Reconstructing High Quality Face-Surfaces using Model Based Stereo

Brian Amberg
University of Basel
brian.amberg@uni-basel.ch

Andrew Blake
Microsoft Research
ablake@microsoft.com

Andrew Fitzgibbon
Microsoft Research
awf@microsoft.com

Sami Romdhani
University of Basel
sami.romdhani@uni-basel.ch

Thomas Vetter
University of Basel
thomas.vetter@uni-basel.ch

Abstract

We present a novel model based stereo system, which accurately extracts the 3D shape and pose of faces from multiple images taken simultaneously. Extracting the 3D shape from images is important in areas such as pose-invariant face recognition and image manipulation.

The method is based on a 3D morphable face model learned from a database of facial scans. The use of a strong face prior allows us to extract high precision surfaces from stereo data of faces, where traditional correlation based stereo methods fail because of the mostly textureless input images. The method uses two or more uncalibrated images of arbitrary baseline, estimating calibration and shape simultaneously. Results using two and three input images are presented.

We replace the lighting and albedo estimation of a monocular method with the use of stereo information, making the system more accurate and robust. We evaluate the method using ground truth data and the standard PIE image dataset. A comparision with the state of the art monocular system shows that the new method has a significantly higher accuracy.

1. Introduction

The accurate estimation of pose and shape of faces is an important precondition for image understanding and manipulation. Detailed 3D models [3] have proven to be among the most accurate methods for tasks as diverse as face recognition [1, 12] and image manipulation [2].

Faces present a difficult problem to low level stereo algorithms. A face is relatively textureless except from a few edges due to wrinkles, the mouth and the eyes. This makes the extraction of high quality surfaces with general correlation based methods virtually impossible. Compare e.g. the recent work of [11] for a state of the art example that still is far too smooth because of the general prior used.

As faces are one of the most important objects of daily life, it is justified to inject specific knowledge into the multiview vision task. This is accomplished here by using a shape prior learned from a database of 3D face surfaces.

Our approach is similar to the monocular methods of [2, 13] but uses stereo information instead of the colour prior of the monocular method. This improves accuracy and robustness, not only because more data is used, but also because no lighting and albedo parameters have to be estimated.

We first give an overview over prior approaches to model based surface extraction, detail the algorithm in Section 4 and compare our results to the state of the art monocular system in Section 5. We show that it is possible to achieve much higher accuracy using multiple images than with one image alone.

2. Prior Work

Deformable 2D face models were introduced by [9, 16] and extended to three dimensions by [3]. These models are generative and are fit to an image using nonlinear optimisation techniques and a suitable distance measure.

Fitting a model to multiple images when point to point correspondences are known has been addressed in [6] and for a different model in [14]. These approaches use relatively coarse manually defined deformable models, for which no prior distribution is known. Our method is different from these approaches, as we fit the model directly to the images instead of going through point to point correspondences. This removes the need to tune a patch-based correspondence detector and allows an accurate handling of occlusions and perspective deformations. As a result higher accuracy is achieved at the cost of a lower processing speed.

The joint estimation of calibration and shape using an active appearance model is addressed by [8]. They show that a calibration accuracy comparable to that of a calibration grid can be achieved. No evaluation of the reconstruction accuracy was given, but the low complexity model used can not represent detailed surfaces as recovered in this paper.

An important part of the algorithm presented here is the simultaneous fitting of the visible contour to multiple images. This is connected to the work of [4, 17] but differs in the use of a detailed shape prior instead of the more general assumption of a closed, continuous surface. Most previous work using silhouettes assumed that the occluding contour could be extracted accurately and completely, while our model based approach is robust enough to allow us to cope with noisily detected edges. [7] uses experiments on synthetic data to assess the accuracy of a surface reconstruction with a detailed shape model from a single contour image. No results on real world data were presented. We extend that work by showing that high quality surfaces can be extracted from real world images when combining a soft edge detection scheme with a derivative based optimisation algorithm.

[3, 18] already used multiple images to fit a detailed model. This was done by applying a moncular shape from shading algorithm to multiple images simultaneously. Our approach improves upon this work by eliminating the lighting and albedo estimation and substituting it with a colour difference cost. No absolute error estimates and only two examples are available for [3, 18], making a quantitative comparision impossible.

Similar to our work are the monocular methods of [3] and [13], which extract a face surfaces from a single image. They work by joint estimation of shape, albedo, pose and lighting. Multiple features are combined in a least squares sense by [13], including a landmark, silhouette, and shape from shading cost. We use this method as a baseline and show how the reconstruction accuracy can be significantlyy improved by eliminating the lighting and albedo estimation and including a stereo colour difference cost. In addition to being more accurate, this method is also more robust against difficult to model albedo effects such as make-up, facial hair, moles and cast shadows.

3. Contributions

We present a novel accurate stereo fitting system using a detailed generative 3D shape model. The accuracy of the method is evaluated on ground-truth data and compared against the state of the art monocular system [13]. Additional experiments on the standard PIE dataset [15] validate the pose invariance of the method on real world face images. We are able to fit a detailed model with a much higher precision than demonstrated in previous papers. This is achieved by combining monocular and stereo cues. The method works directly on the input images, removing the need to first detect correspondences between images and making the method suitable for arbitrarily large baselines. The use of ground-truth data allows the exact quantification of the contribution of each image cue. It is shown how the silhouette information from a few images facilitates the extraction of a surface which is already more accurate than the full monocular method. The additional use of colour differences between images lowers the remaining residual to between 1/2 and 2/3 the residual of the monocular method.

4. Method

Detailed prior knowledge learned from a database of 3D facial scans is used to extract a high quality surface from the input data. The database is used to learn a 3D Morphable Model (3DMM), which is a space of faces spanned by the linear combination of basis-faces. The basis-faces are brought into correspondence in an off-line processing step, and can then be combined to create new faces [3].

The basic paradigm is an analysis by synthesis framework. Hypotheses are generated, their probability given the observed data is evaluated and the parameter set maximising this probability is determined using a nonlinear optimisation procedure. The difficulty lies in measuring the distance from the observed image to the hypothesis. The following sections describe measures contributing to the cost function to be minimised. These terms are combined by computing their weighted sum. The weights depend on the data available and are a tunable setting. As the goal is to present a generally applicable method, all results in this paper were created using the same set of weights, which shows that the choice of the weights is not critical within a certain range.

4.1. The model

The unknowns that need to be determined are the shape of the face and the camera parameters. The cameras are modelled as pinhole cameras and are described by their external and internal parameters $\boldsymbol{\rho}^c = (f^c, \boldsymbol{p}^c, \boldsymbol{t}^c, \boldsymbol{r}^c)$ where f^c denotes the focal length, $\boldsymbol{p}^c$ is the principal point, $\boldsymbol{t}^c$ is the translation and $\boldsymbol{r}^c = (r^c_x, r^c_y, r^c_z)$ are the three rotation angles. The superscript c is the number of the camera. We denote the rotation matrix corresponding to the angles $\boldsymbol{r}^c$ by $\boldsymbol{R}(\boldsymbol{r}^c)$. The action of a camera onto a vertex $\boldsymbol{v}$ is

$$P_{\boldsymbol{\rho}}(\boldsymbol{v}) := \pi_{f,\boldsymbol{p}}(\boldsymbol{R}(\boldsymbol{r})\boldsymbol{v} + \boldsymbol{t}) \tag{1}$$

$$\pi_{f,\boldsymbol{p}}(\boldsymbol{w}) := \left[f\boldsymbol{w}_x/\boldsymbol{w}_z + \boldsymbol{p}_x, f\boldsymbol{w}_y/\boldsymbol{w}_z + \boldsymbol{p}_y\right]^T$$

An additional parameter vector $\boldsymbol{\alpha}$ determines the shape of the face.

A linear morphable model $\mathcal{M} = (\boldsymbol{\mu}, \boldsymbol{D})$ consists of a mean vector $\boldsymbol{\mu}$ and an offset matrix $\boldsymbol{D}$ and maps a parameter vector onto a vector of stacked vertex positions.

$$\mathcal{M}(\boldsymbol{\alpha}) := \boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{\alpha} \tag{2}$$

We denote the rows of the model corresponding to the i'th vertex as $\mathcal{M}^i$, $\boldsymbol{\mu}^i$, and $\boldsymbol{D}^i$ respectively.

While this paper describes the algorithm using the introduced parameters, other parametrisations are better suited for certain problems. An additional global rotation is used when the calibration of the cameras is already known to a relatively high precision. The camera positions are then fixed by a prior distribution on camera parameters, and a global rigid head movement is allowed.

4.2. Shape Prior

To avoid over-fitting and because the face shape is not fully constrained by the images a regularisation of the hypothesis space is needed. Assuming that faces are normally distributed in face space we rotate and scale the face space by performing a PCA such that each parameter has a Gaussian PDF with unit variance ($\mathrm{cov}(\boldsymbol{\alpha}) = \boldsymbol{I}$). After rotation the probability of observing a face with parameters $\boldsymbol{\alpha}$ is proportional to $\exp\{-\|\boldsymbol{\alpha}\|^2\}$. As we have formulated the problem as a minimisation, this contributes with the negative log to our cost function and yields the shape prior term

$$E_p(\boldsymbol{\alpha}) := \|\boldsymbol{\alpha}\|^2 \qquad , \tag{3}$$

which is added with a suitable weight constant to the objective function.

4.3. Landmarks

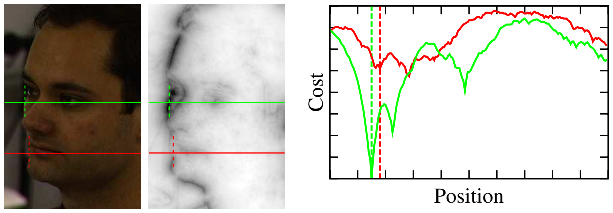
Landmarks are used to provide an initial estimate of the camera position and to constrain the results to a sensible solution. Without landmarks a perfect explanation of a set of images can be achieved by positioning the head such that it is projected onto a single pixel of the same colour in each image. A landmark is a tuple $(l_i, \boldsymbol{p}_i)$ consisting of a vertex index l_i in the mesh and a point $\boldsymbol{p}_i$ in the image plane. The landmark term

$$E_l^c(\boldsymbol{\rho}^c, \boldsymbol{\alpha}) := \sum_i \|P_{\boldsymbol{\rho}^c}(\mathcal{M}^{l_i}(\alpha)) - \boldsymbol{p}_i\|^2 \tag{4}$$

penalises the distance between the projected vertex and its expected position. One landmark term per camera is used.

4.4. Silhouette Term

An accurate, lighting invariant shape cue is the silhouette [4, 17]. The silhouette term measures how well the predicted visible contour matches the edges present in the image. While most work on silhouettes assumes that the contour can be extracted accurately and uniquely, we do not impose this constraint, but are instead able to handle insecurely detected silhouettes and false positives.



Input Image Cost Surface Cut through Cost Surface

Figure 1. **Integration of multiple edge thresholds results in a cost surface encoding not only the direction towards the nearest edge, but also the saliency of each edge**. This allows the algorithm to lock onto weak edges, if they are supported by the whole image. The green line shows a situation where the silhouette edge is the strongest edge, which results in a global minimum at this position. Without careful tuning to the image at hand it is impossible to separate the silhouette at the red line from the texture edges, but the integrated cost surface still shows a local minimum at the right position without being tuned to this specific image.

To evaluate the silhouette cost the points on the visible contour of the current hypothesis and a cost surface defined over the image space of each camera is needed. Assuming that the visible contour points are given as a 3D Morphable Model $\mathcal{M}_s$ and the cost surface is defined as a function $S : \mathbb{R}^2 \mapsto \mathbb{R}$ we define the silhouette term for each camera independently as

$$E_s(\boldsymbol{\rho}^c, \boldsymbol{\alpha}) := \sum_i S(P_{\boldsymbol{\rho}^c}(\mathcal{M}_s^i(\boldsymbol{\alpha})))^2 \tag{5}$$

Determining the silhouette cost surface. No general edge detection method will be able to perfectly find the silhouette in real world images without tuning the settings of the edge detection algorithm to the images at hand. To overcome this problem we propose an image transform that results in a suitable cost surface without committing to a single setting. Similar to [5] we integrate the information over a range of edge thresholds into a smooth cost surface, which has the desirable property of having accentuated minima at edges over the full range of thresholds, while still discerning between strong and weak edges. This allows the optimisation to fix even to weak edges when they are supported by the whole image, while retaining the capability to skip over towards stronger edges when the weak edges are not consistent with the current hypothesis.

To construct the cost surface the following steps are performed

1. Using an edge detector with a sensitivity threshold a a series of binary edge images $E^{a_1}, \ldots, E^{a_n}$ is created.
2. The distance transforms $D^{a_1}, \ldots, D^{a_n}$ of the images are calculated.
3. Using a smoothing constant κ the cost surface is calculated as $S := 1/n \sum_{i=1}^{n} D^{a_i}/(D^{a_i} + \kappa)$. In the experiments a κ of 10 pixels was used. A suitable value for κ is 1/20th of the expected size of the projected head in pixels. κ determines the influence range of an edge in an adaptive manner.

The edge detector used in this paper is gradient magnitude

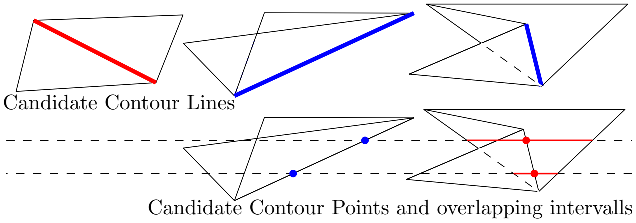


Figure 2. **Efficent determination of an equally spaced set of contour points.** Candidate contour lines (blue) are distinguised from interiour lines (red) by testing if both adjacent triangles project onto the same side of the line. Equally spaced candidate points are created by intersecting a regular grid with the candidate lines. Interior points (red) overlapped by a triangle are then removed, leaving a set of contour points (blue).

thresholding with non-maximum suppression. An example of the resulting cost surface is shown in Figure 1.

Determining the visible contour A set of approximately equally spaced points that project to the visible contour are detected efficiently in a three stage process:

1. First a set of triangle edges that are potentially on the visible contour are selected by testing if the projections of the third point of the two adjacent triangles lie on the same side of the projected line.
2. Next candidate points are created by intersecting equally spaced horizontal and vertical lines with the candidate contour lines. The intersection points are saved per grid-line.
3. Another iteration over the triangles is used to remove all points overlapped by a projected triangle.

Note that this method allows us to find the visible contour accurately working only with the 2D projected topology and without doing a full rasterisation or depths tests. The detected contour points are used to generate a morphable contour model from the full shape model. The resolution of the contour model depends on the density of the intersection grid and not on the resolution of the shape model.

The selection of contour points is kept fixed for some iterations of the optimisation, even though a rotation makes a different set of points to contour points. The difference of the projected point positions when the contour is sliding over the model is small, so this is a justifiable approximation. After some steps of the optimisation the contour model is reinitialised and the minimisation continues.

4.5. Colour Difference Cost

The colour difference term measures the sum of squared differences between colours at corresponding points in two images, where the correspondence is determined by the current shape and camera hypothesis. Given a morphable model $\mathcal{M}_d$ of points on the face surface that are visible in the cameras c and c' the colour cost between two images is

$$E_d(\boldsymbol{\rho}^c, \boldsymbol{\rho}^{c'}, \boldsymbol{\alpha}) := \sum_i \| I^c(P_{\boldsymbol{\rho}^c}(\mathcal{M}_d^i(\alpha))) - I^{c'}(P_{\boldsymbol{\rho}^{c'}}(\mathcal{M}_d^i(\alpha))) \|^2 \quad . \tag{6}$$

When more than two images are to be compared, multiple colour difference terms are used. Which input images are paired depends on the applications. Whenever two images have a large overlap, the colour difference term can be used.

The set of visible points is kept constant for some iterations of the optimisation, even though visibility might change. After the large initial rotation directed by the landmarks not much visibility change is happening and this approximation is save.

Determining the sample points The sample points are chosen to be distributed over the 3D model such that their projections hit approximately each pixel once. This accounts for projected triangle size and distributes the processing load evenly over the image.

Given a hypothesis about cameras c and c' and the current shape we determine a set of sampling points using a procedure that assures that every pixel in the first and in the second camera is assigned at most one sample point, and the maximum number of sample points is created.

1. Create a depth-buffer rendering of the model seen from cameras c and c' and save for each pixel the surface point rendered into the pixel.
2. Initialize an auxiliary depth-buffer with the resolution of camera c'.
3. For each surface point seen in camera c:
   (a) Use a depth comparison to reject points not visible in camera c'.
   (b) Reject points whose depth from camera c' is larger than the value in the auxiliary depth-buffer.
   (c) Overwrite the auxiliary depth-buffer with the depth as seen from camera c' and the index of the pixel in camera c.
4. Use the surface points that made it into the auxiliary depth buffer as sampling points.

4.6. Objective Function

The complete objective function for the stereo system combines these terms into a weighted sum, where the weights w_p, w_l, w_s, w_d define the relative importance of each term.

$$E(\boldsymbol{\rho}^{c_1}, \ldots, \boldsymbol{\rho}^{c_n}, \boldsymbol{\alpha}) = w_p E_p(\boldsymbol{\alpha}) + w_l \sum_i E_l^{c_i}(\boldsymbol{\rho}^{c_i}, \boldsymbol{\alpha}) + w_s \sum_i E_s^{c_i}(\boldsymbol{\rho}^{c_i}, \boldsymbol{\alpha}) + w_d \sum_{i,j} E_d^{c_i,c_j}(\boldsymbol{\rho}^{c_i}, \boldsymbol{\rho}^{c_j}, \boldsymbol{\alpha}) \tag{7}$$

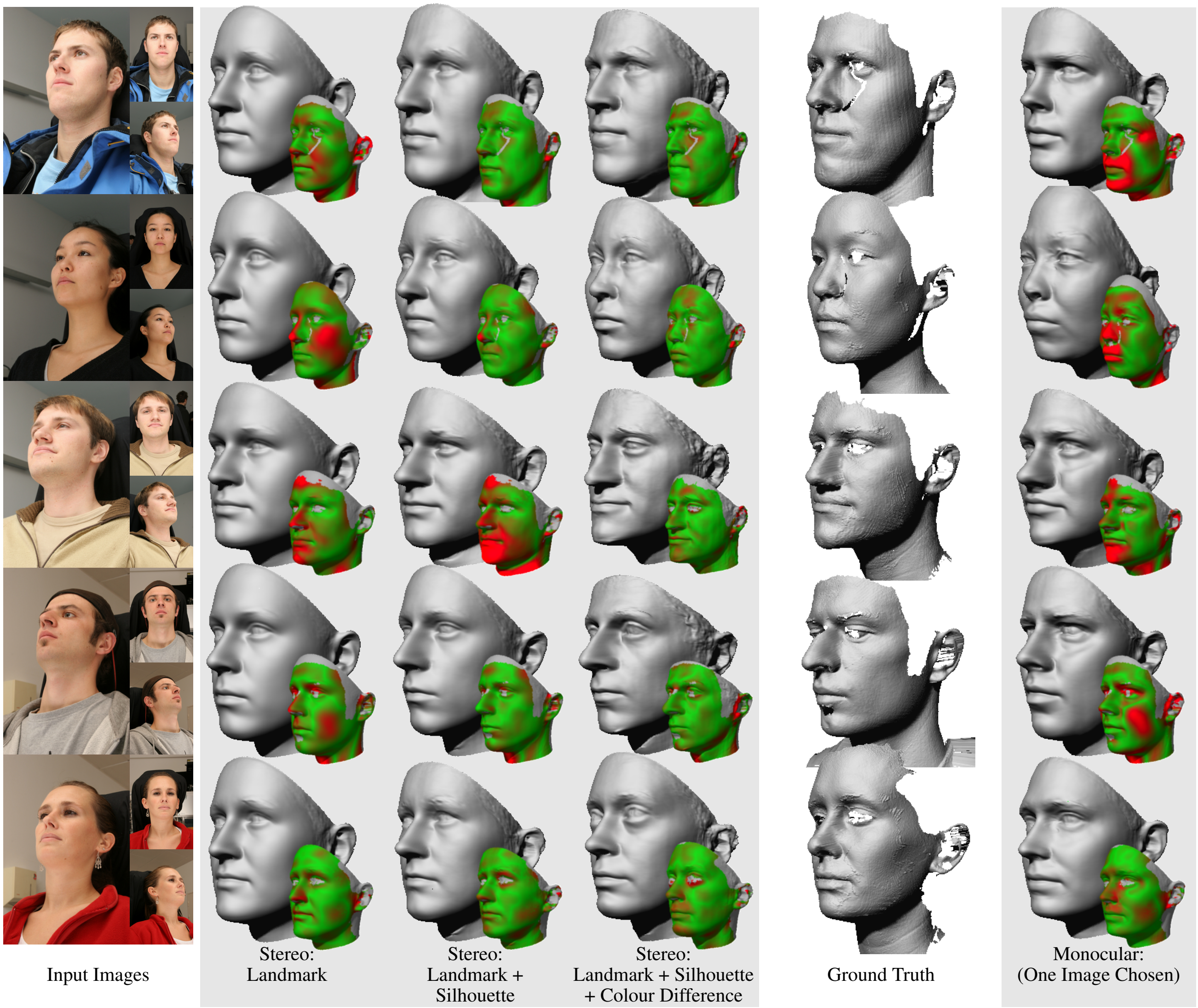

Input Images
Stereo: Landmark
Stereo: Landmark + Silhouette
Stereo: Landmark + Silhouette + Colour Difference
Ground Truth
Monocular: (One Image Chosen)

Figure 3. **Each cue increases the reconstruction accuracy, leading to significantly better result than possible with the monocular system**. Reconstructions of the face surface from three input images are compared to ground truth data acquired with a structured light system. The results are a representative sample showing good fits in the first three rows and a bad fit in the last row. Columns two to four display results for ever more terms in the stereo algorithm, showing the significant contribution of the silhouette and the colour difference term. The results obtained with the state of the art monocular system are shown in the last column. A measure of reconstruction quality is the point wise distance between the reconstructed surface and its closest point on the ground truth. The residual is shown in the inset head renderings. Uncoloured regions have no ground truth data, green is a perfect match, and red denotes a distance of 3mm or more. Originally all input images have the same resolution, they are presented at different sizes only due to space reasons.

4.7. Optimiser

A Levenberg-Marquard (LM) optimiser [10] specialised on nonlinear least squares and the quasi-newton method L-BFGS-B [19] were evaluated. The optimisation routines were provided with fully analytic derivatives. The L-BFGS-B optimiser turned out to be three times faster on our problem, but needs a stronger regularisation to remain stable. The results in this paper are therefore calculated with the LM method. The reason that L-BFGS-B is faster, even though it needs a larger number of function and Jacobian evaluations than LM, is that the runtime of the LM method is dominated by the QR-decomposition of the approximately 250000×150 element Jacobian.

Scaling The parameter space is scaled automatically, to make the optimisation numerically stable. Scaling factors for all parameters are determined on each restart of the optimisation such that a change of one scaled unit corresponds to an average change of one pixel of the projected vertex positions in all images.

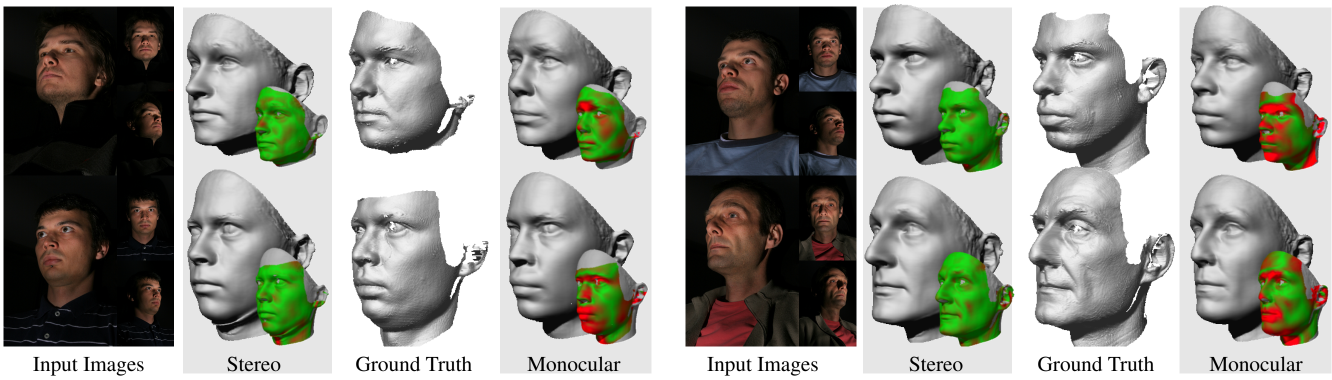


Figure 4. **The new stereo algorithm is robust under directed lighting and yields significantly more accurate surface reconstructions than the monocular algorithm.** The monocular results were created by fitting to the frontal view, the stereo results use three images. The distribution of the distance between the recovered surface and the ground truth is shown in the smaller inset head renderings. Uncoloured regions have no ground truth data, green is a perfect match and red is distance of 3mm or more.

4.8. Loose Ends

To actually make it work some additional details had to be taken care of. When the focal length is not restricted by an additional term, it is estimated very far off for some heads, degrading the reconstruction accuracy. To overcome this a term restricting the focal length to be between one and five times the size of the sensor was added. To make the method faster a multi-stage fitting similar to that of [13] using a pyramid of sampling resolutions and shape parameters is used. Specular highlights adversely influence the performance of the method. At the moment this is overcome by ignoring saturated pixels, but that should be changed to include the rich information available in specular highlights.

5. Evaluation

We evaluate the method using three datasets. The first dataset containts images of 20 subjects from three views with ambient only lighting. This should be a difficult dataset for monocular fitting, as monocular shape estimation depends on shape from shading. The second dataset containts five subjects and two lighting conditions using single directed light sources. For these first two datasets the shape was acquired simultaneously by a structured light scanner, giving us ground truth data for the experimental validation. These datasets are used to determine the influence of each term of the stereo reconstruction method and to compare against the monocular method [13]. We show that the reconstruction accuracy can be greatly improved under all three lighting conditions. The third dataset is the neutral expression part of the CMU PIE image dataset [15]. No ground truth is available for this dataset, but an evaluation of consistency of the results shows that the stereo method is robust against large pose variations.

The Model was learned with a modified optical flow algorithm from a dataset of 100m+100f cyberware scans.

5.1. Ground Truth

The ground truth datasets are used to quantitatively measure the improvement in reconstruction accuracy that can be achieved by using model based stereo instead of a monocular reconstruction. The model did not contain the test examples. Using the monocular system from [13] and the stereo system introduced in this paper we reconstructed the shape from the images in the ground-truth dataset and aligned it rigidly with the scanned surface. The distance the vertices of the reconstruction and their closest points on the scanned surface was calculated. Reconstruction results and the distribution of the residual over the surface is shown for a representative set of examples in Figures 3 and 4. Figure 3 demonstrates that the addition of each term to the cost function results in a significant increase in reconstruction accuracy, resulting in much better surfaces than those recovered by the monocular system. Similar results were obtained for the directed lighting dataset, depicted in Figure 4. Figure 5 sums up the results over all subjects in the dataset, showing that the distance between the reconstructed surface and the ground truth is a lot smaller for the stereo method than for the monocular method. Already fitting the silhouette of three images results in a lower residual than using the full monocular system on a single image.

It is interesting to note that while a smaller residual indicates a better fit, there is an infinite number of surfaces with the same residual but with very different perceptual quality. For many applications it is more important to accurately reconstruct the shape of the mouth, nose, and eyes than to accurately match the size of the cheeks. So when evaluating the reconstruction quality it is important to take a look at the reconstructed shapes. To allow this comparision the ground truth is included in Figure 3 and Figure 4 showing that our method is indeed far more accurate than [13], which is perceptually somewhere between multiview silhouette reconstruction and multiview stereo reconstruction.

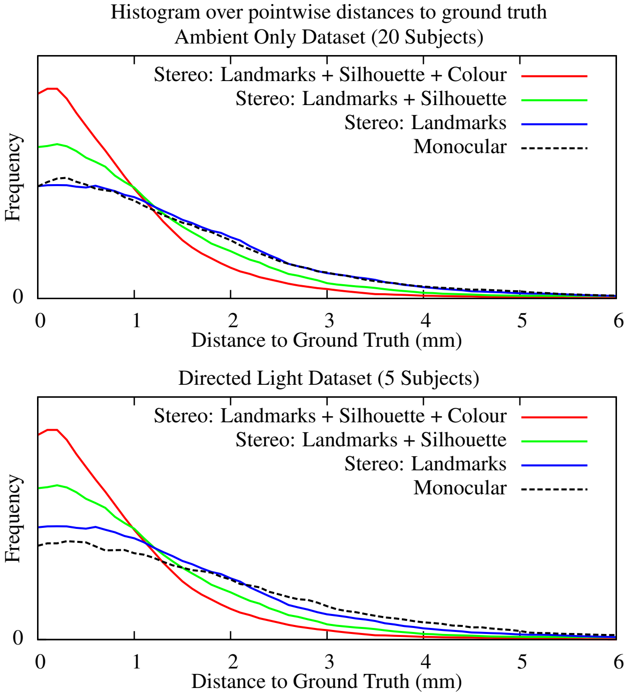


Figure 5. **The use of multi-view information results in a much higher accuracy than achievable by the monocular method.** The histogram over the residual over all subjects shows that already the multi-view silhouette fitting reduces the residual towards the ground truth to a value lower than that of the monocular method. A higher frequency of lower residuals is better.

5.2. Face Recognition

Reconstructing the shape of a face seen under varying pose should always result in the same surface. We use this to test our method on the neutral expression part of the standard PIE dataset [15]. 912 neutral expression images of 68 subjects were manually marked with five landmarks per image to initialise the optimisation. The dataset was split into a gallery and three probe datasets with two to four images. The cameras chosen for the gallery set have the PIE numbers 22, 25, 29, the probe image camera sets are detailed in Figure 7. To quantify the accuracy of the reconstruction we compute the similarity between the shapes, and report how many probe images yield a shape which is closest to the gallery shape. The distance between two shape vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ was defined as in [12] to be

$$d(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \frac{\boldsymbol{\alpha}_1 \cdot \boldsymbol{\alpha}_2}{\|\boldsymbol{\alpha}_1\|_2 \|\boldsymbol{\alpha}_2\|_2} \tag{8}$$

The recognition results tabulated in Figure 7 and exemplified in Figure 8. They show that the system is robust against large variations in pose and can handle narrow and wide baseline setups with a variable number of images. Most of the wrong classifications are caused by occlusion of the face or ears due to hair, which is not yet handled in our system.

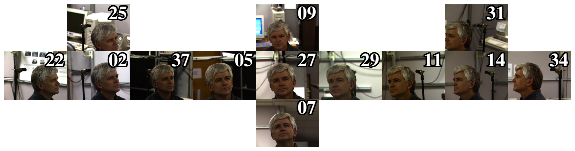


Figure 6. Camera Positions of the PIE dataset

Probe Images	Landmarks		+ Silhouette		+ Colour	
	1st	2nd	1st	2nd	1st	2nd
37 05	10%	18%	50%	68%	63%	82%
37 05 09	7%	18%	62%	74%	74%	85%
37 05 09 14	19%	37%	76%	82%	87%	94%

Figure 7. **The model based stereo system is robust against pose variations.** The system was able to extract similar surfaces for different views of the same subject. The shape based recognition rate increases when more views are used. The columns labelled “1st” show the frequency of correct results, “2nd” is the frequency with which the correct result was within the first two subjects returned.

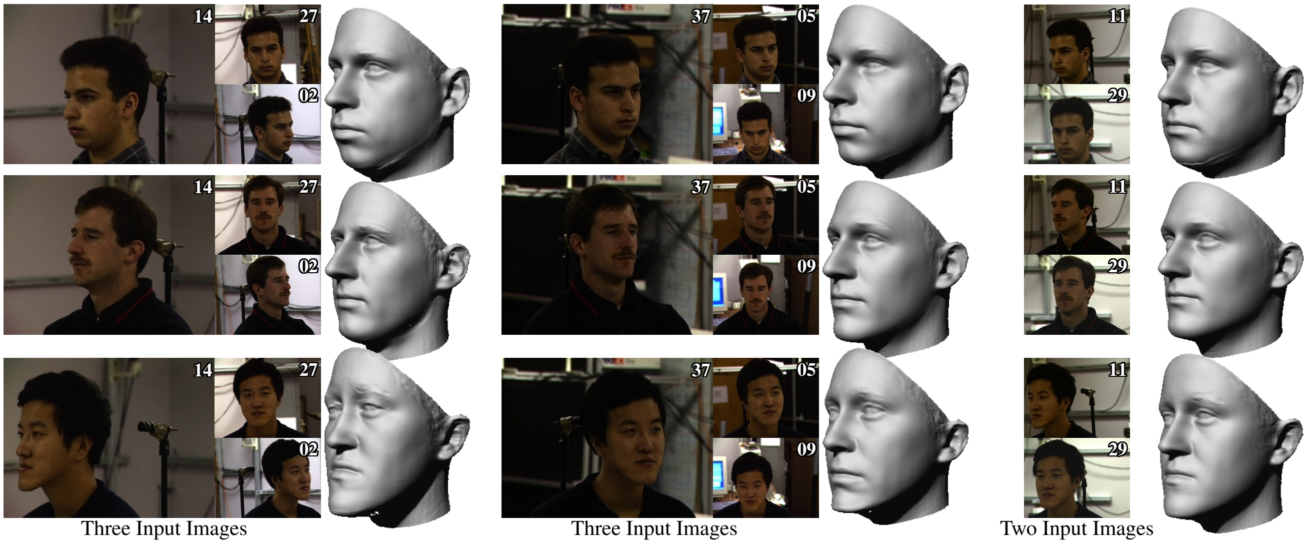
This result should be seen mainly as a measure of the viewpoint invariance of our method, and not as a proposed face recognition algorithm. All images of a subject were taken simultaneously, which would not be the case in a real face-recognition experiment. On the other hand only the shape and not the albedo of the faces was used for recognition which severely limits the discriminative power of the algorithm. Higher recognition rates (99.9% to 75.6%) have been reported [12] for the full PIE dataset for a system incorporating the monocular method of [13].

6. Conclusion and Future Work

When multiple images are available we can replace the estimation of lighting and albedo of the monocular system with stereo information. This results in a significant improvement in reconstruction accuracy. This was demonstrated using ground truth data and the standard PIE image dataset. The method fits the morphable model directly to the images, removing the problems associated with uniqueness, occlusion, and the irregular deformation of patches between images encountered in traditional correlation based stereo and bundle adjustement methods. These advantages makes the method applicable to images of any baseline.

An interesting complement to the presented algorithm would be a method to evolve the recovered surface further, such that surfaces which are not exactly within the span of the model could be recovered.

Using the accurate estimation of surfaces that has been demonstrated, it may be possible to learn a detailed temporal expression model from multi-view video sequences, which could then be used to track, predict, and manipulate expressions in video streams.



Three Input Images
Wide Baseline

Three Input Images
Narrow Baseline

Two Input Images
Narrow Baseline

Figure 8. **The stereo system is robust against large variations in pose and baseline.** Each row shows reconstruction results of the same subjects with different set of cameras. The first two rows show successfull reconstructions while the last row illustrates an example where the shape was not extracted reliably. The input images are annotated with their PIE camera number and are all used at the same resolution.

References

[1] V. Blanz, P. Grother, J. P. Phillips, and T. Vetter. Face recognition based on frontal views generated from non-frontal images. In *CVPR 2005: Proc. - V. 2*, pages 454–461, Washington, DC, USA, 2005. IEEE Computer Society.

[2] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel. Exchanging faces in images. In *Proceedings of EG 2004*, volume 23, pages 669–676, September 2004.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99: Proceedings*, pages 187–194, New York, NY, USA, 1999. ACM Press.

[4] C. H. Esteban and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *Comput. Vis. Image Underst.*, 96(3):367–392, December 2004.

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. Technical report, Cornell Computing and Information Science, September 2004.

[6] P. Fua and C. Miccio. From regular images to animated heads: A least squares approach. In *ECCV '98*, volume 1, pages 188–202, London, UK, 1998. Springer-Verlag.

[7] M. Keller, R. Knothe, and T. Vetter. 3d reconstruction of human faces from occluding contours. In *Proceedings of the Mirage*, INRIA Rocquencourt, France, 2007,. Springer.

[8] S. Koterba, S. Baker, I. Matthews, C. Hu, J. Xiao, J. Cohn, and T. Kanade. Multi-view aam fitting and camera calibration. In *ICCV '05: Proceedings*, volume 1, pages 511–518, Washington, DC, USA, 2005. IEEE Computer Society.

[9] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[10] J. J. Moré, B. S. Garbow, and K. E. Hillstrom. User guide for MINPACK-1. Technical Report ANL-80-74, Argonne National Laboratory, Argonne, IL, USA, Aug. 1980.

[11] D. Murray and J. J. Little. Patchlets: Representing stereo vision data with surface elements. In *WACV-MOTION '05*, volume 1, pages 192–199, Washington, DC, USA, 2005. IEEE Computer Society.

[12] S. Romdhani, J. Ho, T. Vetter, and D. J. Kriegman. Face recognition using 3-d models: Pose and illumination. *Proceedings of the IEEE*, 94(11):1977–1999, 2006.

[13] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR 2005*, volume 2, pages 986–993, 2005.

[14] Y. Shan, Z. Liu, and Z. Zhang. Model-based bundle adjustment with application to face modeling. In *ICCV 2001*, volume 2, pages 644–651, 2001.

[15] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002*, pages 46–51, 2002.

[16] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):733–742, July 1997.

[17] G. Vogiatzis, C. Hernandez, and R. Cipolla. Reconstruction in the round using photometric normals and silhouettes. In *CVPR 2006*, pages 1847–1854, Washington, DC, USA, 2006. IEEE Computer Society.

[18] C. Wallraven, V. Blanz, and T. Vetter. 3d-reconstruction of faces: Combining stereo with class-based knowledge. In *Mustererkennung 1999, 21. DAGM-Symposium*, pages 405–412, London, UK, 1999. Springer-Verlag.

[19] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B. *ACM Trans. On Mathematical Software*, 23(4):550–560, December 1997.