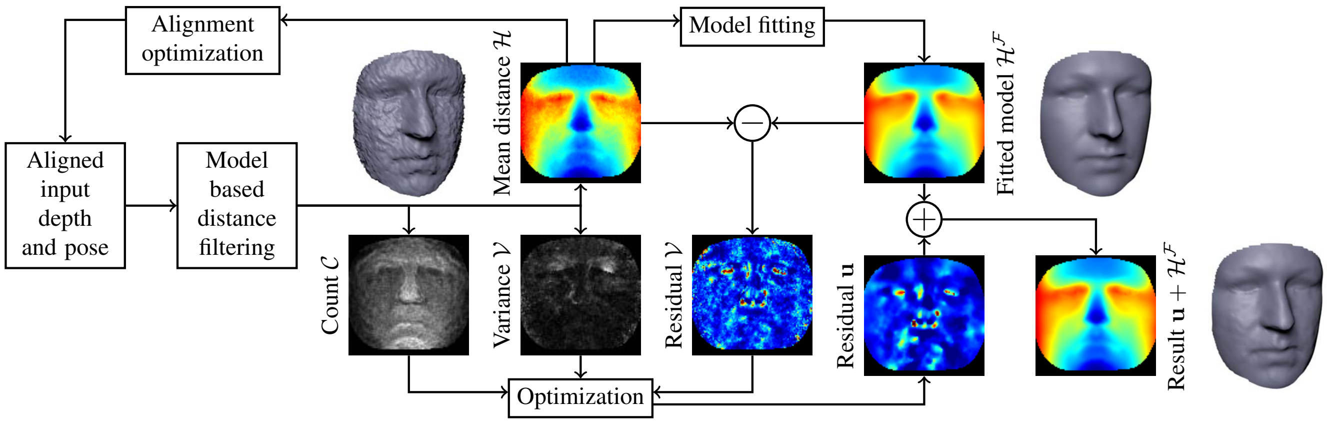

Alignment optimization
Aligned input depth and pose
Model based distance filtering
Mean distance $\mathcal{H}$
Model fitting
Fitted model $\mathcal{H}^{\mathcal{F}}$
Count $\mathcal{C}$
Variance $\mathcal{V}$
Residual $\mathcal{V}$
Residual $\mathbf{u}$
Result $\mathbf{u} + \mathcal{H}^{\mathcal{F}}$
Optimization

Figure 1. Overview of proposed approach.

maps the 3D point $\mathbf{X}_s$ to a point $\mathbf{m}$ on the normalized image plane. The scalar parameter ξ models the mirror. Finally, the image point is given by $\mathbf{p} = K_0\mathbf{m}$, where K_0 denotes the virtual camera intrinsic parameters. Given an image point $\mathbf{p}$ and a height map $\mathcal{H}$ one can obtain the corresponding face point $\mathbf{X}$ as follows

$$\mathbf{X} = \mathcal{H}(\mathbf{p})\hbar^{-1}(K_0^{-1}\mathbf{p}, \xi). \tag{1}$$

For validation of the assumption that each point on the face is visible from the virtual camera center, we conducted an experiment. For a height map resolution of $N = M = 100$ we have computed the number of ray-face intersections with 200 faces randomly sampled from the BFM [33] by shooting one ray per pixel for each camera position. The total ray count per camera position amounts to 8679 as not all pixels in the height map representation have a corresponding face point. All rays with more than one intersection represent a case in which our assumption is violated. Therefore, we seek for a camera center which has a minimal number of violations. The statistical face model is designed in such a way that the YZ-plane is the plane of symmetry of the face, therefore we limited our search to this plane. To clarify the setup we have displayed the mean face of the statistical model and its YZ bounding box in Fig. 2. We considered the camera centers $C_0 \in \{(0, -30+10i, -70+10j) : 0 \leq i \leq 10, 0 \leq j \leq 10\}$ (units in mm) and at each position we computed K_0 and R_0 such that the border of the mean face projects to the boundaries of the height map. Fig. 2 shows a contour plot of the percentage of rays that have two or more intersections averaged over the 200 faces. We highlighted the region in which we get the lowest percentage of multiple intersections. Positions that have a negative Y coordinate have a higher percentage of multiple intersections because they cannot represent the ocular cavity, the nostril area and nose tip whereas points above $Y = 50$ tend to intersect both the upper and lower lip due to the very steep angle especially when close to the mean face. This angle becomes less and less steep as we go further away from the mean face, this is reflected by the generally lower amount of intersections with decreasing Z coordinate values. Since on average only 0.03% of the rays have multiple intersections our assumption is justified and - as shown later in the experiments - the remaining errors are small or negligible.

3. Depth Integration

In this section we will explain how input depth maps $\mathcal{D} = \{D_1, \ldots, D_n\}$ are brought into the height map representation. For each input image I_i, each pixel $\mathbf{x} = (x_x, x_y)$ is unprojected using the corresponding depth value $d = D_i(x_x, x_y)$ to obtain a point in world coordinates $\tilde{\mathbf{X}} = R_i^\top K_i^{-1}[x_x, x_y, d]^\top + C_i$. The point is then transformed into the virtual camera reference frame $\mathbf{X} = R_0\tilde{\mathbf{X}} - R_0^\top C_0$. Then, the 3D point is projected into the height map representation. The position in the height map is given by $\mathbf{p} = \text{proj}_{\mathcal{H}}(\mathbf{X}) := K_0\hbar(\frac{\mathbf{X}}{\|\mathbf{X}\|}, \xi)$ while the distance is simply $\|\mathbf{X}\|$. Since multiple points will project to the same position we compute a weighted mean distance that takes into account the camera viewing direction [30] and the distance to the mean face of the statistical model [33]. Additionally, we also compute the weighted variance $\mathcal{V} \in \mathbb{R}^{N\times M}$ and number of projected points $\mathcal{C} \in \mathbb{R}^{N\times M}$. The final height map value is given by

$$\mathcal{H}(\mathbf{p}) = \tfrac{1}{\mathcal{C}(\mathbf{p})} \sum_{i:\, \mathbf{p}=\text{proj}_{\mathcal{H}}(\mathbf{X}_i)} I(\mathbf{X}_i)W(\mathbf{X}_i)\|\mathbf{X}_i\|. \tag{2}$$

The term $I(\mathbf{X}_i) = \mathbf{1}_{\{|\|\mathbf{X}_i\| - \mathcal{H}^{\boldsymbol{\mu}}(\mathbf{p})| < \tau^I(\mathbf{p})\}}$ is an indicator function that discards points that are further away than $\tau^I(\mathbf{p})$ from the distance map of the mean face $\mathcal{H}^{\boldsymbol{\mu}}$. The threshold is computed from the variation in distance values at the position $\mathbf{p}$ in the height map. The factor

$$W(\mathbf{X}_i) = \left\langle \frac{C_i - \tilde{\mathbf{X}}_i}{\|C_i - \tilde{\mathbf{X}}_i\|}, N^{\boldsymbol{\mu}}(\tilde{\mathbf{X}}_i) \right\rangle \tag{3}$$

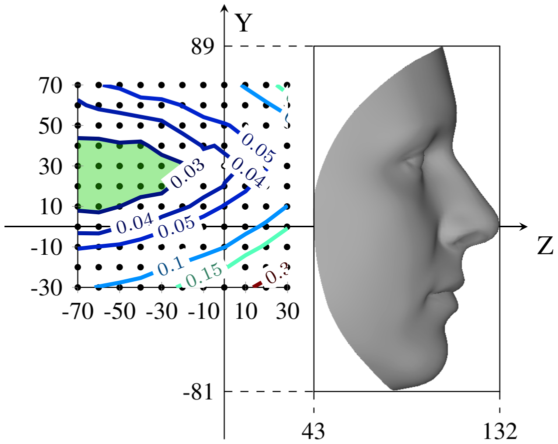


Figure 2. Process for finding the best projection center to map a face to a height map representation. Each dot in the contour plot represents a sampled projection center for which we computed the number of intersections with 200 face samples from the BFM by shooting one ray for each height map pixel. The height map resolution of 100 × 100 pixels gives a total of 8679 rays per height map which all intersect the face. We show the sampled positions relative to the mean face of the statistical model and its bounding box with coordinates in millimetres as a reference. The contour plot shows the average percentage of rays that have intersected a face multiple times. In the optimal region marked in light green we have 0.03% multiple intersections on average (2.6 intersections per face).

weighs the influence of samples based on the cosine of the angle between the camera viewing direction and the normal of the mean face at the point $\tilde{\mathbf{X}}_i$ which is denoted as $N^{\boldsymbol{\mu}}(\tilde{\mathbf{X}}_i)$. The normalization weight

$$\mathcal{C}(\mathbf{p}) = \sum_{i\colon \mathbf{p}=\mathrm{proj}_{\mathcal{H}}(\mathbf{X}_i)} I(\mathbf{X}_i) W(\mathbf{X}_i) \tag{4}$$

corresponds to the weighted number of projected points. The weighted variance is computed as

$$\mathcal{V}(\mathbf{p}) = \frac{1}{\mathcal{C}(\mathbf{p})} \sum_{i\colon \mathbf{p}=\mathrm{proj}_{\mathcal{H}}(\mathbf{X}_i)} I(\mathbf{X}_i) W(\mathbf{X}_i) (\mathcal{H}(\mathbf{p}) - \|\mathbf{X}_i\|)^2. \tag{5}$$

Note that the variance is computed efficiently in an online fashion [40].

4. Alignment

A precise alignment is of great importance when fitting a parametric face model. This step is commonly performed using alternating optimization, which are variants ICP algorithms [35]. We propose to improve the initial landmark based alignment with a refinement that can efficiently be computed in the height map representation. Let $\mathcal{M}^{\mathcal{H}}$ be the mesh corresponding to the height map $\mathcal{H}$. The goal of this step is to align $\mathcal{M}^{\mathcal{H}}$ with the mean face of the statistical model $\boldsymbol{\mu}$. Our height map based method is closely related to registration methods for range images that use a projection to find the corresponding points during the alignment optimization [11, 6]. However, due to the fact that in our case both target and source mesh are represented in a height map that share the same virtual camera, we can evaluate the 3D euclidean distance between points directly in the height map representation. This allows to circumvent the most expensive step of ICP algorithms, namely finding the point correspondences. We propose to minimize

$$E(\boldsymbol{\alpha}) = \sum_{\mathbf{p}\in\mathcal{H}} W^A(\mathbf{p}) \min\left(\left|\mathcal{H}^{\boldsymbol{\alpha}}(\mathbf{p}) - \mathcal{H}^{\boldsymbol{\mu}}(\mathbf{p})\right|, \tau^A\right) \tag{6}$$

with $\mathcal{H}^{\boldsymbol{\alpha}} = \mathrm{proj}_{\mathcal{M}}\left(\mathcal{T}(\mathcal{M}^{\mathcal{H}}, \boldsymbol{\alpha})\right)$. The function $\mathcal{T}$ denotes a similarity transform that depends on a scaling factor, yaw, pitch and roll angles and a translation vector that are stored in $\boldsymbol{\alpha}$. The function $\mathrm{proj}_{\mathcal{M}}(\mathcal{M}^{\mathcal{H}})$ denotes the projection of $\mathcal{M}^{\mathcal{H}}$ into the height map representation. The threshold τ^A clamps the maximal difference to reduce the influence of outliers. Finally, W^A is a weighing matrix that enforces good alignment in the eye, nose and mouth region. The correspondences between points in $\mathcal{M}^{\mathcal{H}}$ and $\mathcal{M}^{\boldsymbol{\mu}}$ are given implicitly by $\mathrm{proj}_{\mathcal{M}}(\cdot)$ whereas taking the difference of height map values gives the signed euclidean distance between corresponding points. One important detail is that W^A does not depend on the similarity transform $\mathcal{T}$. This forces the optimization to find an alignment with some overlap as shrinking the solution to a single point would cost $\tau^A \sum_{\mathbf{p}\in\mathcal{H}} W^A(\mathbf{p})$, which is the maximum over all solutions $\boldsymbol{\alpha}$. Therefore, our energy does not need normalization and overlapping constraints as proposed in [6]. The energy in Eq. (6) is minimized using the gradient descent based, L-BFGS line search approach, implemented in the Ceres solver [1].

5. Model Fitting

The most important step when fitting a statistical model to some data is to find good correspondences between the two. Generally, one has to first align the input data to some reference model, a common choice is the mean shape, and then establish the correspondences between the reference and the data, which is then projected into the model. Statistical models that are metric, such as [33], require an iterative refinement of the fitted model to estimate the right scale. For this purpose the fitted model is iteratively refined by repeating the same procedure that we have described above with the fitted model as a reference for the alignment and correspondence computation until convergence. This procedure has two problems for our application. First, finding correspondences at each iteration is expensive and not suited for a real-time algorithm. Second, we have no notion of scale. Therefore, we have decided to construct a scale-free parametric model directly in the height map representation. The scale is factored out from the model by aligning each face to the mean shape before the statistical model is computed. This allows for a much more efficient fitting approach that consists of an alignment step and a projection into the model without any iterative refinement.

For completeness and to facilitate the understanding of the model fitting approach we will quickly describe the face model presented in [33]. A face is composed of m vertices $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$. Each point is then concatenated into a $3m$ dimensional vector $\left[\mathbf{X}_1^\top, \ldots, \mathbf{X}_m^\top\right]^\top$. The parametric face model

$$\mathcal{F} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{U}) \tag{7}$$

consists of the mean $\boldsymbol{\mu} \in \mathbb{R}^{3m}$, the standard deviation $\boldsymbol{\sigma} \in \mathbb{R}^{n-1}$ and an orthonormal basis of principal components $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_{n-1}] \in \mathbb{R}^{3m \times n-1}$. Faces $\mathbf{f}$ are sampled by computing linear combinations of the principal components

$$\mathbf{f}(\boldsymbol{\beta}) = \boldsymbol{\mu} + \mathbf{U}\operatorname{diag}(\boldsymbol{\sigma})\boldsymbol{\beta} \tag{8}$$

Where each component in $\boldsymbol{\beta} \in \mathbb{R}^{n-1}$ is drawn from a normal distribution with zero mean and unit variance.

To construct a parametric height map face model we sampled $p = 2000$ faces $\mathbf{f}(\boldsymbol{\beta}_1), \ldots, \mathbf{f}(\boldsymbol{\beta}_p)$ from the BFM. Each face is then aligned against the mean face $\boldsymbol{\mu}$ using Eq. (6). We denote the aligned faces as $\mathbf{f}^A(\cdot)$. The aligned face is then projected into the height map representation $\mathcal{H}_i = \operatorname{proj}_{\mathcal{M}}\left(\mathbf{f}^A(\boldsymbol{\beta}_i)\right)$ to obtain the data matrix

$$D = [\operatorname{vec}(\mathcal{H}_1), \ldots, \operatorname{vec}(\mathcal{H}_p)] \in \mathbb{R}^{NM \times p}. \tag{9}$$

We now apply a covariance based PCA to the mean normalized data to obtain the height map face model

$$\mathcal{F}_{\mathcal{H}} = (\boldsymbol{\mu}_{\mathcal{H}}, \boldsymbol{\sigma}_{\mathcal{H}}, \mathbf{U}_{\mathcal{H}}) \tag{10}$$

where $\boldsymbol{\mu}_{\mathcal{H}} = \frac{1}{p}\sum_{i=1}^{p} \mathcal{H}_i \in \mathbb{R}^{NM}$ is the mean face of the statistical model, $\boldsymbol{\sigma}_{\mathcal{H}} \in \mathbb{R}^{p-1}$ is the standard deviation and $\mathbf{U}_{\mathcal{H}} \in \mathbb{R}^{NM \times p-1}$ is an orthonormal basis of principal components as in Eq. (7).

Fitting a parametric height map face model to a height map $\mathcal{H}$ amounts to finding coefficients $\boldsymbol{\beta}$ such that

$$\operatorname{vec}(\mathcal{H}) = \boldsymbol{\mu}_{\mathcal{H}} + \tilde{\mathbf{U}}_{\mathcal{H}}\boldsymbol{\Sigma}_{\mathcal{H}}\boldsymbol{\beta} \tag{11}$$

where $\tilde{\mathbf{U}}_{\mathcal{H}}$ is the matrix that contains the first $q \ll p$ principal components of $\mathbf{U}_{\mathcal{H}}$ and $\boldsymbol{\Sigma}_{\mathcal{H}} = \operatorname{diag}(\boldsymbol{\sigma}_{\mathcal{H}})$. It's easy to see that the least squares solution is given by

$$\boldsymbol{\beta}^* = \boldsymbol{\Sigma}_{\mathcal{H}}^{-1}\tilde{\mathbf{U}}_{\mathcal{H}}^\top(\operatorname{vec}(\mathcal{H}) - \boldsymbol{\mu}_{\mathcal{H}}). \tag{12}$$

This model fitting approach is very sensitive to noise and outliers, therefore we propose an extension that weighs the contribution of every facial point differently. Given a weight matrix $\tilde{W}_F \in \mathbb{R}^{M \times N}$ we want to minimize

$$W_F \operatorname{vec}(\mathcal{H}) = W_F \left(\boldsymbol{\mu}_{\mathcal{H}} + \tilde{\mathbf{U}}_{\mathcal{H}}\boldsymbol{\Sigma}_{\mathcal{H}}\boldsymbol{\beta}\right) \tag{13}$$

where $W_F = \operatorname{diag}(\operatorname{vec}(\tilde{W}_F)) \in \mathbb{R}^{MN \times MN}$. Again one can easily see that the least squares solution is given by

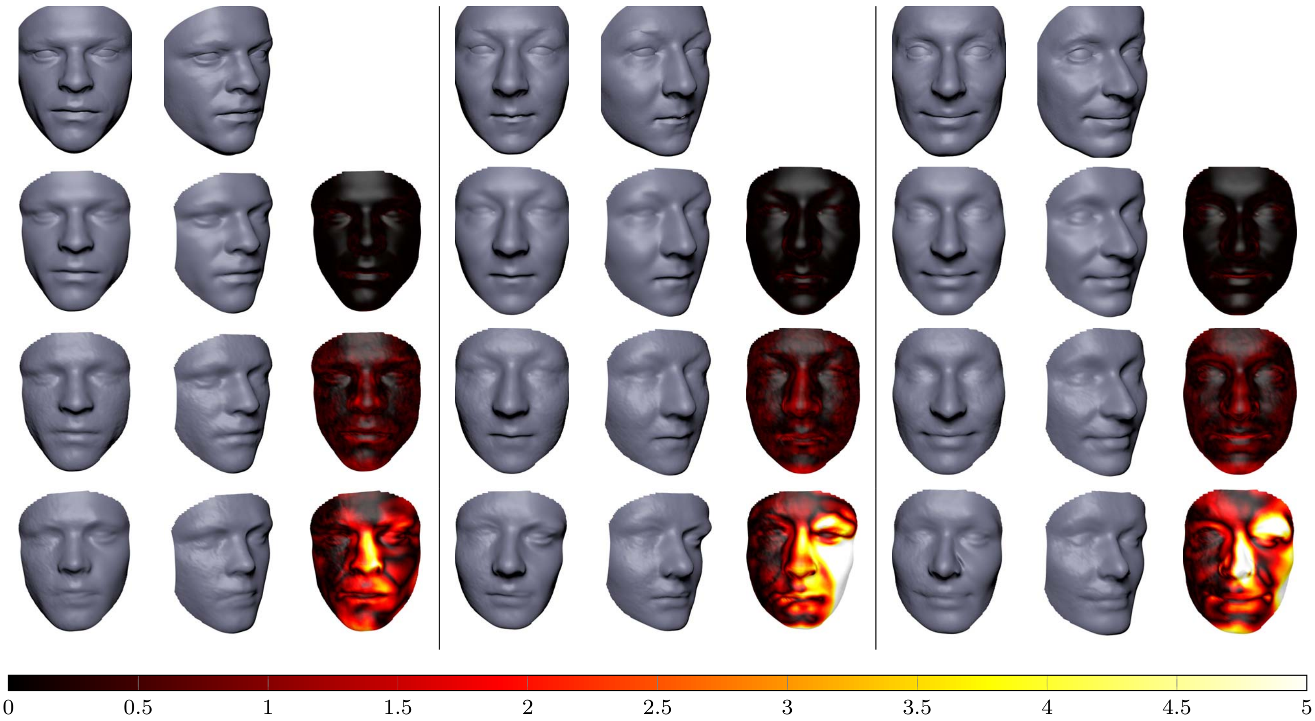
$$\boldsymbol{\beta}^* = \boldsymbol{\Sigma}_{\mathcal{H}}^{-1}(\tilde{\mathbf{U}}_{\mathcal{H}}^\top W_F^2 \tilde{\mathbf{U}}_{\mathcal{H}})^{-1}\tilde{\mathbf{U}}_{\mathcal{H}}^\top W_F^2(\operatorname{vec}(\mathcal{H}) - \boldsymbol{\mu}_{\mathcal{H}}). \tag{14}$$

6. Optimization

Low dimensional parametric face models yield smooth and visually pleasing reconstructions but cannot represent instance specific shape details such as large moles, even if they are observed well in the input data. Especially for tasks such as authentication this is not desirable as such instance specific data is important to distinguish one person from another. The input depth information is very detailed but often quite noisy, especially when computed on mobile devices with limited resources. The optimization procedure proposed in this paper tries to find a good tradeoff between the two afore mentioned extremes. It tries to enforce a smooth result while also preserving facial details that are not present in the face model. This, for example, allows us to get a complete reconstruction of the whole face in cases where only one side of the face is well observed and at the same time the details are still kept in the model for the well observed side (an example is given in Fig. 4 top row, right side). We propose the following method to add the details back to the shape model based reconstruction. From a height map $\mathcal{H}$ of weighted mean distances, cf. Eq. (2), and a fitted model $\mathcal{H}^{\mathcal{F}}$ computed using Eq. (14), we compute the residual $\mathcal{R} = \mathcal{H} - \mathcal{H}^{\mathcal{F}}$. The noise will manifest itself as random variation around zero while errors in the geometry will be visible as consistent positive or negative deviations from zero. This can be exploited by regularizing the residual difference map with a smoothness prior which enforces smooth surfaces but still allows for discontinuities, such as the Huber Total Variation [10]. Taking all these considerations into account we propose to minimize

$$E(\mathbf{u}) = \sum_{i,j} \left\|\nabla \mathbf{u}_{i,j}\right\|_{\epsilon} + \lambda \left\|W^O_{i+jN}(\mathbf{u}_{i,j} - \mathcal{R}_{i,j})\right\|_2^2 \tag{15}$$

where $\mathbf{u} \in \mathbb{R}^{M \times N}$ is the sought solution, $W^O = \operatorname{diag}(\operatorname{vec}(\mathcal{V}))^{-1}\operatorname{vec}(\mathcal{C}) \in \mathbb{R}^{MN}$ is a weight vector that is proportional to the sample count and inversely proportional to the variance. Further, weighting parameter $\lambda \in \mathbb{R}_{\geq 0}$ trades solution smoothness against data fidelity and $\|\cdot\|_\epsilon$ denotes the Huber norm [10]. The rationale behind the choice of W^O is the following. If the variance $\mathcal{V}$ is low and the number of samples $\mathcal{C}$ is high, the mean distance $\mathcal{H}$ should be accurate. Therefore, we want to strongly penalize a deviation from the residual. This is indeed the case as W^O



		No outliers						10% outliers					
	$\sigma =$	0.0	1.0	2.0	3.0	4.0	5.0	0.0	1.0	2.0	3.0	4.0	5.0
# Images	1	0.8/5.6	0.9/3.9	1.2/7.6	1.1/6.0	1.4/7.1	1.6/5.3	1.0/4.0	1.0/5.7	1.2/5.0	1.4/8.1	1.6/7.2	1.8/5.4
	1^*	0.1/1.4	0.1/1.1	0.2/1.4	0.3/1.7	0.4/2.1	0.5/2.7	0.3/1.4	0.3/1.5	0.3/1.5	0.4/2.1	0.4/2.4	0.5/3.2
	2	0.1/1.2	0.1/1.3	0.2/1.8	0.4/2.1	0.5/2.6	0.6/3.2	0.4/1.5	0.4/1.8	0.4/2.2	0.5/2.6	0.5/2.9	0.6/3.4
	5	0.1/0.9	0.1/1.0	0.2/1.4	0.3/1.7	0.4/2.1	0.5/2.5	0.4/1.3	0.4/1.5	0.4/1.7	0.4/2.1	0.4/2.6	0.5/3.0
	11	0.1/0.9	0.1/1.0	0.2/1.7	0.2/1.6	0.4/1.9	0.5/2.5	0.4/1.3	0.4/1.4	0.4/1.6	0.4/2.0	0.4/2.4	0.4/3.0

Figure 3. Experimental evaluation of the reconstruction error for varying number of depth maps, noise and outliers on synthetic data. First row: faces sampled from the BFM [33] that are used as ground truth for the evaluation. For each face we have rendered 1 depth map from $-45°$, 1 frontal depth map (denoted as 1^*), 2 depth maps from $-45°$ and $+45°$ and 5 respectively 11 depth maps sampled uniformly between $-45°$ and $45°$. Each depth map has been corrupted with Gaussian noise with 0 mean and standard deviation $\sigma \in \{0, 1, 2, 3, 4, 5\}$ and up to 10% outliers sampled uniformly from $[0, 10]$. A unit is equivalent to $1mm$. Second row: reconstruction result with 11 depth maps, no noise and no outliers. Third row: reconstruction result with 5 depth maps, $\sigma = 2$ and 10% outliers. Fourth row: reconstruction result with 1 lateral depth map, $\sigma = 5$ and 10% outliers. The table reports the average and maximal error in mm for all possible combinations averaged over 10 faces sampled from the BFM that have not been used to train the height map face model.

will be large. On the other hand, if the variance is high or the number of samples is low, it's likely that the mean distance will not be very accurate and therefore W^O should be small. The final optimized residual $\mathbf{u}$ is added back to the fitted model to obtain the final solution

$$\mathcal{H} = \mathcal{H}^{\mathcal{F}} + \mathbf{u}. \tag{16}$$

The proposed energy is convex in $\mathbf{u}$ and can be efficiently optimized using a first-order primal-dual algorithm [10].

7. Experimental Evaluation

7.1. Reconstruction Accuracy on Synthetic Data

To assess the accuracy and robustness of the proposed method we performed the following experiment. We have sampled 10 faces from the BFM that have not been used to create the height map face model $\mathcal{F}_{\mathcal{H}}$. For each face we have rendered 11 depth maps from positions that see the face at angles between $-45°$ and $45°$, where $0°$ denotes a frontal viewing position. Each depth map has been corrupted with noise sampled from a normal distribution with 0 mean and standard deviation $\sigma \in \{0, 1, 2, 3, 4, 5\}$ and contaminated with up to 10% outliers sampled from a uniform distribution between $[0, 10]$. We reconstructed the face model with a varying number of depth maps, noise and outliers. To compute the average distance in millimetres between the original model and the reconstruction we use [12]. The results are reported in Fig. 3 along with renderings of the reconstructions for a few selected configurations. The second row shows an ideal case with many

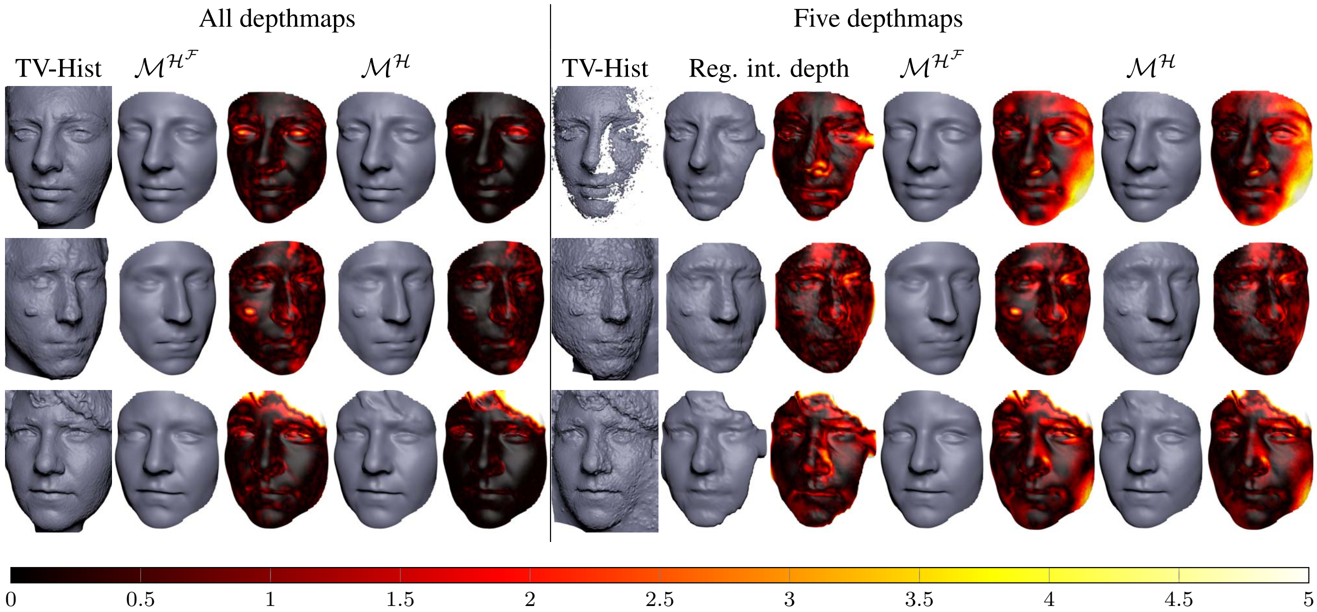

All depthmaps
Five depthmaps
TV-Hist
$\mathcal{M}^{\mathcal{H}^{\mathcal{F}}}$
$\mathcal{M}^{\mathcal{H}}$
TV-Hist
Reg. int. depth
$\mathcal{M}^{\mathcal{H}^{\mathcal{F}}}$
$\mathcal{M}^{\mathcal{H}}$
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Figure 4. Experimental evaluation of the reconstruction error for varying number of depth maps on real data. Left column: results computed using all available depth maps (between 75 and 105). Right column: results computed using 5 depth maps. From left to right in each column: Result computed using TV-Hist [43], Result computed by regularizing directly the integrated depth in the height map representation (only right column), distance between TV-Hist and regularized integrated depth (only right column), fitted height map model ($q = 100$ principal components), distance between TV-Hist result and fitted height map model, proposed approach, distance between TV-Hist and proposed approach. The color map units are in mm.

depth maps, no noise and no outliers which has a very low reconstruction error of only $0.1mm$ on average with a maximal error of $0.9mm$. This shows again that the proposed height map representation yields a good parametrization of the face. The third row shows that even with considerable noise ($\sigma = 2mm$) and outliers (10%) the reconstruction accuracy is still very high when using 5 depth maps which cover all parts of the face. In this case the average and maximal errors amount to $0.4mm$ and $1.7mm$, respectively. In the extreme case, where only a single depth map that sees the face from the side with strong noise $\sigma = 5mm$ and 10% outliers is used, the errors get bigger. However, the reconstruction nicely fills in the missing part thanks to the height map shape model and yields a visually plausible result.

7.2. Reconstruction Accuracy on Real Data

To validate the performance of the proposed approach on real data we have captured images of three subjects with the back camera of a LG Nexus 6P smart phone with locked auto exposure and autofocus at a resolution of 1280×960 pixels. To simulate a big mole we have attached a raisin to the cheek of one of the subjects. We have then computed the extrinsic calibrations using VisualSFM [41]. To get high quality reconstructions, which we consider as the reference solution for the quantitative evaluation, for each subject we have used our implementation of TV-Hist [43], a very accurate volumetric depth map fusion approach, using depth maps computed with the publicly available plane sweeping implementation of [17]. A visual comparison of the reconstruction accuracy of the fitted height map model and the full proposed approach is presented in Fig. 4. In a first experiment we used all the depth maps to get the best possible reconstruction. For the height map face model we have used $q = 100$ components which contain 98.4% of the variation present in the data that has been used to the train the model. Generally, the full proposed approach yields reconstructions that have a smaller distance to the reference solution. The most prominent difference is visible in the model with the mole, which simply cannot be represented using just the height map shape model. Using our proposed approach we recover such instance specific shape details that are strongly seen in the data by optimizing for a smooth residual as explained in Sec. 6. In a second experiment we have taken the first 5 depth maps of each sequence. Those consist of mostly one depth map that sees the face at a close to frontal view and a few more depth maps that see the face with increasing angle from the left side. Here, we immediately observe that a reconstruction without underlying face model does not lead satisfactory results, as parts of the face are not well covered by measurements. To underline this we made an additional experiment where we use the regularization of described in Sec. 6 directly on the integrated depth, i.e. no shape model is used. This leads to inferior results in areas where the data evidence is small. Using our pro-

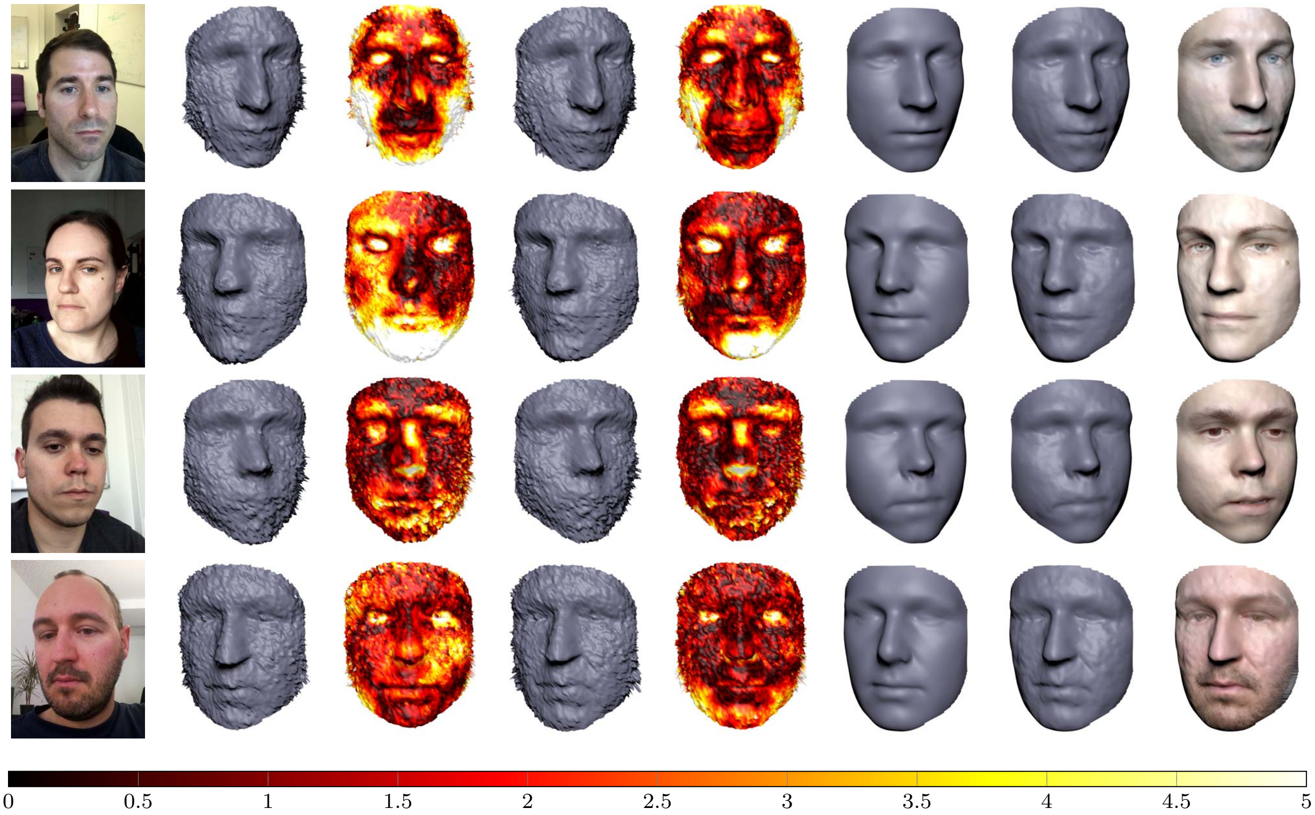


Figure 5. Results computed on a mobile device using the proposed approach. From left to right: example of input image, integrated depth before alignment, distance of integrated depth before alignment to mean face, integrated depth after alignment, distance of integrated depth after alignment to mean face, fitted height map model, proposed approach, proposed approach with texture.

posed formulation we can recover the geometry with high accuracy.

8. Results

If not stated explicitly in the text all the results in the paper have been generated with the following settings. The camera center and the mirror parameter are set to $C_0 = (0, 20, -20)^\top$ and $\xi = 50$, respectively. The height map resolution is set to 100×100 pixels. The alignment threshold is set to $\tau^A = 20$. The number of principal components of the height map face model $\mathcal{F}_\mathcal{H}$ have been set to $q = 35$. The optimization parameters have been set to $\epsilon = 0.5$ and $\lambda = 10$. All models are optimized using 1000 iterations. All the final results presented in Fig. 5 have been computed on a LG Nexus 5 or Motorola Nexus 6 smart phone. The extrinsic calibrations, depth maps and initial landmark based alignment are computed in real-time on the mobile device using the methods presented in [38, 24, 34]. The resolution of the depth maps is 320×240 pixels. Our unoptimized implementation on average requires $40ms$ to integrate a single depth map, $1.3s$ for the alignment, $80ms$ for the model fitting and $1.5s$ for the optimization. Additionally the computation of the depth maps using the method proposed in [24] requires $170ms$ per depth map. The computation of the depth map and the integration into the height map representation can be done online while scanning. The respective runtimes on a commodity PC running an Intel Core i7-2700K CPU at 3.50GHz are $13ms$ for the depth integration of a single depth map, $130ms$ for the alignment, $20ms$ for the model fitting and $150ms$ for the optimization.

9. Conclusion

We presented an efficient and accurate method for reconstructing faces on commodity mobile devices. Our experimental evaluation shows that our model is able to accurately recover the facial geometry and even recovers instance specific shape details. We showed several model of faces which are fully computed on a mobile phone in only a few seconds. Future work could improve the speed and robustness of the method by using discriminatively trained classifiers. Using an atlas of multiple height maps could be a direction to further improve the accuracy of the method.

Acknowledgements This work was supported by Grant 16703.1 PFES-ES of CTI Switzerland.

References

[1] S. Agarwal, K. Mierle, and Others. Ceres solver. http://ceres-solver.org. 4

[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. 1, 2

[3] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid icp algorithms for surface registration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 2

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

[5] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. 29(4):40, 2010. 1

[6] G. Blais and M. D. Levine. Registering multiview range data to create 3d computer objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):820–824, 1995. 4

[7] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, 1999. 1, 2

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

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

[10] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011. 5, 6

[11] Y. Chen and G. Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992. 4

[12] 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. 6

[13] 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

[14] 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 *British Machine Vision Conference (BMVC)*, 2008. 2

[15] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 32(8):1362–1376, 2010. 1, 2

[16] C. Geyer and K. Daniilidis. A unifying theory for central panoramic systems and practical implications. In *European conference on computer vision*, pages 445–461. Springer, 2000. 2

[17] C. Häne, L. Heng, G. H. Lee, A. Sizov, and M. Pollefeys. Real-time direct dense matching on fisheye images using plane-sweeping stereo. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 57–64. IEEE, 2014. 7

[18] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. 2

[19] V. H. Hiep, R. Keriven, P. Labatut, and J.-P. Pons. Towards high-resolution large-scale multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[20] V. Kazemi, C. Keskin, J. Taylor, P. Kohli, and S. Izadi. Real-time face reconstruction from a single depth image. In *International Conference on 3D Vision (3DV)*, 2014. 2

[21] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Eurographics symposium on Geometry processing (SGP)*, 2006. 2

[22] M. Klingensmith, I. Dryanovski, S. Srinivasa, and J. Xiao. Chisel: Real time large scale 3d reconstruction onboard a mobile device using spatially hashed signed distance fields. 2015. 2

[23] R. Koch, M. Pollefeys, and L. Van Gool. Multi viewpoint stereo from uncalibrated video sequences. In *European conference on computer vision (ECCV)*, 1998. 2

[24] K. Kolev, P. Tanskanen, P. Speciale, and M. Pollefeys. Turning mobile phones into 3d scanners. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 8

[25] 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 *IEEE International conference on computer vision (ICCV)*, 2007. 2

[26] V. Lempitsky and Y. Boykov. Global optimization for shape fitting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 2

[27] C. Mei and P. Rives. Single view point omnidirectional camera calibration from planar grids. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3945–3950. IEEE, 2007. 2

[28] R. A. Newcombe and A. J. Davison. Live dense reconstruction with a single moving camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[29] 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 *IEEE international symposium on Mixed and augmented reality (ISMAR)*, 2011. 2

[30] 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 *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pages 127–136. IEEE, 2011. 3

[31] 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

[32] A. Patel and W. A. Smith. 3d morphable face models revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[33] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. 2009. 1, 2, 3, 4, 5, 6

[34] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014. 2, 8

[35] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on*, pages 145–152. IEEE, 2001. 4

[36] 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 *International Conference on 3D Vision (3DV)*, 2015. 2

[37] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *European Conference on Computer Vision (ECCV)*, 2014. 2

[38] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys. Live metric 3d reconstruction on mobile phones. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 2, 8

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

[40] B. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962. 4

[41] C. Wu et al. Visualsfm: A visual structure from motion system. 2011. 7

[42] R. Yang and M. Pollefeys. Multi-resolution real-time stereo on commodity graphics hardware. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 2

[43] C. Zach. Fast and high quality fusion of depth maps. In *International symposium on 3D data processing, visualization and transmission (3DPVT)*, 2008. 2, 7

[44] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust tv-l 1 range image integration. In *IEEE International Conference on Computer Vision (ICCV)*, 2007. 2