Automated Registration of 3D Faces using Dense Surface Models

Tim J. Hutton[1], Bernard F. Buxton[2] and Peter Hammond[1]

[1]Biomedical Informatics Unit,
Eastman Dental Institute, University College London,
256 Gray's Inn Road, London WC1X 8LD

[2]Department of Computer Science, University College London,
Gower Street, London WC1E 6BT

T.Hutton@eastman.ucl.ac.uk

Abstract

Dense surface models can be used to register unseen surfaces, using an algorithm which is a hybrid of iterative closest-point (ICP) and active shape model (ASM) fitting. In this paper we give details of this procedure and show how it can be improved by sequentially extending the transform group over which it operates. We also evaluate it for robustness to the position of the target and to shape variation across a set of unseen examples. The fit was successful on all 21 examples in our test set, with an average RMS error of 3.0mm. An initial comparison of 3 people landmarking the same scans suggests that this is within the normal landmark reproducibility range for 3D face scans.

1 Introduction

The technology for non-invasively acquiring three-dimensional surface scans of biological subjects, especially of the human face, is becoming widespread. Such scans are medically useful as a means of capturing a detailed record of an individual's face at a moment in time, for example to audit the outcome of surgery or to analyse facial growth associated with different genetic conditions.

To analyse these surfaces usefully they must be registered. However, their manual annotation with landmarks can be time-consuming and subject to error. A technique that could accurately and robustly register surface scans of the human face would find ready applications in medical diagnosis, security, human motion tracking and animation.

Many techniques exist for registering surfaces automatically using features such as local curvature [1, 11, 21, 15, 20, 12] but they do not use a model of the range of shape variation expected in the target surface. Where the shape of the target is known to vary significantly, especially for biological objects such as the human face, it seems inevitable that such methods will register unseen surfaces less accurately.

Point distribution models (PDMs) have proved to be a useful way both of capturing shape variation and of using it to regularise the search for matching structures in images

BMVC 2003 doi:10.5244/C.17.45

[7]. Extending PDMs for use with 3D surfaces, however, is not straightforward since a dense correspondence between a set of surfaces must first be established. Methods for doing this automatically have been suggested [3, 4, 8] but have only been demonstrated on closed surfaces. It is not clear how these techniques could be used on open surfaces such as scans of the face that often have holes and an ill-defined boundary. Methods for semi-automatically making a dense correspondence between surfaces, using a sparse set of hand-placed landmarks, have also been proposed [17, 14]. The second of these methods explicitly deals with the question of open surfaces.

A method for using the surface model to register new faces automatically was briefly described in [14]. In this paper we review this procedure in detail and show how it can be improved by sequentially extending the group of transforms over which it operates (section 2.2). We then evaluate its robustness to variation in the position and orientation of the target surface (section 3.1), and to the shape variation across the human population (section 3.2).

2 Materials and Method

A collection of 421 3D face scans of different people was acquired on a DSP400 machine (`www.3dMD.com`). The subjects range in age from 1 to 80 years old, and are of both genders and different ethnic groups. Facial expression is mostly neutral throughout the set. Subjects with moustaches or beards were excluded. The extent of the face surface captured varies significantly across the dataset, some scans included neck and shoulder regions while others stop at the line of the jaw. Small holes and other noise in the surface (mostly caused by hair hanging across the face, eyebrows, eyelashes, or the nostrils) are quite frequent.

For the evaluation of the performance of our method, the full set was split at random into 400 for a training set and 21 for a test set.

2.1 Building the Dense Surface Model

Our starting point is to use hand-placed landmarks across the area of interest. For this study we chose the following ten landmarks: the corners of the eyes and mouth, the tip of the nose, the point underneath the nose (subnasale), soft-tissue nasion (bridge of the nose) and a chin point (gnathion) [10]. This choice is determined by the need for landmarks with high reproducibility that are biologically homologous (i.e. correspond biologically, not just in terms of eg. curvature similarity) and that cover the face.

Points in-between these well-defined landmarks, such as on the cheeks and across the forehead have no precise biological correspondence and yet aspects of their shape contain useful biological information. Our method makes use of this extra information by interpolating the correspondence between the landmarks. For our interpolating function we use the thin-plate spline (TPS) [2] which is often used for medical image analysis because it minimizes the bending energy of the transform, hence keeping distortion of the surface scans to a minimum.

The process of making a dense correspondence between all the surfaces in the training set is illustrated in Fig. 1. Each surface is TPS-warped using the ten landmarks onto the set of mean landmarks (computed using Procrustes alignment [13]). Each warped surface is then resampled using a base mesh and returned to its original position (‘unwarped’).

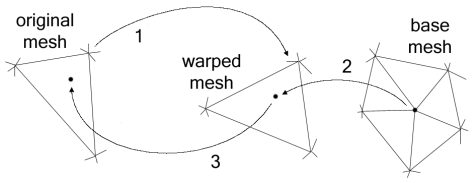
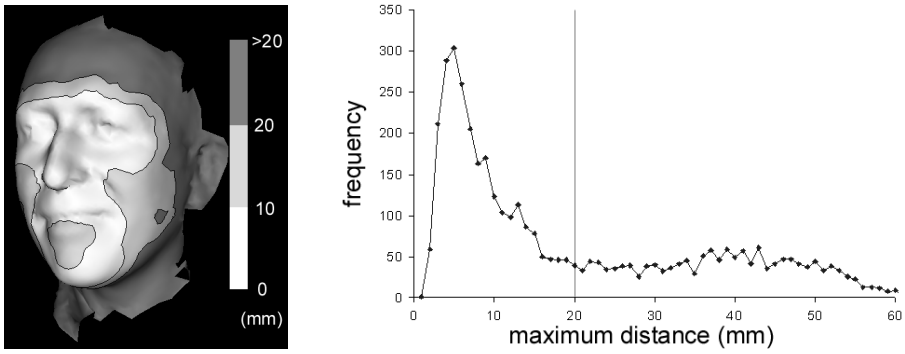


Figure 1: The dense correspondence algorithm illustrated. Each mesh is first TPS-warped (1) into alignment with the base mesh. The warped mesh is then resampled by finding the nearest point to each vertex of the base mesh (2). Finally the resampled vertex is returned to the same relative position within the corresponding triangle on the original mesh (3).



(a) (b)

Figure 2: *(a)* The maximum over the set of faces of the distances from each vertex on the base mesh to the nearest point on the other warped surfaces, shown as intensities on the base mesh. *(b)* A histogram of the same data.

If the distance from any base mesh vertex to the closest point on any warped surface is greater than some threshold (we use 20mm for our data) then that vertex is rejected. This effectively trims the surface to those parts of the face that appear in all the training examples, and is an essential step when the input data varies widely in extent as is common with surface-acquisition devices. Fig. 2 shows the maximum over the set of faces of the distances from each vertex on the base mesh to the nearest point on the other warped surfaces, as intensities on the base mesh *(a)* and as a histogram *(b)*. Since these are *maximum* distances it is clear that the majority of surfaces are brought into a close alignment in the areas of interest by the TPS-warping. This reduces the chance of folds being produced in the surface (caused by the closest-point projection from the base mesh), and indeed we do not see any folding in the final model. By contrast, when *only* closest-point projection is used, without first attempting to reduce the shape differences, folding problems can occur [16].

Now that we have constructed a dense mesh of corresponding vertices on the surface of each of the scans, we can treat all the vertices as landmarks. Following [7], we then apply the Procrustes algorithm to align all the shapes according to these new dense landmark points and produce a mean shape. Because our data is calibrated for size, we do not include scaling in the Procrustes alignment, building instead a size-and-shape model

[9]. This allows us to constrain the size variation when fitting, making it more robust (see Fig. 4).

A statistical model and face-shape-space is constructed by applying principal components analysis (PCA) to the data obtained from the above Procrustes alignment. As usual [7], each example can be represented by a shape vector of concatenated x, y and z coordinates for all n vertices and thus the mean computed:

$$\mathbf{x} = [x_1, y_1, z_1, \ldots, x_n, y_n, z_n]^{\mathrm{T}} \qquad \text{and} \qquad \overline{\mathbf{x}} = \frac{1}{s}\sum_{i=1}^{s} \mathbf{x}_i \,, \tag{1}$$

where $s = 400$ is the number of examples in our training set.

We compute the $3n \times s$ matrix $\mathbf{D}$ defined as

$$\mathbf{D} = [(\mathbf{x}_1 - \overline{\mathbf{x}})|\ldots|(\mathbf{x}_s - \overline{\mathbf{x}})] \tag{2}$$

and then decompose the matrix $\mathbf{T}$ given by

$$\mathbf{T} = \frac{1}{s-1}\mathbf{D}^{\mathrm{T}}\mathbf{D} \tag{3}$$

to compute its eigenvalues $\lambda_1 \ldots \lambda_s$ and corresponding eigenvectors $e_1 \ldots e_s$. The eigenvalues λ_i give the variances of the principal components or modes of shape variation of the faces within our training set, whilst the modes themselves are given by $\phi_i = \mathbf{D}\mathbf{e}_i$. As usual, each mode ϕ_i is then normalized to unit magnitude.

The computed modes ϕ_i can be treated as deformations of the whole mesh and can be added directly to the coordinates of the vertices of the mean mesh to synthesise face templates $\mathbf{x}$ which may either be new faces or members $\mathbf{x}_i$ of the training set. Thus we let

$$\mathbf{x} = \overline{\mathbf{x}} + \Phi\mathbf{W}\mathbf{b} \tag{4}$$

where $\Phi = [\phi_1|\phi_2|\ldots|\phi_t]$ is the matrix of the first t modes of variation and $\mathbf{b} = [b_1, b_2 \ldots b_t]^{\mathrm{T}}$ is a set of parameters (in standard deviations) controlling the amplitude of each mode. The number of parameters, t, is chosen so that we include 98% of the shape variation:

$$\min \;\; t \in \mathbb{N} \;\; \text{such that} \qquad \frac{\sum_{i=1}^{t}\lambda_i}{\sum_{j=1}^{s}\lambda_j} \geq 0.98 \,, \tag{5}$$

although other methods of choosing t are possible (see for example [18]).

It is most convenient, for example to ensure that the face shapes are sampled with correct probability, to measure the components of $\mathbf{b}$ in units of standard deviations. The matrix $\mathbf{W}$ in (4) is then a diagonal "unwhitening" matrix given by

$$\mathbf{W}_{ij} = \begin{cases} \sqrt{\lambda_i} & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The whitened parameters $\mathbf{b}$ for any template $\mathbf{x}$ can be found using:

$$\mathbf{b} = \mathbf{W}^{-1}\Phi^{\mathrm{T}}(\mathbf{x} - \overline{\mathbf{x}}) \tag{7}$$

where $\mathbf{W}^{-1}$ is simply a diagonal matrix with elements $\mathbf{W}_{ii}^{-1} = \frac{1}{\sqrt{\lambda_i}}$.

2.2 Fitting

As with active shape models [7], the search is initiated by placing the mean surface into the scene and iteratively deforming it using the parameters $\mathbf{b}$ to best match what is found locally around each vertex. With 2D and 3D images, a model of the target grey-level profile must be built in order to drive the search but with surfaces the task is much easier, needing only a simple nearest point search.

The key insight of shape model fitting is that the deformable template can be constrained to vary only within a range of shapes defined by the training set. This is step 6 in the sequence of operations below:

1. the initial template is $\mathbf{x}(0)$, the target is $\mathbf{y}$
2. standard ICP procedure [1] is used to fit the template $\mathbf{x}(0)$ to the surface $\mathbf{y}$, giving $\mathbf{x}(1)$
3. closest-point mapping (see below) $\mathbf{x}(1)$ onto $\mathbf{y}$ gives $\mathbf{x}(2)$
4. $\mathbf{x}(2)$ is least-squares (LS) aligned (initially using the Euclidean transform group) with $\overline{\mathbf{x}}$ to give $\mathbf{x}(3)$
5. the t parameters required to model $\mathbf{x}(3)$ are computed:
$$\mathbf{b} = \mathbf{W}^{-1}\Phi^{\mathrm{T}}(\mathbf{x}(3) - \overline{\mathbf{x}}) \tag{8}$$
6. the parameters $\mathbf{b}$ are limited to be legal in some way (see below), giving $\mathbf{b}'$
7. the best-guess surface is therefore computed: $\mathbf{x}(4) = \overline{\mathbf{x}} + \Phi\mathbf{W}\mathbf{b}'$
8. $\mathbf{x}(4)$ is LS-aligned with $\mathbf{x}(2)$ to give a new template $\mathbf{x}(5)$
9. if the RMS vertex difference between $\mathbf{x}(1)$ and $\mathbf{x}(5)$ is greater than some threshold ε then $\mathbf{x}(1) \leftarrow \mathbf{x}(5)$ (i.e. the new template is used for the next iteration) and repeat from step 3
10. otherwise, repeat from step 3 but extend the transform group for the LS-alignment in steps 4 and 8 first to the similarity group and then to the affine group.

By "closest-point mapping" in step 3 we mean moving each vertex in the template to the closest point on the target surface, not necessarily itself a vertex. An efficient implementation of this for polygonal surfaces is available in VTK's vtkCellLocator class [19].

In step 6, the method used for restricting the parameters depends on our model of the distribution. A reasonable assumption of a normal distribution tells us that we should limit $||\mathbf{b}||$ to be less than some k, where k^2 is given by the upper critical value of the χ^2 distribution with t degrees of freedom. For example, with our 400 examples, the model we computed from (5) has $t = 40$. If we use a significance level of $\alpha = 0.025$ (to include approximately 98% of the distribution) then we find $\chi^2(t,\alpha) = 59.34$ and so we should limit $||\mathbf{b}||$ to be less than 7.7 standard deviations. This method is encapsulated in the following rule:

$$\mathbf{b}' = \begin{cases} \frac{\sqrt{\chi^2(t,\alpha)}}{||\mathbf{b}||}\mathbf{b} & \text{if } ||\mathbf{b}|| > \sqrt{\chi^2(t,\alpha)} \\ \mathbf{b} & \text{otherwise} \end{cases} \tag{9}$$

For the threshold change between iterations, ε, we use a small multiple of a measure of the size of the template, defining:

$$\varepsilon = 10^{-4} \times \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{v}_i - \overline{\mathbf{v}}|| \tag{10}$$

where $\mathbf{v}_i$ is a vertex in the template. Our starting template gives $\varepsilon = 0.0051$mm.

When the fit does not correctly locate the target face it may be many iterations before the template change threshold ε is reached. To avoid this situation we additionally impose a limit of 1000 iterations on the fitting process. If this limit is reached we regard the fitting to have failed. The fitting process takes approximately one minute to run.

As described in step 10, the LS-alignments of the template during the fit are initially carried out without scaling, using the Euclidean (rigid-body) transformation group. Since we built a size-and-shape model the template can vary in size to match the target but only to a limited degree (defined by the training set) and only where correlated with the shape of the target. Experiments showed that fitting with the similarity group can give greater accuracy in some cases but is less robust than using the Euclidean group (see Fig. 4). Thus after the template change drops below ε for the first time we extend the transform group, allowing the template to scale to better fit the target surface. When the change drops below ε for the second time we extend to the affine group, which again can give greater accuracy in some cases. The sequential introduction of larger transform groups works well in practice, sacrificing neither robustness nor accuracy.

3 Evaluation

To evaluate the robustness and accuracy of the fitting procedure described above, we tested it on the 21 unseen scans in the test set. The scans were manually landmarked in the same way as the ones in the training set, the 10 landmarks providing the ground-truth for measuring the accuracy of the fit. Ideally we would use cross-validation to obtain a better estimate of the fitting performance over a larger set of test cases but we have not yet done this.

There are several aspects of fitting performance that we need to test. Firstly, we look at how robust the fitting is to the position of the target as regards rotation and translation. We do this by altering the starting position of the template and measuring the accuracy of the final fit to a given target. Secondly, we look at robustness to shape variation by fitting to the different examples in the test set.

3.1 Testing for robustness to position and orientation of the target

To examine the reliability of the fitting with regards to the position of the target, an example was selected at random from the test set to be used as the target surface (#13 in Fig. 4(a)). We then ran the fitting procedure to termination multiple times, each from a different starting position. For the rotation robustness experiments, the mean template was placed in approximately the correct position by LS-aligning it with the known landmarks for the target surface, and was then rotated about one of the x, y or z-axes by a specified amount. The translation robustness experiments were carried out in a similar manner.

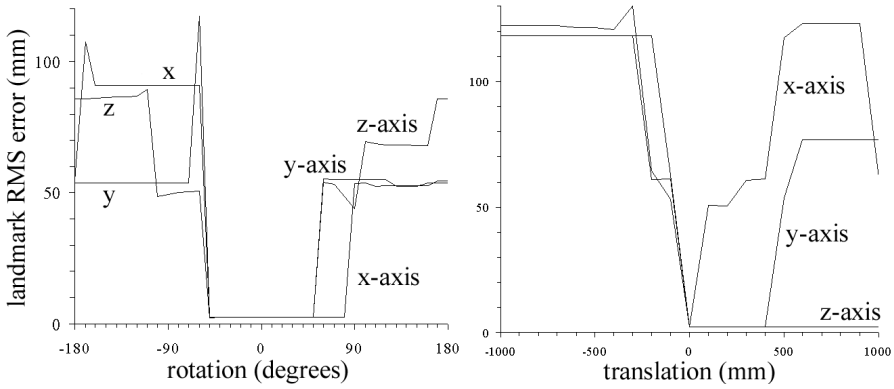


Figure 3: *(a)* Graph showing the accuracy of the fit as a function of the rotation perturbation of the initial template (x, y and z axes). *(b)* The landmark RMS error after fitting from each of the translation perturbation positions.

The locations of the ten landmarks for any template $\mathbf{x}$ can be computed using a similar process to step 3 in Fig. 1. The landmarks on the base mesh are projected onto the template by finding the corresponding position in the same triangle in $\mathbf{x}$, using barycentric coordinates. Comparison of the predicted landmarks with the known target landmarks allows us to compute the RMS error, showing how successful the fit has been.

Figure 3(a) shows how the landmark accuracy after fitting is dependent on the degree of rotation of the initial position of the template. All three axes show clear basins of convergence - if the rotation perturbation is within $\pm 50^\circ$ or thereabouts around any axis then the template converges to the same place, giving an RMS error on this example of 2.4mm. Many of the cases beyond 50° reached the 1000 iteration limit and so would be regarded as failures. Typically the starting positions that gave a low RMS error reached their final positions within 400 iterations.

A similar experiment was run to test the reliability of the fitting to translation of the target. The mean template was again placed in approximately the correct position by LS-alignment of it with the known landmarks for the target surface, and then translated along one of the x, y or z-axes. The fitting results are shown in Fig. 3(b). The results are much more asymmetric than for the rotation experiments. This is due to the closest-point search (step 3) finding mis-correspondences between the face template points and the structure in the scene being fitted. For particular scans of the human face there are often spurious structures to the side of the face such as hair and clothing that make fitting from different sides more or less likely to converge correctly. In general, however, we expect results from fitting to other surfaces to be broadly similar.

3.2 Testing fitting accuracy across a set of faces

The second aspect of the fitting that we tested was the robustness to shape variation. To do this we measured the RMS landmark error after fitting across the test set of 21 faces, from a single starting position given by $\overline{\mathbf{x}}$ in (1), ie. the mean position from the training

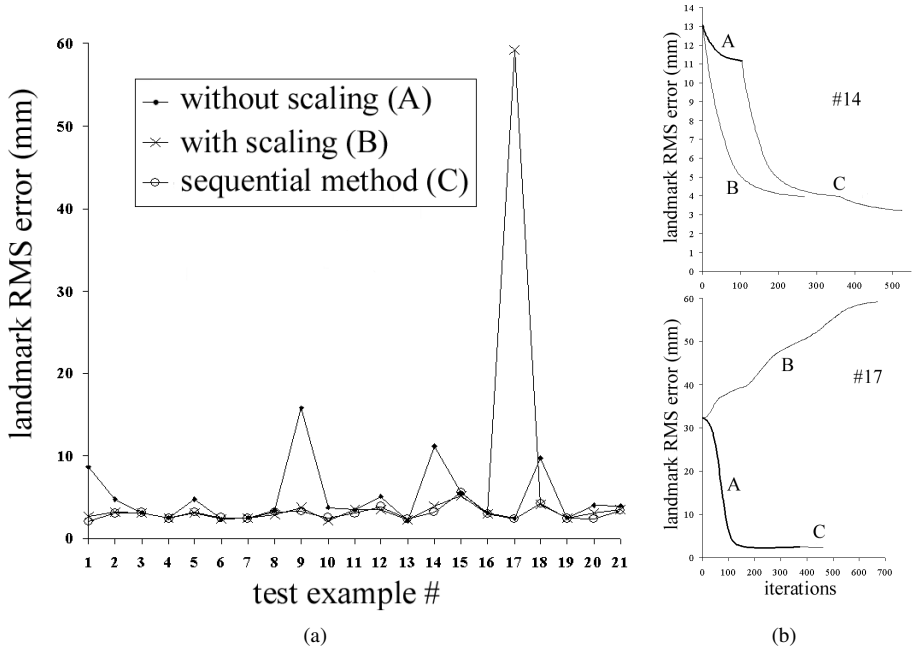


Figure 4: *(a)* RMS landmark error after fitting to the 21 faces in the test set. Three methods are compared: fitting with the Euclidean group only (A), fitting with the similarity group only (B) and fitting with the sequential introduction of Euclidean, similarity and affine groups (C). The sequential method (C) gives both robustness and accuracy. *(b)* The RMS landmark error over the fitting process for two examples in the test set: 14 and 17.

set.

Figure 4(a) shows the error after fitting for each of the 21 faces. The method of sequentially introducing larger transform groups as described in section 2.2 is method (C) in this figure. Fitting with this method gives an average RMS error over these 21 faces of 3.0mm, and a maximum of 5.5mm.

To demonstrate the effect of using other transform groups, we also ran this experiment using only the Euclidean group, method (A), and using only the similarity group, method (B). These methods did substantially worse, as can be seen in Fig. 4(a). Method (A) gave an average error of 5.0mm (15.8mm max), while method (B) gave an average of 5.8mm (59.2mm max). Method (B) gave an especially bad result on example 17. Disregarding this example, the average for method (B) was 3.1mm (5.2mm max).

To illustrate why method (C) is better, in Fig. 4(b) we look at how the template behaved during the fit for two examples (14 and 17). On example 14, method (A) does not get very close to the target, giving an RMS error of 11.2mm. Method (B) does much better, reducing the error very quickly to 3.96mm. Method (C) initially uses the Euclidean group and so follows method (A) to start with. After the introduction of scaling, method (C) reaches the same position as did method (B) but then with the introduction of the affine group the error drops further, finally reaching 3.2mm.

Example 17 shows a different situation. Here method (B) does not reduce the error

but instead causes it to increase, the template is moving away from the correct position. Method (A) converges to the correct solution (the final error is 2.4mm). The subsequent introduction of larger transform groups has little effect on this example (but does not make the results worse).

4 Conclusions

In this paper we have presented an algorithm for fitting a dense surface model to an unseen surface. In summary, the fitting procedure gave consistent results across a set of 21 test faces, with an average landmark RMS error of 3.0mm (min 2.0mm, max 5.5mm). One possible source of this error is the manual landmarking itself. While we have not yet seen or undertaken any thorough studies of the reproducibility of landmarks placed on 3D surfaces, an initial evaluation using 2 additional people to landmark the 21 test scans gave an RMS error of 2.67mm for one and 3.48mm for the other, suggesting that the performance of the fitting procedure is within the limits of human landmark placement variability.

We evaluated the accuracy of the fit to a range of starting positions and found that consistently good results were achieved on a test case from within a rotation of 50° about any of the three coordinate axes. We expect this finding to be true when fitting to other faces but have not tested this. We found that the robustness to translation of the target was highly dependent on the amount of clutter in the scene but this is to be expected since the method is a local optimiser.

The technique of showing convergence basins ([12] and Fig. 3) is very useful for evaluating robustness since it shows the accuracy of the fit and the range of initial conditions over which that accuracy is attainable.

Although we are explicitly modelling the range of face sizes in the training set, it was found that greater accuracy could be achieved on some surfaces by allowing the template to scale as necessary. It was found that sequentially enlarging the transform group gave this improved accuracy without compromising robustness, at the cost of requiring more iterations to converge (Fig. 4(b)). This is an improvement over the method given in [14]. Similar methods for extending the representational ability of the shape model using finite element analysis or other constraints have been previously proposed [5, 6].

The method is not restricted to use on faces and could be usefully applied to surface scans of other biological objects on which landmarks can be placed, such as the human ear-canal [18].

Acknowledgments

The DSP400 scanner was provided with generous funding from the Birth Defects Foundation (`www.birthdefects.co.uk`). Thanks to the many volunteers who allowed their scans to be used for our research. Thanks also to Satheesh Maheswaran for landmarking one set of test cases.

References

[1] P. Besl and N. McKay. A method for registration of 3-D shapes. *IEEE Trans. Pattern Anal. Machine Intell.*, 14:239–256, 1992.

[2] F.L. Bookstein. Shape and the information in medical images: A decade of the morphometric synthesis. *Computer Vision and Image Understanding*, 66(2):97–118, 1997.

[3] A.D. Brett, A. Hill, and C.J. Taylor. A method of 3D surface correspondence for automated landmark generation. In *British Machine Vision Conference*, pages 709–718. BMVA, 1997.

[4] A.D. Brett and C.J. Taylor. A method of automated landmark generation for automated 3D PDM construction. In *British Machine Vision Conference*, pages 914–923. BMVA, 1998.

[5] T.F. Cootes and C.J. Taylor. Combining point distribution models with shape models based on finite element analysis. *Image and Vision Computing*, 13(5):403–309, 1995.

[6] T.F. Cootes and C.J. Taylor. Combining elastic and statistical models of appearance variation. In *Proc. European Conference on Computer Vision, Vol. 1*, pages 149–163, 2000.

[7] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[8] R.H. Davies, C.J. Twining, T.F. Cootes, J.C. Waterton, and C.J. Taylor. 3D statistical shape models using direct optimisation of description length. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proc. 7th European Conference on Computer Vision*, volume 3, pages 3–20, Copenhagen, Denmark, 2002.

[9] I.L. Dryden and K.V. Mardi. *Statistical Shape Analysis*. Wiley, 1998.

[10] L.G. Farkas. *Anthropometry of the head and face*. Raven Press, New York, 1994.

[11] J. Feldmar and N. Ayache. Rigid, affine and locally affine registration of free-form surfaces. Technical report, INRIA, 1994. no. 2220.

[12] A.W. Fitzgibbon. Robust registration of 2D and 3D point sets. In *Proc. British Machine Vision Conference*, pages 411–420, 2001.

[13] J.C. Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975.

[14] T.J. Hutton, B.F. Buxton, and P. Hammond. Dense surface point distribution models of the human face. In *Proc. IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 153–160, Kauai, Hawaii, 2001.

[15] A.E. Johnson and M. Hebert. Using spin-images for efficient object recognition in cluttered 3d scenes. *IEEE Trans. Pattern Anal. Machine Intell.*, 21(5):433–449, 1999.

[16] M. Loizos. Registration of facial expression and animation on 3D face data using point distribution models. Master's thesis, Computer Science Dept., University College London, 2002.

[17] C. Lorenz and N. Krahnstöver. Generation of point-based 3d statistical shape models for anatomical objects. *Computer Vision and Image Understanding*, 77:175–191, 2000.

[18] R.R. Paulsen, R. Larsen, S. Laugesen, C. Nielsen, and B.K. Ersbøll. Building and testing a statistical shape model of the human ear canal. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2002, 5th Int. Conference, Tokyo, Japan*. Springer, 2002.

[19] W. Schroeder, K. Martin, and W. Lorenson. *The Visualization Toolkit (2nd ed.)*. Prentice-Hall, New Jersey, 1997. `http://www.vtk.org`.

[20] Y. Wang, B.S. Peterson, and L.H. Staib. Shape-based 3D surface correspondence using geodesics and local geometry. In *Computer Vision and Pattern Recognition*, volume 2, pages 644–651, Hilton Head Island, South Carolina, 2000.

[21] S.M. Yamany and A.A. Farag. Surface signatures: an orientation independent free-form surface representation scheme for the purpose of objects registration and matching. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(8):1105–1120, 2002.