Automated 3D Face Reconstruction from Multiple Images using Quality Measures

Marcel Piotraschke and Volker Blanz
Institute for Vision and Graphics, University of Siegen, Germany
piotraschke@nt.uni-siegen.de, blanz@informatik.uni-siegen.de

Abstract

Automated 3D reconstruction of faces from images is challenging if the image material is difficult in terms of pose, lighting, occlusions and facial expressions, and if the initial 2D feature positions are inaccurate or unreliable. We propose a method that reconstructs individual 3D shapes from multiple single images of one person, judges their quality and then combines the best of all results. This is done separately for different regions of the face. The core element of this algorithm and the focus of our paper is a quality measure that judges a reconstruction without information about the true shape. We evaluate different quality measures, develop a method for combining results, and present a complete processing pipeline for automated reconstruction.

1. Introduction

Algorithms that reconstruct 3D faces from images by fitting a deformable face model, such as a 3D Morphable Model (3DMM), rely on a relatively precise initial positioning of the face [7] or on a set of feature point coordinates [8]. For an automated procedure, it is straight-forward to combine these algorithms with automatic face and landmark detection, such as the algorithm by Zhu and Ramanan [33] or other feature detectors [11]. In practice, however, this combination has turned out to be more challenging than expected, posing a number of fundamental questions. The feature point detection is a non-trivial task, especially if the image material includes complex lighting, facial expressions, wrinkles, eye glasses or facial hair. Therefore, the features may be inaccurate, and some may even be outliers. Moreover, the optimal set of features for 3DMM fitting includes points that are not easy to detect, such as the facial silhouette and the ears. Those points are necessary for the 3DMM to converge to the correct pose angle, and this in turn affects the shape estimate.

Therefore, a simple combination of existing methods produces results that are substantially worse than those obtained with manually labeled features. Attempts to make 3DMM fitting more robust [10] are promising but still not sufficient. Instead, we argue that in many real-world applications more than one image of a person is available, so an automated algorithm can exploit redundant data from multiple images to gain robustness and reliability. Our algorithm outperforms existing methods of simultaneous 3D reconstruction from multiple images [7] significantly, which may be due to the fact that outliers in feature positions adversely affect the simultaneous least squares solution.

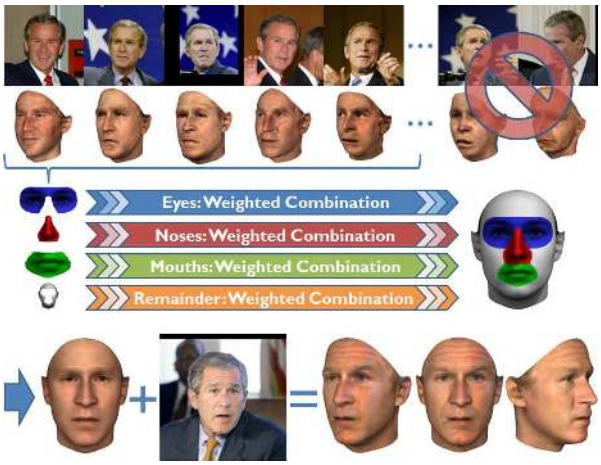


Figure 1: A segment-based, weighted linear combination is used to create the final head shape. The weight decreases with the rank. Implausible segments are discarded. Note that each facial segment is handled separately. Optionally the texture can be extracted from one of the input images.

In contrast, our algorithm calculates separate reconstructions from each input image, and then combines them to an optimal overall solution. We propose a method that selects the most plausible reconstructions, operates on different region of the face separately, and merges them into a single 3D face.

The key component of our algorithm is a new measure

for the visual quality of 3D reconstructions, based on surface normals. Automated assessment of visual quality in computer graphics and vision is a fundamental challenge. Simple image comparisons are insufficient because they are insensitive to small but important errors and artifacts. Euclidean distance in 3D overrates global shape deformations that would be irrelevant to human observers. Mahalanobis distance is also inconsistent with the quality ratings of humans. In an experimental comparison with quality ratings from human subjects, our new, normal based measure outperforms these existing criteria.

In summary, the contributions of this paper are:

- a general measure of the quality (naturalness) of a shape reconstruction,
- an algorithm for selecting and combining reconstructions of different facial regions (segments) from different input images into a single 3D face,
- an automated algorithm that produces 3D shape reconstructions from multiple images of a person, which goes beyond a simple combination of landmark detection and 3DMM fitting.

2. Related Work

Although several approaches have been published related to high quality 3D reconstructions of faces from 2D images, automated reconstruction still remains a challenging task, especially with facial expressions. It is often difficult to find images with a neutral expression, as most people tend to smile in portraits.

In the literature on face modeling several different approaches can be found. For high quality 3D reconstructions of faces which are used in computer games and movies, the state of the art techniques still require 3D scans of the person using laser scanners or multi-view camera setups [3, 9, 4, 5, 12, 21, 2]. Additionally, substantial post processing is required to combine the generated 3D data and to morph between different facial expressions and visemes to realistically animate the subjects face.

Approaches like the one presented in this paper try to obviate the need for special equipment. Instead, they make use of data that can be easily produced with standard equipment or that is already available, such as photo or video data. Multi-view geometry [26, 14, 28, 13] is a common procedure to reconstruct 3D shapes from several single images or video frames. Although these algorithms are quite flexible in usage for different scenarios varying from the reconstruction of buildings, smaller objects and even faces, they cannot sufficiently handle non-rigid transformations (facial expressions) within a series of input images.

Other recent publications have shown promising results by aligning a 3D face to single or multiple images as well as to videos frames. The approaches by Park et. al [22], Aldrian and Smith [1] and Dou et. al [24] reconstruct the 3D shape from a single image. Wang et al. [30] extract the silhouette from several input images to reconstruct the 3D shape, while Roth et. al [27] use an image collection for photometric stereo-based normal estimation which iteratively optimizes the surface reconstruction. By estimating the pose and computing the optical flow, a high detail refinement of the 3D shape is performed, resulting in a 3D to 2D correspondence [20, 18, 15, 29]. Suwajanakorn et. al [29] even captured fine details like wrinkles and in [19] Kemelmacher-Shlizerman and Seitz showed that also 'faces in the wild' can be handled properly. But these approaches lack an additional 3D to 3D correspondence. In this paper we address 3D to 2D as well as 3D to 3D correspondence.

To reconstruct a 3D shape of a face from a 2D image, Blanz and Vetter [7] introduced the 3DMM. With the Basel Face Model [23], a 3DMM has been made available to the public and Zhu et. al [34] presented a discriminative 3DMM based on local features that provides accurate reconstructions. A common and significant drawback of the 3DMM is its lack of robustness in the case of 'faces in the wild', especially if the facial landmarks are not perfectly detected. Although Breuer et al. [11] propose to use a Support Vector Machine for automatic 3D face reconstruction and in [10] an idea is presented to correct misplaced landmarks to some extent, both implementations were not robust enough to handle difficult scenarios caused by facial expressions or complex lighting conditions. With the approach in this paper, we aim to overcome the previous drawbacks of the 3DMM.

Additionally there are approaches which are not aiming at the reconstruction of faces directly, but provide a strong foundation for further processing by detecting faces, estimating poses, localizing feature points or aligning face geometries [33, 32, 25, 31, 17].

3. 3D Morphable Model

The 3D Morphable Model [7] is a vector space of 3D shapes and textures, $\mathbf{S}_i = (X_1, Y_1, Z_1, X_2, \ldots, Z_n)^T$ and $\mathbf{T}_i = (r_1, g_1, b_1, r_2, \ldots, b_n)^T$, with X, Y, Z coordinates and r, g, b colors of $n = 113\,753$ vertices. In our experiments, the 3DMM is constructed from 3D scans of 200 individuals and from 35 additional scans that show facial expressions of a single individual [6]. On the individual shapes, the expressions and the textures, a PCA defines eigenvectors $\mathbf{s}_i$, $\mathbf{u}_i$ and $\mathbf{t}_i$, respectively, and average shapes and textures $\overline{\mathbf{s}}$ and $\overline{\mathbf{t}}$. In this basis, new faces can be approximated by linear combinations

$$\mathbf{S} = \overline{\mathbf{s}} + \sum_{i=1}^{m} \alpha_i \mathbf{s}_i + \sum_{i=1}^{p} \gamma_i \mathbf{u}_i \qquad \mathbf{T} = \overline{\mathbf{t}} + \sum_{i=1}^{m} \beta_i \mathbf{t}_i. \quad (1)$$

We use $m = 100$ eigenvectors for individual variations and $p = 4$ for the most important degrees of freedom of facial

expressions, with a focus on mouth movements.

Please note that a high percentage of images, for example those in the database 'faces in the wild', involve non-neutral facial expressions, so our approach of combining multiple images only makes sense with this additional degree of freedom. We use separate PCAs and basis vectors for shape and expression in order to be able to give the 3D faces neutral expressions ($\gamma_i = 0$) after fitting.

3D shape reconstruction by fitting the model to an image is essentially a minimization of the image distance

$$d_{image} = \sum_{u,v} \|I_{input}(u,v) - I_{model}(u,v)\|^2 \qquad (2)$$

in all 3 color channels, with respect to the linear coefficients $\alpha_i, \gamma_i, \beta_i$ and some imaging parameters ρ_i that control pose, lighting and other parameters (for details see [7]).

Overfitting is avoided by a regularization term that is the Mahalanobis distance from the starting conditions,

$$d_{maha} = \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_i \frac{\gamma_i^2}{\sigma_{S,i}^2} + \sum_i \frac{\beta_i^2}{\sigma_{T,i}^2} + \sum_i \frac{(\rho_i - \overline{\rho}_i)^2}{\sigma_{R,i}^2}, \qquad (3)$$

where $\overline{\rho}_i$ denotes the starting values of the rendering parameters, and σ are the standard deviation from PCA.

A stochastic newton optimization algorithm minimizes the weighted sum of d_{image}, d_{maha} and an additional term

$$d_{features} = \sum_j \| \begin{pmatrix} x_j \\ y_j \end{pmatrix} - \begin{pmatrix} P_x(X_{k_j}, Y_{k_j}, Z_{k_j}) \\ P_y(X_{k_j}, Y_{k_j}, Z_{k_j}) \end{pmatrix} \|^2 \qquad (4)$$

which is the sum of squared distances between 2D feature positions x_j, y_j and the projected positions of the corresponding vertex k_j, with a perspective projection P [8]. $d_{features}$ is only for initialization, with a weight that decreases as the fitting proceeds.

Unlike earlier work on 3DMM fitting [8], we use a feature detection algorithm by Zhu and Ramanan [33] for an automated process. The reduced precision of these features and the suboptimal choice of features (silhouettes, ears) affect the quality of the output significantly. In the remainder of this paper we describe how to select the most successful reconstructions based on a given set of images of a person, and how to combine these to a 3D face.

4. Quality Measures

For a meaningful quality measure, it is important to be independent of facial expression. Therefore, we use "neutralized" facial expressions (with $\gamma_i = 0$) in this section except for image distance. Because the image distance compares the rendered reconstruction with the original input image, it needs to be as close as possible to the original face.

Figure 2: The image distance is computed by subtracting the input image with a modified version where the reconstructed face is rendered on top of the original face.

4.1. Image Distance

In contrast to all others distance functions that are discussed in this paper, the image distance d_{image} Eq. (2) is the only one that penalizes differences between the original face and the reconstruction. The other distance measures will only estimate the plausibility of naturalness of reconstructed faces.

Fig. 2 illustrates one major drawback of this error function: it is not possible to penalize the fact that the projected face does not occlude the complete face in the input image. This is the case for Obama's right ear. In $I_{input} - I_{model}$, the image distance for most pixels of the right ear is zero and therefore the error is quite small. The reconstructed ear is rendered on the cheek, but due to the similar color, this has also little effect on d_{image}. In general, d_{image} fails to capture small but relevant errors and artifacts in the reconstruction.

On the other hand, d_{image} can also be high even though the faces look similar, for example when the overall color tone is wrong or the face is slightly shifted. All of these problems are caused by the fact that d_{image} is a sum of all pixels and that many small errors count more than a few large errors.

Even though d_{image} turns out to be suboptimal for rating the quality or plausibility of the 3D reconstruction, as we will demonstrate in Section 5, it makes sense to use d_{image} in the fitting procedure because, unlike the following criteria, it measures the distance from the input face, and it is easy to compute.

4.2. Mahalanobis Distance

Mahalanobis distance measures the distance of the current solution from the average face using PCA, taking into account the standard deviations observed in the training data. It is directly related to the multivariate Gaussian probability density function which is estimated by PCA. Just as the image distance, the Mahalanobis distance is already integrated in the 3DMM fitting procedure. For the experiments in Section 5, where we only want to rate the quality of the reconstructed shape, we simplified Eq. (3) to

$$d_{maha} = \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2}, \qquad (5)$$

so we measure only the distance of the neutral face shape from the average face, while expressions, texture and rendering parameters are omitted. The motivation is that, unlike neutral shape, the texture and expression of a successful reconstruction may be far from the average if the input image is unusual (hair, facial hair, eye glasses, smile).

4.3. Euclidean Distance

A more direct measure for the distance of a 3DMM shape from the average face is the Euclidean distance between the reconstructed shape vector (with neutralized expression) $\mathbf{S}$, and the average vector $\bar{\mathbf{s}}$:

$$d_{eucl}(\mathbf{S}, \bar{\mathbf{s}}) = \sqrt{\sum_{i=1}^{3n} (\mathbf{S}_i - \bar{\mathbf{s}}_i)^2} = ||\mathbf{S} - \bar{\mathbf{s}}||_2. \qquad (6)$$

Please note that d_{eucl} is sensitive to rigid transformations of the faces. The 3DMM shape vectors are, by construction, aligned in a least-squares sense. In 3DMM fitting, rigid transformations are applied to these externally, and captured by rendering parameters ρ_i (Section 3). Still, a general drawback of d_{eucl} remains with respect to simple, global transformations, e.g. anisotropic scaling, which does not affect naturalness or shape similarity, but has significant effect on d_{eucl}.

Equation (6) tends to overrate outlier vertices in the sum of squared distances. For the evaluation (Section 5), we also considered a modified distance which is the sum of 3D vertex distances (square root on a per-vertex level). But we found no improvement, so Section 5 will refer to Equation (6) only.

4.4. Normal Distance

We have observed that local or even global distortions of the surface are a common feature of failed 3D face reconstructions. This is true for most or perhaps all 3DMM algorithms (see Section 2) and – in a different context – even for 3D shape capture setups such as scanners or stereo and multiview techniques. For shape fitting algorithms, it is unlikely that a failed reconstruction is misaligned and still close to the average, because misalignments tend to have undesired effects on the cost functions of the fitting algorithm and therefore lead away from the set of plausible faces. In our context, misalignments may be caused by inaccurate initial feature positions. Also, other potential reasons for failed reconstructions, such as lighting effects, occlusions or extreme facial expressions, tend to lead the algorithm far away from the average, and a very sensitive measure for this is the deviation of surface normals from the average.

We would like to point out that regularization mechanisms, such as Equation (3) reduce this effect and keep the solution close to the average. Still, for practical purposes, we have observed that (1) if the weight of the regularization is too large, it implies suboptimal results on images that would otherwise be reconstructed successfully, so there is a fundamental tradeoff between quality and robustness, and (2) the regularizer Eq. (3) is not a reliable measure of plausibility of faces, as we will see in Section 5.

(a) (b) (c)

Figure 3: The Normal distance is determined by computing the angle between the normal of the average (Fig. 3a) and the reconstructed face (Fig. 3b) per corresponding vertex pair (see Fig. 3c). These values are averaged per segment (see Fig. 4a) or face to obtain a global distance value.

Based on the dense point-to-point correspondence between vertices i of the 3DMM, the new distance measure d_{normal} analyzes the difference between the surface normals $\mathbf{n}_i$ of the reconstructed face, and the normals $\mathbf{n'}_i$ of the average face:

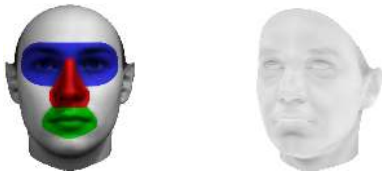
$$d_{normal} = \frac{1}{n} \sum_{i=1}^{n} \arccos \frac{\mathbf{n}_i \cdot \mathbf{n'}_i}{\|\mathbf{n}_i\| \, \|\mathbf{n'}_i\|}. \qquad (7)$$

The idea of this Normal distance is illustrated in Fig. 3. Note that, unlike d_{eucl}, d_{normal} is insensitive to scaling and shifting. By segmenting the full face into distinct facial regions (eyes, mouth, nose and surrounding region, see Fig. 4a), separate distances d_{normal} can be defined that reflect the plausibilities of regions separately. We will use this idea in Section 6.

In human faces, the normals in some vertices on the nose, the eyes or the lips vary more than others. We have analyzed the original 200 3D scans of the 3DMM and created different weight maps ω (see Fig. 4b) which account for these local differences by scaling regions with high normal variation either up (considering them most diagnostic) or down (normalization). In a first step, we computed the average deviation angle $\bar{\phi}_i$ of the normal $\mathbf{n}_i$ from the average normal $\mathbf{n'}_i$ in each vertex i across all 200 faces. Then, we found the best weight map to be defined by $\hat{\omega}_i = 1 - \frac{\bar{\phi}_i - \bar{\phi}_{min}}{\bar{\phi}_{max} - \bar{\phi}_{min}}$, and the weighted Normal distance is

$$d_{normalW} = \frac{1}{n} \sum_{i=1}^{n} \hat{\omega}_i \, \arccos \frac{\mathbf{n}_i \cdot \mathbf{n'}_i}{\|\mathbf{n}_i\| \, \|\mathbf{n'}_i\|}, \qquad (8)$$

for which we obtained experimental results that are summarized in the next section.



(a) (b)

Figure 4: Fig. 4a shows the different face segments. The 3DMM based weight map is shown in Fig. 4b.

5. Evaluating the Distance Measures

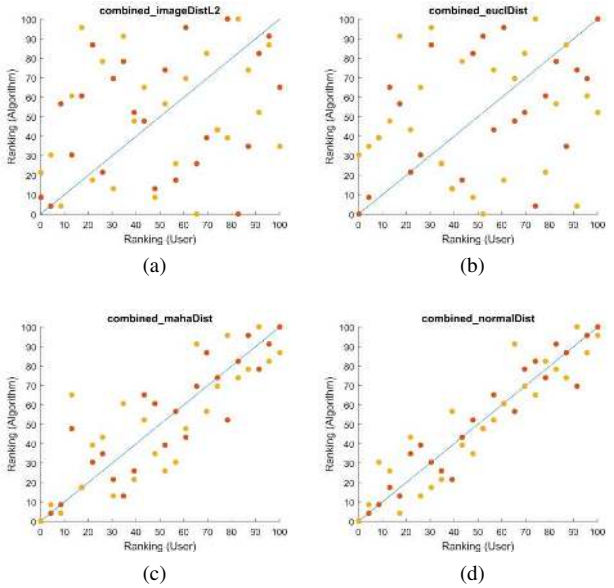
The goal of this evaluation is to find out which quality measure is closest to the ratings that human observers would assign to different reconstructions. For humans, quality may mean how natural and plausible the 3D face looks, but also how similar it is to the person in the image. For failed reconstructions, both criteria are usually violated at the same time, so the distance measures from Section 4 are good candidates even though most do not measure similarity to the input face.

5.1. Evaluation 1

The first ranking was performed on 24 3D reconstructions from pictures of Barack Obama based on automatically detected landmarks. The automatic detection of landmarks is based on the approach of Zhu and Ramanan [33]. An additional set of 24 reconstructions was created by using manually selected landmarks on the same input images. Again the algorithmic distance measures introduced in Section 4 were used to perform a ranking. All reconstructions were created from a single image as described in Section 3.

We asked four naive participants to create a ranking in each of the two sets of 24 reconstructions, based on the perceived quality of the reconstruction. The individual user rankings were combined to define an overall ranking list, which was compared to the ranking of each distance measure. As can be seen in Table 1, the mean and max errors (difference of ranks assigned to each reconstruction) of Mahalanobis and Normal distance are much less than the ones based on Euclidean and image distance. Furthermore, based on the numbers for $d_{normalW}$ (see Eq. 8), it can be noted that the influence of the weight map is not very strong compared to the ranking based on d_{normal} (see Eq. 7).

In Fig. 5 the correlation of each distance measure is visualized: The horizontal axis describes the average user ranking, while each distance measure is mapped to the vertical axis. If a distance measure correlates perfectly with the user ranking, the dots of the scatter diagram are aligned along the diagonal. As can be seen in Fig. 5a, for the image distance the dots are widely scattered. The same can be observed for the Euclidean distance in Fig. 5b. Consequently,



(a) (b) (c) (d)

Figure 5: Visualization of the correlation between the average user ranking and each distant measure (100 $\widehat{=}$ very good, 0 $\widehat{=}$ very bad) for reconstructions based on automatic *(red)* and manual *(yellow)* landmark selection.

both measures are not useful to distinguish plausible from implausible reconstructions in a way that correlates to the opinion of users. For the Mahalanobis (see Fig. 5c) and the Normal distance (see Fig. 5d), the correlation between the user rating and the rating based on the algorithmic distance measures is clearly visible. Especially the Normal distance predicts much of the quality judgments of our participants. The correlations for all $2 \cdot 24$ reconstructions (automatic and manual) are for d_{image}: 0.27, for $d_{euclidean}$: 0.27, for d_{maha}: 0.85 and for d_{normal}: 0.94.

5.2. Evaluation 2

In a second evaluation, 3D reconstructions based on images of Obama (24 images), Lawrence (32 images), Annan

	Obama dataset			
	automatic landmarks		manual landmarks	
	mean error	max error	mean error	max error
d_{image}	6.92	19	6.83	18
d_{eucl}	5.67	16	8.08	20
d_{maha}	2.25	8	3.42	12
d_{normal}	1.33	5	2.25	6
$d_{normalW}$	1.33	5	2.17	6

Table 1: Mean and max difference of ranks for 24 reconstructions with automatically and 24 with manually selected landmarks based on the perceived quality of four naive participants (see Section 5.1).

(32 images), Watson (46 images) and Carell (28 images) were rated. Again two distinct sets were created, but this time only automatically selected landmarks were utilized. The first set was created by fitting to a single image, while for the second set a simultaneous fit to two images was performed by applying the multifit approach of Blanz and Vetter [7]. A fixed reference image was selected and was then combined with each other image of the collection for the person. Please note that the facial landmarks differ from to the ones in Section 5.1. Thus, although the same input images are used for the Obama dataset, the reconstructions are different.

For each dataset, the distance measures were used to create a ranking list. Then we asked seven naive participants to rate each 3D reconstruction. Possible ratings were 'very good', 'good', 'acceptable' or 'failed'. The individual ratings were averaged and then used to create a ranking. Many reconstructions obtained the same average ratings and therefore many positions in the ranking are shared. This implies higher discrepancies between the rank list derived from humans, and the rank list from distance measures than in Evaluation 1, where we asked participants to create a unique ranking directly. Still, the Normal distance matches the user rating best, as can be seen in Table 2 and 3.

	Obama	Lawrence	Annan	Watson	Carell
d_{image}	5.29 (13)	7.56 (20)	9.59 (30)	13.15 (29)	8.04 (18)
d_{eucl}	8.38 (20)	8.38 (22)	10.84 (23)	13.54 (35)	8.82 (20)
d_{maha}	5.79 (13)	7.00 (18)	5.22 (16)	11.94 (27)	3.46 (9)
d_{normal}	5.21 (13)	5.44 (14)	4.97 (12)	11.50 (27)	2.61 (8)
$d_{normalW}$	5.21 (13)	5.31 (14)	4.97 (12)	11.41 (27)	2.46 (8)

Table 2: Mean and max *(in brackets)* difference of ranks for reconstructions from a single image based on the perceived quality of seven naive participants (see Section 5.2).

	Obama	Lawrence	Annan	Watson	Carell
d_{image}	6.75 (14)	8.34 (23)	8.75 (24)	10.80 (33)	6.32 (19)
d_{eucl}	5.58 (16)	8.47 (24)	6.56 (19)	10.02 (34)	9.32 (24)
d_{maha}	4.33 (11)	5.47 (14)	5.25 (19)	10.07 (28)	6.82 (19)
d_{normal}	4.08 (10)	4.41 (14)	4.31 (15)	7.85 (25)	4.96 (17)
$d_{normalW}$	4.08 (10)	4.41 (14)	4.31 (15)	7.80 (25)	4.96 (17)

Table 3: Mean and max *(in brackets)* difference of ranks for reconstructions from multiple images based on the perceived quality of seven naive participants (see Section 5.2).

6. Weighted Linear Combination per Segment

The automated 3D reconstruction that we propose in this paper compensates the reduced precision and reliability of automatically detected feature positions by using more than a single image of the face. Note that, unlike stereo and multiview algorithms, we allow for nonrigid deformations due to facial expressions, and large differences in the (unknown) imaging conditions.

Figure 6: Plausibility rating of the single image based reconstructions using Normal distance with subsequent ordering.

Our strategy is to apply single image 3DMM fitting (Section 3) on each of the input images of the person separately, based on landmarks detected by the algorithm by Zhu and Ramanan [33], select the m best results (Fig. 1 and 6) on each segment (Fig. 4a) using d_{normal}, compute weighted linear combinations of these and merge them into a single 3D face.

The shape for each segment is determined by a weighted linear combination of corresponding segments based on the ranking list order. The weight decreases with the the rank. Thus the combined shape for each individual segment

$$\mathbf{S}_{seg} = \sum_{i=0}^{m-1} \alpha_i \, \mathbf{S}_{seg,i} \tag{9}$$

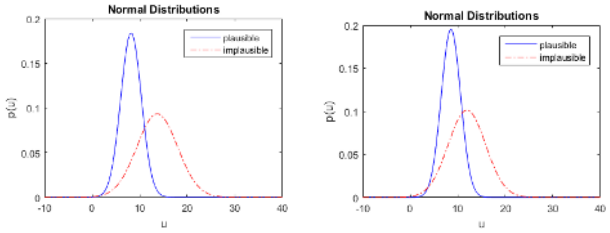
is determined by m individual reconstructions of corresponding segments $\mathbf{S}_{seg,i}$ weighted by

$$\alpha_i = \frac{1 - (i \cdot \frac{1}{m})}{\sum_{c=0}^{m-1} 1 - (c \cdot \frac{1}{m})}. \tag{10}$$

The algorithm is summarized in Fig. 1. Note that for illustration, Fig. 6 and 1 refer to the shape of the entire face, and not for separate segments as in our algorithm.

An important element of our algorithm is to define a threshold quality value that determines which reconstructions are considered in the weighted sum. Based on the data from Section 5, we estimated a threshold that separates plausible from implausible reconstructions. In Evaluation 1, participants were also asked which faces are still plausible and which are not. In Evaluation 2, the threshold is supposed to be between ratings "acceptable" and "failed". For both data sets, we estimated Gaussian Distributions $p_0(u)$ and $p_1(u)$ for plausible and implausible reconstructions using the arithmetic mean and the estimated standard deviations of d_{normal} in either set.

In a maximum likelihood approach, the threshold equals the intersection point of the Gaussian distributions $p_0(u)$ and $p_1(u)$. Based on our data, this threshold equals $u = 11$ as is shown in Fig. 7 and can be computed by solving $p_0(u) = p_1(u)$.



(a) (b)

Figure 7: Gaussian distribution of d_{normal} for plausible *(blue)* and implausible *(red)* 3D reconstructions. For Evaluation 1 (Fig. 7a) the intersection is in $u = 11.08$ and for Evaluation 2 (Fig. 7b) it is in $u = 10.95$.

Now, all segments with Normal distances larger than this threshold are discarded, as illustrated in Fig. 1. After the shape for each segment has been reconstructed using a weighted linear combination based on the ranking order for the remaining segments, all independent segments are combined to build the shape of the complete face using the method described in [7]. One of the input images, for example the one with minimum d_{normal}, can be used for texture transfer as in [7], so the texture is not just a linear combination of all input images, but captured from a single input image with inverse projection and lighting.

7. Results

We compared our approach with an existing method of simultaneous 3D reconstruction from multiple images [7]. To that end we used sets of 8 to 15 images showing the same face from different angles. A subset of these images is shown in the second column of Fig. 8 and 9. The first column in each of these figures show the results of the existing approach, whereas the results of our approach are presented in Col. 3 with a uniform color and in Col. 4 with the combined texture colors. Therefore the textures of each individual segment have been linearly combined in exactly the same way as has been described for the shape in Section 6. In respect of shape estimation it outperforms the existing method if a fully automated approach is demanded and if the landmark locations may not fit the input image perfectly.

While the input images in Fig. 8 and 9 are taken under controlled lighting conditions and lack facial expressions, we also tested our approach with images from the Labeled Faces in the Wild [16] database. Here pose, expression and lighting differ in each image and the resolution is only 250*250px. The results are shown in Fig 10. From left to right the columns contain one image per dataset and person and two views of the reconstructed shape. At first with an uniform coloring and then with the extracted texture from

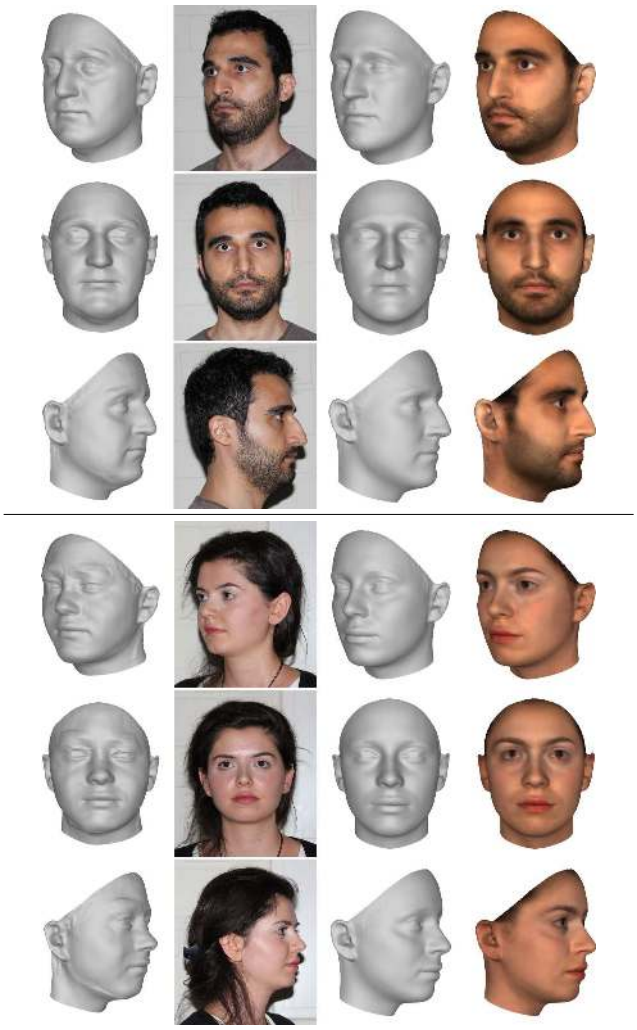


Figure 8: Two face reconstructions from multiple images using the existing 3DMM approach (Col. 1), the proposed method with Normal distance (Col. 3) plus combined color (Col. 4). A subset of the input images is shown in Col. 2.

the input image on the left. To retain a fully automated process, the input image belongs to the most plausible single image reconstruction (Fig. 6). As is shown in Fig. 1, any other input image can also be used for texture transfer, because the 3DMM enables a 2D to 3D correspondence between the image and each reconstruction as well as a 3D to 3D correspondence between all reconstructions.

8. Conclusions

We have proposed an algorithm that reconstructs a 3D face from a set of arbitrary images of a person. The core idea is to perform separate reconstructions on each image and combine the best of all reconstructions into the final shape. An important element of our work is to evaluate dif-

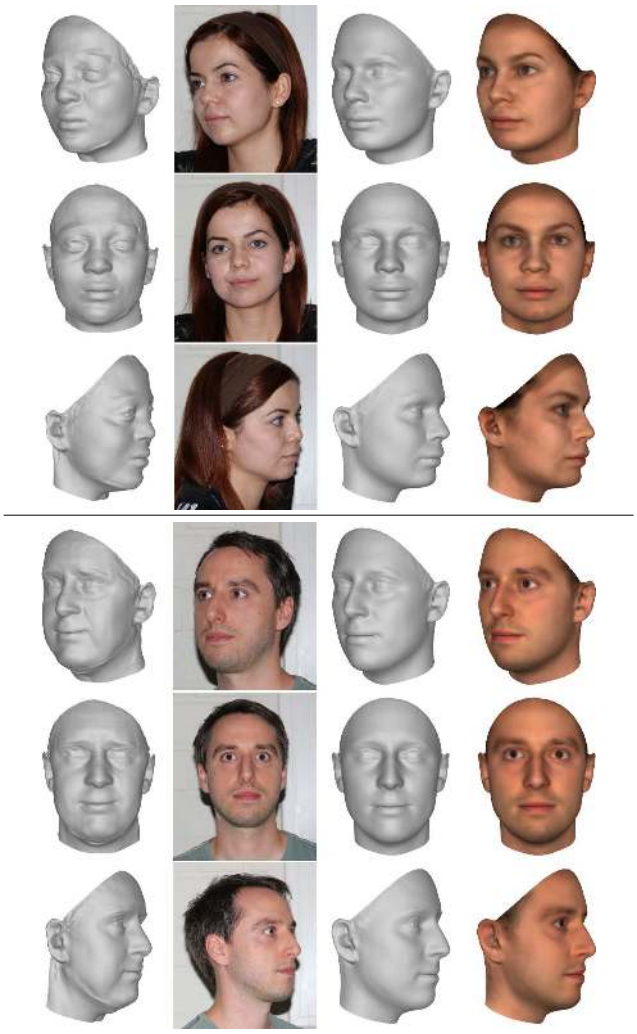


Figure 9: Two face reconstructions from multiple images using the existing 3DMM approach (Col. 1), the proposed method with Normal distance (Col. 3) plus combined color (Col. 4). A subset of the input images is shown in Col. 2.

ferent quality measures of 3D reconstructions. Combined with a feature point detector, we obtain an automated algorithm for 3D reconstruction that accounts for errors in the feature coordinates. Our method is modular, scalable and flexible, and it overcomes some of the problems that have restricted 3DMMs so far.

On a more fundamental level, it is the combination of results (multiple images, multiple segments) which makes our algorithm robust, and this is an alternative strategy to combining all input data into a single optimization problem.

It is a non-trivial result that multiple suboptimal 3D faces can be combined into a single, much more appealing one, and that this result is not just the average face. Another non-trivial result is that the reconstruction quality can be assessed without knowing the ground truth shape.

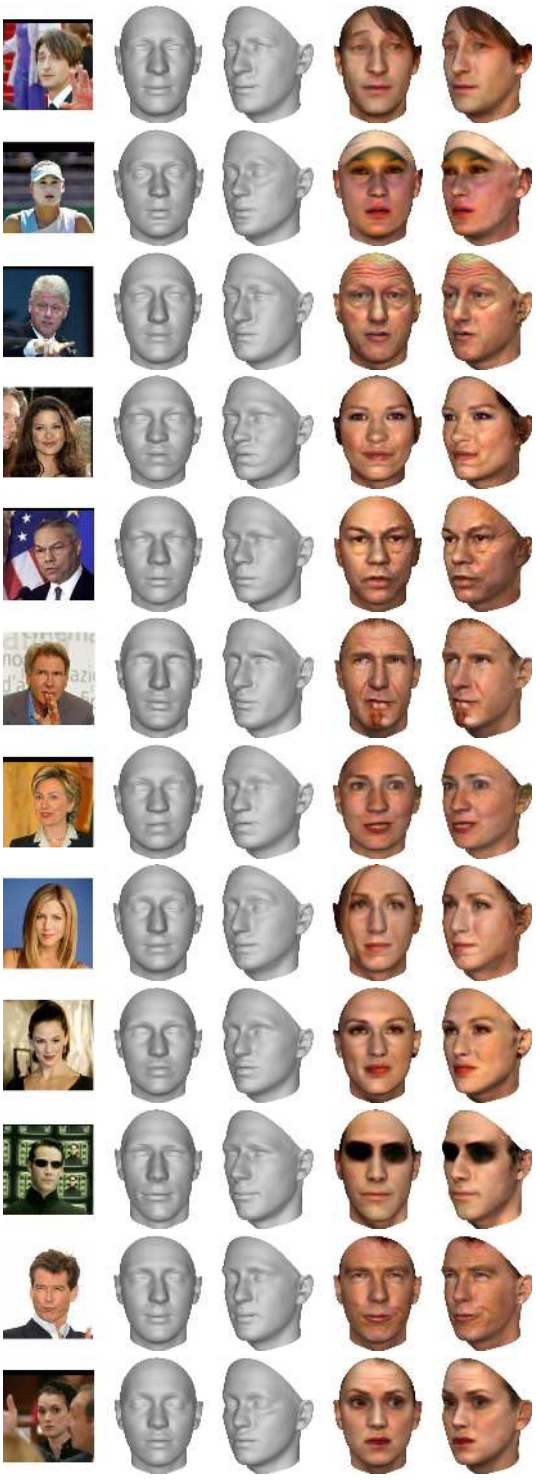


Figure 10: 3D face reconstructions for image sets from the LFW [16] database. From left to right the columns contain an example image from the image set, two views of the reconstructed shape and two views with extracted textures.

References

[1] O. Aldrian and W. Smith. A Linear Approach to Face Shape and Texture Recovery using a 3D Morphable Model. In *British Machine Vision Conference*, pages 75.1–75.10. BMVA Press, 2010.

[2] O. Alexander, G. Fyffe, J. Busch, X. Yu, R. Ichikari, A. Jones, P. Debevec, J. Jimenez, E. Danvoye, B. Antionazzi, M. Eheler, Z. Kysela, and von der Pahlen, Javier. Digital Ira: Creating a Real-Time Photoreal Digital Actor. *ACM SIGGRAPH Posters*, pages 1:1–1:1, 2013.

[3] O. Alexander, M. Rogers, W. Lambeth, M. Chiang, and P. Debevec. The Digital Emily Project: Photoreal Facial Modeling and Animation. *ACM SIGGRAPH Courses*, pages 12:1–12:15, 2009.

[4] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-Quality Single-Shot Capture of Facial Geometry. *ACM Transactions on Graphics*, 29(3):40:1–40:9, 2010.

[5] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. W. Sumner, and M. Gross. High-quality Passive Facial Performance Capture using Anchor Frames. *ACM Transactions on Graphics*, 30(4):75:1–75:10, 2011.

[6] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating Faces in Images and Video. *Computer Graphics Forum (EUROGRAPHICS)*, 22(3):641–650, 2003.

[7] V. Blanz and T. Vetter. A Morphable Model for the Synthesis of 3D Faces. *Proceedings of SIGGRAPH*, pages 187–194, 1999.

[8] V. Blanz and T. Vetter. Face Recognition Based on Fitting a 3D Morphable Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(9):1063–1074, 2003.

[9] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer. High Resolution Passive Facial Performance Capture. *ACM Transactions on Graphics*, 29(4):41:1–41:10, 2010.

[10] P. Breuer and V. Blanz. Self-Adapting Feature Layers. *European Conference on Computer Vision (ECCV)*, pages 299–312, 2010.

[11] P. Breuer, K.-I. Kim, W. Kienzle, B. Schölkopf, and V. Blanz. Automatic 3D Face Reconstruction from Single Images or Video. *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2008.

[12] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec. Multiview Face Capture using Polarized Spherical Gradient Illumination. *ACM Transactions on Graphics*, 30(6):129:1–129:10, 2011.

[13] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-View Stereo for Community Photo Collections. *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

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

[15] T. Hassner. Viewing Real-World Faces in 3D. *IEEE International Conference on Computer Vision (ICCV)*, pages 3607–3614, 2013.

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.

[17] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3D Face Alignment from 2D Videos in Real-Time. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2015.

[18] I. Kemelmacher-Shlizerman. Internet-based Morphable Model. *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[19] I. Kemelmacher-Shlizerman and S. M. Seitz. Face Reconstruction in the Wild. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 1746–1753, Washington, DC, USA, 2011. IEEE Computer Society.

[20] I. Kemelmacher-Shlizerman and S. M. Seitz. Collection Flow. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1792–1799, 2012.

[21] M. Klaudiny and A. Hilton. High-Detail 3D Capture and Non-sequential Alignment of Facial Performance. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 17–24, 2012.

[22] S. W. Park, Jingu Heo, and M. Savvides. 3D Face Reconstruction from a Single 2D Face Image. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 1–8, 2008.

[23] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D Face Model for Pose and Illumination Invariant Face Recognition. *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009.

[24] Pengfei Dou, Yuhang Wu, Shishir K. Shah, and Ioannis A. Kakadiaris. Robust 3D Face Shape Reconstruction from Single Images via Two-Fold Coupled Structure Learning and Off-the-Shelf Landmark Detectors. In *British Machine Vision Conference*, pages 1–5. BMVA Press, 2014.

[25] P. Perakis, G. Passalis, T. Theoharis, and I. A. Kakadiaris. 3D Facial Landmark Detection under Large Yaw and Expression Variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1552–1564, 2013.

[26] M. Pollefeys, R. Koch, and L. van Gool. Self-Calibration and Metric Reconstruction Inspite of Varying and Unknown Intrinsic Camera Parameters. *International Journal of Computer Vision*, 32(1):7–25, 1999.

[27] J. Roth, Y. Tong, and X. Liu. Unconstrained 3D Face Reconstruction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[28] S. N. Sinha, P. Mordohai, and M. Pollefeys. Multi-View Stereo via Graph Cuts on the Dual of an Adaptive Tetrahedral Mesh. *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[29] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total Moving Face Reconstruction. *European Conference on Computer Vision (ECCV)*, pages 796–812, 2014.

[30] S. Wang, L. Zhang, and D. Samaras. Face Reconstruction Across Different Poses and Arbitrary Illumination Conditions. *Lecture Notes in Computer Science*, pages 91–101, 2005.

[31] X. Xiong and De la Torre, Fernando. Supervised Descent Method and Its Applications to Face Alignment. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013.

[32] F. Zhou, J. Brandt, and Z. Lin. Exemplar-based Graph Matching for Robust Facial Landmark Localization. *IEEE International Conference on Computer Vision (ICCV)*, pages 1025–1032, 2013.

[33] X. Zhu and D. Ramanan. Face Detection, Pose Estimation, and Landmark Localization in the Wild. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886, 2012.

[34] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li. Discriminative 3D Morphable Model Fitting. *Automatic Face and Gesture Recognition (FG)*, 2015.