Expression Invariant 3D Face Recognition with a Morphable Model

Brian Amberg
brian.amberg@unibas.ch

Reinhard Knothe
reinhard.knothe@unibas.ch

Thomas Vetter
thomas.vetter@unibas.ch

Abstract

We describe an expression-invariant method for face recognition by fitting an identity/expression separated 3D Morphable Model to shape data. The expression model greatly improves recognition and retrieval rates in the uncooperative setting, while achieving recognition rates on par with the best recognition algorithms in the face recognition great vendor test. The fitting is performed with a robust nonrigid ICP algorithm. It is able to perform face recognition in a fully automated scenario and on noisy data. The system was evaluated on two datasets, one with a high noise level and strong expressions, and the standard UND range scan database, showing that while expression invariance increases recognition and retrieval performance for the expression dataset, it does not decrease performance on the neutral dataset. The high recognition rates are achieved even with a purely shape based method, without taking image data into account.

1. Introduction

We present a system which is using shape information from a 3D scanner to perform automated face recognition. The main novelty of the system is its invariance to expressions. The system is tested on two public datasets. It is fully automatic and can handle the typical artifacts of 3D scanners, namely outliers and missing regions. Face recognition in this setting is a difficult task, and difficult tasks benefit from strong prior knowledge. To introduce the prior knowledge we use a 3D Morphable Model (3DMM) [5], which is a generative statistical model of 3D faces. 3DMMs have been applied successfully for face recognition on different modalities. The most challenging setting is recognition from single images under varying light and illumination. This was adressed by [6, 12]. There, a 3DMM with shape, texture and illumination model was fit to probe and gallery images. As the model separates shape and albedo parameters from pose and lighting, it enables pose and lighting-invariant recognition. We use the same idea for expression-invariant face recognition from 3D shape. We fit an identity/expression separating 3DMM [3] to shape data and normalize the resulting face by removing the pose and expression components. See Figure 1 for an example of expression normalization. The expression and pose normalized data allows then efficient and effective recognition. A 3D MM has been fitted to range data before [4, 14] and the results were even evaluated on part of the UND database. Our approach differs from this work in the fitting method employed, which is independent of the acquisition device, and in the use of an expression model to improve face recognition. Additionally, our method is fully automatic, needing only a single easy to detect directed landmark, while [4] needed seven manually selected landmarks.

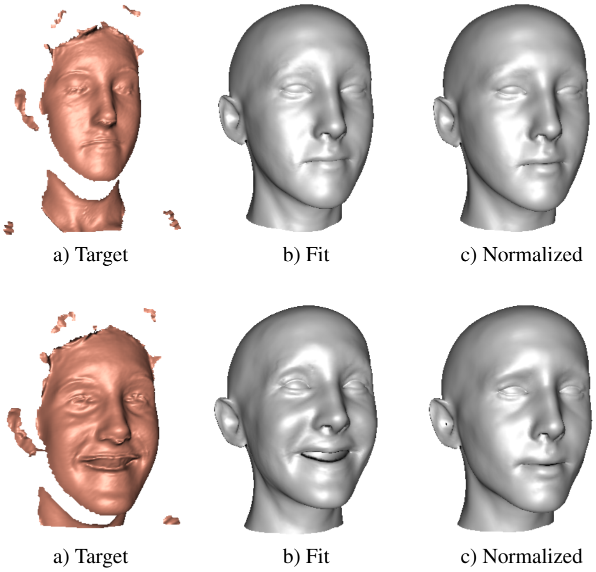


Figure 1. Expression normalisation for two scans of the same individual. The robust fitting gives a good estimate (b) of the true face surface given the noisy measurement (a). It fills in holes and removes artifacts using prior knowledge from the face model. The pose and expression normalized faces (c) are used for face recognition.

Expression-invariant recognition for shape data was also approached in [9], where a person specific 3D Morphable Expression Model was learned for each subject in the gallery. In contrast, we are using a general 3DMM learned from an independent database of face shapes which can be

978-1-4244-2154-1/08/$25.00 ©2008 IEEE

applied without any relearning to a new scan. This makes the enrollment phase trivial and the recognition phase effectively constant in the size of the gallery while still being accurate. We have to fit just one model to the probe, which can then be compared efficiently to the enrolled subjects, by comparing their coefficients in the low dimensional face space. While the number of comparisions is still at most linear in the number of examples (and can be made sublinear with an indexing method) the time it takes to compare coefficients in face space is neglectible compared to fitting time. Model-less approaches which align the probe to each example in the database using e.g. ICP [15] suffer from the same problem as [9]. Because the probe has to be aligned with each gallery scan these methods scale linearly in the gallery size, while our model based approach needs only a single fit to the probe.

Another interesting model-less approach [7] compares surfaces by the distribution of geodesics, which stays constant for nonrigidly deforming (but not stretching or tearing) objects. This approach is difficult to apply in this setting though, as the scanning produces holes, disconnected regions and strong noise, which can best be handled by a method which uses specific information about the object class.

2. Model

A PCA model [5] built from 270 subjects was used. It was build from one neutral expression face scan per identity and 135 expression scans of a subset of the subjects. The data was registered with a modification of [2]. The identity model consists of a mean shape $\boldsymbol{\mu}$ and a matrix of offset vectors $\mathbf{M}_n$ such that a new face instance $\boldsymbol{f}$ is generated from a vector of coefficients $\boldsymbol{\alpha}_n$ as

$$\boldsymbol{f}(\boldsymbol{\alpha}_n) = \boldsymbol{\mu} + \mathbf{M}_n\boldsymbol{\alpha}_n \qquad . \tag{1}$$

The model is constructed such that the α_i are independently normally distributed with zero mean and unit variance under the standard assumption of a Gaussian distribution of the data. This was done by performing PCA on the data matrix built from the mean free shape vectors. Additionally, for each of the 50 expression scans, we calculated an expression vector as the difference between the expression scan and the corresponding neutral scan of that subject. This data is already mode-centered, if we regard the neutral expression as the natural mode of expression data. On these offset vectors again PCA was applied to get an expression matrix $\mathbf{M}_e$ and expression coefficients $\boldsymbol{\alpha}_e$, such that the complete expression model is

$$\boldsymbol{f}(\boldsymbol{\alpha}_n, \boldsymbol{\alpha}_e) = \boldsymbol{\mu} + \mathbf{M}_n\boldsymbol{\alpha}_n + \mathbf{M}_e\boldsymbol{\alpha}_e = \boldsymbol{\mu} + \mathbf{M}\boldsymbol{\alpha} \quad , \tag{2}$$

$$\mathbf{M} = \begin{bmatrix}\mathbf{M}_n & | & \mathbf{M}_e\end{bmatrix} \qquad \boldsymbol{\alpha} = \begin{bmatrix}\boldsymbol{\alpha}_n \\ \boldsymbol{\alpha}_e\end{bmatrix} \quad . \tag{3}$$

This model assumes, that it is possible to transfer the expression deformation from one face to another. Even if this should not be strictly true, which is what authors using e.g. tensor models assume, it is a good enough assumption to make the method *invariant* to expressions. And a big advantage of this independence assumption is that we can train on far less data than in a tensor framework, because we do not need the full cartesian product of expressions and identities. In fact we have not even had any expression scan available for most of the training subjects, but were able to learn useful statistics from the available scans.

The basic assumption of this paper is, that the face and expression space are linearly independent, such that each face is represented by a unique set of coefficients. Independence of the identiy and expression spaces is not enforced in this work, but seems to be inherent in face space. We have observerd, that the identity space contains a bit of expressions, mainly smiles, which is due to the difficulty of acquiring perfectly neutral expressions.

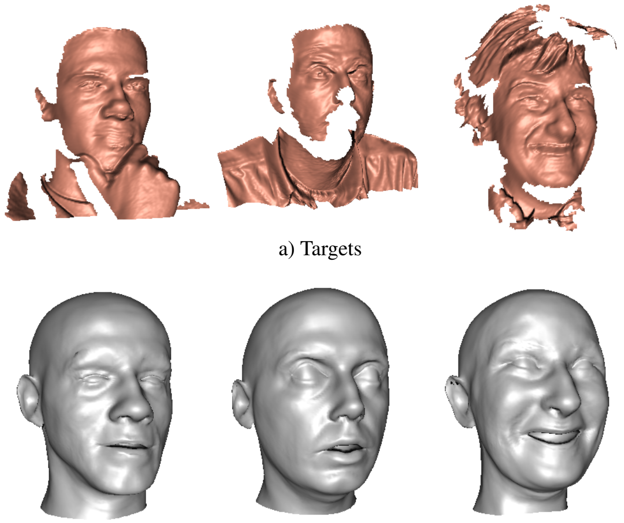
We use the registered scans and a mirrored version of each registered scan to increase the variability of the model. This allows us to calculate a model with more than 175 neutral coefficients.

3. Fitting

The fitting algorithm used in this paper is a variant of the nonrigid ICP work in [2]. The main difference, is that the deformation model is a statistical model and the optimisation in each step is an iterative method, which finds the minimum of a convex function. Additionally, as it is applied on noisy data (see Figure 2), we included a more elaborate robust weighting term. Like other ICP methods, it is a local optimization method, which does not guarantee convergence to the global mimimum, but is dependent on the initialization. It consists of the following steps

- Iterate over regularization values $\theta_1 > \cdots > \theta_N$:
  - Repeat until convergence:
    1. Find candidate correspondences by searching for the closest compatible point for each model vertex.
    2. Weight the correspondences by their distance using a robust estimator.
    3. Fit the 3DMM to these correspondences using a regularization strength of θ_i.
    4. Continue with the lower θ_{i+1} if the median change in vertex position is smaller than a threshold.

The search for the closest compatible point takes only points into account which have conforming normals, are closer than a threshold, and are not on or close to the border of the



a) Targets

b) Fits

Figure 2. The reconstruction (b) is robust against scans (a) with artifacts, noise, and holes.

scan. This has the effect of removing many outliers. The search is sped up by organizing the target scan in a space partitioning tree made up of spheres. The correspondences are then weighted with a robust function by their residual distance. The robust function is linear for distances smaller than 2mm, behaves like $1/x$ between 2mm and 20mm, and is zero for a distance larger than 20mm. Note, that it is necessary to balance robustness and regularization, as the right balance depends on the noise characteristic of the data. Suitable values were determined manually from a few scans of the GavabDB database and kept constant for all experiments as well on the GavaDB as on the UND database.

In step 3 the 3DMM is fit to 3D-3D point correspondences. This is done with a gauss-newton least squares optimization, using an analytic Jacobian and Gauss-Newton Hessian approximation. Denote the correspondence points by $\mathbf{u} = \left[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\right]$ and the rows of the model which correspond to the ith vertex by subscript i, then we can write the cost function mimized in this step as

$$f(\mathbf{R}, \boldsymbol{t}, \boldsymbol{\alpha}) = \sum_i \left\|\mathbf{R}(\boldsymbol{\mu}_i + \mathbf{M}_i\boldsymbol{\alpha}) + \boldsymbol{t} - \boldsymbol{u}_i\right\|^2 + \lambda \left\|\boldsymbol{\alpha}\right\|^2 \quad . \tag{4}$$

This can be minimized more efficiently by changing the direction of the rigid transform to

$$\begin{aligned} f(\mathbf{R}, \boldsymbol{t}, \boldsymbol{\alpha}) &= \sum_i \left\|\boldsymbol{\mu}_i + \mathbf{M}_i\boldsymbol{\alpha} + \boldsymbol{t}' - \mathbf{R}'\boldsymbol{u}_i\right\|^2 + \lambda \left\|\boldsymbol{\alpha}\right\|^2 \\ \boldsymbol{t}' = \mathbf{R}^{-1}\boldsymbol{t} & \qquad \mathbf{R}' = \mathbf{R}^{-1} \qquad . \end{aligned} \tag{5}$$

because then the Jacobian consists of a large constant part and three columns which depend on the iteration.

$$F_i = \boldsymbol{\mu}_i + \mathbf{M}_i\boldsymbol{\alpha} + \boldsymbol{t}' - \mathbf{R}'_{r_1,r_2,r_3}\boldsymbol{u}_i \tag{6}$$

$$\frac{\partial F_i}{\partial \boldsymbol{\alpha}} = \mathbf{M}_i \qquad \frac{\partial F_i}{\partial \boldsymbol{t}'} = \mathbf{I}_3 \qquad \frac{\partial F_i}{\partial r_i} = \frac{\partial \mathbf{R}'_{r_1,r_2,r_3}}{\partial r_i}\boldsymbol{u}_i \tag{7}$$

$$\mathbf{J} = \left[\mathbf{J}_c \quad | \quad \mathbf{J}_d\right] \tag{8}$$

$$\mathbf{J}_c = \begin{bmatrix} \mathbf{M} & \mathbf{1} \otimes \mathbf{I}_3 \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \tag{9}$$

$$\mathbf{J}_d = \begin{bmatrix} (\mathbf{I} \otimes \frac{\partial \mathbf{R}'}{\partial r_1})\mathbf{u}^T & (\mathbf{I} \otimes \frac{\partial \mathbf{R}'}{\partial r_2})\mathbf{u}^T & (\mathbf{I} \otimes \frac{\partial \mathbf{R}'}{\partial r_3})\mathbf{u}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{10}$$

Accordingly, the Hessian can be approximated as

$$\mathbf{H} = \begin{bmatrix} \mathbf{J}_c^T\mathbf{J}_c & (\mathbf{J}_c^T\mathbf{J}_d)^T \\ \mathbf{J}_c^T\mathbf{J}_d & \mathbf{J}_d^T\mathbf{J}_d \end{bmatrix} \qquad . \tag{11}$$

By precalculating the constant parts of the matrices we can remove most of the computation time, making step 3 very fast.

We initialize the registration by locating the tip of the nose with the method of [13]. This initialization is good enough to for a fully automatic fit, as the fitting behaves like rigid ICP in the beginning, and rigid ICP is known to have a large basin of convergence.

4. Experiments

We evaluated the system on two databases with and without the expression model. We used the GavabDB [10] database and the UND [8] database. For both databases, only the shape information was used. The GavabDB database contains 427 scans, with seven scans per ID, three neutral and four expressions. The expressions in this dataset vary considerably, including sticking out the tongue and strong facial distortions. Additionally it has strong artifacts due to facial hair, motion and the bad scanner quality. This dataset is typical for a non-cooperative environment. The UND database was used in the face recognition grand challenge [11] and consists of 953 scans, with one to eight scans per ID. It is of better quality and contains only slight expression variations. It represents a cooperative scenario.

The fitting was initialized by detecting the nose, and assuming that the face is upright and looking along the z-axis. The nose was detected with the method of [13]. The GavabDB database has the scans already aligned and the tip of the nose is at the origin. We used this information for the GavabDB experiments. The same regularisation parameters were used for all experiments, even though the GavabDB data is more noisy than the UND data. The parameters were set manually based on a few scans from the GavabDB database. We used 250 principal identity components and 60 expression components for all experiments.

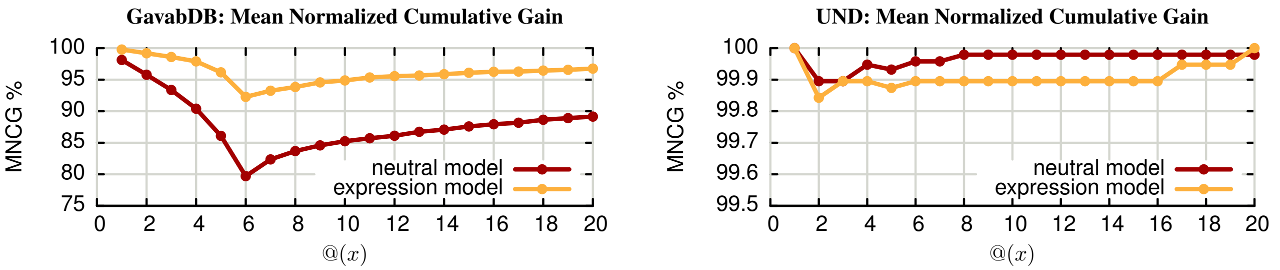


Figure 3. For the expression dataset the retrieval rate is improved by including the expression model, while for the neutral expression dataset the performance does not decrease. Plotted is the mean normalized cumulative gain, which is the number of retrieved correct answers divided by the number of possible correct answers. Note also the different scales of the MNCG curves for the two datasets. Our approach has a high accuracy on the neutral (UND) dataset.

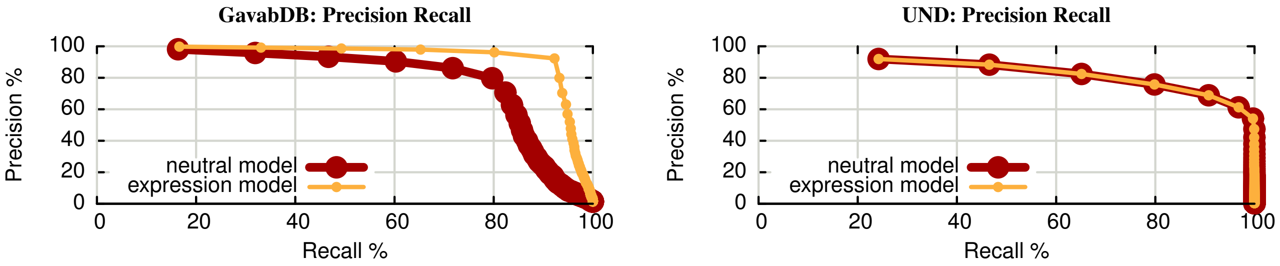


Figure 4. Use of the expression model improves retrieval performance. Plotted are precision and recall for different retrieval depths. The lower precision of the UND database is due to the fact that some queries have no correct answers. For the UND database we achieve total recall when querying nine answers, while the maximal number of scans per individual is eight, while for the GavabDB database the expression model gives a strong improvement in recall rate but full recall can not be achieved.

In the experiments the distances between all scans were calculated, and we measured recognition and retrieval rates by treating every scan once as the probe and all other scans as the gallery. Both databases were used independently.

4.1. Retrieval Measures

We measure similarity between faces as the angle between the face parameters in Mahalanobis space, which has proven to have high recognition rates [6]. The distance measure is

$$s(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \arccos\left(\frac{\boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_2}{\|\boldsymbol{\alpha}_1\|\|\boldsymbol{\alpha}_2\|}\right) \quad . \tag{12}$$

We observed that the angular measure gives slightly larger recognition rates than the Mahalanobis distance. The Mahalanobis angle has the effect of regarding all caricatures of a face, which lie on a ray from the origin towards any identity, as the same identity. We also evaluated other measures, but found them to be consistently worse than the Mahalanobis angle.

4.2. Results

As expected, the two datasets behave differently because of the presence of expressions in the examples. We first describe the results for the cooperative and then for the uncooperative setting.

4.2.1 UND

For the UND database we have good recognition rates with the neutral model. The mean cumulative normalized gain curve in Figure 3 shows for varying retrieval depth the number of correctly retrieved scans divided by the maximal number of scans that could be retrieved at this level. From this it can be seen that the first match is always the correct match, if there is any match in the database. But for some probes no example is in the gallery. Therefore for face recognition we have to threshold the maximum allowed distance to be able to reject impostors. Varying the distance threshold leads to varying false acceptance rates (FAR) and false rejection rates (FRR), which are shown in Figure 5. Even though we have been tuning the model to the GavabDB dataset and not the UND dataset our recognition rates at any FAR rate are as good or better than the best results from the face recognition vendor test. This shows, that our basic face recognition method without expression modelling gives convincing results. Now we analyze how the expression modelling impacts recognition results on this expression-less database. If face and expression space are not independent, then adding invariance towards expres-

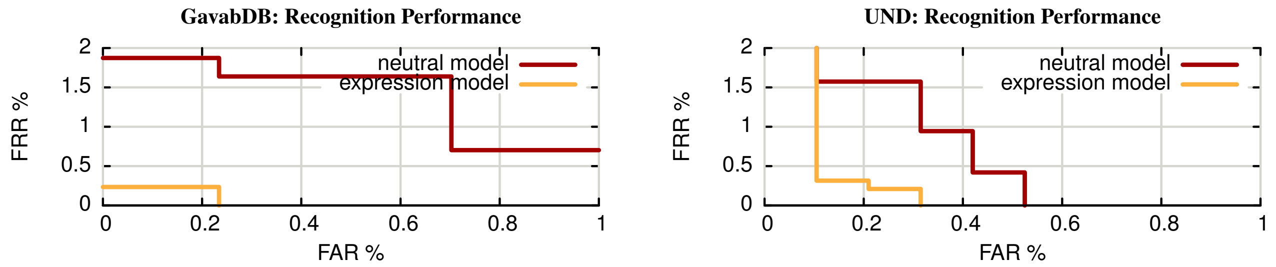


Figure 5. Impostor detection is reliable, as the minimum distance to a match is smaller than the minimum distance to a nonmatch. Note the vast increase in recognition performance with the expression model on the expression database, and the fact that the recognition rate is not decreasing on the neutral database, even though we added expression invariance. Already for 0.5% false acceptance rate we can operate ot 0% false rejection rate. false acceptance rate with less than 4% false rejection rate, or less than 0.5% FAR with less than 0.5% FRR.

sions should make the recognition rates decrease. In fact, while we find no significant increase in recognition and retrieval rates, the results are also not worse when including expression variance. Let us now turn towards the expression database, where we expect to see an increase in recognition rate due to the expression model.

4.2.2 GavabDB

The recognition rates on the GavabDB without expression model are not quite as good as for the expression-less UND dataset, so here we hope to find some improvement by using expression normalization. And indeed, the closest point recognition rate with only the neutral model is 98.1% which can be improved to 99.7% by adding the expression model. Also the FAR/FRR values decrease considerably. The largest improvement can be seen in retrieval performance, displayed in the precision recall curves in Figure 4 and mean cumulative normalized gain curves in Figure 3. This is because there are multiple examples in the gallery, so finding a single match is relatively easy. But retrieving all examples from the database, even those with strong expressions, is only made possible by the expression model.

5. Speed

Though the method as presented operates at only approximately 40 to 90 seconds per query, depending on the number of coefficients that are estimated, it has the potential for speedup. It is possible to parallelize the closest point estimation and the optimisation, and more elaborate fitting algorithms including multiresolution schemes can be developed. The speed also depends on the number of vertices and components. The results presented here used 11.000 vertices and 250 neutral plus 60 expression components.

6. Conclusion

We have used a 3D Morphable Model with a separating expression model to develop an expression-invariant face recognition algorithm. We have shown, that the system has excellent recognition rates as well on data with expressions as on data taken in a cooperative environment. The introduction of expression invariance did not incur a significant loss of precision on easier neutral data. The strong prior knowledge of the 3DMM allows robust handling of noisy data and allowed us to build a fully automatic face recognition system. We also introduced a relatively efficient fitting algorithm, which, as it has the potential for parallelisation, could be made even faster.

As we do establish correspondence between the model and the scans, it is trivial to add image based classification for datasets where a calibrated photo is available. This can be done by comparing the rectified textures, which should result in even higher recognition rates. It is also important to note that the expression normalization described here for range data can be applied equally well to other modalities, using any of the proposed 3DMM fitting algorithms.

In the future we plan to include the additional texture cues and make the method faster, such that it is applicable in real world scenarios where a processing time of 40 seconds per probe is still a problem. Furthermore we would like to investigate more sophisticated fitting algorithms and a morphable model with a larger expression space.

Acknowledgement

The authors wish to thank Pascal Paysan for the data acquisition and Frank ter Haar for the pose normalisation results. This work was supported in part by a grant from Microsoft Research and the Swiss National Science Foundation (200021-103814 and NCCR COME project 5005-66380).

References

[1] B. Amberg, R. Knothe, and T. Vetter. SHREC'08 entry: Shape based face recognition with a morphable model. In *SMI'08, Shape Modeling International*, pages 1–2, New York, NY, USA, 2008.

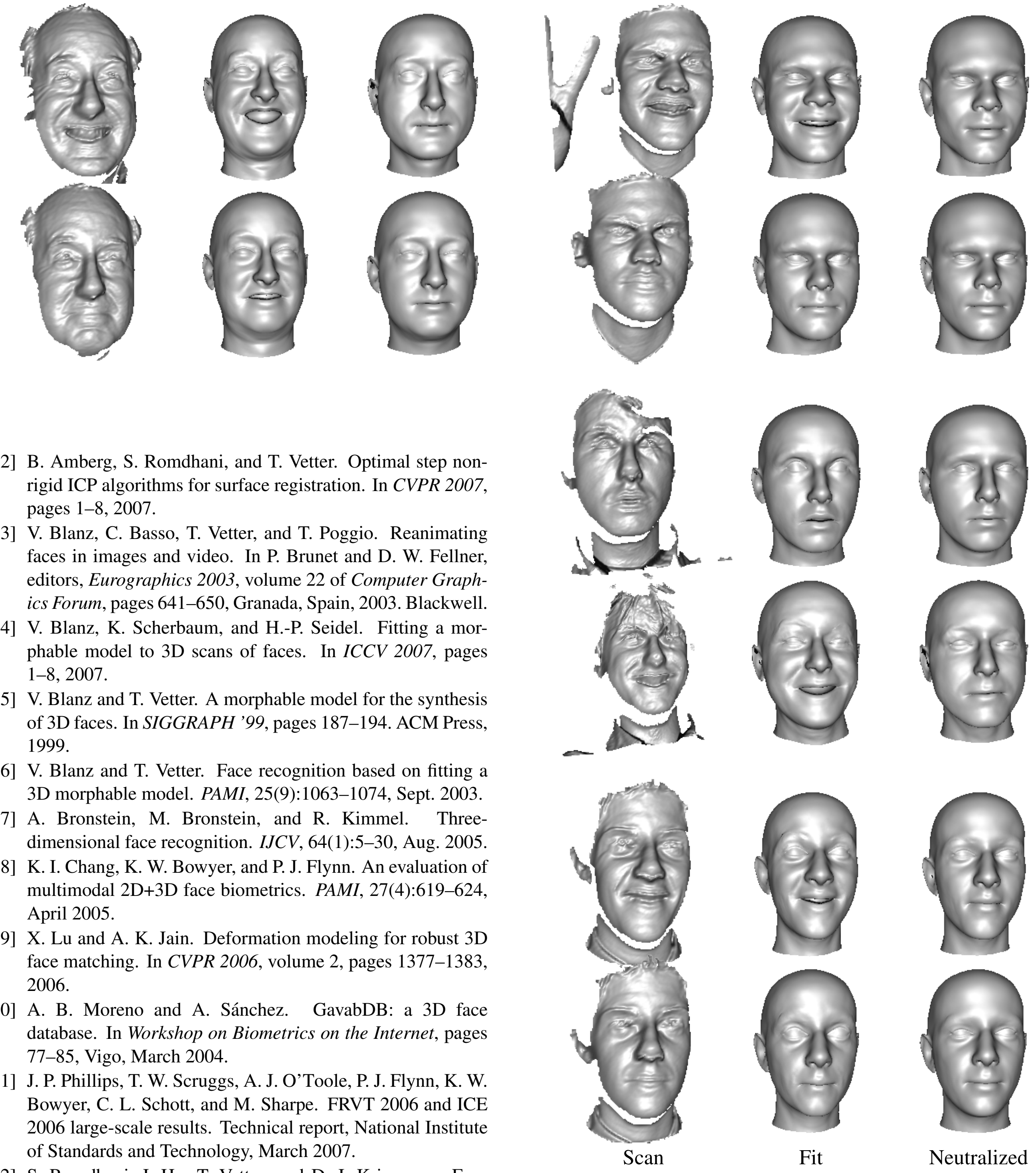


Figure 6. More expression normalisation results. Shown are pairs of results from the same subject.

[2] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid ICP algorithms for surface registration. In *CVPR 2007*, pages 1–8, 2007.

[3] V. Blanz, C. Basso, T. Vetter, and T. Poggio. Reanimating faces in images and video. In P. Brunet and D. W. Fellner, editors, *Eurographics 2003*, volume 22 of *Computer Graphics Forum*, pages 641–650, Granada, Spain, 2003. Blackwell.

[4] V. Blanz, K. Scherbaum, and H.-P. Seidel. Fitting a morphable model to 3D scans of faces. In *ICCV 2007*, pages 1–8, 2007.

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH '99*, pages 187–194. ACM Press, 1999.

[6] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *PAMI*, 25(9):1063–1074, Sept. 2003.

[7] A. Bronstein, M. Bronstein, and R. Kimmel. Three-dimensional face recognition. *IJCV*, 64(1):5–30, Aug. 2005.

[8] K. I. Chang, K. W. Bowyer, and P. J. Flynn. An evaluation of multimodal 2D+3D face biometrics. *PAMI*, 27(4):619–624, April 2005.

[9] X. Lu and A. K. Jain. Deformation modeling for robust 3D face matching. In *CVPR 2006*, volume 2, pages 1377–1383, 2006.

[10] A. B. Moreno and A. Sánchez. GavabDB: a 3D face database. In *Workshop on Biometrics on the Internet*, pages 77–85, Vigo, March 2004.

[11] J. P. Phillips, T. W. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale results. Technical report, National Institute of Standards and Technology, March 2007.

[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] F. B. ter Haar and R. C. Veltkamp. A 3D Face Matching Framework. In *Proc. Shape Modeling International (SMI'08)*, pages 103–110.

[14] F. B. ter Haar and R. C. Veltkamp. 3D Face Model Fitting for Recognition. In *ECCV 2008*, 2008.

[15] P. Yan and K. W. Bowyer. A fast algorithm for ICP-based 3D shape biometrics. In *Automatic Identification Advanced Technologies, 2005.*, pages 213–218, 2005.