Model-Based Teeth Reconstruction

Chenglei Wu[2] Derek Bradley[1] Pablo Garrido[3] Michael Zollhöfer[3]
Christian Theobalt[3] Markus Gross[1,2] Thabo Beeler[1]
1) Disney Research 2) ETH Zurich 3) Max Planck Institute for Informatics

Figure 1: *We present an algorithm to reconstruct teeth (right) from photographs (left). The reconstructed teeth accurately fit the input data as can be seen when overlaid over the input images (center).*

Abstract

In recent years, sophisticated image-based reconstruction methods for the human face have been developed. These methods capture highly detailed static and dynamic geometry of the whole face, or specific models of face regions, such as hair, eyes or eye lids. Unfortunately, image-based methods to capture the mouth cavity in general, and the teeth in particular, have received very little attention. The accurate rendering of teeth, however, is crucial for the realistic display of facial expressions, and currently high quality face animations resort to tooth row models created by tedious manual work. In dentistry, special intra-oral scanners for teeth were developed, but they are invasive, expensive, cumbersome to use, and not readily available. In this paper, we therefore present the first approach for non-invasive reconstruction of an entire person-specific tooth row from just a sparse set of photographs of the mouth region. The basis of our approach is a new parametric tooth row prior learned from high quality dental scans. A new model-based reconstruction approach fits teeth to the photographs such that visible teeth are accurately matched and occluded teeth plausibly synthesized. Our approach seamlessly integrates into photogrammetric multi-camera reconstruction setups for entire faces, but also enables high quality teeth modeling from normal uncalibrated photographs and even short videos captured with a mobile phone.

Keywords: Teeth Capture, Face Reconstruction, Teeth Modeling

Concepts: •**Computing methodologies** → **Reconstruction; Computer graphics; Shape modeling;**

SA '16 Technical Papers,, December 05 - 08, 2016, , Macao
ISBN: 978-1-4503-4514-9/16/12
DOI: http://dx.doi.org/10.1145/2980179.2980233

ACM Reference Format
Wu, C., Bradley, D., Garrido, P., Zollhöfer, M., Theobalt, C., Gross, M., Beeler, T. 2016. Model-Based Teeth Reconstruction. ACM Trans. Graph. 35, 6, Article 220 (November 2016), 13 pages.
DOI = 10.1145/2980179.2980233 http://doi.acm.org/10.1145/2980179.2980233.

1 Introduction

Digital humans have become ubiquitous in our everyday lives, from digital actors in visual effects and games to virtual patients in the medical field. Over the past decades, both research and industry have made tremendous progress when it comes to the creation of digital faces, mostly focusing on appearance, shape and deformation of skin. More recently, some work has emerged which also targets other facial features, such as the eyes and facial hair. However, capturing the mouth cavity in general, and teeth in particular, has only received very little attention so far. Teeth contribute substantially to the appearance of a face, and as can be seen from Fig. 11, expressions may convey a very different intent if teeth are not explicitly modeled. Furthermore, teeth can be an invaluable cue for rigid head pose estimation and are essential for physical simulation, where they serve as a collision boundary. In addition, in medical dentistry, digital teeth models have long become a central asset, since they allow to virtually plan a patient's procedure.

Not surprisingly, most of the effort to capture teeth stems from the medical dental field. While acquired with plaster-cast imprints in the past, more and more intra-oral scanners are making their way into clinics. While these devices can capture the shape of the teeth at high quality, the capture procedure is very invasive and the devices themselves are costly and not readily available. Minimally invasive systems such as photogrammetric camera rigs, which have become the de facto standard for facial capture, have so far not been able to faithfully reconstruct the teeth at high quality. This is mainly due to the complex appearance properties of teeth. Teeth are both extremely specular due to the translucent enamel coating and highly diffuse due to the underlying dentine, both of which exhibit strong subsurface scattering. Consequently, teeth have only few visible features, the strongest being the boundary between individual teeth, which is not even a feature on the surface, thus reconstructing teeth using photogrammetric approaches is very challenging. On the upside, teeth are rigid and their shape variation from subject to subject is manageable, and as such teeth render themselves well to statistical modeling. Camera-based reconstruction of the mouth interior is further complicated by non-trivial occlusions. It is often hard for people to open the mouth sufficiently wide without the use of dedicated lip spreading devices, and even then the entire mouth cavity is typically not visible from a single pose.

In this paper, we therefore propose the first method for non-invasive reconstruction of a detailed person-specific geometric model of an entire tooth row from a sparse set of normal photographs of the

mouth region. In these images, the person can make natural mouth expressions, without the requirement to uncomfortably spread the mouth with mechanical support. Our algorithm is based on several contributions. First, we contribute a new parametric prior model of an entire tooth row that is learned from a digitized database of high-quality plaster casts. This tooth model encodes the local shape variation of each individual tooth, the pose variation of each tooth within the entire tooth row, as well as the global position and scale of a tooth row. It also encodes prior distributions of the model parameters. Second, we contribute a new image-based approach that reconstructs a person-specific tooth row, which matches the visible teeth in the input images and synthesizes plausible geometry for partially occluded and completely hidden teeth on the basis of the prior model. Our algorithm only requires minimal user interaction and can operate on a set of individual, uncalibrated images, making teeth capture as easy and convenient as taking a few pictures or even a short video clip recorded with a standard mobile phone. Furthermore, our method naturally extends traditional photogrammetric face capture systems, increasing the quality of digital face scanning.

2 Related Work

Capturing the face, and teeth in particular, has been a topic of interest in both digital media for entertainment as well as medical dentistry. In the following we highlight related work in both fields.

Face Capture for Digital Media. Nowadays, photo-real digital characters [Alexander et al. 2009; Alexander et al. 2013] are extensively used in digital media production, i.e. as virtual actors in movies and games, or avatars in virtual reality or telepresence applications. Realistic rendering of animated human heads is of particular importance. In the past, graphics and vision research developed sophisticated methods to reconstruct shape and appearance models of static and dynamic human heads. Current state-of-the-art approaches allow to capture facial geometry [Beeler et al. 2010; Ghosh et al. 2011] and the skin's reflectance properties [Weyrich et al. 2006] in high detail based on photogrammetric approaches with controlled multi-camera and light setups. Using similar capture setups, even dynamic face geometry [Beeler et al. 2011; Valgaerts et al. 2012] and up-to micro-scale skin deformations [Nagano et al. 2015] can be captured. Some methods also work on monocular video [Garrido et al. 2013; Suwajanakorn et al. 2014; Shi et al. 2014; Wu et al. 2016], and in real-time [Cao et al. 2014; Cao et al. 2015]. Other works developed dedicated image-based methods to reconstruct head regions that are hard to capture with the aforementioned setups, such as the moving eyelids [Bermano et al. 2015], the eyeball [Bérard et al. 2014], or hair [Beeler et al. 2012; Echevarria et al. 2014; Hu et al. 2015]. In combination, these approaches enable capturing high quality models of almost all directly visible parts of a head. However, reconstruction of one important face region that is essential for realistic rendering has been largely neglected so far, the mouth interior. Especially, rendering a realistic tooth row is crucial for believable animation of the mouth region. Some works reconstruct parts of the mouth interior, such as the tongue or palate surface using sophisticated sensor modalities like MRI [Hewer et al. 2014; Hewer et al. 2015]. A much more practical image-based reconstruction of the mouth interior in general, and the tooth row in particular, is highly challenging due to non-trivial self occlusions, drastic lighting changes in the mouth interior, and complex material properties, such as specularities and subsurface scattering. Up to now, there is no passive photogrammetric approach for reconstructing detailed actor-specific tooth rows, and in most practical animation settings, the mouth interior is modeled by artists in a tedious manual process.

Recently, methods for face reenactment and face replacement in monocular video were proposed [Dale et al. 2011; Garrido et al. 2015; Suwajanakorn et al. 2015]. These often employ some form of parametric face model, but are challenged by their inability to reconstruct a tooth row which would greatly facilitate mouth re-rendering. Some methods resort to image-based retrieval strategies to resynthesize the mouth from similar video frames or a dedicated image-database [Dale et al. 2011; Garrido et al. 2014; Kawai et al. 2014; Suwajanakorn et al. 2015; Kim et al. 2015; Thies et al. 2016], which often leads to ghosting and temporal aliasing artifacts. The virtual dubbing method by Garrido et al. [2015] and the real-time RGB-D face reenactment by Thies et al. [2015] resort to a generic tooth row model aligned with a blendshape face. Ichim et al. [2015] also use a generic tooth row model for reconstructing computer game avatars from multi-view imagery. This yields plausible re-rendering results, but the true person-specific mouth appearance, in particular under extreme facial expressions, is not reproduced.

All these methods would profit from our method's ability to reconstruct personalized high-quality teeth from a few images or a short video clip.

Tooth Reconstruction in Medical Dentistry. Reconstruction of individual teeth plays an important role in dentistry to identify medical problems and plan surgical procedures such as tooth restoration. Under these controlled conditions, more invasive approaches can be afforded. Models of a patient's teeth are frequently created from plaster cast imprints, which are digitized using a laser scanner, or extracted from CT recordings [Omachi et al. 2007; Yanagisawa et al. 2014]. Abdelmunim et al. [2011] has build a database of individual teeth from such data and Binh Huy et al. [2009] has proposed an interactive approach for segmentation of upper and lower teeth from CT data. These methods yield high-quality models, but are time consuming, expensive, and may not be without health risks through radiation. New intra-oral scanners, such as the *3M™ True Definition* scanner, the *iTero* scanner and the *3Shape TRIOS®* pod scanner simplify individual tooth scanning[1]. They all use variants of active structured light scanning. Unfortunately, they are highly expensive and are only available to medical experts.

Sadly, due to the aforementioned non-trivial occlusions in the mouth interior and the difficult appearance properties of teeth, camera-based photogrammetric reconstruction, even of individual teeth, is very challenging. Teeth are featureless in general (except for a few occlusion boundaries and creases), and they are highly specular and exhibit strong subsurface scattering. Some previous work tried to exploit these reflectance properties by using variants of shape-from-shading [Carter et al. 2010; Farag et al. 2013; Mostafa et al. 2014]. These methods are still challenged by occlusions and often capture incomplete medium-quality models of a single tooth or the occlusal (i.e. biting) surface. A different category of approaches use contour data in combination with a small set of feature points to reconstruct a single tooth's occlusal surface [Zheng et al. 2011]. High-quality reconstruction of a complete teeth row, from just image data, has yet to be demonstrated.

To alleviate the problem of incomplete tooth reconstruction (due to occlusions), statistical models that represent shape variation of individual teeth in a low-dimensional subspace were studied [Mehl et al. 2005], inspired by similar models of faces [Blanz and Vetter 1999]. Some image-based reconstruction approaches using parametric models employ shading cues and adapted reflectance models [Carter et al. 2010; Farag et al. 2013; Mostafa et al. 2014], others expect scan point clouds as input [Munim et al. 2007]. A closed form solution for reconstruction has also been proposed [El Munim and Farag 2007]. Statistical models can also predict the original shape of a damaged tooth and thus be used for tooth restoration and inlay recon-

[1] www.3m.com;www.itero.com;www.3shapedental.com

struction [Blanz et al. 2004; Mehl and Blanz 2005; Buchaillard et al. 2007], again for a single tooth. The use of a statistical model for an entire tooth row was shown by Farag et al. [2013] in their statistical shape-from-shading approach, however their method requires a very specific oral cavity input image that must be taken from inside the mouth, which is far more invasive than the external hand-held photos that our method can operate on. We present the first statistical tooth model that encodes global and local variation in tooth shape and position in a unified framework, allowing to reconstruct plausible teeth from extra-oral imagery.

All previous approaches share one or more of the following three limitations: (1) they are invasive and unpleasant for the patient; (2) capturing is a tedious, expensive, and time-consuming process that may require expert knowledge; and (3) reconstruction is limited to a single tooth. In this paper, we propose the first end-to-end approach to reconstruct detailed geometric models of all teeth in an entire row (upper and lower tooth set) from normal photographs of the mouth region. Our versatile and easy method opens up new possibilities in face capture for media production, and could enable novel means of doctor-patient communication in medical fields.

3 Teeth Prior Model

We wish to reconstruct teeth given sparse image data with large occlusions, which can be made tractable with a statistical teeth prior. To this end, in this section, we define a parametric model of teeth and train a prior on the model, given 3D data from real subjects. Our model can be used later on to fit to image data and recover complete upper and lower tooth rows, even under occlusions.

Humans typically have 32 individual teeth (28 if wisdom teeth are removed), separated into two rows (top and bottom) with basic symmetry. The teeth are divided into four categories: incisors, canines, premolars and molars, as illustrated in Fig. 2. We wish to define a parametric model of the teeth that encodes 1 - the local shape variation of each individual tooth, 2 - the pose variation of each tooth within a row of teeth, and 3 - the global position and scale of a row of teeth. Furthermore, we wish to learn a prior on the parameters of the model given a set of high-quality 3D teeth scans from dentistry. In the following, we first describe how we create a teeth database from the dentistry data (Section 3.1), then provide the definition of the teeth model (Section 3.2), and finally train the model to learn a prior for fitting teeth to new subjects, given the teeth database (Section 3.3). Throughout, we will describe the process for only the top row of teeth, as the bottom row is completely analogous.

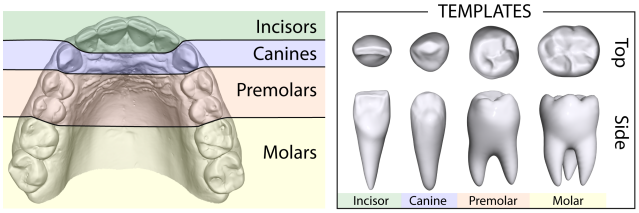

Figure 2: *Human teeth are divided into four main categories: incisors, canines, premolars and molars. We manually create a 3D template mesh for each category.*

3.1 Data Preparation

To build a teeth database we obtained high resolution plaster cast 3D scans of 86 different teeth rows from the field of medical dentistry, with a mixture of upper and lower teeth. A subset of the scans are shown in Fig. 3. While the scans contain detailed tooth geometry, they also contain the surrounding gums and are not in correspondence across subjects. Furthermore, there is no semantic segmentation of the meshes into individual teeth.

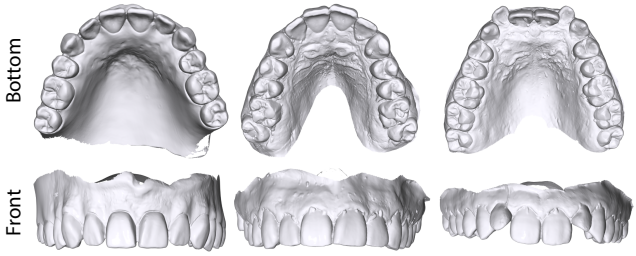

Figure 3: *We obtain a set of high resolution 3D scans for training a teeth model, however the teeth are not segmented nor are they in correspondence across subjects.*

Teeth Templates. In order to build and train a model, we need a teeth database with separate per-tooth geometry, which is in correspondence across subjects. We start by artistically creating a tooth template mesh. Since the four categories of teeth are quite different in shape, we create four separate template meshes (see Fig. 2). For completeness, we model the teeth all the way to the roots.

Template Fitting. We now wish to fit instances of the teeth template meshes to the individual teeth in the plaster cast scans. This will simultaneously solve the segmentation problem and place the teeth in vertex correspondence across subjects. We devise a semi-automatic template fitting approach. First, a user defines a segmentation contour for each tooth by clicking a few points at the inter-tooth boundaries and the boundaries between teeth and gums. This is illustrated in Fig. 4.a, where the contour for the left incisor is highlighted in purple and the other contours are shown in green. The segmentation contours are computed automatically by following high curvature paths between the selected points. In addition, the user selects a few predefined landmarks per tooth (three for incisors and canines, and five for premolars and molars), which will guide the registration and seed the segmentation. Fig. 4.a shows the landmarks for one of the incisors in red. Segmentation is performed by flood-filling from the selected landmarks until the segmentation contours are reached (Fig. 4.b). Finally, for each tooth the appropriate template mesh is first rigidly aligned to the tooth given the selected landmarks, and then non-rigidly deformed to tightly fit the segmented tooth region using iterative Laplacian deformation [Sorkine et al. 2004], with soft vertex constraints computed as the closest surface point along the normal direction in each iteration (Fig. 4.c). Once registered, we additionally compute a mask indicating which part of the template corresponds to the segmented tooth, and also mark the line of vertices corresponding to the gum boundary on the aligned template. Since the remainder of the template has just been deformed as rigidly as possible, we will not consider it when computing our teeth prior.

The result after fitting to all scans is a database of tooth rows with per-tooth mesh correspondence. Note that although several of the steps above required manual interactions, building the database is a one-time investment.

3.2 Parametric Teeth Model

We now describe a parametric model that defines a row of teeth. Since we plan to fit this model to new subjects it is important that our model can account for the variation of local shape and arrangements of teeth. At the same time, we expect that when fitting later, several

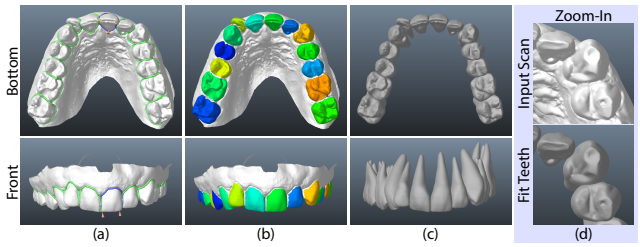

Figure 4: *We align the template meshes to the teeth scans by segmenting the individual teeth and then non-rigidly deforming a template to each tooth. (a) shows the user interaction which generates the segmentation in (b). The final fitted templates shown in (c) closely match the input scans as can be seen in the zoom-in (d).*

teeth will be at least partially occluded, and so the model must also be able to plausibly fill in under-constrained regions. For these reasons, the model is defined for a row of teeth and we separate the global parameters that define the general shape of a row from the local parameters that define the local placement and shape of each tooth.

Mathematically speaking, the model encodes the deviation in shape and pose from a *canonical tooth row*, which is computed as the mean of the database, shown in Fig. 5. Specifically, for every tooth τ, the canonical parameters include the average shape per tooth $\mathcal{S}_\tau^c$ as well as the average pose of the tooth T_τ^c in the tooth row. In addition, we compute a shape-subspace $\mathcal{B}_\tau$ per tooth that encodes its variation in shape. These properties are extracted from our teeth database as described in Section 3.3 and remain invariant during fitting later on. Thus, the degrees of freedom during fitting are parametrized by the shape coefficients $\mathbf{a}_\tau$ as well as a rigid transformation matrix T_τ that encodes the relative pose variation per tooth. In addition, we include global parameters that affect the teeth row as a whole, namely a rigid transformation T as well as anisotropic scaling along all axes Φ. With these parameters, the tooth row model is evaluated for each tooth τ as follows

$$\mathcal{Z}_\tau = T\ \Phi\ T_\tau\ T_\tau^c \left(\mathcal{S}_\tau^c + \sum_i^{|\mathcal{B}_\tau|} \mathbf{a}_\tau^i \mathcal{B}_\tau^i \right), \qquad (1)$$

where $\mathcal{Z}_\tau$ is the final shape and pose reconstruction of a synthesized tooth in the row, under the model parameters $\{T, \Phi, T_\tau, \mathbf{a}_\tau\}$.

3.3 Teeth Prior Training

Evaluating the model defined in Eq. 1 for a set of teeth provides an instance of a tooth row that represents either the upper or lower teeth for an individual. We now wish to train the model to obtain a prior that will yield plausible reconstructions of teeth. Training is performed on the teeth database computed in Section 3.1.

Tooth Row Model. The tooth row model is defined by the global parameters T and Φ, as well as the canonical positions T_τ^c and local transformations T_τ of each tooth. To train the tooth row model, we require all the database tooth rows to be globally aligned (rigidly plus anisotropic scale). To this end, we arbitrarily choose one tooth row as a reference and align all other tooth rows to the reference by computing the rigid transformations and anisotropic scales. Once aligned, we compute the mean tooth row by averaging the corresponding vertices of every sample. We manually define the coordinate system of the mean row by placing the origin between the two frontmost incisors as follows: we set the y-axis to point in the direction from the teeth roots to the crowns, the z-axis to point towards the mouth cavity, and the x-axis to form a right-hand coordinate system with the other axes.

We now again compute global rigid transformations and anisotropic scale to align all samples to the mean tooth row. Unlike the rigid transformation that accounts for global pose, the anisotropic scale is due to anatomical differences in the population. Consequently, we wish to quantify this variation such that it can be employed as a prior during fitting later on. As it is reasonable to assume the population follows a normal distribution, we model the prior on global anisotropic scale by a multivariate Gaussian distribution $\mathcal{N}_\Phi$ over the three degrees of freedom.

Now for an individual tooth τ, we rigidly align the corresponding tooth template to the mean tooth to obtain the canonical tooth transformations T_τ^c. We can then express the corresponding tooth samples of the aligned database tooth rows in the local coordinate frame T_τ^c of the tooth. The pose variation within this local coordinate frame represents the remaining pose residual for this tooth type, which we wish to quantify mathematically so that it can be employed as a prior in the fitting step later on. Again, assuming a normal distribution of our samples and a dependency of the individual variables in the rigid transformation, we construct a multivariate Gaussian distribution $\mathcal{N}_{T_\tau}$ over the six degrees of freedom of the local transformation (three for translation and three for rotation). Note that we learn a separate Gaussian distribution for the local transformation of each tooth. Furthermore, we can now remove the local rigid transformations and align all samples of a tooth class, such that the only residual left is due to shape variation, which we will capture in our Local Shape Model in the next paragraph.

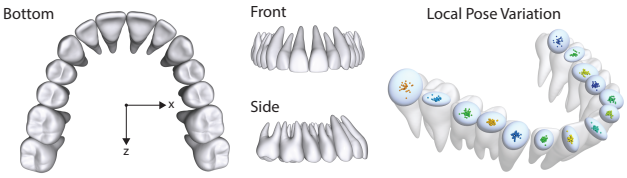

Figure 5: *The canonical tooth row shown on the left is computed as the mean of our database. Aligning all samples of the database globally to the canonical model, allows to quantify the remaining local pose variation as shown on the right. The colored dots correspond to the centers of the aligned database samples, and the large ellipses visualize the computed Gaussian prior distribution at 2σ.*

Local Shape Model. With all samples of a tooth type aligned as described above, the only variation left is due to the shape which we will capture through subspace analysis. We again assume the samples follow a normal distribution and employ Principal Component Analysis (PCA) to determine the optimal subspace. PCA provides the mean tooth shape, represented in our model as the canonical shape $\mathcal{S}_\tau^c$, as well as an orthogonal shape basis $\mathcal{B}_\tau$. To avoid overfitting, we truncate the basis to include 95 percent of the energy, which is approximately 4 components for incisors, 6 for canines, and 10 for both premolars and molars, on average. For each tooth, we also mark the ring of vertices corresponding to the average gum line observed in the scans (Fig. 9).

As shown in Fig. 6, the two major eigenmodes are consistent across teeth. The first one intuitively encodes variation in tooth length, whereas the second one captures thickness changes. Higher modes capture smaller shape changes such as local asymmetries, which differ for different teeth classes. The computed eigenvalues define a zero-mean normal distribution $\mathcal{N}_{\mathbf{a}_\tau}$ for each shape parameter.

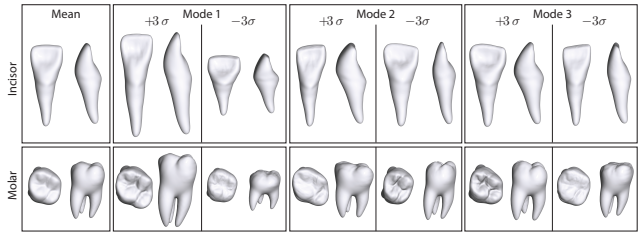

Figure 6: *We show the major modes of our shape subspaces for an incisor and a molar tooth. The modes are displayed at $\pm 3\sigma$. The first two modes roughly correspond between all teeth, with the first mode corresponding to length, and the second roughly to overall thickness. Higher modes account for tooth specific shape details, e.g. how pronounced the crown is in the case of a molar.*

As a final step, we replace the teeth in the mean tooth row with the canonical shapes $\{\mathcal{S}_\tau^c\}$ by fitting the shape model to the individual teeth, leading to the canonical tooth row shown in Fig. 5. The resulting teeth prior is a parametric model that defines the position and shape of a tooth row with both global and local control. In addition, it provides a trained statistical prior on the variation of model parameters based on a database of high-quality dentistry scans.

4 Teeth Fitting

We now describe our image-based fitting method for reconstructing teeth. Our approach involves automatic teeth boundary extraction (Section 4.1), estimating the teeth model parameters from the boundaries (Section 4.2 and Section 4.3), fine-scale out-of-model deformation to reconstruct exact teeth shape (Section 4.4), and finally recovering teeth color textures from the imagery with an option to incorporate 3D gums for visualization (Section 4.5).

4.1 Teeth Boundary Extraction

As discussed earlier, teeth exhibit very few visual features due to their complex appearance properties. We found the most reliable feature to be the silhouette, based on which we will formulate our optimization as described in the following sections.

Teeth silhouettes could be obtained by manually labeling the input imagery. While this is feasible and our algorithm can operate on such data, it can quickly become very cumbersome with an increasing number of images. In particular, for the use case of video-based teeth reconstruction demonstrated in Section 5, manual annotation is impractical.

Figure 7: *To train the Boosted Edge Learning (BEL) detector, we manually label a set of training images. We distinguish between three classes, according to the occluding object: teeth (red), gums (green), and lips (blue). Only teeth boundaries are real silhouettes of at least one tooth.*

Instead, we wish to automatically detect the silhouettes in the input images. As shown in Fig. 7, we define three different types of teeth boundaries based on the type of occlusion: teeth, gums and lips. Note that we will only use the teeth and gum boundaries for fitting the tooth row and a simple gum model. Explicitly differentiating between gums and lips, however, is crucial to prevent erroneous alignment of the tooth model gum lines to the lip boundaries. To identify the boundaries, we employ the Boosted Edge Learning (BEL) algorithm proposed by Dollár et al. [2006]. BEL is a general-purpose supervised learning algorithm for edge and object boundary detection that classifies image pixels as boundaries based on a large set of generic fast features over a small image patch, including gradients, histograms of filter responses, and Haar wavelets at different scales. We train three separate detectors on a set of hand-labeled input images, each corresponding to one particular tooth boundary. A few examples of our training data are shown in Fig. 7. The output of BEL is a likelihood map $\mathcal{P}^*$ that encodes the probability that a pixel belongs to an edge, as can be seen in Fig. 8.b.

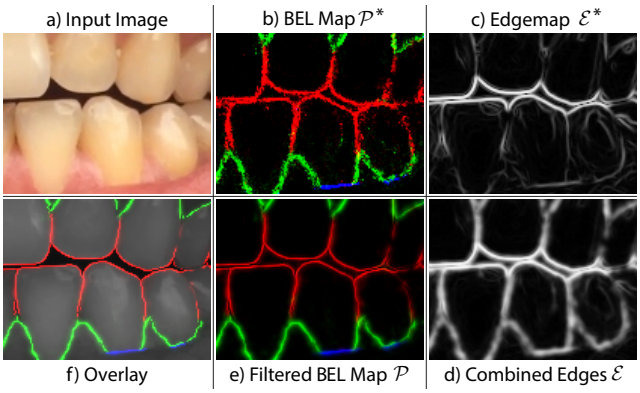

Figure 8: *Our goal is to automatically extract accurate teeth boundaries from the input images (a). A trained edge detector (BEL) provides robust, but coarse estimates of teeth boundaries (b). Accurate, but less robust edges are estimated using Gabor filters (c). These two maps are complementary, since BEL is robust even if teeth boundaries are very faint, such as in the upper right image, and Gabor is accurate if there are strong edges, e.g. between upper and lower teeth. The combined edge map (d) is used to refine the BEL probability (e) to yield accurate tooth boundaries (f) uniquely assigned to one of the three classes defined earlier; teeth (red), gums (green), and lips (blue).*

While BEL robustly detects the boundaries, the localization of the edge exhibits some uncertainty, especially when training on larger contours (we found 2px contours to provide adequate results). In the example provided in Fig. 8, BEL merges upper and lower teeth in places. Thus, we require an additional filtering step to reduce the noise and sharpen the edges. To better localize the correct position of an edge, we first convolve the original input images with a bank of Gabor filters [Grigorescu et al. 2002] at different orientations to obtain an edge map $\mathcal{E}^*$ (Fig. 8.c) and an estimate of the orientation of the edge $\vec{\mathcal{O}}^*$.

The edge detector performs very well where clear edges are visible in the image, but oftentimes fails to detect edges between teeth that are very smooth due to their appearance properties (see for example the edge between the two teeth in the upper right corner in Fig. 8). In such cases, the BEL output is superior and so we combine the two as

$$\mathcal{E} = \mathcal{E}^* + \gamma \mathcal{N}_\sigma * \mathcal{P}^*, \tag{2}$$

where $\mathcal{N}_\sigma * \mathcal{P}^*$ is a smoothed version of the BEL map, and γ is a scale factor, which we set to 1 for all our experiments. Similarly, we also augment the edge orientation map $\vec{\mathcal{O}}^*$ with orientations estimated from the smoothed BEL map. To further increase robustness,

this combined orientation map is then smoothed with a Gaussian ($\sigma = 1.4$px, weighted by $\mathcal{E}$) yielding $\vec{\mathcal{O}}$.

Based on the combined edge map $\mathcal{E}$ and smoothed orientation map $\vec{\mathcal{O}}$, the goal is now to diffuse the BEL probability map $\mathcal{P}^*$ such that the probabilities accumulate at the detected edges, thus improving the silhouette detection. Diffusion takes place in the directions orthogonal to the estimated orientation of the edge denoted by $\vec{\mathcal{O}}_\updownarrow$, where the specific direction is indicated by an up- or down-arrow, respectively. The diffusion speed is governed by the gradient of the edge map along the diffusion directions $\nabla_{\vec{\mathcal{O}}_\updownarrow}\mathcal{E}$. Mathematically speaking, the diffusion can be formulated as

$$\mathcal{P}^{k+1} = \mathcal{P}^k + \lambda\left(\nabla_{\vec{\mathcal{O}}_\uparrow}\mathcal{E}\cdot\mathcal{P}^k_{\vec{\mathcal{O}}_\uparrow} + \nabla_{\vec{\mathcal{O}}_\downarrow}\mathcal{E}\cdot\mathcal{P}^k_{\vec{\mathcal{O}}_\downarrow}\right), \tag{3}$$

starting with $\mathcal{P}^0 = \mathcal{P}^*$ and where $\mathcal{P}^k_{\vec{\mathcal{O}}_\updownarrow}$ denotes the probabilities along the diffusion directions. The diffused BEL probability map is renormalized to the range [0,1] and thresholded using hysteresis to further remove outliers, yielding the final probability map $\mathcal{P}$ shown in Fig. 8.e. Overlaying $\mathcal{P}$ over the original image (see Fig. 8.f) shows that our filtering step yields accurate teeth boundaries, which are uniquely assigned to one of the three classes defined earlier; teeth (red), gums (green), and lips (blue).

4.2 Teeth Model Parameter Estimation

We use the detected tooth and gum boundaries in the edge maps $\mathcal{P}$ to find the parameters of our teeth model that best explain the input. For convenience, we define the set of all parameters as:

$$\mathcal{X} = \{T, \Phi\} \cup \{T_\tau\}_{\tau=1}^N \cup \{\mathbf{a}_\tau\}_{\tau=1}^N . \tag{4}$$

Here, $N = 14$ is the total number of teeth in a given row, either upper or lower. The set consists of the global rigid transformation T, the global anisotropic scaling Φ, the local per-tooth rigid transform T_τ and the local per-tooth shape coefficients $\mathbf{a}_\tau$. We formulate the problem of finding the optimal model parameters that best explain a given set of input observations as a *Maximum A Posteriori* (MAP) estimation problem in the unknown parameters $\mathcal{X}^*$. To this end, we maximize the following posterior probability distribution:

$$\mathcal{X}^* = \underset{\mathcal{X}}{\operatorname{argmax}}\, p(\mathcal{X}|\mathcal{P}) = \underset{\mathcal{X}}{\operatorname{argmax}}\, p(\mathcal{P}|\mathcal{X})p(\mathcal{X}) . \tag{5}$$

Here, $p(\mathcal{P}|\mathcal{X})$ is the likelihood of observing the edge map $\mathcal{P}$ given the teeth row parameterized by $\mathcal{X}$. $p(\mathcal{X})$ is a prior on the distribution of teeth. In the following, we describe in detail how we formulate the teeth edge likelihood $p(\mathcal{P}|\mathcal{X})$ and the teeth prior $p(\mathcal{X})$.

Teeth Edge Likelihood. The teeth edge likelihood $p(\mathcal{P}|\mathcal{X})$ measures the probability that the edge map $\mathcal{P}$ has been generated by the teeth with parameters $\mathcal{X}$. We assume that the data points $\mathbf{c}_i \in \mathcal{P}$ were generated by independent random processes and thus $p(\mathcal{P}|\mathcal{X})$ is a product of per-sample point likelihoods:

$$p(\mathcal{P}|\mathcal{X}) = \prod_{i=1}^{|\mathcal{P}|} p(\mathbf{c}_i|\mathcal{X}) . \tag{6}$$

Note that since we define the per-sample likelihoods over the input edge-maps, we implicitly make use of the encoded visibility information. We split the per-sample point likelihoods into a product of two different likelihood functions:

$$p(\mathbf{c}_i|\mathcal{X}) = p_{point}(\mathbf{c}_i|\mathcal{X})\cdot p_{plane}(\mathbf{c}_i|\mathcal{X}). \tag{7}$$

Here, p_{point} models the distance and p_{plane} the tangential noise of the image formation process. This combination of point-to-point with point-to-plane likelihood functions has been demonstrated to exhibit good performance in other contexts, such as non-rigid registration [Zollhöfer et al. 2014]. We assume zero-mean and normal distributed noise models $\mathcal{N}(0, \sigma^2_{point})$ and $\mathcal{N}(0, \sigma^2_{plane})$ for the two likelihoods. The distance likelihood p_{point} encodes the Euclidean distance in the image domain between a detected contour point $\mathbf{c}_i$ and the projection $\hat{\mathbf{c}}_i$ of the corresponding silhouette point, and is defined as

$$p_{point}(\mathbf{c}_i|\mathcal{X}) \propto \exp\left[-w_i\frac{1}{2}\left(\frac{||\mathbf{c}_i - \hat{\mathbf{c}}_i||_2}{\sigma_{point}}\right)^2\right] . \tag{8}$$

The weight w_i specifies the confidence in the contour detection $\mathbf{c}_i$ and is given by $\mathcal{P}(\mathbf{c}_i)$, and $\sigma_{point} = \sqrt{500}$. The tangential likelihood p_{plane} tolerates sliding along the silhouette, while penalizing deviation from the detected contour point $\mathbf{c}_i$ along the projection of the corresponding silhouette normal $\hat{\mathbf{n}}_i$. The tangential likelihood is given as

$$p_{plane}(\mathbf{c}_i|\mathcal{X}) \propto \exp\left[-w_i\frac{1}{2}\left(\frac{\langle\hat{\mathbf{n}}_i, (\mathbf{c}_i - \hat{\mathbf{c}}_i)\rangle}{\sigma_{plane}}\right)^2\right] , \tag{9}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product, the weight w_i corresponds to the confidence of the detected edge and $\sigma_{plane} = \sqrt{10}$.

Teeth Prior. Fitting the tooth row model to the set of detected edges is a highly underconstrained inverse problem, since many instances of the tooth row may satisfy the constraints given above. In addition, the problem is highly non-linear in the unknown model parameters $\mathcal{X}$. We tackle both problems with a statistical prior learned from the captured high-quality training data, as described in Section 3. The prior guides the optimizer through the highly complex energy landscape and allows discriminating between likely and unlikely teeth configurations. The four types of model parameters $\{T, \Phi, T_\tau, \mathbf{a}_\tau\}$ and all teeth are assumed to be statistically independent, thus dividing the prior into:

$$p(\mathcal{X}) \propto p_{scale}(\Phi)\cdot\prod_{\tau=1}^{N}\left[p_{local}(T_\tau)\cdot p_{shape}(\mathbf{a}_\tau)\right] . \tag{10}$$

We model the individual priors of p_{scale}, p_{local} and p_{shape} based on multivariate normal distributions (please refer to Sec. 3.2 for further details). Note, we do not use a prior on the global transformation T, since all poses are assumed to be equally likely.

Extension: Camera Parameter Estimation. For an uncalibrated camera setup, instead of solving for a single global rigid transformation T, we solve for the set of per-camera transformations $\{T_\nu\}_{\nu=1}^{N_\nu}$, where N_ν denotes the number of viewpoints. The intrinsics of the camera(s) are assumed to be fixed and calibrated in advance. This can be considered a special case of rigid structure-from-motion [Webb and Aggarwal 1982], where the structure is regularized by the proposed teeth prior, which allows to accurately and robustly calibrate the viewpoint positions at the same time as optimizing for the model parameters.

4.3 Optimization

In the proposed MAP estimation problem, the model data ($\hat{\mathbf{c}}_i$) that corresponds to a sample point $\mathbf{c}_i$ is a hidden variable. Therefore, we alternate k times between computing model correspondences $\hat{\mathbf{c}}_i$ for the detected edge points $\mathbf{c}_i$ (both teeth and gum boundaries) and

optimizing for the model parameters. These two steps are iterated until convergence or for a predefined number of iterations. The convergence criterion is met if the change in the residual is lower than a small threshold. This can be interpreted as a variant of the Expectation Maximization (EM) algorithm [Dempster et al. 1977].

E-Step. To determine the correspondences, we project the visible silhouette points and the gum line points of the teeth into the viewpoints. The teeth are evaluated for the latest guess $\mathcal{X}^k$ of the model parameters from EM iteration k. For every detected contour point $\mathbf{c}_i$, we find the best correspondence by finding the projected silhouette point $\hat{\mathbf{c}}_i$ for the teeth contours (and analogously the projected gum line point for gum contours) that minimizes

$$\hat{\mathbf{c}}_i = \underset{\hat{\mathbf{c}}_j}{\operatorname{argmin}} ||\mathbf{c}_i - \hat{\mathbf{c}}_j||_2^2 \cdot \exp\left[-\left(\frac{\langle \mathbf{n}_i, \hat{\mathbf{n}}_j \rangle}{\sigma_{angle}}\right)^2\right], \qquad (11)$$

where $\mathbf{n}_i$ is the image-space normal at the contour point $\mathbf{c}_i$ and $\sigma_{angle} = 0.3$. Intuitively, this equation finds the closest point with similar orientation to the detected edge point.

M-Step. After updating the correspondences, we optimize for updated model parameters $\mathcal{X}^{k+1}$ assuming the correspondences remain fixed. We tackle this optimization by minimizing the negative log-likelihood as follows:

$$\mathcal{X}^{k+1} = \underset{\mathcal{X}}{\operatorname{argmax}}\, p(\mathcal{X}|\mathcal{P}) = \underset{\mathcal{X}}{\operatorname{argmin}} \left[-\log p(\mathcal{X}|\mathcal{P}) \right]. \qquad (12)$$

Since the teeth edge likelihood and the teeth prior are governed by normal distributions, the resulting optimization problem reduces to a standard non-linear least-squares problem, which we solve using the Gauss-Newton (GN) method. Since the optimization problem is highly non-linear, a simultaneous optimization of all parameters is unlikely to converge to the global optimum, especially if the initialization is far from the detected input contours. To alleviate this problem, we apply a progressive coarse-to-fine optimization strategy. First we start optimizing only the global model parameters, namely T and Φ. After a few iterations (10 for all performed experiments), we add the local transformations T_τ to the parameter set. Finally, the shape coefficients $\mathbf{a}_\tau$ are included in the parameter set and we optimize the full model. Due to noise and outliers in the detected edge maps, we also employ a robust correspondence filtering strategy based on the median absolute deviation δ_i (see also Rousseeuw and Lerow [1987]):

$$\delta_i = \epsilon \cdot \underset{j \in \mathbf{B}_i}{\operatorname{median}} \left|\left|\mathbf{c}_j - \hat{\mathbf{c}}_j\right|\right|_1. \qquad (13)$$

Here, $\mathbf{B}_i$ is the set of edge pixels captured by the same camera and belonging to the same tooth as $\mathbf{c}_i$. The scalar $\epsilon = 1.4826$ is a theoretical correction factor. A detected edge pixel $\mathbf{c}_i$ is flagged as an outlier if $\left|\left|\hat{\mathbf{c}}_i - \mathbf{c}_i\right|\right|_1 > \eta \cdot \delta_i$. In all our experiments, we use the penalization threshold $\eta = 2.5$.

Initialization and Implementation. We need to start our highly non-linear optimization from a reasonable initial guess $\mathcal{X}^0$ of the model parameters to succeed. Parameters are initialized to the mean of their corresponding normal distribution. The only parameter that is not governed by a normal distribution is the global pose T of the model. To this end, we devise a lightweight and intuitive approach to initialize T by asking the user to identify two teeth per row in at least one viewpoint. Depending on the use case, more viewpoints might be required (see Section 5). From the two strokes and preselected vertices on the tooth model, we get four very rough 2D-3D correspondences which are sufficient to compute a rough initial guess for T. Note that this guess might still be very inaccurate as the user is allowed to draw the strokes freely on the teeth. We thus refine the initialization by finding the closest points in the edge map $\mathcal{P}$ to the two strokes and assigning them to the respective teeth. We now run our optimization restricting the E-step to select only correspondences from this set for the two teeth, which will provide the final initial guess. This is the only manual step of our pipeline, and the proposed method to identify two teeth per row is fast, intuitive and can be conducted by non-technical operators with no CG background, e.g. medical personnel, in a standard image viewer. In practice, we found the gum line constraint is less reliable than the tooth boundary constraint. Therefore, the gum line constraint is drastically downweighted in the first EM iterations, and only in the last iteration is treated equally with the tooth boundary constraint.

4.4 Out-of-Model Deformation

The optimization will yield teeth geometry that matches the detected silhouettes as closely as possible within the shape subspace defined by our tooth model. Some teeth might have shapes that are slightly outside of this subspace, which has been designed to only capture the major tooth variation to serve as a robust prior. To overcome this slight mismatch, we conclude tooth shape reconstruction with an out-of-model deformation step. For each tooth contour point $\mathbf{c}_i$ we perform another E-Step to find corresponding vertices on the tooth silhouette and compute the closest point along the ray through $\mathbf{c}_i$ as a target position for deforming the tooth out-of-model. Based on these correspondences, we perform Laplacian deformation, while keeping the root vertices fixed. The out-of-model deformation provides very close fits to tooth edge maps.

4.5 Color and Gums

To produce compelling visual reproductions of the teeth, we additionally compute their color and incorporate the gums. Color is computed per vertex from the input images, after segmenting the teeth using color filtering. We filter the images conservatively to avoid adding any skin color to the teeth. For every tooth that has been partially colored, we solve Laplace's equation on the surface to smoothly fill in the rest. Teeth that are completely occluded are colored as the average of their neighboring (partially visible) teeth.

We finalize the tooth row by fitting a generic 3D model of gums, which makes the reconstruction more realistic and complete. Starting with a template gum mesh created by an artist, we label the edge-loops on the mesh that correspond to boundaries with the individual teeth, as shown in Fig. 9.a. Then, given a reconstructed tooth row we can use the average gum lines computed from the database in Section 3.3 (Fig. 9.b) as constraints for an iterative non-rigid Laplacian deformation scheme [Sorkine et al. 2004]. In each iteration, correspondences are computed from the labeled gum vertices to the closest corresponding tooth gum line point, which slowly deforms the gums into position (Fig. 9.c). This approach has the advantage that even if the gums are completely occluded (which is oftentimes the case) we will reconstruct a plausible gum shape. But on the other hand, if the gums are in fact visible we can use the gums component of the filtered BEL Map $\mathcal{P}$ as constraints for fitting, which we show for several results in Section 5.

5 Results

We demonstrate the applicability of the proposed method on three different use cases, 1 - integration with an existing photogrammetric face scanning system, 2 - reconstruction from an unstructured set of uncalibrated input images, and 3 - progressive teeth recon-

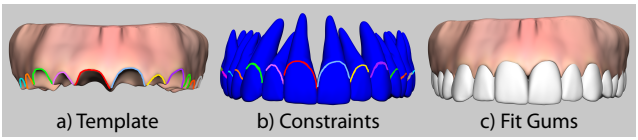

Figure 9: *We complete the tooth row by fitting a generic gum template (a), to either the detected gum line, when available, or otherwise the computed average gum line per tooth (b), using a non-rigid deformation scheme (c).*

struction from handheld video capture. Finally, we also provide evaluations of our method by comparing quantitatively to ground truth teeth acquired using state-of-the-art intra-oral acquisition, comparing against a simple retrieval-based method, and evaluating the effect of the training database size on reconstruction accuracy.

Calibrated Multi-View Capture. Passive photogrammetric face capture has become widely used in several areas, including the entertainment industry. These systems offer a fast and convenient means to acquire the face geometry at high resolution using a set of cameras. We demonstrate that the proposed method seamlessly integrates into these systems without requiring any modification of the existing hardware. To this end, we show integration into the system proposed by Beeler et al. [2010], but other systems would work just as well. Their system simultaneously captures a set of eight calibrated viewpoints of the person to be scanned. In a preprocessing step, we label 40 images randomly selected from different viewpoints, which serve as input data to train the edge detector (see Section 4.1). This training is a one-time upfront effort, and no retraining is required for future subjects. When capturing a new subject, the operator identifies two teeth per row in the frontal two viewpoints to initialize the system, leading to a total of 8 strokes.

Figure 11: *The proposed method can be seamlessly integrated into existing facial capture setups, without the need to change any hardware. Teeth are highly important for the perception of expressions.*

Fig. 10 and Fig. 1 show several captured subjects plus reconstructed teeth. Most people tend to only expose their upper teeth when smiling naturally. Fig. 11 shows the reconstructed teeth of one person combined with a scan of her face, all computed from the same eight input images. A nice feature of the proposed method is that it can fit teeth rows to partially incomplete data. For example, the molars are typically not visible, yet still our method produces plausible teeth reconstructions. This capability is further explored in Fig. 12, where we simulate missing teeth by manually removing their detected silhouettes. As can be seen, the system succeeds at suggesting plausible teeth, since it propagates model coefficients, e.g. overall scale, from visible teeth as part of the optimization.

Uncalibrated Multi-View Capture. Since teeth are rigid, we do not require that all viewpoints are captured simultaneously, but instead can acquire them one by one. While taking longer than using

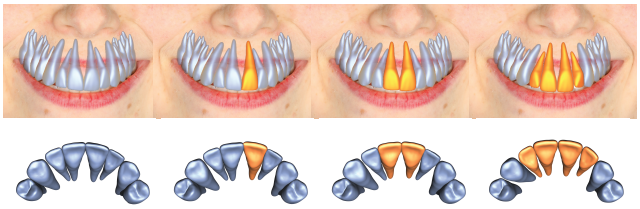

Figure 12: *A nice feature of our method is its capability to fit to partial data. In this example, we manually removed the silhouettes of 1, 2 and 4 frontal teeth to simulate such data. Our method is not only robust, but also suggests plausible teeth.*

a rig such as in the previous use case, this approach has the advantage that a single handheld camera is sufficient and the subjects can move their lips to expose more of the teeth when taking an image from a particular viewpoint. We employ a single Canon Rebel T5 Camera with 60mm macro lens, on which we mount a flash. Flash and lens are cross-polarized to reduce specularities and produce comparable images as in the previous use case. This allows us to re-use the same edge detector without re-training. We pre-calibrate the camera intrinsics, which is again a one-time upfront effort only, required once per camera. Unlike the previous use case, the extrinsics of the viewpoints are unknown and thus we ask the operator to identify two teeth per row and viewpoint for initialization. These can be different teeth for every viewpoint, leading to a total of $4N_\nu$ input strokes. Fig. 13 shows the estimated camera positions (right) as well as the computed teeth model (center) overlaid over one of the 11 input image (left).

Figure 13: *For uncalibrated setups, our algorithm can jointly optimize for shape (center) and camera parameters (right).*

To evaluate the accuracy of our camera estimation results, one option would be to apply a structure-from-motion (SfM) algorithm on the input images and compare to the resulting camera positions. However, the input images are captured without requiring the actor to hold the expression, which is completely suitable for our teeth capture method, but will introduce additional errors in the SfM results. Therefore, we evaluate our camera estimation accuracy on one of our calibrated multi-view datasets, simply by ignoring the camera calibration during reconstruction. For a set of 8 images capturing the face region from front, left, right and below at approximately 1 meter away, our estimated cameras contained an average positional error of 4 cm and average rotation error of 2.1 degrees.

Progressive Reconstruction from Video. As a last use case, we present the progressive reconstruction from input video captured by a handheld device, such as an iPhone. Video has the advantage that we can leverage thousands of viewpoints, which is great for silhouette based reconstruction methods. While we could treat this use case as a special case of uncalibrated multi-view capture, the requirement to label every viewpoint would require several thousands of user input strokes, which is not practical. Instead, we leverage the temporal coherence of the video footage, which guarantees that adjacent viewpoints exhibit only a small baseline. Therefore, we can initialize a viewpoint with the extrinsic of the previous time step and afterwards jointly optimize for camera and model parameters.

Figure 10: *We reconstructed the teeth of several subjects captured in a multi-view photogrammetric system (8 cameras). Most people tend to only expose their upper teeth, unless explicitly asked to show both upper and lower teeth as in the bottom row.*

Therefore, user input is only required for a single frame, yielding only 4 input strokes to reconstruct both upper and lower teeth.

While the strategy to re-optimize teeth and camera parameters for every frame works, it is computationally expensive. We found that the teeth are only refined slightly at every frame, and that co-optimization is only required when viewpoints add substantially new information. Therefore, we propose an adaptive scheme, that keeps the teeth fixed and only optimizes camera parameters, as long as the final residual of the optimization is below a given threshold. In these cases, the current estimation of the teeth model matches the observed data reasonably well. Once the model fails to sufficiently explain the observed data for a given frame, we select $\tilde{N}_\nu$ viewpoints from the already optimized cameras by sampling uniformly over the solid angle covered by the views and refine the teeth model by co-optimizing camera and model parameters. Then, we continue to process the remaining views, again optimizing for camera parameters only, until the residual exceeds the predefined threshold. Once all frames have been optimized, we run a final co-optimization step by again selecting $\tilde{N}_\nu$ viewpoints as representatives.

We found this approach produces accurate teeth reconstructions, while requiring only very few co-optimizations (less than 10, compared to thousands using the naïve approach), providing a drastic speed-up. Fig. 14 gives a quantitative illustration of the process. The first estimate (blue curve) of the model, reconstruced from a front view, is sufficient to track up to frame 173, where the residual exceeds the preset threshold of 1.5 pixels for the first time. The method then selects 10 views (red stars) to estimate an updated model. The new model is already sufficient to track to all frames (red curve). One more co-optimization step (yellow stars) produces the final teeth model, which exhibits the lowest overall tracking error (yellow curve). For illustration, we computed the residuals for all frames to show the overall improvement. In practice, tracking would take place in one go with progressively improved models. The remaining residual is due to noise and inaccuracies in the edge detection and the lack of expressive power of the model to exactly fit the detected silhouettes. This last inaccuracy can be further reduced using the out-of-model step described in Section 4.4. Fig. 15 shows

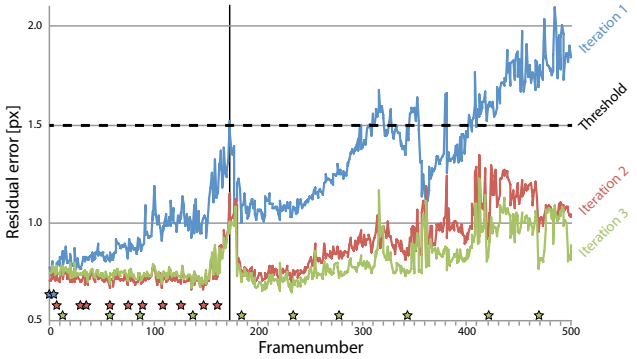

Figure 14: *For video data we propose a progressive reconstruction scheme, where teeth parameters are kept fixed when solving for camera extrinsics until the mean residual per detected edge point is higher than a threshold (1.5px in this case). At that point, we select a subset of already tracked cameras sampled uniformly based on their viewing angle (stars). We then jointly optimize for teeth and camera parameters to update the model and continue tracking. Above, the model is updated three times (blue, red, yellow) - once at the beginning, once after exceeding the threshold at frame 173, and a final time after all frames have been processed. This approach is both efficient and robust, as well as accurate, see Fig. 15.*

the result obtained by progressive tracking of a video acquired with a handheld device (iPhone 6). The video was captured outdoors without special equipment, such as polarizers used in the previous use cases. Therefore, the imagery is visually quite different, which is why we retrain the edge detector on five frames of the video. Also, we pre-calibrate the intrinsics of the device and undistort the images before processing. The resulting teeth are very accurate, since our method benefits from the large number of viewpoints that constrain the silhouette based optimization.

Comparison with Retrieval-Based Teeth Fitting. The majority of healthy people have the same number of teeth, arranged in the same order with incisors in front followed by canines, premolars and then molars. This apparent global similarity across subjects might suggest that a simple retrieval-based method for finding the most similar teeth row in the database is sufficient for estimating the teeth of an unseen individual. However, the shape and structure of teeth across subjects are in fact very unique (one reason why dental records are used for person-identification in forensics), and thus a retrieval-based method does not lead to accurate results. Given an incredibly large database, the likelihood of finding a match with acceptable error would increase, but creating such a database would be impractical. Our statistical teeth row model can in a unified way well-explain the individual local tooth variation as well as the global variation of the whole teeth row. This enables a compact yet highly expressive model that requires far fewer training samples than any retrieval-based method for comparable accuracy. We demonstrate this point in Fig. 16, where we compare our reconstruction to the closest matching teeth row from our training database after optimizing for the global rigid transformation that best aligns the teeth to the input images. This shows that a retrieval-based method cannot fit the teeth well, in particular due to the limited size of the database, but our model built from the same training data is visibly more accurate.

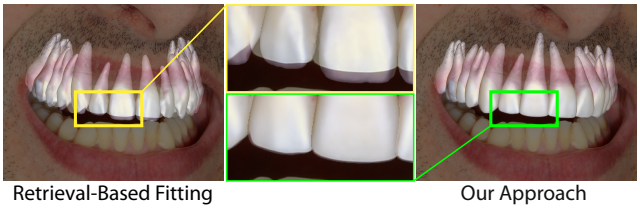

Figure 16: *We compare our model-based fitting method to retrieval-based fitting, which finds the closest fitting teeth row from the database. Here we clearly see that retrieval-based fitting cannot accurately recover the teeth of an unseen individual.*

Quantitative Evaluation. Finally, we assess the accuracy of our extra-oral teeth reconstruction technique based on 8 images with ground-truth data acquired at a dental clinic using an intra-oral scanner. Obviously, the two modalities have very different properties. Where the intra-oral scanner provides highly accurate reconstructions, it is invasive and not easily accessible. Our proposed method reconstructs teeth that closely match the ground-truth data within a few millimeters of accuracy, as can be seen in Fig. 17, yet requires only a few photographs from afar, which can be captured quickly and easily.

In order to evaluate the robustness of our system, we further assess the accuracy of our reconstruction while varying the size of the teeth database that is used to train the teeth prior model. Table 1 shows reconstruction errors for database sizes of 50, 40, 30, 20 and 10 training teeth rows. The reconstruction example is the same as in Fig. 17, and the reported accuracy is the average Euclidean error over all non-root vertices across all teeth. This evaluation shows that

Figure 15: *We reconstructed the teeth from a short videoclip (1000 frames) captured with a handheld device (iPhone 6) outdoors. The algorithm recovers both camera parameters and accurate teeth geometry by progressively processing the data.*

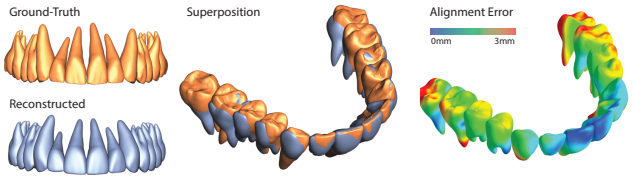

Figure 17: *We assess the accuracy of the proposed approach by comparing the reconstructed teeth of a subject captured in the calibrated multi-view setup (blue) with an intra-oral scan (orange). Since only the frontmost teeth are visible, we rigidly align the two models using the frontmost 6 teeth (incisors and canines). Our method is capable of reconstructing the teeth within a few millimeters accuracy, from purely extra-oral images.*

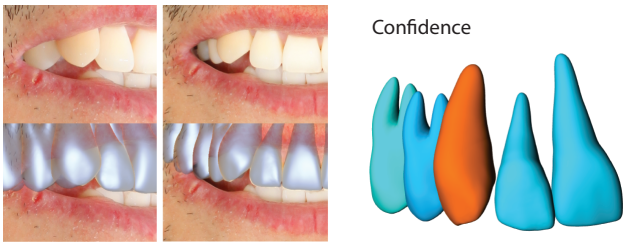

Figure 18: *The canine of this person has a shape far outside the prior (left). While our method can not accurately reconstruct the shape of such outliers, it provides a confidence measure (right) of the reconstruction, using the remaining per-tooth residual. This can be used to flag problematic teeth that require special attention.*

our method is robust to varying database sizes as quality gracefully degrades when the amount of training data is reduced.

Database Size	50	40	30	20	10
Average Error (mm)	0.86	0.93	0.95	0.97	1.05

Table 1: *Our method is robust to the database size, as we show by varying the amount of training data and computing the average Euclidean error of the (non-root) teeth vertices for the reconstruction example shown in Fig. 17.*

Limitations and Future Work. Our approach is the first to reconstruct a personalized teeth model of high quality based on a lightweight capture setup, but it still has several limitations. Our model is based on a statistical prior and as such limited to the variation present in the database. Most inaccuracies are small thanks to the out-of-model deformation step, but we typically miss some of the high frequency details, such as sharp edges or creases. In one case, shown in Fig. 18, our method was not able to faithfully reconstruct a complete tooth, since the particular tooth shape is far outside the prior. Still, our method can provide an indication of confidence through the remaining per tooth residual, displayed on the right of Fig. 18. This could be used to flag problematic teeth, which could be remodeled and added to the training data to extend the expressiveness of the prior. Even though casually captured uncalibrated photographs are handled, teeth have to be sufficiently exposed to allow for reliable contour detection. As this is a silhouette based method, more views provide significantly more constraints. Note, unnatural mouth spreading with a mechanical support, as in some previous methods, or intra-oral mirrors are not required, but could be employed to increase the number of teeth that can be reconstructed. The automatic silhouette detection requires teeth to be captured in an environment that resembles the one it has been trained on. This could be alleviated by training on a larger set with varying environments. Currently, initialization of our model requires simple manual interaction, which could easily be replaced by an automatically trained tooth detector or augmenting a parametric face model with a generic tooth row. The latter could also improve video-based tooth reconstruction under fast motion.

In the future, additional constraints could be considered to improve teeth reconstruction. Geometric matching of occlusal surfaces of corresponding upper and lower teeth (in particular molars) or explicit collision handling can provide valuable constraints. Also, shape-from-specularity could be employed to acquire high-frequency geometric detail and we would also like to estimate a more sophisticated personalized tooth appearance model that captures the specularity and subsurface scattering of teeth due to their two layer structure. Additional parts of the mouth and mouth cavity, such as lips, the tongue or the wall of the cavity could be captured. This not only enables reconstruction of even more complete face models, but also provides additional constraints to our teeth reconstruction approach, such as collisions between teeth and lips. A more complex tooth row prior model, for example one that uses Markov chains, could better encode the relationship of neighboring teeth (or opposing teeth in upper and lower rows). Finally, the reconstruction of teeth from a single uncalibrated image, like a selfie, and the incremental reconstruction, while providing live visual feedback, are still challenging unsolved problems.

6 Conclusion

We proposed the first model-based approach for reconstructing a personalized high-quality 3D teeth model, given just a sparse set of (uncalibrated) images or a short monocular video sequence as

input. Unlike related approaches in the medical field, ours is non-invasive and reconstructs a geometric model of the entire tooth row including the gums from photographs captured from afar and potentially simultaneously. To this end, we leverage the statistical information of a novel parametric tooth prior learned from high-quality 3D dental scans that models the global deformations of an entire tooth row as well as the individual variation of each single tooth. Fitting is based on a new optimization approach that leverages the visual contour information in the images to accurately align the model to photographs, where the teeth are visible, and also plausibly synthesizes partially occluded teeth. Apart from a few input strokes to identify two teeth, the proposed pipeline is fully automatic.

Our approach opens up new ways to reconstruct personalized teeth at high quality and brings teeth reconstruction within reach of commodity use, without requiring expensive specialized equipment. We believe that our versatile and easy-to-use approach will be useful in medical dentistry for previsualization and to enable novel ways of doctor-patient communication, as well as in the entertainment industry, where the proposed method to reconstruct teeth and gum models seamlessly integrates into existing photogrammetric multi-camera setups for face capture.

Acknowledgements

We thank the reviewers for their valuable comments and our test subjects for their time and patience. We are also grateful to Angiels Diaz for labeling the 2D lip contours to train the BEL detector, and to Maurizio Nitti for creating the 3D teeth templates. This work would also not be possible without the teeth scans provided by Prof. Dr. Irena Sailer and Vincent Fehmer. Finally, this work was supported by the ERC Starting Grant CapReal (335545).

References

ABDELMUNIM, H., CHEN, D., FARAG, A., PUSATERI, R., CARTER, C., MILLER, M., FARMAN, A., AND TASMAN, D. 2011. A 3d human teeth database construction based on a point-based shape registration. In *IEEE ICIP*, 1617–1620.

ALEXANDER, O., ROGERS, M., LAMBETH, W., CHIANG, M., AND DEBEVEC, P. 2009. The digital emily project: Photoreal facial modeling and animation. In *ACM SIGGRAPH 2009 Courses*, 12:1–12:15.

ALEXANDER, O., FYFFE, G., BUSCH, J., YU, X., ICHIKARI, R., JONES, A., DEBEVEC, P., JIMENEZ, J., DANVOYE, E., ANTIONAZZI, B., EHELER, M., KYSELA, Z., AND VON DER PAHLEN, J. 2013. Digital ira: Creating a real-time photoreal digital actor. In *ACM SIGGRAPH 2013 Posters*, 1:1–1:1.

BEELER, T., BICKEL, B., BEARDSLEY, P., SUMNER, B., AND GROSS, M. 2010. High-quality single-shot capture of facial geometry. *ACM Trans. Graphics (Proc. SIGGRAPH) 29*, 4, 40:1–40:9.

BEELER, T., HAHN, F., BRADLEY, D., BICKEL, B., BEARDSLEY, P., GOTSMAN, C., SUMNER, R. W., AND GROSS, M. 2011. High-quality passive facial performance capture using anchor frames. *ACM Trans. Graphics (Proc. SIGGRAPH) 30*, 4, 75:1–75:10.

BEELER, T., BICKEL, B., NORIS, G., MARSCHNER, S., BEARDSLEY, P., SUMNER, R. W., AND GROSS, M. 2012. Coupled 3d reconstruction of sparse facial hair and skin. *ACM Trans. Graphics (Proc. SIGGRAPH) 31*, 117:1–117:10.

BÉRARD, P., BRADLEY, D., NITTI, M., BEELER, T., AND GROSS, M. 2014. High-quality capture of eyes. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 33*, 6, 223:1–223:12.

BERMANO, A., BEELER, T., KOZLOV, Y., BRADLEY, D., BICKEL, B., AND GROSS, M. 2015. Detailed spatio-temporal reconstruction of eyelids. *ACM Trans. Graphics (Proc. SIGGRAPH) 34*, 4, 44:1–44:11.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99*, 187–194.

BLANZ, V., MEHL, A., VETTER, T., AND SEIDEL, H.-P. 2004. A statistical method for robust 3d surface reconstruction from sparse data. In *3DPVT*, 293–300.

BUCHAILLARD, S. I., ONG, S. H., PAYAN, Y., AND FOONG, K. 2007. 3d statistical models for tooth surface reconstruction. *Comput. Biol. Med. 37*, 10, 1461–1471.

CAO, C., HOU, Q., AND ZHOU, K. 2014. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graphics (Proc. SIGGRAPH) 33*, 4, 43:1–43:10.

CAO, C., BRADLEY, D., ZHOU, K., AND BEELER, T. 2015. Real-time high-fidelity facial performance capture. *ACM Trans. Graphics (Proc. SIGGRAPH) 34*, 4, 46:1–46:9.

CARTER, C., PUSATERI, R., CHEN, D., AHMED, A., AND FARAG, A. 2010. Shape from shading for hybrid surfaces as applied to tooth reconstruction. In *IEEE ICIP*, 4049–4052.

DALE, K., SUNKAVALLI, K., JOHNSON, M. K., VLASIC, D., MATUSIK, W., AND PFISTER, H. 2011. Video face replacement. In *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 30, 130:1–130:10.

DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY 39*, 1, 1–38.

DOLLÁR, P., TU, Z., AND BELONGIE, S. 2006. Supervised learning of edges and object boundaries. In *IEEE CVPR*, 1964–1971.

ECHEVARRIA, J. I., BRADLEY, D., GUTIERREZ, D., AND BEELER, T. 2014. Capturing and stylizing hair for 3d fabrication. *ACM Trans. Graphics (Proc. SIGGRAPH) 33*, 4, 125:1–125:11.

EL MUNIM, H., AND FARAG, A. 2007. Shape representation and registration using vector distance functions. In *IEEE CVPR*, 1–8.

FARAG, A., ELHABIAN, S., ABDELREHIM, A., ABOELMAATY, W., FARMAN, A., AND TASMAN, D. 2013. Model-based human teeth shape recovery from a single optical image with unknown illumination. In *Medical Computer Vision: Recognition Techniques and Applications in Medical Imaging (MCV '12)*, 263–272.

GARRIDO, P., VALGAERT, L., WU, C., AND THEOBALT, C. 2013. Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 32*, 6, 158:1–158:10.

GARRIDO, P., VALGAERTS, L., REHMSEN, O., THORMÄHLEN, T., PEREZ, P., AND THEOBALT, C. 2014. Automatic face reenactment. In *IEEE CVPR*, 4217–4224.

GARRIDO, P., VALGAERTS, L., SARMADI, H., STEINER, I., VARANASI, K., PEREZ, P., AND THEOBALT, C. 2015. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *Comput. Graph. Forum 34*, 2, 193–204.

GHOSH, A., FYFFE, G., TUNWATTANAPONG, B., BUSCH, J., YU, X., AND DEBEVEC, P. 2011. Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 30*, 6, 129:1–129:10.

GRIGORESCU, S., PETKOV, N., AND KRUIZINGA, P. 2002. Comparison of texture features based on gabor filters. *IEEE Trans. Image Proc. 11*, 10 (Oct), 1160–1167.

HEWER, A., STEINER, I., AND WUHRER, S. 2014. A hybrid approach to 3D tongue modeling from vocal tract MRI using unsupervised image segmentation and mesh deformation. In *Interspeech*, 418–421.

HEWER, A., STEINER, I., BOLKART, T., WUHRER, S., AND RICHMOND, K. 2015. A statistical shape space model of the palate surface trained on 3D MRI scans of the vocal tract. In *18th International Congress of Phonetic Sciences (ICPhS)*.

HU, L., MA, C., LUO, L., AND LI, H. 2015. Single-view hair modeling using a hairstyle database. *ACM Trans. Graphics (Proc. SIGGRAPH) 34*, 4.

ICHIM, A. E., BOUAZIZ, S., AND PAULY, M. 2015. Dynamic 3d avatar creation from hand-held video input. *ACM Trans. Graphics (Proc. SIGGRAPH) 34*, 4, 45:1–45:14.

KAWAI, M., IWAO, T., MIMA, D., MAEJIMA, A., AND MORISHIMA, S. 2014. Data-driven speech animation synthesis focusing on realistic inside of the mouth. *Journal of Information Processing 22*, 2, 401–409.

KIM, T., YUE, Y., TAYLOR, S., AND MATTHEWS, I. 2015. A decision tree framework for spatiotemporal sequence prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 577–586.

LE, B. H., DENG, Z., XIA, J., CHANG, Y.-B., AND ZHOU, X. 2009. An interactive geometric technique for upper and lower teeth segmentation. In *Proceedings of the 12th International Conference on Medical Image Computing and Computer-Assisted Intervention*, MICCAI '09, 968–975.

MEHL, A., AND BLANZ, V. 2005. A new approach for automatic reconstruction of occlusal surfaces with the biogeneric tooth model. *Int. J. Comput. Dent. 8*, 13–25.

MEHL, A., BLANZ, V., AND HICKEL, R. 2005. Biogeneric tooth: a new mathematical representation for tooth morphology in lower first molars. *Eur. J. Oral. Sci. 113*, 333–340.

MOSTAFA, E., ELHABIAN, S., ABDELRAHIM, A., ELSHAZLY, S., AND FARAG, A. 2014. Statistical morphable model for human teeth restoration. In *IEEE ICIP*, 4285–4288.

MUNIM, H. E. A. E., FARAG, A., AND FARMAN, A. 2007. A new variational approach for 3d shape registration. In *International Symposium on Biomedical Imaging: From Nano to Macro (ISBI '07)*, 1324–1327.

NAGANO, K., FYFFE, G., ALEXANDER, O., BARBIČ, J., LI, H., GHOSH, A., AND DEBEVEC, P. 2015. Skin microstructure deformation with displacement map convolution. *ACM Trans. Graphics (Proc. SIGGRAPH) 34*, 4.

OMACHI, S., SAITO, K., ASO, H., KASAHARA, S., YAMADA, S., AND KIMURA, K. 2007. Tooth shape reconstruction from ct images using spline curves. In *Wavelet Analysis and Pattern Recognition*, vol. 1, 393–396.

ROUSSEEUW, P. J., AND LEROY, A. M. 1987. *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., New York, NY.

SHI, F., WU, H.-T., TONG, X., AND CHAI, J. 2014. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 33*, 6, 222:1–222:13.

SORKINE, O., COHEN-OR, D., LIPMAN, Y., ALEXA, M., RÖSSL, C., AND SEIDEL, H.-P. 2004. Laplacian surface editing. In *Proc. SGP*, 179–188.

SUWAJANAKORN, S., KEMELMACHER-SHLIZERMAN, I., AND SEITZ, S. M. 2014. Total moving face reconstruction. In *ECCV*.

SUWAJANAKORN, S., SEITZ, S. M., AND KEMELMACHER-SHLIZERMAN, I. 2015. What makes tom hanks look like tom hanks. In *Proc. ICCV*.

THIES, J., ZOLLHÖFER, M., NIESSNER, M., VALGAERTS, L., STAMMINGER, M., AND THEOBALT, C. 2015. Real-time expression transfer for facial reenactment. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 34*, 6.

THIES, J., ZOLLHÖFER, M., STAMMINGER, M., THEOBALT, C., AND NIESSNER, M. 2016. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.

VALGAERTS, L., WU, C., BRUHN, A., SEIDEL, H.-P., AND THEOBALT, C. 2012. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 31*, 6, 187:1–187:11.

WEBB, J. A., AND AGGARWAL, J. K. 1982. Structure from motion of rigid and jointed objects. *Artificial Intelligence 19*, 1, 107–130.

WEYRICH, T., MATUSIK, W., PFISTER, H., BICKEL, B., DONNER, C., TU, C., MCANDLESS, J., LEE, J., NGAN, A., JENSEN, H. W., AND GROSS, M. 2006. Analysis of human faces using a measurement-based skin reflectance model. In *ACM SIGGRAPH '06*, 1013–1024.

WU, C., BRADLEY, D., GROSS, M., AND BEELER, T. 2016. An anatomically-constrained local deformation model for monocular face capture. *ACM Trans. Graph. 35*, 4 (July), 115:1–115:12.

YANAGISAWA, R., SUGAYA, Y., KASAHARA, S., AND OMACHI, S. 2014. Tooth shape reconstruction from dental ct images with the region-growing method. *Dentomaxillofacial Radiology 43*, 6, 20140080.

ZHENG, S.-X., LI, J., AND SUN, Q.-F. 2011. A novel 3d morphing approach for tooth occlusal surface reconstruction. *Comput. Aided Des. 43*, 3, 293–302.

ZOLLHÖFER, M., NIESSNER, M., IZADI, S., REHMANN, C., ZACH, C., FISHER, M., WU, C., FITZGIBBON, A., LOOP, C., THEOBALT, C., ET AL. 2014. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (TOG) 33*, 4, 156.