A Linear Initialization for Automatic Fitting of 3D Morphable Models

Nathan Faggian, Andrew P. Paplinski
Clayton School of Information Technology
Monash University, Victoria, Australia
{nathanf,app}@mail.csse.monash.edu.au

Abstract

Morphable Models are dense 3D models built using PCA of aligned laser scanned data which is then fitted to images. During fitting, variations in illumination, pose and identity can be addressed by the rendering process; Morphable Models are therefore extremely useful for object modeling. One problem however is that Morphable Model fitting is slow since gradient descent methods are generally used. This paper outlines a combination of both a constrained Active Appearance Model and closed form 2D point based fitting method. We demonstrate preliminary results of the method and a simplified exterior orientation solution for our face model case.

1. Introduction

A Morphable Model (MM) is a statistical representation of both the shape and texture of an object in a certain domain. Most commonly MMs are applied in the face domain [9, 7]. A MM is built from 3D scans of data, which is then put into dense correspondence[2]. Using the aligned data the MM is composed of two models built using PCA for shape and texture variation:

$$\hat{\mathrm{s}} = \bar{\mathrm{s}} + \mathrm{S} \cdot \mathrm{diag}(\sigma_\mathrm{s})\mathrm{c}_\mathrm{s} \quad \hat{\mathrm{t}} = \bar{\mathrm{t}} + \mathrm{T} \cdot \mathrm{diag}(\sigma_\mathrm{t})\mathrm{c}_\mathrm{t} \qquad (1)$$

where $\hat{\mathrm{s}}$ and $\hat{\mathrm{t}}$ are novel $(3\mathrm{N} \times 1)$ shape and texture vectors, $\bar{\mathrm{s}}$ and $\bar{\mathrm{t}}$ are the $(3\mathrm{N} \times 1)$ mean shape and texture vectors, S and T are the $(3\mathrm{N} \times \mathrm{M})$ column (eigenvectors) spaces of the shapes and textures, σ are the corresponding eigenvalues, c_s and c_t are shape and texture coefficients. The linear equations describe the variation of shape within the span of the training heads. The coefficients $(\mathrm{c}_\mathrm{s}, \mathrm{c}_\mathrm{t})$ are scaled by the corresponding eigenvalue (σ) (dominance) of the eigenvector (S, T). Using varied coefficients it is possible to render different heads within the span of the original 3D scans. In the context of facial modelling, MMs are not unlike simpler 2D models such as the Active Appearance Model [3].

Cootes et al [3] presented the Active Appearance Model (AAM) as a method to model objects in images. It is a modeling by synthesis approach to image analysis and is a popular technique that has been broadly used in the field of computer vision; like the MM its use is dominant in the domain of facial modeling. An AAM represents an encoding of both the shape and texture information of the object, where the goal is to be able to estimate any valid instance of the object using PCA:

$$\hat{\mathrm{s}} = \bar{\mathrm{s}} + \mathrm{S}\alpha \quad \hat{\mathrm{t}} = \bar{\mathrm{t}} + \mathrm{T}\beta \qquad (2)$$

As with the MM the shape and texture models encode the modes of variation, although it is the variation of hand-labeled training samples. A new shape $\hat{\mathrm{s}}$ or texture $\hat{\mathrm{t}}$ can be constructed as a linear combination (α or β) of the principal components (column space) of the measurement matrices for shape and texture. Both the MM and the AAM can generate novel instances of the object through rendering.

$$\mathrm{I}(\alpha, \beta) = \mathrm{F}(\bar{\mathrm{s}} + \mathrm{S}\alpha, \bar{\mathrm{t}} + \mathrm{T}\beta) \qquad (3)$$

Rendering is the process of transforming a generated texture into a desired shape; the texture exists in a shape free representation and is warped ($\mathbf{F}$) to the desired shape. Being able to render a novel images is important. It is an integral component of analysis by synthesis fitting methods that are currently used in MM and AAM fitting.

2. MM Fitting

Currently the most accurate fitting methods for fitting MMs involve iterative gradient descent approach[9]. These use an analysis by synthesis approach where the coefficients are inferred from a difference between a rendered head (image) and the input image. Effectively it is the minimization of the cost function:

$$||\mathrm{F}(\bar{\mathrm{s}} + \mathrm{S} \cdot \mathrm{diag}(\sigma)\mathrm{c}_\mathrm{s}, \bar{\mathrm{t}} + \mathrm{T} \cdot \mathrm{diag}(\sigma)\mathrm{c}_\mathrm{t}) - \mathrm{I}||^2_\mathrm{F} \qquad (4)$$

where F is a rendering function that when provided with shape and texture coefficients produces an image that is

aligned with the input image, I. The only problem with such methods is the speed at which they can determine model coefficients. Romdhani et al [7] introduced a fast method that takes advantage of an inverse mapping relationship. It uses a multi-feature fitting strategy to reduce the cost function quickly and robustly, although the solution is not real-time and requires hand initialization.

2.1. Fast MM Fitting

If it possible to ignore the constraint of a photo-realistic result then a viable alternative to standard gradient descent methods is available. This is the recently proposed method by Blanz et al [1] which is a concise and mathematically optimal method to reconstruct a MM from a sparse set of either 2D or 3D feature points. The method has two advantages 1) it relies on only linear operators and 2) operates in real-time (at the expense of model accuracy). Using the assumption that only a small set of corresponding points are available the method minimizes the cost function:

$$||\mathrm{L}\cdot\mathrm{V}\cdot\mathrm{S}\cdot\mathrm{diag}(\sigma)\mathrm{c_s}-\mathrm{r}||_{\mathrm{F}}^2 \tag{5}$$

where L is a camera matrix, containing the full set of intrinsic and extrinsic parameters, V is a subset selection matrix, S is the $(3\mathrm{N}\times\mathrm{M})$ column (eigenvectors) space of the training shapes, σ are the corresponding eigenvalues, $\mathrm{c_s}$ are shape coefficients and r is a $(2\mathrm{P}\times 1)$ set of feature points.

3. AAMs

The problem with the Blanz et al method is the determination of V. The matrix V maps the smaller set of 2D coordinates in an image to corresponding 3D vertices in the MM. It maps from the high-dimensional MM shape vector to the smaller (2D or 3D) shape vector, r. The key contribution of this paper is an automatic way to deduce the mapping V and the shape vector, r. We make use of AAMs trained within the span of the MM 3D head basis. Specifically the AAMs are constructed using the 2D projection of the MM data and inherently provide the important 2D to 3D mapping. This was identified in [4] and is exploited in this work for fully automatic fitting of MMs.

3.1. AAM Fitting

The fitting algorithm used in our implementation is based on the Inverse Compositional Image Alignment (ICIA) AAM fitting method [6]. ICIA is a fitting method that is based on the earlier forwards-additive method [5] where its major importance is that the roles of the template and the image are reversed. This step allows the computation related to the Jacobian (which defines how the image pixels move with respect to a transformation) to be precomputed. This modification results in a very efficient AAM fitting algorithm. When applied to AAMs, ICIA, optimizes for two components that define the change in shape of the AAM. The local $\mathrm{W}(\bar{\mathrm{s}};\alpha)$ and global $\mathrm{N}(\bar{\mathrm{s}};\gamma)$ transformation. These transforms describe the variation of shape in two different cases: 1) α defines a shift of individual vertices's from the normalized model 2) γ is a similarity transformation that moves the AAM in a image, allowing the AAM to express different scales, rotations and translations.

$$\mathrm{W}(\bar{\mathrm{s}};\alpha)\simeq\hat{\mathrm{s}}=\bar{\mathrm{s}}+\mathrm{S}\alpha,\quad \mathrm{N}(\bar{\mathrm{s}};\gamma)\simeq\hat{\mathrm{s}}=\bar{\mathrm{s}}+\mathrm{S}^*\gamma$$

where (W,N) are local and similarity transforms, $(\bar{\mathrm{s}})$ is the mean shape, (α,γ) are parameter vectors and $(\mathrm{S},\mathrm{S}^*)$ define the column space of the local and similarity transforms. Fitting AAMs using ICIA is then the minimization of:

$$\sum_x\left[\mathrm{I}(\mathrm{N}(\mathrm{W}(\bar{\mathrm{s}};\alpha));\gamma)-\mathrm{T}(\mathrm{x})\right]^2 \tag{6}$$

where $\mathrm{N}(\mathrm{I}(\mathrm{W}(\bar{\mathrm{s}};\alpha));\gamma)$ is the pixels in an image sampled under a globally and locally transformed overlayed AAM and $\mathrm{T}(\mathrm{x})$ is a rendered AAM using the mean shape and mean texture. The cost function is minimized by estimating the correct update to the α and γ parameters, $\triangle\alpha,\triangle\gamma$ which are then inverted and composed with the previously updated α,γ.

3.2. MM fitting with AAM+MM

Blanz's non-regularized form [1] for fitting a given set of points y, provided by the AAM, is the problem of estimating the coefficients for shape $\mathrm{c_s}$, where L (multiplied with V) is a mapping function:

$$\mathrm{y}=\mathrm{L}\mathrm{c_s} \tag{7}$$

where $\mathrm{c_s}$ is a $(\mathrm{M}\times 1)$ vector of shape coefficients and y is a $(2\mathrm{P}\times 1)$ vector of demeaned AAM features in image coordinates:

$$\mathrm{y}=\mathrm{r}-\mathrm{L}\bar{\mathrm{v}} \tag{8}$$

such that L is a $(2\mathrm{P}\times 3\mathrm{N})$ mapping matrix and $\bar{\mathrm{v}}$ is the mean $(3\mathrm{N}\times 1)$ shape vector. When it is assumed that L is known it might seem obvious to solve for $\mathrm{c_s}$ (equation 7) with an application of the pseudo-inverse of L:

$$\mathrm{c_s}=\mathrm{L}^{+}\mathrm{y} \tag{9}$$

Although the significant problem with this straightforward method is that it does not represent a meaningful result, it minimizes the wrong error. This is because the derived vector $\mathrm{c_s}$ is not restricted to the span of the face model (MM) and thus cannot be used in (1). A more appropriate error to minimize is one within the span of the MM:

$$\mathrm{y}=\mathrm{L}\cdot\mathrm{S}\cdot\mathrm{diag}(\sigma)\mathrm{c_s} \tag{10}$$

where S is the $(3\mathrm{N} \times \mathrm{M})$ eigenvector matrix and σ is the $(\mathrm{M} \times \mathrm{M})$ corresponding eigenvalue matrix. The solution to this can be derived through the use of SVD and a simple restating of the equation to match:

$$\mathrm{y} = \mathrm{Q}\mathrm{c}_\mathrm{s} \tag{11}$$

where Q is the $(2\mathrm{P} \times \mathrm{M})$ image plane projection of the subset of scaled eigenvectors. SVD can be used to compute the inverse of an arbitrary $(\mathrm{M} \times \mathrm{N})$ matrix,this is applied directly as the solution to (10):

$$\mathrm{c}_\mathrm{s} = \mathrm{U}^\mathrm{T}\mathrm{S}^{-1}\mathrm{V}\mathrm{y} \tag{12}$$

this provides a concise and closed form solution to the problem of estimating a dense 3D shape from a sparse set of 2D points. In our implementation of Blanz's fitting we chose the more complicated regularized solution, refer to [1] for details.

4. AAM+MM Exterior Orientation

To apply the direct solution shown in the previous section the rigid motion of the MM with respect to the image must be estimated. It is a search for the optimal rotation, scaling and translation of the model that minimizes the difference between the projected 3D model (X) and the AAM points $(\hat{\mathrm{x}})$. A rigid body transformation of the model is defined like so:

$$\begin{bmatrix}\hat{\mathrm{X}}\\ 1\end{bmatrix} = \begin{bmatrix}\mathrm{R} & \mathrm{t}\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mathrm{X}\\ 1\end{bmatrix} \tag{13}$$

where R is the (3×3) rotation matrix, t is the (3×1) translation vector and $\hat{\mathrm{X}}$ is the (3×1) rotated and translated X vector. Using a scaled orthographic projection we can describe the observed 2D AAM points as projected model points that have undergone an unknown rigid transformation.

$$\hat{\mathrm{x}} = \begin{bmatrix}1\mathrm{s} & 0 & 0\\ 0 & 1\mathrm{s} & 0\end{bmatrix}\begin{bmatrix}\mathrm{R} & \mathrm{t}\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mathrm{X}\\ 1\end{bmatrix} \tag{14}$$

where s is the scaling constant and $\hat{\mathrm{x}}$ is an observed 2D AAM point. Under a scaled orthographic projection the equation for the elements of $\hat{\mathrm{x}}$ become:

$$\begin{bmatrix}\hat{\mathrm{x}_\mathrm{x}}\\ \hat{\mathrm{x}_\mathrm{y}}\end{bmatrix} = \begin{bmatrix}\mathrm{s}\cdot\mathrm{r}_1 & \mathrm{s}\cdot\mathrm{t}_\mathrm{x}\\ \mathrm{s}\cdot\mathrm{r}_2 & \mathrm{s}\cdot\mathrm{t}_\mathrm{y}\end{bmatrix}\mathrm{X} \tag{15}$$

where $\mathrm{r}_1, \mathrm{r}_2$ are the first and second rows of the rotation matrix and $\mathrm{t}_\mathrm{x}, \mathrm{t}_\mathrm{y}$ are the x, y translations. The solution to the rigid motion of the model and the scale is coupled (equation 15), it is not easily solved directly and should be derived in a least mean squared sense. Fortunately a least squares solution exists which attempts to linearize the non-linear problem. Rotation is estimated by directly estimating the parameters for the canonical exponential form, show as the Rodriguez equation:

$$\mathrm{R} = \mathrm{I}_3 + \hat{\mathrm{v}}\sin\theta + \hat{\mathrm{v}}^2(\cos\theta - 1) \tag{16}$$

where $\hat{\mathrm{w}}$ is a skew symmetric matrix:

$$\hat{\mathrm{v}} = \begin{bmatrix}0 & -\mathrm{v}_\mathrm{z} & \mathrm{v}_\mathrm{y}\\ \mathrm{v}_\mathrm{z} & 0 & -\mathrm{v}_\mathrm{x}\\ -\mathrm{v}_\mathrm{y} & \mathrm{v}_\mathrm{x} & 0\end{bmatrix} \tag{17}$$

If it is assumed that rotations are relatively small then the first order approximation for rotation becomes:

$$\mathrm{R} = \begin{bmatrix}1 & -\mathrm{v}_\mathrm{z}\cdot\sin\theta & \mathrm{v}_\mathrm{y}\cdot\sin\theta\\ \mathrm{v}_\mathrm{z}\cdot\sin\theta & 1 & -\mathrm{v}_\mathrm{x}\cdot\sin\theta\\ -\mathrm{v}_\mathrm{y}\cdot\sin\theta & \mathrm{v}_\mathrm{x}\cdot\sin\theta & 1\end{bmatrix} \tag{18}$$

For a least squares form we first examine the expanded exponential canonical form. For now ignoring the effects of scale and translation and solving for $\hat{\mathrm{x}} = \mathrm{R}\mathrm{X}$:

$$\begin{bmatrix}\hat{\mathrm{x}}_\mathrm{x}\\ \hat{\mathrm{x}}_\mathrm{y}\end{bmatrix} = \begin{bmatrix}1 & -\mathrm{v}_\mathrm{z}\cdot\sin\theta & \mathrm{v}_\mathrm{y}\cdot\sin\theta\\ \mathrm{v}_\mathrm{z}\cdot\sin\theta & 1 & -\mathrm{v}_\mathrm{x}\cdot\sin\theta\end{bmatrix}\mathrm{X} \tag{19}$$

this can be re-arranged into the least squares equations:

$$\begin{bmatrix}-\mathrm{X}_\mathrm{y} & \mathrm{X}_\mathrm{z} & 0\\ \mathrm{X}_\mathrm{x} & 0 & \mathrm{X}_\mathrm{z}\end{bmatrix}\begin{bmatrix}\mathrm{v}_\mathrm{z}\cdot\sin\theta\\ \mathrm{v}_\mathrm{y}\cdot\sin\theta\\ \mathrm{v}_\mathrm{x}\cdot\sin\theta\end{bmatrix} = \begin{bmatrix}\hat{\mathrm{x}}_\mathrm{x}\\ \hat{\mathrm{x}}_\mathrm{y}\end{bmatrix} \tag{20}$$

Given equation 20 it is possible to estimate (in a first order sense) rotation between the 3D model and 2D feature points. This is applied iteratively to estimate the true rigid rotation of the feature points. Once rotation is determined it is possible to address translation and scale. Ignoring the effects of rotation the equation for translation and scaling $(\hat{\mathrm{x}} = \mathrm{s}\mathrm{X} + \mathrm{t})$ is:

$$\begin{bmatrix}\hat{\mathrm{x}}_\mathrm{x}\\ \hat{\mathrm{x}}_\mathrm{y}\end{bmatrix} = \begin{bmatrix}\mathrm{s} & 0 & 0 & \mathrm{t}_\mathrm{x}\\ 0 & \mathrm{s} & 0 & \mathrm{t}_\mathrm{y}\end{bmatrix}\mathrm{X} \tag{21}$$

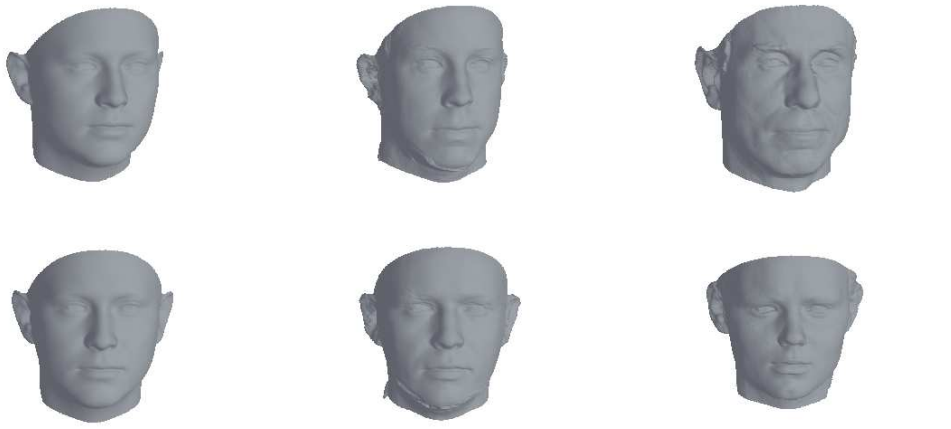
this can also take the form of a least squares equation:

$$\begin{bmatrix}\mathrm{X}_\mathrm{x} & 1 & 0\\ \mathrm{X}_\mathrm{y} & 0 & 1\end{bmatrix}\begin{bmatrix}\mathrm{s}\\ \mathrm{t}_\mathrm{x}\\ \mathrm{t}_\mathrm{y}\end{bmatrix} = \begin{bmatrix}\hat{\mathrm{x}}_\mathrm{x}\\ \hat{\mathrm{x}}_\mathrm{y}\end{bmatrix} \tag{22}$$

and from this $\mathrm{t}_\mathrm{x}, \mathrm{t}_\mathrm{y}, \mathrm{s}$ can be calculated.

5. Results

Two results of the new fitting technique are shown in Figure 1. Using the same methods as [4] (Section 3) a multi-pose AAM was constructed and fitted to MM data. By fitting the AAM to the input image (shown above) the pose



(a) AAM+MM Pose (b) AAM+MM Identity+Pose (c) Ground Truth, from USF [8] data

Figure 1. AAM+MM fittings

was first estimated and then passed to the fitting method [1] and the results observed. The technique was completely automatic and required no labeling of feature points. Figure 1.(a) demonstrates the multi-pose AAM fitting of input images. Figure 1.(b) demonstrates that the simple pose estimation algorithm effectively determined the model poses. Figure 1.(c) shows the result of the identity fitting. Figure 1.(d) shows the USF ground truth data [8]. In both cases the re-projection error of the 2D points between the USF data and the MM fitting was below 1 pixel.

6. Conclusion

We have proposed a combination of the AAM and MM and shown preliminary results of the combination. The key benefit is that no human labeling of feature points is required for the MM fitting. We have automated the correspondence computation provided by AAM. The resulting MM fitting provides a good initial estimate of the 3D structure from a subset of 2D features, quickly.

7. Acknowledgments

Thanks are sent to Sami Romdhani from the faculty of Computer Science at UniBas for putting the USF data into correspondence and the Australian Research Council for its continued funding.

References

[1] V. Blanz, A. Mehl, T. Vetter, and H.-P. Seidel. A statistical method for robust 3D surface reconstruction from sparse data. In *Second International Symposium on 3D Data Processing, Visualization and Transmission*, pages 293–300, 2004.

[2] C.Basso, T. Vetter, and V. Blanz. Regularized 3D morphable models. In *Workshop on: Higher-Level Knowledge in 3D Modeling and Motion Analysis*, 2003.

[3] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *Proc. European Conference on Computer Vision*, volume 2, pages 484–498. Springer, 1998.

[4] N. Faggian, S. Romdhani, J. Sherrah, and A. Paplinski. Color active appearance model analysis using a 3D morphable model. In *Digital Image Computing: Techniques and Applications*, December 2005.

[5] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[6] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 2000.

[7] S. Romdhani and T. Vetter. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[8] P. S. Sarkar. USF DARPA humanID 3D face database. University of South Florida, Tampa, FL.

[9] T. Vetter and V. Blanz. A morphable model for the synthesis of 3D faces. In *Siggraph 1999, Computer Graphics Proceedings*, pages 187–194, 1999.