View-Based Active Appearance Models

T.F. Cootes, K. Walker, C.J. Taylor
Dept. Imaging Science and Biomedical Engineering
University of Manchester, Manchester M13 9PT U.K.
t.cootes@man.ac.uk

Abstract

We demonstrate that a small number of 2D statistical models are sufficient to capture the shape and appearance of a face from any viewpoint (full profile to fronto-parallel). Each model is linear and can be matched rapidly to new images using the Active Appearance Model algorithm. We show how such a set of models can be used to estimate head pose, to track faces through large angles of head rotation and to synthesize faces from unseen viewpoints.

1 Introduction

The appearance of a face in a 2D image can change dramatically as the viewing angle changes. The majority of work on face tracking and recognition assumes near fronto-parallel views, and tends to break down when presented with large rotations or profile views. Three general approaches have been used to deal with this; a) use a full 3D model [15], b) introduce non-linearities into a 2D model [6] and c) use a set of models to represent appearance from different view points [11]. In this paper we explore the last approach, using statistical models of shape and appearance to represent the variations in appearance from a particular viewpoint.

These appearance models are trained on example images labelled with sets of landmarks to define the correspondences between images [1]. Lanitis *et. al.*[9] showed that a linear model was sufficient to simulate considerable changes in viewpoint, as long as all the modelled features (the landmarks) remained visible. A model trained on near fronto-parallel face images can cope with pose variations of up to 45^o either side. For much larger angle displacements, some features become occluded, and the assumptions of the model break down.

We demonstrate that to deal with full 180^o rotation (from left profile to right profile), we need only 5 models, roughly centred on viewpoints at -90^o,-45^o,0^o,45^o,90^o (where 0^o corresponds to fronto-parallel). The pairs of models at $\pm 90^o$ (full profile) and $\pm 45^o$(half profile) are simply reflections of each other, so there are only 3 distinct models. We can use these models for estimating head pose, for tracking faces through wide changes in orientation and for synthesizing new views of a subject given a single view.

Each model is trained on labelled images of a variety of people with a range of orientations chosen so none of the features for that model become occluded. The different models use different sets of features (see Figure 1). Each example view can then be approximated using the appropriate appearance model with a vector of parameters, $\mathbf{c}$. We assume that as the orientation changes, the parameters, $\mathbf{c}$, trace out an approximately elliptical path. We can learn the relationship between $\mathbf{c}$ and head orientation, allowing us to both estimate the orientation of any head and to be able to synthesize a face at any orientation.

By using the Active Appearance Model algorithm [4, 1] we can match any of the individual models to a new image rapidly. If we know in advance the approximate pose, we can easily select the most suitable model. If we do not know, we can search with each of the five models and choose the one which achieves the best match. Once a model is selected and matched, we can estimate the head pose, and thus track the face, switching to a new model if the head pose varies significantly.

Given a single image of a new person, we can match the models to estimate the pose. We can then use the best fitting model to generate new views from angles similar to that of the original image. We can also exploit correlations across models of different views to estimate the appearance of the subject in a completely different view. Though this can perhaps be done most effectively with a full 3D model [15], we demonstrate that good results can be achieved just with a set of 2D models.

In the following we describe the techniques in more detail and give examples of the model, its ability to estimate pose, to track faces and to synthesize unseen views.

2 Background

Statistical models of shape and texture have been widely used for recognition, tracking and synthesis [7, 9, 4, 14], but have tended to only be used with near fronto-parallel images.

Moghaddam and Pentland [11] describe using view-based eigenface models to represent a wide variety of viewpoints. Our work is similar to this, but by including shape variation (rather than the rigid eigenpatches), we require fewer models and can obtain better reconstructions with fewer model modes.

Maurer and von der Malsburg [10] demonstrated tracking heads through wide angles by tracking graphs whose nodes are facial features, located with Gabor jets. The system is effective for tracking, but is not able to synthesize the appearance of the face being tracked.

Murase and Nayar [6] showed that the projections of multiple views of a rigid object into an eigenspace fell on a 2D manifold in that space. By modelling this manifold they could recognise objects from arbitrary views. A similar approach has been taken by Gong *et. al.*[13, 8] who use non-linear representations of the projections into an eigen-face space for tracking and pose estimation, and by Graham and Allinson [5] who use it for recognition from unfamiliar viewpoints.

Romdhani *et. al.*[12] has extended the Active Shape Model to deal with full 180^o rotation of a face using a non-linear model. However, the non-linearities mean the method is slow to match to a new image.

Vetter [15] has demonstrated how a 3D statistical model of face shape and texture can be used to generate new views given a single view. The model can be matched to a new image from more or less any viewpoint using a general optimisation scheme, though this is slow. By explicitly taking into account the 3D nature of the problem, this approach is likely to yield better reconstructions than the purely 2D method described below. However, the view based models we propose could be used to drive the parameters of the 3D head model, speeding up matching times.

3 Statistical Models of Appearance

An appearance model can represent both the shape and texture variability seen in a training set. The training set consists of labelled images, where key landmark points are marked on each example object. The training set is usually labelled manually, though automatic methods are being developed. For instance, Figure 1 shows examples of labelled images used to train the view-based face models.

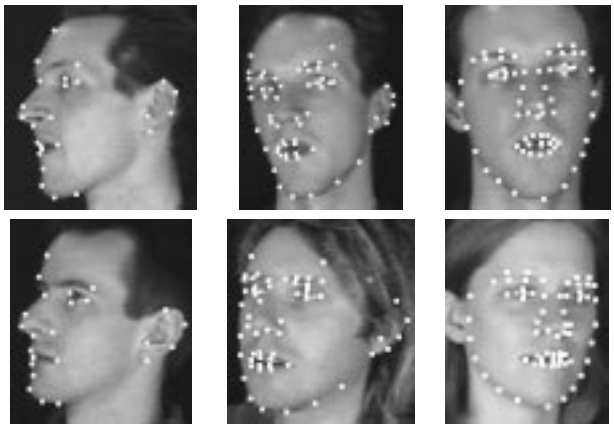


Figure 1. Examples from the training sets for the models

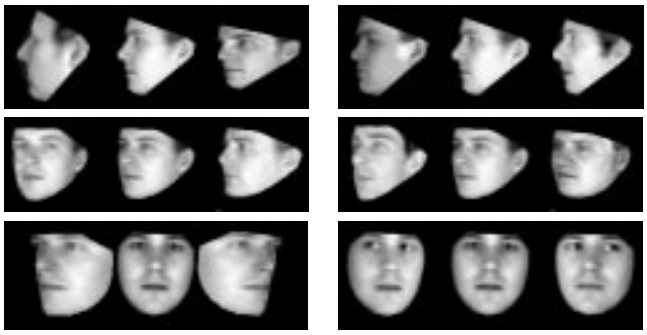
Given such a set we can generate a statistical models of shape and texture variation (see [1, 4] for details). The shape of an object can be represented as a vector $\mathbf{x}$ and the texture (grey-levels or colour values) represented as a vector $\mathbf{g}$. The appearance model has parameters, $\mathbf{c}$, controlling the shape and texture according to

$$\begin{aligned} \mathbf{x} &= \bar{\mathbf{x}} + \mathbf{Q}_s \mathbf{c} \\ \mathbf{g} &= \bar{\mathbf{g}} + \mathbf{Q}_g \mathbf{c} \end{aligned} \tag{1}$$

where $\bar{\mathbf{x}}$ is the mean shape, $\bar{\mathbf{g}}$ the mean texture and $\mathbf{Q}_s$,$\mathbf{Q}_s$ are matrices describing the modes of variation derived from the training set.

We trained three distinct models on data similar to that shown in Figure 1. The profile model was trained on 234 landmarked images taken of 15 individuals from different orientations. The half-profile model was trained on 82 images, and the frontal model on 294 images.

An example image can be synthesised for a given $\mathbf{c}$ by generating a texture image from the vector $\mathbf{g}$ and warping it using the control points described by $\mathbf{x}$. For instance, Figure 2 shows the effects of varying the first two appearance model parameters, c_1, c_2, of models trained on a set of face images, labelled as shown in Figure 1. These change both the shape and the texture component of the synthesised image.



c_1 varies ± 2 s.d.s c_2 varies ± 2 s.d.s

Figure 2. First two modes of the face models (top to bottom: profile, half-profile and frontal)

4 Predicting Pose

We assume that the model parameters are related to the viewing angle, θ, approximately as

$$\mathbf{c} = \mathbf{c}_0 + \mathbf{c}_c \cos(\theta) + \mathbf{c}_s \sin(\theta) \quad (2)$$

where $\mathbf{c}_0$, $\mathbf{c}_c$ and $\mathbf{c}_s$ are vectors estimated from training data (see below).

(Here we consider only rotation about a vertical axis - head turning. Nodding can be dealt with in a similar way.)

This is an accurate representation of the relationship between the shape, $\mathbf{x}$, and orientation angle under an affine projection (the landmarks trace circles in 3D which are projected to ellipses in 2D), but our experiments suggest it is also an acceptable approximation for the appearance model parameters, $\mathbf{c}$.

In order to learn the relationship for a given model, we must know the orientation of each of our training examples. We do not yet have access to a system which can measure it accurately, such as that used by [12, 8, 13]. However, we are able to estimate the angle by finding the frames in our training sequences at full profile and fronto-parallel by eye, then assuming a constant rate of rotation across the frames between. This leads to images labelled with orientations, θ_i, accurate to about $\pm 10^o$. For each such image we find the best fitting model parameters, $\mathbf{c}_i$. We then perform regression between $\{\mathbf{c}_i\}$ and the vectors $\{(1, \cos(\theta_i), \sin(\theta_i))'\}$ to learn $\mathbf{c}_0$,$\mathbf{c}_c$ and $\mathbf{c}_s$.

Figure 3 shows reconstructions in which the orientation, θ, is varied in Equation 2.

Given a new example with parameters $\mathbf{c}$, we can estimate its orientation as follows. Let $\mathbf{R}_c^{-1}$ be the left pseudo-inverse of the matrix $(\mathbf{c}_c|\mathbf{c}_s)$ (thus $\mathbf{R}_c^{-1}(\mathbf{c}_c|\mathbf{c}_s) = \mathbf{I}_2$).

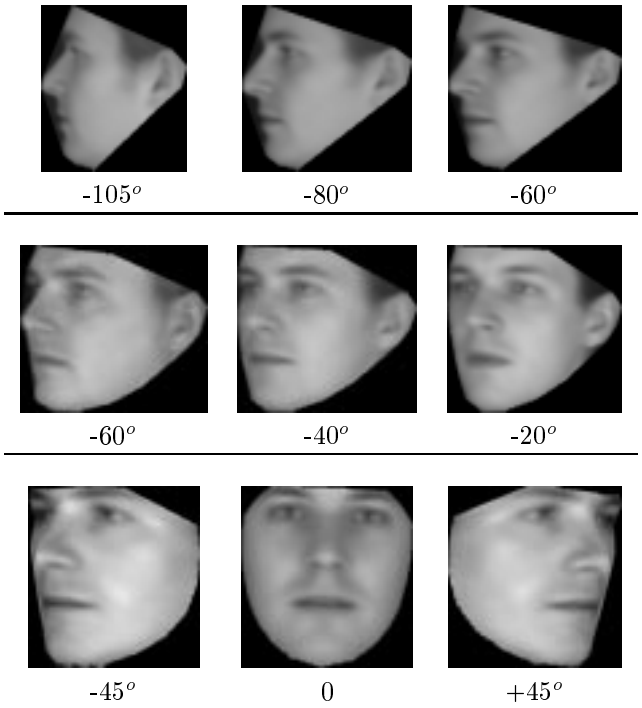


Figure 3. Rotation modes of three face models

Let

$$(x_a, y_a)' = \mathbf{R}_c^{-1}(\mathbf{c} - \mathbf{c}_0) \quad (3)$$

then the best estimate of the orientation is $\tan^{-1}(y_a/x_a)$.

Figure 4 shows the predicted orientations vs the actual orientations for the training sets for each of the models. It demonstrates that equation 2 is an acceptable model of parameter variation under rotation.

5 Tracking through wide angles

We can use the set of models to track faces through wide angle changes (full left profile to full right profile). We use a simple scheme in which we keep an estimate of the current head orientation and use it to choose which model should be used to match to the next image.

To track a face through a sequence we locate it in the first frame using a global search scheme similar to that described in [3]. This involves placing a model instance centred on each point on a grid across the image, then running a few iterations of the AAM algorithm. Poor fits are discarded and good ones retained for more iterations. This is repeated for each model, and the best fitting model is used to estimate the position and orientation of the head.

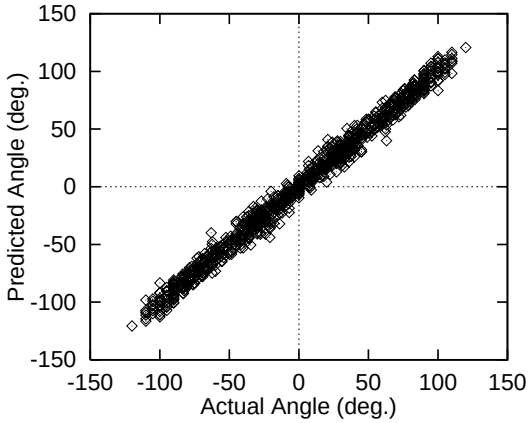


Figure 4. Prediction vs actual angle across training set

Model	Angle Range
Left Profile	-110^o - -60^o
Left Half-Profile	-60^o - -40^o
Frontal	-40^o - 40^o
Right Half-Profile	40^o - 60^o
Right Profile	60^o - 110^o

Table 1. Valid angle ranges for each model

We then project the current best model instance into the next frame and run a multi-resolution seach with the AAM. We estimate the head orientation from the results of the search, as described above. We then use the orientation to choose the most appropriate model with which to continue. Each model is valid over a particular range of angles, determined from its training set (see Table 1). If the orientation suggests changing to a new model, we estimate the parameters of the new model from those of the current best fit. We then perform an AAM search to match the new model more accurately. This process is repeated for each subsequent frame, switching to new models as the angle estimate dictates.

When switching to a new model we must estimate the image pose (position, within image orientation and scale) and model parameters of the new example from those of the old. We assume linear relationships which can be determined from the training sets for each model, as long as there are some images (with intermediate head orientations) which belong to the training sets for both models.

Figure 7 shows the results of using the models to track the face in a new test sequence (in this case a previously unseen sequence of a person who is in the training set). The model reconstruction is shown superimposed on frames from the sequence. The methods appears to track well, and is able to reconstruct a convincing simulation of the sequence.

We used this system to track 15 new sequences of the people in the training set. Each sequence contained between 20 and 30 frames. Figure 5 shows the estimate of the angle from tracking against the actual angle. In all but one case the tracking succeeded, and a good estimate of the angle is obtained. In one case the models lost track and were unable to recover.

The system currently works off-line, loading sequences from disk. On a 450MHz Pentium III it runs at about 3 frames per second, though so far little work has been done to optimise this.

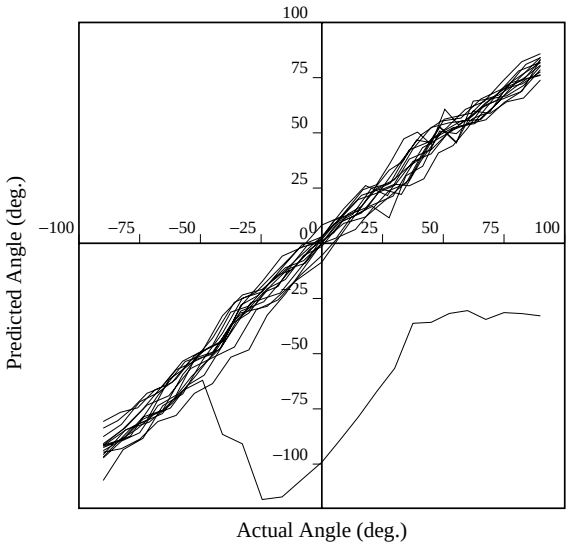


Figure 5. Comparison of angle derived from AAM tracking with actual angle (15 sequences)

6 Predicting Unseen Views

Given a single view of a new person, we can find the best model match and determine their head orientation. We can then use the best model to synthesize new views at any orientation that can be represented by the model. If the best matching parameters are $\mathbf{c}$, we use equation 3 to estimate the angle, θ. Let $\mathbf{c}_{res}$ be the residual vector not explained by the rotation model, ie

$$\mathbf{c}_{res} = \mathbf{c} - (\mathbf{c}_0 + \mathbf{c}_c \cos(\theta) + \mathbf{c}_s \sin(\theta)) \qquad (4)$$

To reconstruct at a new angle, α, we simply use the parameters

$$\mathbf{c}(\alpha) = \mathbf{c}_0 + \mathbf{c}_c \cos(\alpha) + \mathbf{c}_s \sin(\alpha) + \mathbf{c}_{res} \quad (5)$$

This only allows us to vary the angle in the range defined by the closest model. Since the models all represent the same 3D structure, we anticipate that there will be correlations between parameters for different views of the same individual. To do this effectively we must first project out the effects of pose, lighting etc. A principled approach to this is described in [2]. However, for our experiments, since there is little lighting or expression change in the training set, it is sufficient just to remove the orientation components.

In order to learn the relationship between parameters in one model and those in another, we perform the following steps. For each frame in the training set we use equation 4 to determine the orientation independent component of the parameters for each model. We then compute the mean of such residuals for each person. Let $\hat{\mathbf{c}}_{i,j}$ be the mean of such residuals in the i^{th} model for the j^{th} person. By applying PCA to the means for a given model, we can find the projection, $\mathbf{P}_j$, into an 'identity' sub-space.

Let the projection of each mean in the subspace be

$$\mathbf{b}_{ij} = \mathbf{P}_j^T (\hat{\mathbf{c}}_{i,j} - \hat{\mathbf{c}}_j) \quad (6)$$

where $\hat{\mathbf{c}}_j$ is the mean of the means.

We can use linear regression to learn the relationship which maps each $\mathbf{b}_{ij}$ in the identity space of the j^{th} model to the corresponding mean $\mathbf{b}_{ik}$ in the identity space of the k^{th} model,

$$\mathbf{b}_{ij} = \mathbf{r}_{jk} + \mathbf{R}_{jk} \mathbf{b}_{ik} \quad (7)$$

Thus to reconstruct a new view of a person given a match in a different view;

1. remove the effects of orientation (Eq.4),
2. project into the identity sub-space for the model (Eq.6),
3. project across into the subspace of the target model (Eq.7),
4. project that into the residual space (inverting Eq.6)
5. add the appropriate orientation (Eq. 5).

Figure 6 demonstrates this. Models were built on the data for all but one person. The profile model was then matched to a profile image of the missing person (the reconstruction is shown). The method described above is then used to predict the appearance using the frontal model at two different angles. For comparison, corresponding images of the person at similar angles are shown. Given the small nature of the training set (in this case only 14 people, yielding a 13-D identity space), the results are encouraging.

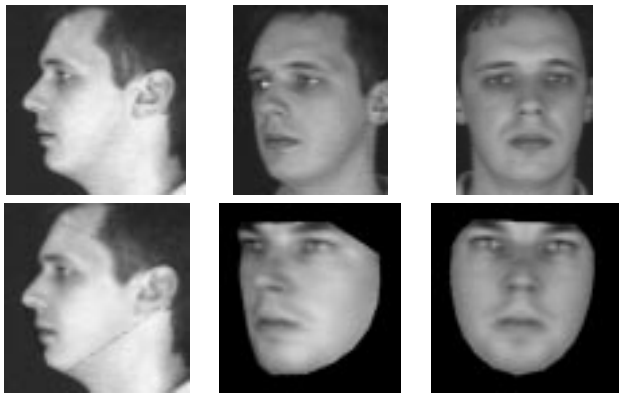


Figure 6. The best fit with a profile model is projected to the frontal model to predict new views

7 Discussion and Conclusions

We have demonstrated that a small number of view-based statistical models of appearance can represent the face from a wide range of viewing angles. Although we have concentrated on rotation about a vertical axis, rotation about a horizontal axis (nodding) could easily be included (and probably wouldn't require any extra models for modest rotations). We have shown that the models can be used to track faces through wide angle changes, and that they can be used to predict appearance from new viewpoints given a single image of a person.

So far we have only tested the methods on a relatively small and clean data set. We intend to gather more data in order to obtain better generalisation ability, to include expression and lighting changes and to investigate its performance on more cluttered backgrounds. We hope to obtain better calibrated training images in order to obtain more accurate angle estimates.

We anticipate the approach will be useful in many applications, including driving animated avatars, calculating head pose and making face recognition systems more invariant to viewing angle.

References

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In H.Burkhardt and B. Neumann, editors, 5^{th} *European Conference on Computer Vision*, volume 2, pages 484–498. Springer, Berlin, 1998.

[2] N. Costen, T. F. Cootes, G. J. Edwards, and C. J. Taylor. Automatic extraction of the face identity sub-

space. In T. Pridmore and D. Elliman, editors, 10^{th} *British Machine Vison Conference*, volume 1, pages 513–522, Nottingham, UK, Sept. 1999. BMVA Press.
[3] G. Edwards, T. F. Cootes, and C. J. Taylor. Advances in active appearance models. In 7^{th} *International Conference on Computer Vision*, pages 137–142, 1999.
[4] G. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In 3^{rd} *International Conference on Automatic Face and Gesture Recognition 1998*, pages 300–305, Japan, 1998.
[5] D. Graham and N. Allinson. Face recognition from unfamiliar views: Subspace methods and pose dependency. In 3^{rd} *International Conference on Automatic Face and Gesture Recognition 1998*, pages 348–353, Japan, 1998.
[6] H.Murase and S. Nayar. Learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, pages 5–25, Jan. 1995.
[7] M. J. Jones and T. Poggio. Multidimensional morphable models : A framework for representing and matching object classes. *International Journal of Computer Vision*, 2(29):107–131, 1998.
[8] J. Kwong and S. Gong. Learning support vector machines for a multi-view face model. In T. Pridmore and D. Elliman, editors, 10^{th} *British Machine Vison Conference*, volume 2, pages 503–512, Nottingham, UK, Sept. 1999. BMVA Press.
[9] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.
[10] T. Maurer and C. von der Malsburg. Tracking and learning graphs and pose on image sequences of faces. In 2^{nd} *International Conference on Automatic Face and Gesture Recognition 1997*, pages 176–181, Los Alamitos, California, Oct. 1996. IEEE Computer Society Press.
[11] B. Moghaddam and A. Pentland. Face recognition using view-based and modular eigenspaces. In *SPIE*, volume 2277, pages 12–21, 1994.
[12] S. Romdhani, S. Gong, and A. Psarrou. A multi-view non-linear active shape model using kernel pca. In T. Pridmore and D. Elliman, editors, 10^{th} *British Machine Vison Conference*, volume 2, pages 483–492, Nottingham, UK, Sept. 1999. BMVA Press.
[13] J. Sherrah, S. Gong, and E. Ong. Understanding pose discrimination in similarity space. In T. Pridmore and D. Elliman, editors, 10^{th} *British Machine Vison Conference*, volume 2, pages 523–532, Nottingham, UK, Sept. 1999. BMVA Press.
[14] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.
[15] T. Vetter. Learning novel views to a single face image. In 2^{nd} *International Conference on Automatic Face and Gesture Recognition 1997*, pages 22–27, Los Alamitos, California, Oct. 1996. IEEE Computer Society Press.

Figure 7. Reconstruction of tracked faces superimposed on sequences