Model-based Stereo with Occlusions

Fabiano Romeiro and Todd Zickler

School of Engineering and Applied Sciences
Harvard University
romeiro@fas.harvard.edu zickler@eecs.harvard.edu

Abstract. This paper addresses the recovery of face models from stereo pairs of images in the presence of foreign-body occlusions. In the proposed approach, a 3D morphable model (3DMM) for faces is augmented by an occlusion map defined on the model shape, and occlusion is detected with minimal computational overhead by incorporating robust estimators in the fitting process. Additionally, the method uses an explicit model for texture (or reflectance) in addition to shape, which is in contrast to most existing multi-view methods that use a shape model alone. We argue that both model components are required to handle certain classes of occluders, and we present empirical results to support this claim. In fact, the empirical results in this paper suggest that even in the absence of occlusions, stereo reconstruction using existing shape-only face models can perform poorly by some measures, and that the inclusion of an explicit texture model may be worth its computational expense.

1 Introduction

Being able to automatically recognize faces, track them, and estimate their expression and pose are important for many applications. Performing these tasks reliably requires the ability to represent the appearance of faces over large variations in illumination and viewpoint. It also requires the ability to model the effects of occlusions—both self-occlusions caused by the face itself and occlusions caused by "foreign bodies" (eye glasses, long facial hair, clothing, hands and limbs, etc.) in the environment.

Illumination effects can often be well-represented using purely image-based methods (e.g. [1–4]), but to effectively handle extreme changes in 3D pose, one typically requires a mechanism for "warping" 2D images. 3D morphable models (3DMMs), which are parametric models of shape and reflectance, are useful for this purpose because they explicitly represent 3D shape and therefore handle self-occlusions in a natural way.

In a 3D model-based approach, one is faced with the problem of finding the parameters of the model that best explain the input data. The estimated model parameters can then be used to perform recognition, track the face, detect expressions, synthesize new images, etc. The fitting problem is complicated in the presence of foreign-body occluders, because unlike self-occlusions, the image effects induced by foreign bodies cannot be explained by the face model.

In this paper we present a 3D model-based method for face reconstruction and recognition that exploits stereo imaging to handle foreign body occlusions. In the proposed approach, occlusion is represented using a single occlusion map defined on the 3D shape model, and this occlusion map is recovered efficiently by incorporating robust estimators in the fitting process.

In addition to including an occlusion map, we differentiate between two types of constraints for fitting a model to multiple views. According to the first constraint, each image should agree with a given model's shape and reflectance; and according to the second, the images should agree with each other given the model's shape. We find that the importance of these two constraints (roughly speaking, the "texture match" and the "stereo match") varies depending on the type of foreign body occluders that are present. We also find that even in the absence of occluders, explicitly enforcing the texture match constraint significantly improves fitting accuracy in comparison to an approach that uses the stereo match constraint alone (suggested in [5]).

1.1 Related Work

3D Morphable Models (3DMMs) [6] use high resolution linear 3D shape and texture models to represent faces. Typically, this model is fit to an input image by minimizing an energy function that measures the difference between intensities in the observed image and those predicted by the model. Recognition can be performed based on the model parameters [7] or by using the model to synthesize new views of the face in a canonical pose and lighting configuration [8].

Using a stereo pair for the fitting of a 3DMM imposes additional geometric constraints on the face shape, which can improve the quality of results. Also, by imposing a stereo matching constraint, the fitting of the shape and texture parameters can be decoupled [5]. According to this approach, the shape parameters are recovered by minimizing the per vertex intensity differences between two calibrated views, and the texture is estimated separately using this shape. While the decoupling of shape and texture is appealing from an efficiency standpoint, the results we show here suggest that there are significant benefits to estimating both components jointly.

Explicit handling of foreign-body occlusions has been addressed for the case of monocular fitting of 3DMMs in [9], where a generalized EM algorithm is used to alternate between the estimation of a visibility map given the model and the model parameters given the visibility map. To account for spatial coherence of occluders the visibility map is modeled by a Markov random field (MRF) on the image plane. In contrast, we model occlusions using a visibility map on *the surface*, and approximate the occlusion process using a robust estimator. While it gives up the preference for spatial coherence, the proposed approach can be implemented with little computational overhead. In addition, it can be easily extended to more views, since the occlusion map is on the surface.

Also related to this work are 2D active appearance models (AAMs), which trade precision for speed and are often used for tracking. 2D AAMs [10] typically use low-resolution 2D deformable shapes along with linear texture mod-

els. The fitting is done by matching a warped face image (with the warping being given by the linear shape model) against the linear texture model, and solving for the shape and texture parameters that give the best fit. Performance can be improved using an extension to the inverse compositional image alignment algorithm [11], by including 3D constraints [12], or by using multiple views [13, 14]. Fitting AAMs in the presence of occlusions can also be approached using robust estimators [15]. The main advantages of the 3D approach over 2D AAMs are the ability to directly model lighting effects because it has access to surface normals and to more easily handle self-occlusions.

2 Background

2.1 3D Morphable Models for Faces

As a 3D morphable model for faces, we use the shape and texture bases (3DFS-100) made available by the University of Freiburg [6]. These bases were obtained by first concatenating the N vertices (or RGB color values in the case of texture) of each scan i of a large set of high resolution 3D face scans into vectors (FS_i for shape, and FT_i for texture), and putting them into correspondence. That is, the vectors are made such that the same entry in each vector corresponds to the same facial feature [16–18]. These vectors are denoted:

$$FS_i = [[X_1^i Y_1^i Z_1^i]...[X_N^i Y_N^i Z_N^i]], \quad FT_i = [[R_1^i G_1^i B_1^i]...[R_N^i G_N^i B_N^i]]$$

Principal component analysis (PCA) is performed on this set of vectors, and the most significant eigenvectors are used as bases for shape and texture. Shape and texture are then expressed as linear combinations of these basis elements:

$$S = S_0 + \sum_{i=1}^{m} \alpha_i S_i, \quad T = T_0 + \sum_{i=1}^{m} \beta_i T_i,$$

where S_0 and T_0 are the average face shape and texture and (S_1,...,S_m) and (T_1, ..., T_m) are the eigenvectors of shape and texture respectively. Here, $S_i, T_i \in \mathbb{R}^{3N}$. Thus, in this model, faces are represented by the set of coefficients $\alpha = (\alpha_1, ..., \alpha_m)$ and $\beta = (\beta_1, ..., \beta_m)$ that correspond to their shape and texture.

If one assumes the coefficients are drawn from independent normal distributions, PCA also gives an estimate of their probability distributions;

$$P(\alpha) \propto exp(-\frac{1}{2}\sum_{i=1}^{m}\frac{\alpha_i^2}{\sigma_i^2}), \quad P(\beta) \propto exp(-\frac{1}{2}\sum_{i=1}^{m}\frac{\beta_i^2}{\gamma_i^2}), \tag{1}$$

where σ_i and γ_i are determined by the respective eigenvalues of the covariance matrices of $\{FS_i\}$ and $\{FT_i\}$.

2.2 Image Formation Model

We assume faces to be in or close to the space spanned by the shape and texture bases of Sect. 2.1. Then, given a face's shape parameters α and a suitable rigid

body transformation (rotation R and translation t, that align the face model with the actual face), the true color value ($\gamma(k)$) of the face at the position corresponding to the face model's vertex k will equal that predicted by the model:

$$\gamma(k) \approx I_m(k), \tag{2}$$

where $I_m(k)$ is the RGB value of the texture at v_k as given by the texture parameters β, and a suitable set of lighting parameters.

For a lighting model, we assume the surface is Lambertian, and use (R_{amb}, G_{amb}, B_{amb}) for the ambient light color, (R_{dir}, G_{dir}, B_{dir}) for the directional light color, (R_{offset}, G_{offset}, B_{offset}) for the color channels offsets, and l for the directional light direction. Then we have:

$$I_m(k)_R = R_{\text{offset}} + t_{kR} \cdot (R_{\text{amb}} + R_{\text{dir}} \cdot (n_k \cdot l)), \tag{3}$$

with similar definitions for the G,B channels. The symbol t_k represents the k^{th} RGB value in the face model's texture vector representation given the texture coefficients β, and n_k represents the surface normal at v_k.

Assume we are given a stereo pair (I_1, I_2) of face images captured from a pair of calibrated cameras. Letting P_1 and P_2 denote the two camera projection matrices, and assuming we are given the shape parameters α and rigid body transformation parameters (R,t), we have two available measurements of $\gamma(k)$. These can be written $I_1(P_1(R(v_k - c) + c + t))$ and $I_2(P_2(R(v_k - c) + c + t))$, where c is the centroid of the average face shape. Assuming that the cameras are radiometrically calibrated (i.e., have the same exposure, white balance, etc.) with additive Gaussian noise, a reasonable estimator for $\gamma(k)$ is:

$$\hat{\gamma}(k) = \bar{I}(v_k, R, t) \triangleq \frac{I_1(P_1(R(v_k - c) + c + t)) + I_2(P_2(R(v_k - c) + c + t))}{2}. \tag{4}$$

Thus a simple approximation for the distribution of $I_m(k)$, given I_1, I_2, α, R, t is a normal distribution with mean $\bar{I}$ and standard deviation σ_t (say):

$$I_m(k) \sim N(\bar{I}(v_k, R, t), \sigma_t). \tag{5}$$

In addition, when α, I_2, R, t are given, and again assuming that the cameras are radiometrically calibrated, we can use the following model for the noisy observation in I_1 of a vertex v_k that is visible in both images:

$$I_1(P_1(R(v_k - c) + c + t)) \sim N(I_2(P_2(R(v_k - c) + c + t)), \sigma_s). \tag{6}$$

Note that if the cameras are not radiometrically calibrated, this can be generalized by incorporating camera-dependent gains and offsets into I_1 and I_2.

For simplicity, we make use of the following notation in the next section:

- ρ - the 6 parameters of the rigid body transformation (3 for R, 3 for t).
- τ - the 11 lighting parameters (3 for i_{amb}, 3 for i_{dir}, 3 for i_{offset}, i={R,G,B}, and 2 for l).
- s_k - the position of the k^{th} model vertex given pose parameters (R,t) and shape parameters α; $s_k = R(v_k - c) + c + t$.

3 Robust Stereo Fitting of 3DMMs

3.1 Joint shape and texture stereo fitting

We use an energy function that incorporates both a shape model and a texture model by combining terms derived from Eqs. 5 and 6, with regularization:

$$E = \underbrace{\sum_{k|v_k \in V} \frac{||I_1(P_1 s_k) - I_2(P_2 s_k)||^2}{\sigma_s^2}}_{\text{Stereo Match}} + \underbrace{\sum_{i=1}^{m} \frac{\alpha_i^2}{\sigma_i^2}}_{\text{Shape Prior}} + \underbrace{\sum_{k|v_k \in V} \frac{||I_m(k) - \bar{I}(s_k)||^2}{\sigma_t^2}}_{\text{Texture Model Match}} + \underbrace{\sum_{i=1}^{m} \frac{\beta_i^2}{\gamma_i^2}}_{\text{Texture Prior}} . \tag{7}$$

Here, the symbol V is used to denote the set of vertices v_k of the face model with parameters (α, ρ) that are visible in both I_1 and I_2.

Model-fitting is performed by finding parameters $\alpha, \beta, \rho, \tau$ that minimize E. This can be interpreted in a MAP framework as a search for parameters (α,β,ρ,τ) for which the posterior $P(\alpha,\beta,\rho,\tau|I_1,I_2)$ is maximal, and such an interpretation highlights the assumptions underlying our approach. First, we expand the posterior as $P(\alpha,\beta,\rho,\tau|I_1,I_2) = P(\alpha,\rho|I_1,I_2) \cdot P(\beta,\tau|I_1,I_2,\alpha,\rho)$. The first term is then rewritten $P(\alpha,\rho|I_1,I_2) \propto P(I_1|\alpha,\rho,I_2) \cdot P(\alpha)$, which by Bayes' rule, assumes that α, ρ, I_2 are mutually independent and that the distribution of face poses (ρ) is uniform. The assumption that shape (α) and pose (ρ) are independent from I_2 may seem non-trivial. But without knowledge of face texture (β), little can be inferred about I_2, because any image I_2 can be explained by a suitably selected texture.

Using Eq. 6 we write:

$$P(I_1|\alpha,\rho,I_2) \propto \prod_{k|v_k \in V} exp\left(-\frac{1}{2}\frac{||I_1(P_1 s_k) - I_2(P_2 s_k)||^2}{\sigma_s^2}\right). \tag{8}$$

and using Eq. 5 (assuming the texture (β) and scene lighting (τ) independent, and τ uniformly distributed), we write:

$$P(\beta,\tau|I_1,I_2,\alpha,\rho) \propto P(\beta) \cdot \prod_{k|v_k \in V} exp\left(-\frac{1}{2}\frac{||I_m(k) - \bar{I}(s_k)||^2}{\sigma_t^2}\right). \tag{9}$$

Finally, we obtain the energy E by substituting Eqs. 1,8 and 9 into our expression for the posterior, taking the logarithm, negating it and ignoring constant factors.

One can make the following observations about this energy function. First, suppose one were to include only the last three terms in Eq. 7, which would correspond to maximizing $P(I_1|\alpha,\beta,\rho,\tau) \cdot P(I_2|\alpha,\beta,\rho,\tau) \cdot P(\alpha,\beta)$. This approach

would not account for the correlation between I_1 and I_2. The two images are not independent given (α,β,ρ,τ) because the true appearance of the face deviates from that given by the face model, and consequently, the two prediction errors are correlated.

Second, suppose we were to ignore the third and the fourth terms in Eq. 7. This is the approach taken in [5], and it corresponds to maximizing $P(\alpha,\rho|I_1,I_2)$ without including a texture model. As we will show experimentally in Sect. 4, this approach can perform poorly because it does not necessarily ensure that important features (eyes, eyebrows, lips) are properly aligned.

Finally, we can compare our approach to an uncalibrated case in which one has no information about the stereo cameras. In this case, separate pose parameters (ρ_1,ρ_2) could be used for each image, and one might seek to maximize $P(\alpha,\beta,\tau,\rho_1,\rho_2|I_1,I_2)$. In this case, by the same argument as in the first observation, I_1 and I_2 are still not independent given $\alpha,\beta,\tau,\rho_1,\rho_2$, therefore maximizing $P(I_1|\alpha,\beta,\tau,\rho_1)\cdot P(I_2|\alpha,\beta,\tau,\rho_2)\cdot P(\alpha,\beta)$ (which would be the trivial extension of the monocular fitting case to two images [6]) does not necessarily maximize $P(\alpha,\beta,\tau,\rho_1,\rho_2|I_1,I_2)$.

3.2 Handling Occlusion

While the approach in the previous section correctly handles cases of self-occlusion (where one part of the face occludes another), it does not account for the possibility of foreign-body occlusions. To handle such situations, we use a modified version of the energy function in Eq. 7, introducing a robust estimator h_a:

$$E' = \sum_{k|v_k\in V} h_a\left(\frac{||I_1(P_1 s_k) - I_2(P_2 s_k)||^2}{\sigma_s^2} + \frac{||I_m(k) - \bar{I}(s_k)||^2}{\sigma_t^2}\right) + \sum_{i=1}^{m}\frac{\alpha_i^2}{\sigma_i^2} + \sum_{i=1}^{m}\frac{\beta_i^2}{\gamma_i^2} \quad (10)$$

This modification requires little change in the optimization procedure, and allows the fitting to be significantly more robust to foreign-body occlusions (see Sect. 4.2). Intuitively, by introducing the robust estimator we are limiting the impact in the energy function of vertices whose stereo matching term or texture matching term are high. More formally, this approach can be justified by introducing a binary occlusion map $O:\{1,..,N\}\rightarrow\{0,1\}^N$, defined on the set of all vertices of the face model. This map dictates whether a vertex of the face model is occluded by a foreign-body in at least one of the images ($O(k)=1$) or not occluded in either ($O(k)=0$). Thus, the image formation model is altered so that the visible parts of the face present in the images are generated only by vertices v_k for which $O(k)=0$.

In this setting, it can be shown that minimizing E' corresponds to searching for α,β,ρ,τ,O for which $P(\alpha,\beta,\rho,\tau,O|I_1,I_2)$ is maximal. Again, we can write $P(\alpha,\beta,\rho,\tau,O|I_1,I_2) = P(\alpha,\rho,O|I_1,I_2)\cdot P(\beta,\tau|I_1,I_2,\alpha,\rho,O)$. We expand the first term by making the same assumptions as those used in the previous section, obtaining $P(\alpha,\rho,O|I_1,I_2)\propto P(I_1|\alpha,\rho,O,I_2)\cdot P(\alpha,O)$. The term $P(I_1|\alpha,\rho,O,I_2)$ is then approximated as in Eq. 8, where the product is now over $\{k|v_k\in V, O(k)=$

0}. In favor of simplicity and efficiency, we ignore spatial coherence of occlusions, and assume $O(k) \sim$ i.i.d. Bernoulli, obtaining the following prior on O:

$$P(O) \propto \prod_{k|v_k \in V} exp(-\eta_o \cdot O(k)). \tag{11}$$

Using this prior avoids the trivial labeling of all vertices being occluded during the optimization process.

Combining these terms and assuming the shape (α) and occlusion map (O) to be independent, we obtain an expression for $P(\alpha, \rho, O|I_1, I_2)$. Substituting this expression into the posterior along with an expression for the posterior's second term similar to Eq. 9 (but with the product over $\{k|v_k \in V, O(k) = 0\}$, one sees that maximizing the posterior corresponds to minimizing:

$$E'' = \sum_{k|v_k \in V} f(\alpha, \beta, \rho, \tau, O, k) + \sum_{i=1}^{m} \frac{\alpha_i^2}{\sigma_i^2} + \sum_{i=1}^{m} \frac{\beta_i^2}{\gamma_i^2}, \tag{12}$$

where

$$f(\alpha, \beta, \rho, \tau, O, k) = g(\alpha, \beta, \rho, \tau, k) \cdot (1 - O(k)) + 2\eta_o \cdot O(k), \tag{13}$$

and

$$g(\alpha, \beta, \rho, \tau, k) = \frac{||I_1(P_1 s_k) - I_2(P_2 s_k)||^2}{\sigma_s^2} + \frac{||I_m(k) - \bar{I}(s_k)||^2}{\sigma_t^2}. \tag{14}$$

The minimization of E'' can be rearranged as:

$$\min_{\alpha,\beta,\rho,\tau,O} E'' = \min_{\alpha,\beta,\rho,\tau} \{\min_{O} \{ \sum_{k|v_k \in V} f(\alpha, \beta, \rho, \tau, O, k)\} + \sum_{i=1}^{m} \frac{\alpha_i^2}{\sigma_i^2} + \sum_{i=1}^{m} \frac{\beta_i^2}{\gamma_i^2}\} \tag{15}$$

$$= \min_{\alpha,\beta,\rho,\tau} \{ \sum_{k|v_k \in V} h(g(\alpha, \beta, \rho, \tau, k), k) + \sum_{i=1}^{m} \frac{\alpha_i^2}{\sigma_i^2} + \sum_{i=1}^{m} \frac{\beta_i^2}{\gamma_i^2}\} \tag{16}$$

where

$$h(g(\alpha, \beta, \rho, \tau, k), k) = \min_{O(k)} \{g(\alpha, \beta, \rho, \tau, k) \cdot (1 - O(k)) + 2\eta_o \cdot O(k)\}. \tag{17}$$

Relaxing the binary process $O(k)$ to an outlier process that varies continuously $0 \leq O_a(k) \leq 1$, we can approximate $h(g, k)$ by a robust function h_a,

$$h_a(g) = -\sigma_o \cdot ln((1 - exp(-\frac{e_o}{\sigma_o})) \cdot exp(-\frac{g}{\sigma_o}) + exp(-\frac{e_o}{\sigma_o})) \tag{18}$$

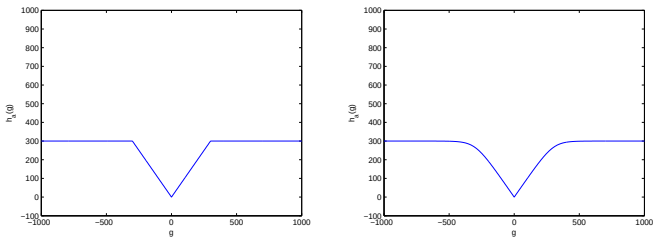
with suitable parameters e_o and σ_o. These parameters are determined empirically to provide a smooth approximation of the min function (see Fig. 1). This leads to E' as in Eq. 10, where the minimization is over α,β,ρ,τ.

Following optimization, the occlusion map is recovered from (for $v_k \in V$):

$$O^*(k) = 1, \text{ if } h_a(g(\alpha^*, \beta^*, \rho^*, \tau^*, k)) \geq 2\eta_o - \varepsilon$$
$$O^*(k) = 0, \text{ if } h_a(g(\alpha^*, \beta^*, \rho^*, \tau^*, k)) < 2\eta_o - \varepsilon,$$

where

$$(\alpha^*, \beta^*, \rho^*, \tau^*) = arg \min_{\alpha,\beta,\rho,\tau} E'. \tag{19}$$



(a) **(b)**

Fig. 1. Robust estimator $h_a(g)$ (Eq. 18) used to handle foreign-body occlusions in the fitting process: (a) $e_o = 300, \sigma_o = 1$ (b) $e_o = 300, \sigma_o = 50$

3.3 On Foreign Body Occlusions

In a stereo setup, there can be several cases of foreign-body occlusion of a vertex of the face model. We can classify these cases with respect to the positioning of the occluder in (see Fig. 2): half-occlusion (HO), where the vertex is occluded in one of I_1 or I_2; full-occlusion-near (FO_n), where the vertex is occluded in both I_1 and I_2 and the occluding object is close to the face; and full-occlusion-far (FO_f), where the occluder is far from the face relative to the face size. We can also classify occluders with respect to their texture, which can be one of: texture-less (non-skincolor); texture-less (skincolor); and textured.

Depending on the type of occlusion, we expect either the stereo match term or the texture match term to play a more prominent role in the fitting process (see Table 1). For example, in the case of half-occlusion (HO) by a non-skinlike surface, one can expect the stereo match term to provide an important cue as to whether a vertex is occluded. This is because the observed intensities at the projections of a half-occluded vertex correspond to observations of two very different surfaces. When the occlusion is of type full-occlusion-near (FO_n) on the other hand, the stereo match term will not provide much help in determining an occlusion because the two observed intensities will come from nearby locations on the occluder and will be very similar. In this case, provided that the occluder has non-skinlike color, the texture match will be the most helpful in determining its presence. Of course, when the occluder lacks texture and is skinlike, there is little visual information to discriminate between it and the face.

Experimental results are shown in Sect. 4.2.

Occluder classification	HO	FO_n	FO_f
texture-less (non-skincolor)	S	T	T
texture-less (skincolor)	X	X	X
textured	S	T	S+T

Table 1. Most relevant terms in the energy function for each of the occlusion cases: S for stereo match term and T for texture match term (see Fig. 2).

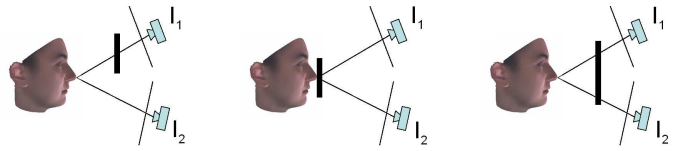


Fig. 2. Categories of foreign-body occlusions. From left to right, occlusions can be one of: half-occlusion (HO), full-occlusion-near (FO_n), full-occlusion-far (FO_f). The stereo and texture terms play different roles in each case (see Table 1).

3.4 Optimization Procedure

Initial Fit Like previous approaches [5, 17], we assume that either by user selection, or by means of an automated detection process, image coordinates of a subset of specific feature points of the face (e.g. corners of the eyes, corners of the mouth, tip of the nose, corners of the ears) in both I_1 and I_2 are available. (Some of the feature points may be occluded in one or both images).

Let $j_1, .., j_p$ denote the indices of the vertices in the face model corresponding to these feature points. Starting from the average shape parameters ($\alpha = 0$), we use a quasi-newton gradient descent method to minimize

$$E_f = \sum_{i=1,..,p} \delta_{1i}||P_1 s_{j_i} - p_{1i}||^2 + \delta_{2i}||P_2 s_{j_i} - p_{2i}||^2, \tag{20}$$

and obtain a rough initial estimate of the shape and rigid body transformation parameters. Here, $\delta_{1i} = 1$ if the i^{th} feature is visible in image I_1 and 0 otherwise (and similarly for δ_{2i} and I_2), p_{1i} is the image coordinate of the i^{th} feature in image I_1, and p_{2i} is the image coordinate of the i^{th} feature in image I_2.

Optimization For comparison purposes we evaluate the fitting performance of E and E' with and without the texture model terms. In experiments where we utilize only the stereo terms in E (or E'), we start with model parameters α, ρ from the initial fit. In experiments that include texture we also start with the average texture parameters ($\beta = 0$), and lighting parameters τ such that $i_{\text{amb}} = 1$, $i_{\text{dir}} = 1$ (i.e., white ambient and directional lights), and $i_{\text{offset}} = 0$ (zero offset), where $i = R, G, B$. The lighting direction l is initialized to be the bisector of the two cameras viewing directions.

We minimize:

$$E + \lambda \cdot E_f \tag{21}$$

with respect to the suitable parameters, using a stochastic quasi-newton gradient descent method.

To avoid local minima, we use a coarse-to-fine approach, with 3 levels of resolution. At the coarsest resolution, we use versions of I_1 and I_2 that are downsampled by a factor of four, together with a corresponding low resolution version of the 3D face model. As we progress toward the finest level of resolution, we use smaller and smaller values for λ, σ_s and σ_t, which gives smaller weights

to the feature term and the shape and texture priors. At regular intervals (more frequently at coarser levels), we recompute the self-occluded vertices (and thus V) as well as the normals (n_k). Instead of computing the energy using all the vertices $v_k \in V$, at each iteration we randomly select a sub-set of these vertices on which to compute the energy (we use 1000, 2000 and 3000, at each level of resolution). In this selection process, we select vertices with probability proportional to the average (over the stereo pair) foreshortened area of the patch around them. When we utilize the complete E or E', we sample at the baricenters of the triangles of the mesh instead of the vertices because that allows for easier computation of the gradient of the energy. In this case, both V and the occlusion map are defined over the set of triangles, and k indexes the triangles that compose the model.

4 Experimental Results

We evaluated the procedure of Sect. 3.4 using the original energy (E) and the robust energy (E'), along with modifications of these energies obtained by excluding the texture terms. Throughout this section, we refer to these as stereo+texture, stereo, robust stereo+texture, and robust stereo, respectively. To ensure a valid comparison between the different cases, we used equivalent parameters for the feature match weight (λ) and the model priors (σ_s and σ_t) in each experiment. Only the first 40 shape and texture basis vectors were used, since this was found to provide adequate results.

4.1 Accuracy in the Absence of Occlusions

To evaluate the benefits of incorporating a texture model in the absence of occlusions, testing was performed on a subset of sixty individuals from the K.U. Leuven stereo face database [5], which contains stereo pairs of each individual in eight different positions. We obtained fitting results using the stereo and the stereo+texture methods for all eight poses in each of the sixty people, for a total of 480 model fits. Note that the stereo fitting approach is that proposed in [5].

Figures 3 and 4 exemplify the differences between the fits obtained using stereo (first two terms of E) and stereo+texture (E). At first glance, the results in Fig. 3 suggest that the shape estimates using both methods are quite similar. The stereo matching cost ($\sum_{k|v_k \in V} \frac{||I_1(P_1 s_k) - I_2(P_2 s_k)||^2}{|V|}$) was computed to be 280.77 for the stereo method and 340.17 for the stereo+texture method, so the shape obtained using only the stereo term is better in terms of the per-vertex stereo intensity match. However, from Fig. 4 it is clear that the eye, eyebrow and mouth alignment between the model and the images is significantly more accurate when the texture model is included.

These results suggest that either approach may be sufficient if the desired output is a depth map or 3D model for image synthesis. For recognition, however, where one links shape parameters to identity, it is important for features in

the fitted model to be aligned with the features in the database models. Our experiments suggest that one way to ensure this alignment is to include a texture model in the fitting procedure.

The same effect can be observed by studying the distribution of the 480 recovered shape models (60 individuals under 8 poses) in the forty-dimensional whitened shape parameter space. Two statistics relate to the quality of the fitting procedure from a recognition standpoint. First, for a single individual, we would like the difference between the fits for different poses to be small. Second, we would like the difference between fits for distinct individuals to be large. These can be measured based on the within-class (within-subject) scatter matrix (S_w) and the between-class scatter matrix (S_b). Roughly speaking, the larger the determinant and trace of ($S_w^{-1}S_b$) are, the more accurate a classifier based on these fits will be. Using results from the 480 fits we found the determinants of $S_w^{-1}S_b$ to be $2.9640e^{-5}$ and $1.3418e^{-11}$ and the traces of $S_w^{-1}S_b$ to be 104.0478 and 69.4101 for the stereo+texture method and the stereo method, respectively. These quantitative results support the qualitative observations in Figs. 3 and 4 and suggest that fits obtained with the inclusion of the texture model are significantly more robust to pose changes.

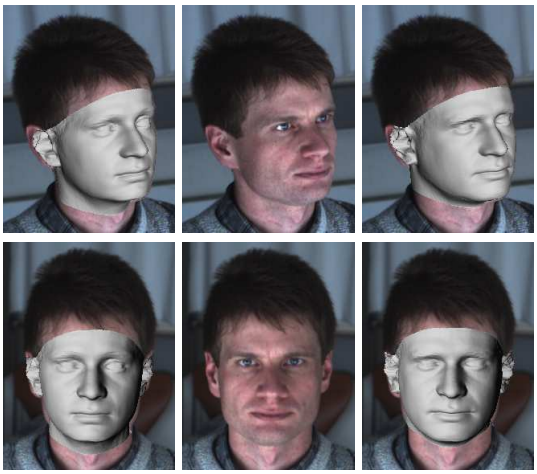


Fig. 3. Comparison of a fit using both stereo and texture to that obtained using stereo alone. Rows indicate left and right images of the stereo pair. First column: shape estimate using stereo, second column: input images, third column: shape estimate using stereo and texture.

4.2 Accuracy with Occlusions

We also tested the occlusion cases described in Sect. 3.3 by applying the robust fitting process to captured data. For these fitting results, a value of $n_o = 250$ was

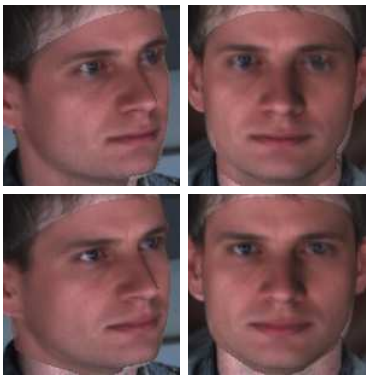


Fig. 4. Same comparison as that in Fig. 3, but mapped with estimated textures and rendered semi-transparently over input images. While both the shape obtained using stereo (top) and that obtained using stereo and texture (bottom) provide reasonable depth maps for the input stereo pair (Fig. 3), only the joint use of stereo and texture ensures feature alignment.

used for the robust stereo method, and a value of $n_o = 800$ was used for the robust stereo+texture method.

Figure 5 shows results obtained using the robust stereo and robust stereo+ texture method in the case of half-occlusion (case HO) by a textureless foreign body. As described in Sect. 3.3, in this case we expect the results for both methods to be similar because the stereo cue is sufficient to detect the occluder. As shown in the figure, this is indeed the case. Notice that the occlusion map captures not only the occluder, but also artifacts that are not predicted by the model, including specular highlights and cast shadows.

Figure 6 shows similar results for the case of a textured occluder that is close to the surface (case FO_n). In this case, the stereo constraint is insufficient for detecting the occluder, and the addition of a texture term provides substantial improvement.

The results from the two occlusion cases are compared to the 'ground truth' shape obtained in the absence of occlusion in Fig. 7. The results obtained by the robust stereo+texture method are relatively consistent over all cases, but the same cannot be said for those obtained using the stereo match alone. Notice that in all cases, the recovered models deviate from the unoccluded model in the unobserved regions of the face. This is to be expected, since there is no shape or texture information available in these regions.

5 Conclusions

We have presented a method for the recovery of face models from stereo pairs of images in the presence of foreign-body occlusions. In this approach, a face model (a 3DMM) is augmented by an occlusion map defined on the model shape, and foreign-body occlusions are detected efficiently using robust esti-

Fig. 5. Models are fit to an input stereo pair (top row) using robust stereo (left columns) and robust stereo and texture (right columns). Here, the face is half-occluded (occluder type *HO*) by a textureless object. The results from the two methods are very similar, showing that the stereo match term alone suffices for detecting the occluder. The bottom row shows the estimated occlusion map with black indicating foreign-body occlusion ($O(k) = 1$), white indicating visible vertices ($O(k) = 0$ and $v_k \in V$), and red indicating self-occlusion ($v_k \notin V$).

mators. The approach uses an explicit model for texture in addition to shape in an energy-based stereo fitting process.

Experimental results demonstrate robustness to occlusions, and they highlight the relative importance of the stereo match term and the texture match term in the energy. They suggest that both shape and texture components of a 3DMM should be incorporated if one seeks to detect general classes of occluders. The results also suggest that even in the absence of foreign-body occlusions, an explicit texture model can significantly improve stereo fitting results. The texture model provides one way of ensuring proper alignment of features (eyes, eyebrows, lips, etc) in the fitted model.

Another possible approach to achieve alignment, and one we plan to explore in the future, is to use only shape in the stereo fitting process and to incorporate a stereo matching term that is more sophisticated than simple per-vertex intensity differences. This is the approach taken in [19], for example, where window-based matching is employed. One may also look at other feature spaces for fitting (e.g. [20]), as well as better models for the distribution of the error in the modeling of texture (Eq. 5).

Finally, if one is to perform recognition based on models obtained in the presence of occlusions, one would likely want a second model refinement step

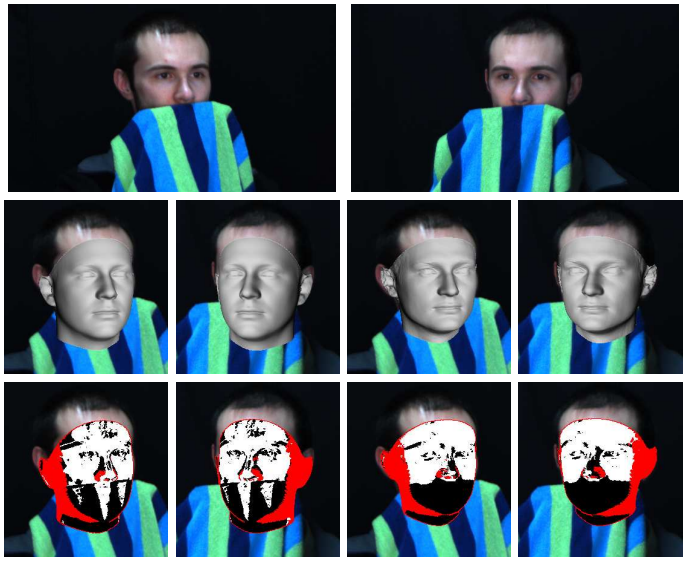


Fig. 6. Same as in Fig. 5, but for the case of a textured foreign-body occluder that is close to the face (occluder of type FO_n). In this case, as evidenced by the occlusion map on the bottom left, the stereo match term alone is not enough to detect the occluder, and the recovered model is inaccurate. Including the texture model (bottom right) significantly improves the result.

in which one breaks the initial model into segments [6] in a way that respects the occlusion boundaries. The goal would then be to infer identity using only the unoccluded segments of the model.

Acknowledgements

This work was supported by an NSF CAREER award, IIS-0546408.

References

1. Georghiades, A., Kriegman, D., Belhumeur, P.: Illumination cones for recognition under variable lighting: Faces. CVPR (1998) 52–59
2. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE TPAMI **19**(7) (1997) 711–720
3. Basri, R., Jacobs, D.: Lambertian reflectance and linear subspaces. IEEE TPAMI **25**(2) (2003) 218–233
4. Lee, K.C., Ho, J., Kriegman, D.J.: Nine points of light: acquiring subspaces for face recognition under variable lighting. CVPR (**1**) 519–526
5. Fransens, R., Strecha, C., Van Gool, L.: Parametric Stereo for Multi-Pose Face Recognition and 3D-Face Modeling. AMFG (2005) 108–123
6. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. Proceedings of ACM SIGGRAPH (1999) 187–194

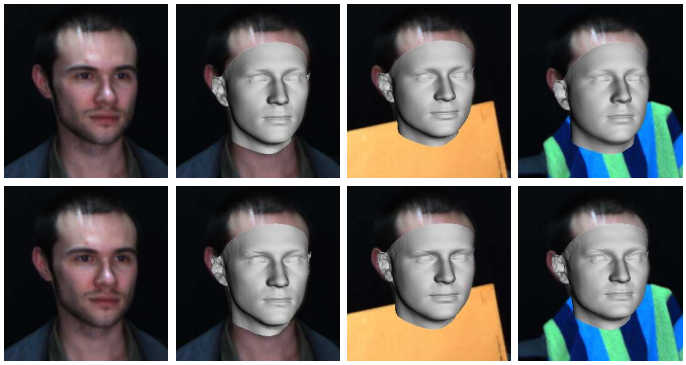


Fig. 7. Comparison of the shapes recovered using robust stereo (first row) and robust stereo and texture (second row) in cases of (from left to right) no occlusion, half-occlusion, and full-occlusion-near. Here, the estimates are overlayed on top of one of their input images. While stereo handles the half occlusion case reasonably well, only combined use of stereo and texture ensures that the recovered model is close to the 'ground truth' shape—at least in its visible regions—in both occlusion cases.

7. Blanz, V., Vetter, T.: Face recognition based on fitting a 3 D morphable model. IEEE TPAMI **25**(9) (2003) 1063–1074
8. Blanz, V., Grother, P., Phillips, P., Vetter, T.: Face recognition based on frontal views generated from non-frontal images. CVPR **2** (2005)
9. De Smet, M., Fransens, R., Van Gool, L., ESAT-PSI, K.: A Generalized EM Approach for 3D Model Based Face Recognition under Occlusions. CVPR **2** (2006) 1423–1430
10. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE TPAMI **23**(6) (2001) 681–685
11. Baker, S., Matthews, I.: Lucas-Kanade 20 Years On: A Unifying Framework. IJCV **56**(3) (2004) 221–255
12. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time combined 2D+ 3D active appearance models. CVPR **2** (2004) 535–542
13. Hu, C., Xiao, J., Matthews, I., Baker, S., Cohn, J., Kanade, T.: Fitting a single active appearance model simultaneously to multiple images. Proc. British Machine Vision Conference (2004)
14. Koterba, S.C., Baker, S., Matthews, I., Hu, C., Xiao, J., Cohn, J., Kanade, T.: Multi-View AAM Fitting and Camera Calibration. In: ICCV. Volume 1. (2005) 511 – 518
15. Gross, R., Matthews, I., Baker, S.: Active Appearance Models with Occlusion. Image and Vision Computing **24**(6) (2006) 593–604
16. Basso, C., Vetter, T., Blanz, V.: Regularized 3D morphable models. IEEE Int. Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis (2003) 3–10
17. Blanz, V., Basso, C., Poggio, T., Vetter, T.: Reanimating Faces in Images and Video. Computer Graphics Forum **22**(3) (2003) 641–650
18. Basso, C., Paysan, P., Vetter, T.: Registration of Expressions Data using a 3D Morphable Model. FGR (2006) 205–210
19. Dimitrijevic, M., Ilic, S., Fua, P.: Accurate face models from uncalibrated and ill-lit video sequences. CVPR **2** (2004)
20. Romdhani, S., Vetter, T.: Estimating 3D Shape and Texture Using Pixel Intensity, Edges, Specular Highlights, Texture Constraints and a Prior. CVPR (2005)