Realistic Texture Extraction for 3D Face Models
Robust to Self-Occlusion

Chengchao Qu

Vision and Fusion Laboratory
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology (KIT), Germany
qu@kit.edu

Abstract: In the context of face modeling, probably the most well-known approach to represent 3D faces is the 3D Morphable Model (3DMM). When 3DMM is fitted to a 2D image, the shape as well as the texture and illumination parameters are simultaneously estimated. However, if real facial texture is needed, texture extraction from the 2D image is necessary. This paper addresses the problems in texture extraction of a single image caused by self-occlusion. Unlike common approaches that leverage the symmetric property of the face by mirroring the visible facial part, which is sensitive to inhomogeneous illumination, this work first generates a virtual texture map for the skin area iteratively by averaging the color of neighbored vertices. Although this step creates unrealistic, overly smoothed texture, illumination stays constant between the real and virtual texture. In the second pass, the mirrored texture is gradually blended with the real or generated texture according to the visibility. This scheme ensures a gentle handling of illumination and yet yields realistic texture. Because the blending area only relates to non-informative area, main facial features still have unique appearance in different face halves. Evaluation results reveal realistic rendering in novel poses robust to challenging illumination conditions and small registration errors.

1 Introduction

Facial analysis has attracted increasing attention in the computer vision and pattern recognition community despite decades of research and application. 3D face modeling has been widely used and proven to be effective in face recognition [BV03] and animation [BBPV03]. Probably the most well-known approach to represent 3D faces is the 3D Morphable Model (3DMM) proposed by Blanz and Vetter
Separate linear subspaces for shape and texture are learned by Principal Component Analysis (PCA) for a compact representation of the respective variations. When 3DMM is fitted to a 2D image, not only the shape coefficients, but also the texture and illumination parameters are simultaneously estimated, resulting in a huge parameter space.

The complete 3DMM fitting takes account of the learned PCA texture model and Phong illumination to minimize the fitting error of shape and appearance. To address the efficiency problem, a series of optimization efforts are made to allow for faster convergence [RV03] and larger convexity basin to avoid local minima [RV05] by introducing more image features (e.g., specular highlight and edge constraint) than just pixel intensity. These methods can reconstruct the texture parameters including the occluded facial part according to the 3DMM dataset and the estimated illumination at the cost of computational time, making these algorithms inappropriate for online applications. From another perspective, when artificially rendered texture is not desired, e.g., for forensic analysis, facial texture extraction from the 2D image is necessary.

As an alternative that is tailored to the above requirements, dependency on the facial appearance in the fitting can be completely abandoned and the 3D shape can be inferred solely based on a few dozens of sparse 2D feature points, allowing for real-time reconstruction. In the approach of Blanz et al. [BMVS04], using several manually annotated 2D feature points, the complete 3D shape and camera projection are reconstructed in closed-form by least squares fitting. Prior knowledge from the 3DMM dataset is utilized to solve this otherwise ill-posed problem. Since the texture is yet to be extracted from the image afterwards, the color values of the image are mapped to the vertices on the 3D model. However, the complex geometry of human faces results in self-occluded facial regions even for a frontal pose.

To alleviate the self-occlusion problem after texture extraction, a posterior step to estimate the global 3DMM texture parameters by assuming photometric invariance [AS10] is still applicable. For real texture extraction, Blanz et al. [BMVS04] reflect the visible part to fill the missing color values. At places where both parts are occluded, the average texture of 3DMM is applied. The strict constraints of constant skin and illumination can impose severe artifacts on the occluded textured face model. Similarly, the 3/4 profile face (±45°) is regarded as the most representative pose for texture extraction by Roy-Chowdhury et al. [RCCG05] and the mirroring approach is employed for occlusion handling. On the other hand, Jiang et al. [JHY+05] interpolate the blank area by averaging the intensity of the connected vertices. Because the algorithm only accepts frontal faces, the smearing effect introduced by interpolation only appears near the neck and ears, which is
not crucial to the overall visual quality. Combining multiple images from different viewpoints yields more realistic results than interpolation \cite{PHL+98}. Although the seamless blending constraint is carefully designed to handle inhomogeneous illumination, only results of studio face images in controlled environment are shown \cite{PHL+98, ZCS06}.

This paper addresses the possible problems in texture extraction and proposes a straightforward solution to generate realistic facial texture for self-occluded regions. This approach assumes the 3D shape of the face is already recovered. After registering the shape back to the 2D image, small displacements caused by the limited 3DMM subspace are dealt with by triangulation and warping. To overcome the self-occlusion problem, unlike the above mentioned approaches in the literature, which either is sensitive to illumination or generates interpolation artifacts, in this work, we combine the advantages of both approaches. The “bad” half of the face is first determined. Starting from the cheek near the nose area, a virtual texture map for the homogeneous area is created iteratively by averaging the color of neighboring visible vertices until the whole face is filled artificially. Although this step creates unrealistic, overly smoothed texture, illumination stays constant between the real and virtual texture. In the second pass, the mirrored texture is gradually blended with increasing weight. At places where the original vertices are visible, the real texture is used, otherwise, the generated texture is taken for blending. This scheme ensures a gentle handling of illumination and yet yields realistic texture. Because the blending area only relates to non-informative area, main facial features, \textit{i.e.}, eyes and mouth, still have unique appearance in different face halves, which is proven to be crucial to face recognition \cite{LSCM03}. The effectiveness of our approach is evaluated on several “in the wild” images and videos containing diverse pose and illumination variations.

The remainder of this paper is organized as follows. A brief introduction to 3DMM and efficient fitting using 2D feature landmarks is given in §2 and §3 respectively. The procedure of the proposed facial texture extraction framework is discussed in detail in §4. The qualitative results are demonstrated in §5. Finally, we conclude our work in §6.

## 2 3D Morphable Model

The 3D Morphable Model, introduced by Blanz and Vetter \cite{BV99}, is a class-specific model to describe 3D objects, especially human faces. 3DMM is composed of 3D geometry and texture constructed from 3D laser scans of human heads. After preprocessing to fill the holes from laser scans, the dense set
with a fixed size of $p$ vertices is put into full point-to-point correspondence by optical flow, so that morphing between faces is possible. As an example, the first shape and texture entry corresponds to the tip of the nose across all 3D faces in Fig. 2.1(a). The shape is represented in a vectorized form $s = \{x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_p, y_p, z_p\} \in \mathbb{R}^{3p}$. Different from its sibling Active Appearance Model (AAM) [CET98], where the texture is defined as the 2D image inside the convex hull of the feature points, the 3D texture is modeled on each of the $p$ vertices as $t = \{r_1, g_1, b_1, r_2, g_2, b_2, \ldots, r_p, g_p, b_p\} \in \mathbb{R}^{3p}$. By applying PCA on all face scans, the shape and texture can be expressed as a convex combination of mean vectors $\bar{s}$ and $\bar{t}$ and the eigenvectors

$$s = \bar{s} + S \text{ diag}(\sigma) \alpha,$$
$$t = \bar{t} + T \text{ diag}(\tau) \beta.$$

The columns of matrices $S \in \mathbb{R}^{3p \times m}$ and $T \in \mathbb{R}^{3p \times m}$ are $m$ eigenvectors and $\sigma \in \mathbb{R}^m$ and $\tau \in \mathbb{R}^m$ are the eigenvalues of the shape and texture respectively. Thus, given the registered 3D scans, 3DMM is mathematically represented as the set $\{\bar{s}, S, \sigma, \bar{t}, T, \tau\}$ and a novel face can be described using the 3DMM coefficients $\{\alpha, \beta\}$ inside the spanned subspace of the training data.

Due to the heavy workload for acquiring, processing and annotating 3D data, there are few public 3DMM datasets available. In this work, we utilize the Basel Face Model (BFM) from Paysan et al. within the group of Prof. Vetter [PKA09], who is the originator of 3DMM [BV99]. In BFM, besides the usual trained 3DMM parameters $\{\bar{s}, S, \sigma, \bar{t}, T, \tau\}$, a manually annotated mask to separate the area of the two eyes, the nose, the mouth and the rest skin region is also included, which is highlighted in Fig. 2.1(b). We make full use of this segmentation mask in our texture extraction method in §4.

3 3D Shape Reconstruction

3D shape reconstruction based on only a few 2D feature points offers a computationally efficient alternative compared to the complete 3DMM fitting with regard to the albedo, illumination and other image features. Blanz et al. [BMVS04] propose a non-iterative solution to recover the non-rigid shape deformation and the rigid motion simultaneously. For a set of $f \ll p$ sparse facial landmarks, the 2D coordinates on the image plane $y \in \mathbb{R}^2$ can be expressed as a linear combination of the projected 3D shape variations of the 3DMM. Assuming that the measurements are subject to uncorrelated Gaussian noise, the error function of 2D and 3D projection is equivalent to

$$\epsilon = ||Qc - y||^2 + \eta||c||^2$$
in Bayesian Maximum a Posteriori (MAP) formulation, where $Q$ combines the PCA eigenvectors of 3DMM, the known 3D-2D mapping and the projection, while $c$ contains the 3DMM shape coefficients $\alpha$. Blanz et al. [BMVS04] linearize the rotation, scaling and translation in the form of extra eigenvectors and shape coefficients, which are attached to $Q$ and $c$ respectively. In this way, a straightforward closed-form solution

$$c = (Q^\top Q + \eta I)^\dagger Q^\top y$$

in ridge regression is made possible and an initial pose estimation is not necessary.

The 2D sparse feature points are either manually annotated or localized by face alignment methods. Faggian et al. [FPS06] first integrate a generative person-specific AAM to localize the 2D coordinates of the feature landmarks to reconstruct the 3D dense face shape automatically. We build on our previous work [QMSB14], where the inconsistent 2D AAM and 3DMM landmark position emerging in the self-occluded area is addressed and 3D face reconstruction robust to pose changes for both static images and videos is proposed. The 66-point AAM landmarks are mapped to the BFM mesh as the prior 3D-2D correspondence for the automatic process. Fig. 2.1(b) illustrates the feature point scheme used in this paper. The reader is referred to [BMVS04, QMSB14] for details.
4 Texture Extraction

After 3D shape reconstruction in §3, the 3DMM shape coefficients $\alpha$ as well as the pose are recovered. Since the texture parameters $\beta$ are not available compared to complete 3DMM fitting, we extract the texture from the image directly. All interim stages for generating realistic texture under non-frontal poses are detailed in this section. As a graphical example, the George Clooney image downloaded from the Internet1 in Fig. 4.1(a) is reconstructed step by step.

4.1 Extraction of Visible Texture

Existing 3D face reconstruction methods mostly focus on the quality of reconstructed shape and do not elaborate on the texture extraction stage, which is only roughly mentioned within a few words, e.g., “the color values of the image are mapped as a texture on the surface” [BMVS04] or “the 2D image is directly mapped to the 3D geometry” [JHY+05]. However, we find that it is nontrivial to generate realistic texture for 3D models. To start from scratch, there are several problems to be solved. The first one is the limitation of the linear 3DMM subspace.

A 3DMM is usually trained with a few hundred 3D face scans, e.g., 100 for the original work [BV99], 200 for BFM [PKA+09], hence, the spanned PCA subspace has only a limited power to describe novel faces. As a result, the reconstructed 3D shape cannot always fit the 2D image perfectly. The deviation can be measured at the aligned 2D landmarks and the projected 3D correspondence. In Figs. 4.1(b)

\[\text{http://img.timeinc.net/time/photoessays/2007/george_clooney/george_clooney_01.jpg}\]
3D Facial Texture Extraction Robust to Self-Occlusion

Figure 4.2: Steps of texture extraction and generation: (a) Visible texture after extraction (b) Texture after erasing vertices near the visible boundary (c) Filled texture with interpolation (d) Initial texture for iterative blending

and (4.1(c)), the lower points of the nose and the left face contour are slightly different. To compensate the small offset, auxiliary points are added around the face to perform Delaunay triangulation and subsequently piecewise warping. Thanks to the dense 3D registration for shape and pose, only little correction is needed, hence, we are free of the unnatural visual affect by warping sparse AAM meshes [GES09]. The result is demonstrated in Fig. 4.1(c).

On account of the large number of vertices, it is common that multiple vertices are projected to the same image pixel. Visibility detection thus determines to which vertex or vertices the color value of the certain image pixel should be assigned. One choice is to use the z-buffer algorithm [PHL+98]. After rendering the 3D shape with the recovered pose parameters, the depth map of the scene is generated by comparing the depth of each vertex. Only the ones closest to the camera are set to visible. Alternatively, we utilize a simple and efficient hidden point detection algorithm by Katz et al. [KTB07]. Occlusion is determined with the vertices alone without the knowledge of surface and normal, etc. Afterwards, the color values on the visible vertices are extracted with non-uniform interpolation. The result in Fig. 4.2(a) shows few errors and outliers.

4.2 Inference of Occluded Texture

Existing research is already aware of the self-occlusion problem of the AAM landmarks for 3D shape reconstruction [QMSB14, LLP+12]. However, to the best of our knowledge, no prior work is dedicated to generating natural texture for the hidden facial area. To make better understanding of our approach, we first introduce a few useful operations. In order to infer the missing texture iteratively, we need to
find out the invisible vertices that are adjacent to the visible ones. Formally, given a set of visible and occluded vertices \( \{ \mathcal{V}, \overline{\mathcal{V}} \} \) and a set of edges \( \mathcal{E} \) in the 3D mesh, these candidates are defined as

\[
\mathcal{V}_\oplus = \{ v | v \in \overline{\mathcal{V}} \land \exists v' \in \mathcal{V} : (v, v') \in \mathcal{E} \}. \tag{4.1}
\]

On the contrary, the set of vertices to be removed are denoted as

\[
\mathcal{V}_\ominus = \{ v | v \in \mathcal{V} \land \exists v' \in \overline{\mathcal{V}} : (v, v') \in \mathcal{E} \}. \tag{4.2}
\]

To interpolate a hidden vertex \( v \in \overline{\mathcal{V}} \), its color value \( g(v) \) is the average color of the adjacent visible vertices

\[
g(v) = \frac{\sum_{v' \in \Omega} g(v')}{|\Omega|}, \text{ where } \Omega = \{ v' | v' \in \mathcal{V} \land (v, v') \in \mathcal{E} \}. \tag{4.3}
\]

Back to Fig. 4.2(a), the boundary texture near the occluded region shows some smearing effect. Comparison with the localized 2D landmarks in Fig. 4.1(b) suggest that the landmark noise and the nearly perpendicular normal direction in this area make the mapped texture prone to quality degradation. Therefore, the boundary region of the “bad” half of face is eliminated by a sequence of removal operations applied to the vertex set \( \mathcal{V}_\ominus \) in Eq. (4.2). Note that we preserve the facial features, i.e., eyes, nose and mouth, by applying the BFM segmentation mask (see Fig. 2.1(b)) and only clear the skin texture, yielding Fig. 4.2(b).

Some approaches leverage the symmetric property of the face by just mirroring the visible part of the face and provide some good-looking results on well illuminated images \cite{BMVS04, RCCG05}. Obviously, the example image Fig. 4.3(c) highlights two drawbacks of this simple strategy. First, even minor illumination difference between the mirrored parts will result in severe inhomogeneous intensity and poor visual effect. On the other hand, some facial regions, e.g., between the chin and neck, are often invisible in both face halves. To deal with these problems, a virtual texture map is first generated by filling the missing color values by iteratively interpolating \( \mathcal{V}_\oplus \) according to Eqs. (4.1) and (4.3). As can be seen in Fig. 4.2(c), although the filled texture is overly smoothed and lacks high-frequency details, the illumination remains constant.

In the final stage, this virtual texture map is blended with the mirrored visible half of face. The basic principle of blending is to maximize the usage of the “good” half of the face while keeping the unique features, e.g., the left and right eye, distinguishable. For this purpose, again the BFM segmentation mask (see Fig. 2.1(b)) offers an effective way to preserve only the best and most important facial texture. An example texture before the fusion process is given in Fig. 4.2(d). Starting
Figure 4.3: Final result of the proposed method for Fig. 4.1(a) in (a) frontal view and (b) novel view compared to (c) texture mirroring and (d) interpolation from this initial map, texture of new vertices selected by Eq. (4.1) are fused in an iterative manner with increasing weight

\[ g(v) = \lambda(i)g_{\text{mirror}}(v) + (1 - \lambda(i))g_{\text{interpolation}}(v), \]  

where \( \lambda(i) = \min\{\frac{i}{N}, 1\} \).

\( i \in \mathbb{N} \) denotes the iteration counter and \( N \) is the length of the blending process. The final result in Figs. 4.3(a) and 4.3(b) demonstrates that both smooth illumination transition and texture details even the beard are well preserved. The simple mirroring approach in Fig. 4.3(c) lacks a texture region under the ear, which is also invisible in the original face half. Moreover, abrupt texture transition and artifacts caused by registration error further degrade the visual quality. Fig. 4.3(d) looks overly smoothed and shows smearing artifact. A boundary of different intensity can be seen, indicating the opposite paths of the iterative interpolation process from the dark region in ear and the middle.

5 Experiments

In this section, the proposed texture extraction approach is compared with the baseline methods on the static Labeled Face Parts in the Wild (LFPW) image dataset [BJKK11] and the YouTube Celebrities video dataset [KKPR08].

LFPW is a relatively new face image dataset for testing face alignment algorithms with annotated landmarks. The images are downloaded from the Internet and contain large variations in pose, illumination, etc. As our example in the previous sections, BFM is utilized as our 3DMM for reconstruction. Because of the lack of diverse expressions when capturing BFM, only images that have approximately
Table 5.1: Comparison of the proposed work against baseline methods with regard to illumination consistency and details of the extracted texture

<table>
<thead>
<tr>
<th></th>
<th>Mirroring</th>
<th>Interpolation</th>
<th>Proposed work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumination consistency</td>
<td>−</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Texture details</td>
<td>+</td>
<td>−</td>
<td>+</td>
</tr>
</tbody>
</table>

neutral expressions are chosen in this experiment. In Fig. 5.1, the frontal view and a novel view of the 3D faces show very realistic texture. Even the originally occluded region is very naturally “hallucinated”. Especially, the skin texture of second image is of high resolution, which is transferred and blended from the visible area. No sign of transition between the authentic and the mirrored texture is visible. Furthermore, the shadow on the face contour is neatly neutralized. Similarly, the challenging light casted on the right face half in the third image is well taken care of, too. Contrarily, the mirroring approach is very sensible to the lighting difference between the two face halves. A hard boundary line is seen in most examples. For the interpolation-based method, only overly smoothed texture without any useful details is generated, which is only applicable to fill small blank areas [JHY05]. Last but not least, the dark or color stripes near the face contour are erroneously extracted from the background, which is inevitable for vertices of nearly perpendicular normal direction in the case of minimal landmark localization error. Our approach that first removes this region before generating virtual texture (see Fig. 4.2(b)) is proven to be quite effective.

Another qualitative evaluation on the YouTube Celebrities video dataset [KKPR08] is performed and the results are illustrated in Fig. 5.2. The dataset is composed of short interviews and TV shows of celebrities and contains low-resolution faces and typical video artifacts. Nevertheless, the well-known facial characteristics of these celebrities are still clearly recognizable in the person-specific 3D models. We summarize the advantages of the proposed approach over the baseline methods in Tab. 5.1.

6 Conclusions

The problem of single image real texture mapping for 3D face models is addressed in this paper. A general extraction workflow for visible texture is first introduced. As the main contribution of this work, a novel iterative blending scheme to draw the advantages of the conventional mirroring and interpolation methods for the occluded face area is proposed, which generates detailed and realistic facial texture
Figure 5.1: Experimental results of the proposed method in frontal (2nd row) and novel view (3rd row) compared to texture mirroring (4th row) and interpolation (5th row) on the LFPW dataset [BJKKT11].
with smooth illumination transition. The effectiveness of the framework is justified on the publicly available “in the wild” images and video data in challenging uncontrolled pose and illumination conditions. An extended algorithm to fuse images of multiple views with careful registration and constraints \cite{PHL98} can be done as future work.

**Bibliography**


3D Facial Texture Extraction Robust to Self-Occlusion


