Accurate Face Models from Uncalibrated and Ill-Lit Video Sequences

M. Dimitrijevic, S. Ilic and P. Fua [*]
Computer Vision Laboratory
Swiss Federal Institute of Technology
1015 Lausanne, Switzerland

Technical Report No: IC/2004/05

Abstract

In this paper, we propose a face reconstruction technique that produces models that not only look good when texture mapped, but are also metrically accurate. Our method is designed to work with short uncalibrated video or movie sequences, even when the lighting is poor resulting in specularities and shadows that complicate the algorithm's task.

Our approach relies on optimizing the shape parameters of a sophisticated PCA based model given pairwise image correspondences as input. All that is required is enough relative motion between camera and subject so that we can derive structure from motion. By matching the results against laser scanning data, we will show that its precision is excellent and can be predicted as a function of the number and quality of the correspondences. This is important if one wishes to obtain the appropriate compromise between processing speed and quality of the results.

Furthermore, our method is in fact not specific to faces and could equally be applied to any shape for which a shape model controlled with relatively small number of parameters exists.

1 Introduction

In recent years, the movie industry has produced such realistic 3–D face models from images that we have come to take them for granted. However, a quick look at the credits at the end of a movie such as "The Matrix Reloaded" and at the budgets that are involved, should alert the careful scientist to the fact that this is a misperception. As explained in a recent SIGGRAPH sketch [1], the extraordinary quality of the models shown in that movie required the use of a studio with five calibrated high resolution cameras, carefully controlled lighting, and an untold number of hours of work. Furthermore the 3–D shapes were obtained not directly from the images but by laser-scanning a plaster cast of the actors' faces.

A few years ago, Blanz & Vetter [2] have proposed an extremely impressive appearance-based approach that addresses this issue using a sophisticated statistical head model. It includes shape and texture

[*] This work was supported in part by the Swiss National Science Foundation.

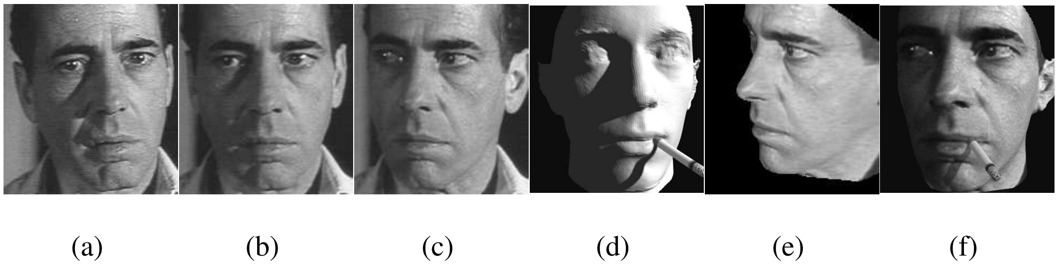


Figure 1: (a,b,c) Three grainy 160×236 images from a sequence of five, digitized from an old celluloid film. (d) Reconstructed model shaded using a light source estimate. Note that the pattern of the shadows corresponds quite accurately to those that appear in (b), which indicates that both the shape the light source estimate are correct. This actor is known to have been a chain smoker. We therefore took the liberty of adding a virtual cigarette, which casts a shadow at the appropriate place. (e,f) Texture-mapped views of the model without and with the virtual cigarette.

components that have been learned from a large database of human heads. It allows reconstruction from a single image and uses the Phong illumination model to handle illumination effects, but the shape and texture recovery may be perturbed by large cast shadows or specularities [3, 4]. In this paper, we propose a technique that reduces the sensitivity to illumination by replacing the texture component of the model by information provided by 2–D point correspondences in all pairs of consecutive images. This helps because such correspondences tend to be affected comparatively little by illumination changes given proper normalization. Furthermore, this approach has the potential for increased automation by eliminating the need for 3–D feature points whose projections are known.

The main insight of this paper is that, given enough such correspondences in all consecutive pairs in the sequence and a linear shape model, recovering both the shape and the pose of *all* cameras with respect to it can be formulated as a least-squares problem that is both close to being quadratic and well conditioned. We can therefore do it accurately even though many correspondences may be erroneous and the others are only precise to the nearest pixel. This results in a method that

- Allows the automated construction of models that not only look good when texture mapped but are also metrically accurate.
- Deals with uncalibrated sequences acquired with uncontrolled scene illumination that may produce cast shadows and specularities.

Figs. 1 and 2 depict two such sequences that our algorithm handles well, even though they create major problems for some of the best current structure-from-motion methods such as graph-cuts stereo algorithms [5, 6]. Our approach is related to the Model-Based Bundle adjustment technique proposed by Shan et al. [7] but we do away with the requirement for a set of 3–D feature points whose projections are known, which, in practice, results in a relatively complex processing chain [8].

This leads to a streamlined algorithm that only requires an approximate 3–D pose estimation in one image and possibly noisy 2–D correspondences between frames to produce high-quality models. We will use laser scanning data to show that, given enough correspondences, they are metrically accurate and that the precision can be estimated from the recovered camera configuration and number of correspondences. As a result, if speed is more important than precision, one can be traded for the other with predictable results by reducing the number of correspondences.

Our contribution is therefore a face reconstruction algorithm that is metrically accurate, predictable, amenable to full automation, and can handle the kind of images one is likely to obtain in uncontrolled

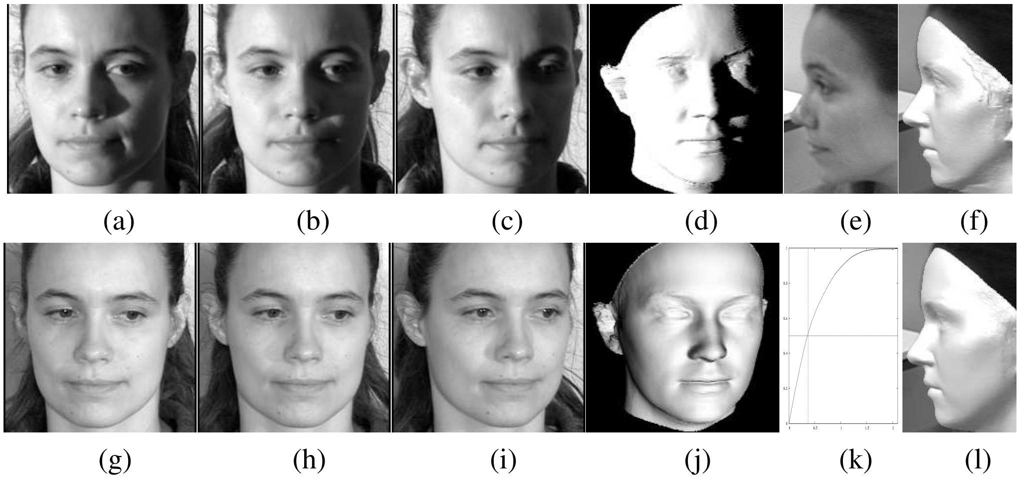


Figure 2: (a,b,c) Three images of a face lighted by a single directional light source. Note the strong shadows that can create problems for traditional stereo algorithms. (d) Shaded model shaded using the estimated light source vector for image (c). The regions in shadow are shown in black and their pattern corresponds to that of the shadows in the image, indicating that both the shape and the light source direction are approximately correct. (e,f) A profile image that has not been used to perform the reconstruction and the model seen from the same viewpoint. (g,h,i) Three images of the same face under a much more diffuse lighting but with some specularities on the forehead and nose. (j) Shaded view of the model reconstructed using these images, whose shape is almost exactly the same as that shown above. (k) Cumulative function of distances between the two models. The median distance is 0.35mm, which corresponds to reprojection errors smaller than 0.5 pixels on average. (l) Profile view that is almost indistinguishable from the one shown in (f).

environments. Furthermore, our method is in fact not specific to faces and could equally well be applied to any shape for which a shape model controlled by a relatively small number of parameters exists. We chose to implement and evaluate our approach in the context of head-modeling because heads are one of the most complex objects for which the appropriate low-dimensional models have been developed.

2 Related Work

Our approach incorporates a sophisticated face model into a bundle-adjustment framework. We briefly review related techniques.

2.1 Bundle-Adjustment and Autocalibration

Bundle-adjustment is a well established technique in the photogrammetric community. Lately, it has been increasingly used in the computer vision community but most results have been demonstrated in man-made environments where feature points can be reliably extracted and matched across images [9]. One cannot assume that those results carry over directly in the case of ill-textured objects for which correspondences cannot be established with great precision.

In earlier work [10], we have shown that a generic facetized model could be used to derive shape constraints that adequately regularize the problem. Similarly, Shan et al. [7] introduced a model-based bundle adjustment technique in which the optimization variables are the camera positions and the model shape parameters. However, both these approaches rely on face models that are too simple to be accurate and involve the use of 3–D features points whose projection is known. Our new approach does away with both of these limitations.

2.2 Using Head Models

In recent years, a great many approaches to modeling faces from image and range data have been proposed. They rely on user-supplied correspondences [11], stereo [12], shading [13, 14], structured light [15], silhouettes [16] or low-intensity lasers.

Successful approaches to automating the fitting process have involved the use of optical flow [17] or appearance based techniques [18] to overcome the fact that faces have little texture and that, as a result, automatically and reliably establishing correspondences is difficult. This latter technique is closely related to ours because head shape and camera motion are recovered simultaneously. However, the optical flow approach avoids the "correspondence problem" at the cost of making assumptions about constant illumination of the face that may be violated as the head moves. This tends to limit the range of images that can be used, especially if the lighting is not diffuse.

The original Blanz & Vetter approach [2] uses a sophisticated statistical head model that includes shape and texture components that have been learned from a large database of human heads. Its shape and texture parameters can be adjusted so that it can synthesize images that closely resemble the input image or images. While excellent results have been obtained using a single image, their quality can degrade in the presence of large specularities and cast shadows, which can be mistakenly interpreted as changes in shape. We therefore chose to use only the shape component of the model and to replace the appearance-based approach by one that relies on pairwise image-point correspondences. Because these can be obtained using normalized cross correlation, which minimizes the influence of illumination changes, our method is robust to such changes, at the cost of having to use a short sequence as opposed to a single image.

3 Approach

In this section, we introduce our approach to head modeling from uncalibrated image sequences that may contain strong specularities and shadows. We use 2–D point correspondences in pairs of consecutive images as our main source of information because the disruptive effect of illumination changes are minimized for images whose viewpoints are close and can be further attenuated by normalization.

In theory, given one image for which the 3–D head pose is roughly and a second one seen from a relatively similar viewpoint, we could recover shape and camera position as follows:

1. Sample the face area in the first image.
2. Use a straightforward normalized cross correlation technique to find corresponding points in the second image.
3. Minimize in the least-squares sense the image distance between these corresponding points and those obtained by backprojecting the points in the first image to the model and reprojecting them into the second image.

In practice, because correspondences can be expected to be noisy, we use an iterative reweighted least square technique and, more importantly, we work with more than two images simultaneously. In the next section, we will show that using in this manner the two consecutive pairs of a three image sequence allows us to formulate a least-squares problem that is well conditioned and, therefore, noise resistant. Our complete approach therefore iteratively adds images at both ends of the sequence by establishing correspondences between the first and last images that have already been processed and neighboring ones that have not been considered yet.

At each step of this process, we perform a least-squares minimization that progressively refines model shape and pose for *all* cameras. Unlike earlier approaches[7], our method has no notion of a *reference* image and the pose in the first image does not need to be precise because it will be refined as all the others.

In the remainder of this Section, we introduce the state variables of our models. We then describe in more detail our optimization scheme using a single image pair and then extend it to the processing of a complete sequence.

3.1 Shape and Pose Parameters

We use the face models developed by Blanz & Vetter[2], which are stored as triangulated meshes with 75292 vertices. Given such a mesh, let us consider S, the *shape vector* obtained by concatenating the X,Y and Z coordinates of all its vertices. A database of 3–D faces was used to compute the shape vectors for 200 people and Principal Component Analysis to approximate them as.

$$S = \overline{S} + \sum_{i=1}^{99} \alpha_i S_i \tag{1}$$

where $\overline{S}$ represents an average face model, the S_i are orthogonal vectors, and the α_i are weights. By varying the α_i one can create a whole range of new faces and we treat them as the state variables that control shape. These models also include texture descriptors but we do not use them in order to increase our method's robustness to illumination changes. In our optimization scheme, we take the position and orientation of the model to be fixed and compute the position and orientation of each cameras with respect to it. This entails no loss of generality and allows us to put all the images on an equal footing.

We assume that the intrinsic camera parameters remain constant throughout the sequence. In theory, given high precision matches, bundle-adjustment can recover both intrinsic parameters and camera motion [9]. In practice, however, we must be prepared to deal with the potentially poor quality of the point matches. Therefore, we have chosen to roughly estimate the intrinsic parameters and to concentrate on computing the extrinsic ones using bundle-adjustment: We use an approximate value for the focal length and assume that the principal point remains in the center of the image.

In short, given an m image sequence, the state of our model is defined by $6 * m + n$ parameters where m is the number of images and n the number of α shape coefficients.

3.2 Transfer Function in a Single Image Pair

Given images $i_1$and i_2, let us assume that approximate camera pose parameters are roughly known for the first one. This lets us project the model into image i_1 and sample the face area more or less densely depending on the amount of computational power and processing time available. Given one such sample $p^j_{i_1}$, we use normalized cross correlation to find a corresponding point $p^j_{i_2}$ in the second image. We also compute a 3–D point by intersecting the line of sight defined by $p^j_{i_1}$ with the 3–D model, and project it into image i_2, which yields the 2–D point $\hat{p}^j_{i_2}$. The function Ψ that maps $p^j_{i_1}$into $\hat{p}^j_{i_2}$ is known as the transfer function. It depends on both cameras pose parameters and on model shape. If they are correct and if the correspondences are perfect, we should have $p^j_{i_2} = \hat{p}^j_{i_2} \quad \forall 1 \leq j \leq m_{i_1}$ and the *reprojection error* for point j can be expressed as $\triangle p^j_{i_1,i_2} = \left(p^j_{i_2} - \hat{p}^j_{i_1}\right)$. Given a large enough set of samples $Q_{i_1} = \{p^j_{i_1}, 1 \leq j \leq m_{i_1}\}$, we can therefore recover the shape and position parameters by minimizing

$$F_2(A, C_{i_1}, C_{i_2}) \quad = \quad \sum_{j \in Q_{i_1}} \left\| \triangle p^j_{i_1 i_2} \right\|^2 \tag{2}$$

where A is the vector of α shape parameters, and C_{i_1} and C_{i_2} are the extrinsic parameters for both cameras. We do not assume either camera to be fixed, which is important because we do not want to rely on perfect initialization in any given image. Our formulation of F_2 introduces a slight bias because we do not treat the two images symmetrically. This could be corrected by using a slightly more sophisticated criterion[7]. In our case, however, because we use several pairs the biases cancel each other and, as a result, have not noticeable influence.

3.3 Complete Sequences

The approach outlined above is not limited to an image pair and extends naturally to triplets of images $i-1, i, i+1$ in the sequence, given an an approximate value for C_i. We a create Q_i set of samples in image i, compute correspondences in the other two, and form the three-image objective function

$$F_3(A, C_{i-1}, C_i, C_{i+1}) \quad = \quad \sum_{j \in Q_i} \left\| \triangle p^j_{i,i-1} \right\|^2 + \left\| \triangle p^j_{i,i+1} \right\|^2$$

In Section4.2, we will argue that minimizing F_3 is a well conditioned least-squares problem if we use enough correspondences and can therefore be used to derive reliable estimates of both camera and shape parameters. We can further refine this estimate by using additional images: We sample *independently* images $i-1$ and $i+1$ to create sample sets Q_{i-1} and Q_{i+1}, compute correspondences in images $i-2$ and $i+2$ and form the objective function

$$\begin{aligned} F_5(A, C_{i-2}, C_{i-1}, C_i, C_{i+1}, C_{i+2}) \quad &= \quad F_3(A, C_{i-1}, C_i, C_{i+1}) \\ + \sum_{j \in Q_{i-1}} \left\| \triangle p^j_{i-1,i-2} \right\|^2 \quad &+ \quad \sum_{j \in Q_{i+1}} \left\| \triangle p^j_{i+1,i+2} \right\|^2 \end{aligned}$$

that we minimize with respect to all the parameters. This process can then be repeated recursively for the whole sequence and the objective function we end up minimizing is

$$F_N(A, C_1, ..., C_N) \quad = \quad \sum_{i=i_o}^{N-1} \sum_{j \in Q_i} \left\| \triangle p^j_{i,i+1} \right\|^2 \tag{3}$$

$$+ \quad \sum_{i=2}^{i0} \sum_{j \in Q_i} \left\| \triangle p^j_{i,i-1} \right\|^2 , \tag{4}$$

where i_o is the index of the one image for which we need an initial pose estimate. Note, that because $C_{i_o}$is optimized at the same time as the other extrinsic camera parameters, that estimate need not be exact. This is made possible by the fact that in our approach, there is never an explicit association between 2–D sample points in the images and specific vertices or facets of the 3–D models. Instead, these associations are computed dynamically during the minimization and can change.

In practice, we developed an optimization schedule in which the number of shape parameters that are allowed to vary progressively increases. Because the correspondences are noisy, we perform iterative reweighted least-squares and add a small regularization term. The function we actually minimize is therefore:

$$F = \frac{F_N}{\sigma_N^2} + \sum_{i=1}^{99} \frac{\alpha_i^2}{\sigma_{S_i}^2}$$

where the σ_{S_i} are the eigen values of the shape covariance matrix provided with the model [2] and σ_N an initially large constant that progressively decreases.

4 Analysis and Validation

In practice, as shown in the example of Fig. 2, we get similar results even when we use different sets of images of the same subject and allow both shape and camera parameters to vary freely. In this section, we will analyze the theoretical underpinning of this desirable behavior.

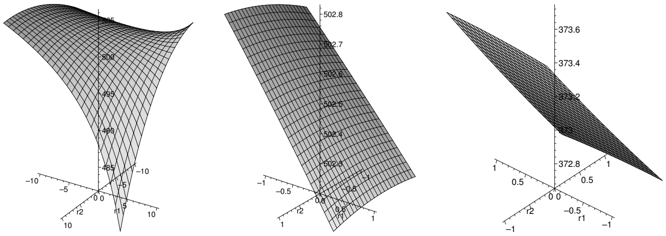
4.1 Transfer Function

The transfer function Ψ introduced in Section 3 back-projects a point $p_{i_1}^j$ in image i_1 to a triangulation facet κ, and from there to a point $\hat{p}_{i_2}^j$ in image i_2. The position of κ can be expressed in terms of its normal $\mathbf{n}^T$ and the distance d of the plane it defines to the optical center of camera i_1. For all points that back-project to the same facet, we can write $\hat{p}_{i_2}^j = H_\kappa p_{i_1}^j$where

$$H_\kappa = K_{i_2}(R_{i_1 i_2} - \frac{\mathbf{t}_{i_1 i_2}\mathbf{n}^T}{d})K_{i_1}^{-1} \quad , \tag{5}$$

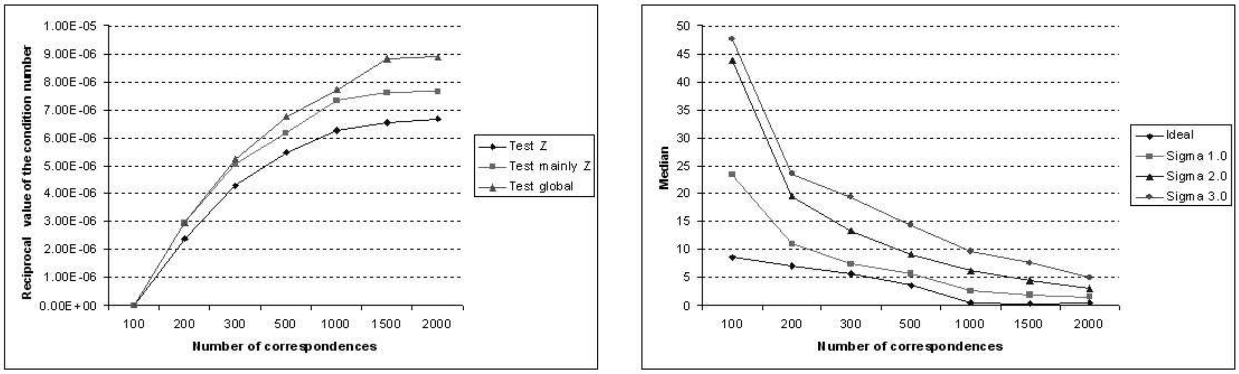
$R_{i_1 i_2}$and $\mathbf{t}_{i_1 i_2}$are the rotation and translation from one camera to the other, and K_{i_1} and $K_{i_2}$are the 3×3 matrices representing the cameras' intrinsic parameters.

If the shape parameters are fixed, the $R_{i_1 i_2}$ and $\mathbf{t}_{i_1 i_2}$ can be estimated uniquely given correspondences on at least 4 non-coplanar facets. In practice, we are dealing with thousands of facets that span all the orientations on a half sphere. We can therefore assume that reliable estimates of these rotations and translations are available. Since deformation from one head to another induces displacements that are small with respect to the distance to the cameras, their true values are relatively close and, therefore, the influence of the rotations can be linearized. The main potential source of non-linearities during minimization is the $\frac{\mathbf{n}^T}{d}$ term in Eq. 5. Given the specific parametrization of our model, it can be written as a fraction of two polynomial in the α_i shape parameters, which in general yields the kind of highly nonlinear behavior depicted by Fig. 3(a). However, explicitly computing a Taylor expansion of this nonlinear term shows that the linear coefficients are of larger magnitudes than the others. We suspect that this is a direct consequence of the fact that the normals cannot vary arbitrarily since the model must keep resembling a face for a wide range of values of the α shape parameters but have not been able to prove it formally yet. In practice, these parameters tend to remain in the range $-1, 1$, and as a result, we observe quasi-linear behaviors such as those shown in Fig. 3(b). As a result, at least near its minimum, the objective function F_N of Eq. 4 is close to being quadratic. It is therefore well suited for Levenberg-Marquardt style minimization we can analyze its optimization behavior of this procedure by simply looking at its Jacobian, which we do in the following section.



(a) (b) (c)

Figure 3: One coordinate of the transfer function for a particular facet as a function of the first two α shape parameters. (a) In the range $-10, 10$ it is non-linear. (b) However, in the range of "meaningful" shape parameters $-1, 1$ it is very close to being linear. (c) Similar linear behavior for another facet with a different orientation.



(a) (b)

Figure 4: (a) Number of correspondences vs. reciprocal condition number for three cases of head motion. (b) Number of correspondences vs. median error for the case of major rotation around the head's vertical axis and small rotations around the other axis for ideal case and for different amount of Gaussian noise ($sigma = 1, 2, 3$) added to the established correspondencies.

4.2 Monte-Carlo Validation

Let $S = [A, C_1, ..., C_N]$, where A is the vector of shape coefficients the C_i are the pose parameters of each camera, be the state vector with respect to which we optimize the function F_N of Eq. 4. For a given S, let $Obs(S)$ be the *observation vector* whose components are the individual $\left\|\triangle p_{i,i+1}^{j}\right\|$terms that appear in Eq. 4 and $J(S)$ be its Jacobian with respect to the state variables. Minimizing F_N using the Levenberg-Marquardt algorithm involves iteratively incrementing S by vectors proportional to dS such that

$$J^T(S)J(S)dS = -J(S)Obs(S)$$

Because F_N is close to being linear, the eigenvalues of J^TJ can be used to evaluate how well conditioned the system is. In particular, if they are strictly positive, the system has a single solution in the vicinity of the minimum we find. In that case robustness to noise should increase with the reciprocial value of the condition number, that it the square root of the ratio of the smallest to the largest eigenvalue. When computing in double-precision, it should be at least bigger then 10^{-12}.

To test this, we performed Monte-Carlo experiments using three different type of head motions and three views each time:

- Pure rotation around the head's vertical axis, without translation. (Test Z in Fig. 4(a))
- Rotation around the head's vertical axis with small rotations around the other axes and small translations. (Test main Z in Fig. 4(a))
- Significant rotations around all three axes and translations. (Test global in Fig. 4(a))

Fig. 4(a) depicts the average condition number over many trials as a function of the number of correspondences being used. The different curves correspond to the three different types of head motions discussed above. In all cases, when there are not enough correspondences, the system is under-constrained, which results in one or more eigenvalues being zero. However, as soon as the number of correspondences is large enough, the system becomes over-constrained and the condition numbers increase, which implies that the optimizer can be expected to converge towards a single minimum. As could be expected from self-calibration work [19, 20], pure rotation around a single axis yields the lowest condition numbers while rotations around all three axis yields the best.

Dominant rotation around the vertical axis yields intermediate numbers and is therefore the most difficult configuration one is likely to encounter in practice, as people will rotate their heads around their necks but would be hard pressed to completely avoid any rotation around the other two axes. It is therefore the configuration we use in the majority of examples shown in Section 5. Fig. 4(b) depicts the median error in that configuration when using synthetic correspondences that are corrupted by increasing amounts of noise and allowing *all* α shape parameters to vary. As expected, as the number of correspondences increases and with it the condition numbers, so does the system's accuracy. These graphs can be used to predict the required number of 2–D correspondences needed to achieve a certain level of precision given the accuracy of the algorithm used to establish them. This is important if one is interested in increasing speed by using as few of them as possible without compromising the quality of the results.

5 Results

In all cases shown here, we initialized the system by roughly specifying the 2–D projection of 5 points in *one* single image of each sequence to compute an approximate initial pose estimate [21]. These points, however, were never used again and that this initial estimate was refined at the same as the pose estimates in all the other images.

Figs. 1 and 2 depict difficult images acquired under challenging lighting conditions. Those of Fig. 1 are small and were digitized from an old celluloid version of the film while those in the first row of Fig. 2 were acquired with a strong directional source that creates cast shadows and specularities. To illustrate those difficulties, in Fig. 5, we show the output of Roy's [5] maximum-flow stereo algorithm that we ran on those images after having registered them using an earlier technique [10]. To increase its robustness, we modified the code that is available on the web by introducing normalized cross correlation into the objective function. The results are good for the images in the second row of Fig. 2 because the lighting

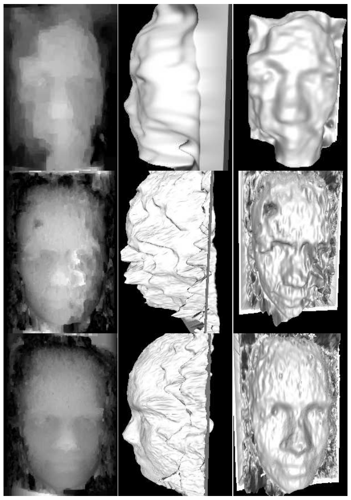


Figure 5: Results of graph cut algorithm. Top row: Disparity map and corresponding shaded surface for the images of Fig. 1. Middle and bottom row: Similar results for the images of Fig. 2.

is closer to being diffuse but much less so for the other two sequences. In other words, our method does a good job of using the model to pool noisy information. Note, in particular, that the shapes of the two recovered models of Fig. 2 are very similar even though they have been acquired under very different illumination conditions. In other words, our approach is relatively insensitive to such changes.

To quantify the accuracy of our method we used the two seven image video sequences depicted by Fig. 6 and the laser scans shown in Fig. 7(e). They were acquired using a Minolta^{tm} laser scanner, whose theoretical precision is approximately 0.3 millimeters but which evidently also produces some completely erroneous points. As discussed in Section 3, because we do not use the true intrinsic parameters but estimated ones, our reconstruction can only be expected to be correct up to an affine transform. To evaluate our results, we therefore compute the affine transform that best maps them onto the laser-scanner data. In Fig. 7(f), we plot the median distances between the affine-transformed model and the scanner data. As expected, they are inversely proportional to the number of images used and are in the range predicted by the synthetic experiments of Section 4 given that accuracy of the correspondences in the order of a pixel and the accuracy of the “ground truth” is somewhat less than 0.3 mm. As in the case of the images of Fig. 2, observed median distances of 0.55mm correspond to reprojection errors that are on average smaller than 0.5 pixel, which again indicates that information from the various images has been correctly merged to achieve an overall accuracy that is greater than that of the individual correspondences while rejecting erroneous ones. This also justifies our choice of not attempting to recover the implicit camera parameters: In theory it is possible, but, in practice, it would require establishing much more accurate 2–D correspondences and therefore using much higher resolution images than the ones shown here.

6 Conclusion

We have presented a model-based structure-from-motion approach to reconstructing faces from uncalibrated video sequences that is robust to uncontrolled scene illumination. Furthermore, it is amenable to full automation because it can be initialized by simply providing a rough initial pose estimate in one view, which many algorithms can now do, and unlike earlier approaches does not require the use of 3–D feature points whose projections are known.

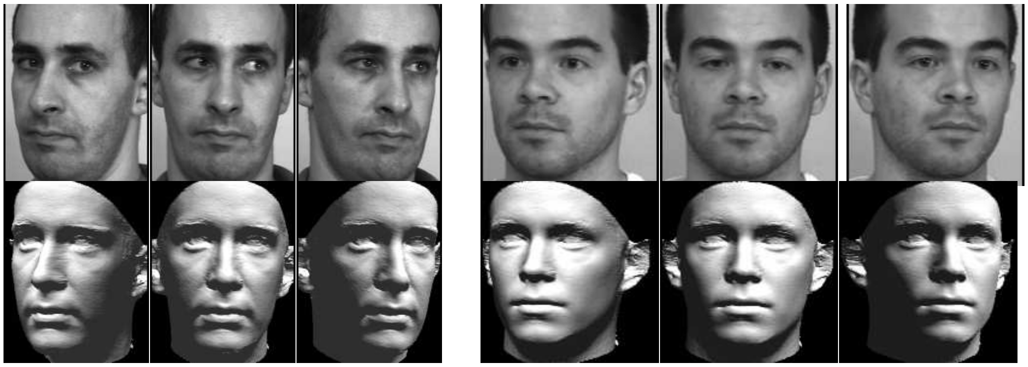


Figure 6: First Row: Three images of short sequences of two different people. Second row: Shaded views of the reconstructed models.

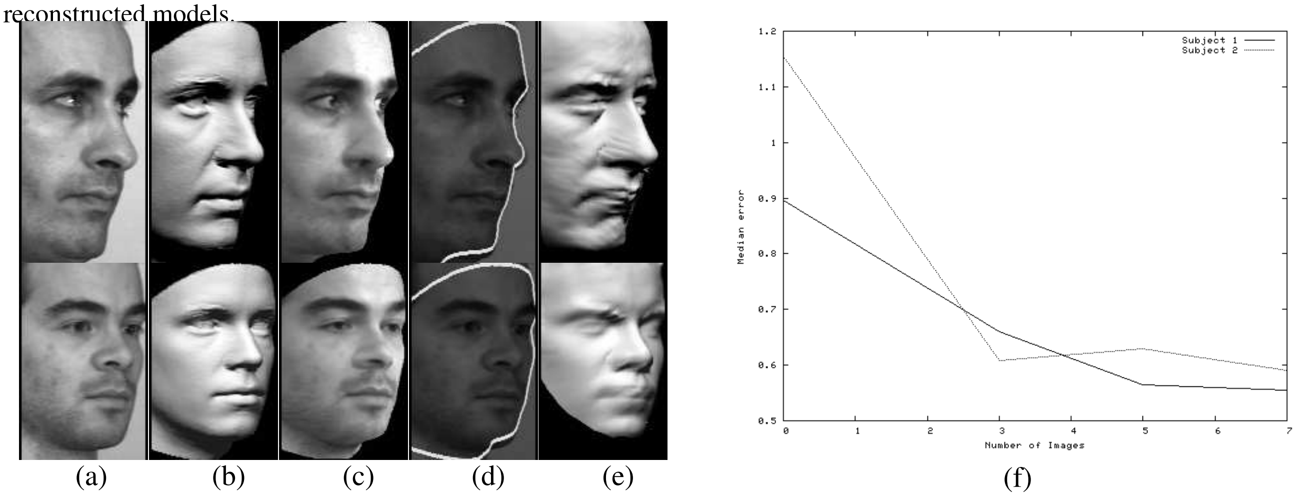


Figure 7: Comparison against profile views and laser scans. (a) Profile views of the two subjects of Fig 6 that have not been used for reconstruction purposes. (b,c) Shaded and texture-mapped views of the models in a similar pose. (d) Occluding contour of the model overlaid on the profile image. (e) Laser scans. (f) Median distance between the two reconstructed models and the two corresponding laser scans as a function of the number of images used to perform the reconstruction.

The main ingredient of this approach, a linear model of a deformable objects that can be used to pool information though a bundle-adjustment procedure, is in fact not specific to faces. It could be used for all manner of deformable objects for which a geometric models may be available but whose texture, unlike that of a face, may be arbitrary. This will require defining the kinds of shapes for which the transfer function has the quasi-linear behavior we have observed in the of faces and, in future work, we will focus formalizing this issue further with a view to developing fast algorithms that exploit this property.

References

[1] G. Borshukov, D. Piponi, O. Larsen, J.P. Lewis, and C. Tempelaar-Lietz, “The matrix revealed,” SIGGRAPH Sketch, 2003.

[2] V. Blanz and T. Vetter, “A Morphable Model for The Synthesis of 3–D Faces,” in *Computer Graphics, SIGGRAPH Proceedings*, Los Angeles, CA, August 1999.

[3] S. Romdhani, Volker Blanz, and Thomas Vetter, “Face identification by fitting a 3d morphable model using linear shape and texture error functions,” in *European Conference on Computer Vision*, Copenhagen, Denmark, 2002, vol. 4, pp. 3–19.

[4] V. Blanz and T. Vetter, “Face Recognition Based on Fitting a 3D Morphable Model,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, 2003.

[5] S. Roy and I.J. Cox, “A Maximum-Flow Formulation of the N-camera Stereo Correspondence Problem,” in *International Conference on Computer Vision*, Bombay, India, 1988, pp. 492–499.

[6] V. Kolmogorov and R. Zabih, “Multi-Camera Scene Reconstruction via Graph Cuts,” in *European Conference on Computer Vision*, Copenhagen, Denmark, May 2002.

[7] Y. Shan, Z. Liu, and Z. Zhang, “Model-Based Bundle Adjustment with Application to Face Modeling,” in *International Conference on Computer Vision*, Vancouver, Canada, July 2001.

[8] Z. Zhang, Z. Liu, D. Adler, M.F. Cohen, E. Hanson, and Y. Shan, “Robust and Rapid Generation of Animated Faces From Video Images: A Model-Based Modeling Approach.,” Technical Report MSR-TR-01-101, Microsoft Research, October 2001.

[9] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.

[10] P. Fua, “Regularized Bundle-Adjustment to Model Heads from Image Sequences without Calibration Data,” *International Journal of Computer Vision*, vol. 38, no. 2, pp. 153–171, July 2000.

[11] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D.H. Salesin, “Synthesizing Realistic Facial Expressions from Photographs,” in *Computer Graphics, SIGGRAPH Proceedings*, July 1998, vol. 26, pp. 75–84.

[12] F. Devernay and O. D. Faugeras, “Computing Differential Properties of 3–D Shapes from Stereoscopic Images without 3–D Models,” in *Conference on Computer Vision and Pattern Recognition*, Seattle, WA, June 1994, pp. 208–213.

[13] Y. G. Leclerc and A. F. Bobick, “The Direct Computation of Height from Shading,” in *Conference on Computer Vision and Pattern Recognition*, Lahaina, Maui, Hawaii, June 1991.

[14] D. Samaras and D. Metaxas, “Incorporating Illumination Constraints in Deformable Models,” in *Conference on Computer Vision and Pattern Recognition*, Santa Barbara, June 1998, pp. 322–329.

[15] M. Proesmans, L. Van Gool, and A. Oosterlinck, “Active acquisition of 3D shape for Moving Objects,” in *International Conference on Image Processing*, Lausanne, Switzerland, September 1996.

[16] L. Tang and T.S. Huang, “Analysis-based facial expression synthesis,” *International Conference on Image Processing*, vol. 94, pp. 98–102, 1996.

[17] D. DeCarlo and D. Metaxas, “Deformable Model-Based Shape and Motion Analysis from Images using Motion Residual Error,” in *International Conference on Computer Vision*, Bombay, India, 1998, pp. 113–119.

[18] S. B. Kang, “A Structure from Motion Approach using Constrained Deformable Models and Apperance Prediction,” Technical Report CRL 97/6, Digital, Cambridge Research Laboratory, October 1997.

[19] P. Sturm, “Critical Motion Sequences for Monocular Self-Calibration and Uncalibrated Euclidean Reconstruction,” in *Conference on Computer Vision and Pattern Recognition*, Puerto Rico, June 1997, pp. 1100–1105.

[20] A. Zisserman, D. Liebowitz, and M. Armstrong, “Resolving ambiguities in auto-calibration,” *Philosophical Transactions: Mathematical, Physical and Engineering Sciences (A)*, vol. 356, no. 1740, pp. 1193 – 1211, May 1998.

[21] D. DeMenthon and L. S. Davis, “Model-based object pose in 25 lines of code,” in *European Conference on Computer Vision*, 1992, pp. 335–343.