A Generic Method for Evaluating Appearance Models

Abstract

Generative models of appearance have been studied extensively as a basis for image *interpretation by synthesis*. Typically, these models are statistical, learnt from sets of training images. Different methods of representation and training have been proposed, but little attention has been paid to evaluating the resulting models. We propose a method of evaluation that is independent of the form of model, relying only on the generative property. The evaluation is based on measures of model *specificity* and model *generalisation ability*. These are calculated from sets of distances between synthetic images generated by the model and those in the training set. We have validated the approach using Active Appearance Models (AAMs) of face and brain images, showing that both measures worsen monotonically as the models are progressively degraded. Finally, we compare three distinct automatic methods of constructing appearance models, and show that we can detect significant differences between them.

1 Introduction

Interpretation by synthesis has become a popular approach to image interpretation, because it provides a systematic framework for applying rich knowledge of the problem domain. Active Appearance Models (AAMs) [1, 2] are typical of this approach. There are two essential components: a generative model of appearance, and a method for searching the model space for the instance that best matches a given target image. In this paper we concentrate on the first of these.

Many generative models of appearance are statistical in nature, derived from sets of training images. AAMs use models that are linear in both shape and texture. Their construction relies on finding a dense correspondence between images in the training set, which can be based on manual annotation or on an automated approach (see below). Other approaches to constructing appearance models include methods based on non-linear manifolds in appearance space [3] and kernel PCA [4]. In the remainder of the paper we restrict our attention to AAMs, but the methods presented could be applied to any generative appearance model.

There has been relatively little previous work on model evaluation. One approach is to test a complete interpretation-by-synthesis framework, providing an implicit evaluation of the models themselves. This requires access to ground truth, allowing interpretation errors to be quantified [1, 8]. The most serious weakness of this approach is that it confounds the effects of model quality and the behaviour of the search algorithm. The need for ground truth data is also undesirable, because it is labour intensive to provide and can introduce subjective error.

We propose a method for evaluating appearance models, that uses just the training set and the model to be evaluated. This builds on the work of Davies et al [6], who tackled the simpler problem of evaluating shape models. Our approach is to measure, directly, the similarity between the distribution of images generated by the model, and the distribution of training images. We define two measures: *specificity* – the overlap of the distribution of model-generated images with the distribution of training images, and *generalisation ability* – the overlap of the distribution of training images with the distribution of model-generated images. We validate the approach by generating progressively degraded models, demonstrating that both specificity and generalisation also degrade, monotonically. We compute the sensitivity of the two measures, showing that sensitivity is a better measure of model quality and then apply the method to a real model evaluation problem.

2 Background

2.1 Statistical Models of Appearance

Statistical models of shape and appearance (combined appearance models) were introduced by Cootes, Edwards, Lanitis and Taylor [1, 2], and have since been applied extensively (eg [14, 11, 10]). The construction of an appearance model depends on establishing a dense correspondence across a training set of images using a set of landmark points marked consistently on each training image.

Using the notation of Cootes [2], the shape (configuration of landmark points) can be represented as a vector $\mathbf{x}$ and the texture (intensity values) in a shape-normalised frame represented as a vector $\mathbf{g}$.

The shape and texture are controlled by statistical models of the form:

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

where $\mathbf{b}_s$ are shape parameters, $\mathbf{b}_g$ are texture parameters, $\bar{\mathbf{x}}$ and $\bar{\mathbf{g}}$ are the mean shape and texture, and $\mathbf{P}_s$ and $\mathbf{P}_g$ are the principal modes of shape and texture variation respectively. Since shape and texture are often correlated, this can be taken into account in a combined statistical model of the form:

$$\begin{aligned} \mathbf{x} &= \bar{\mathbf{x}} + \mathbf{Q}_s \mathbf{c} \\ \mathbf{g} &= \bar{\mathbf{g}} + \mathbf{Q}_g \mathbf{c} \end{aligned} \tag{2}$$

where the model parameters $\mathbf{c}$ control the shape and texture simultaneously and $\mathbf{Q}_s$, $\mathbf{Q}_g$ are matrices describing the modes of variation derived from the training set. The effect of varying one element of $\mathbf{c}$ for a model built from a set of face images is shown in Figure 1.

2.2 The Correspondence Problem

A key step in building a combined appearance model is that of establishing a dense correspondence across the set of training images. In practice, this is often achieved by marking up the training set manually with a set of key landmarks and interpolating between them. Recently there has been considerable interest in automating this process. One approach

Figure 1: The effect of varying the first model parameter of a facial appearance model by ± 2.5 standard deviations.

is to use non-rigid registration methods, developed for use in medical image analysis, to align the images by optimising a measure of image similarity [11, 14]. An alternative approach refines an initial estimate of correspondence so as to code the training set of images as efficiently as possible [5]. Twining et al have recently described an approach based on optimising the total description length of the training set, using the model [16].

In section 4.1 we validate our approach to model evaluation by deliberately perturbing the correspondences in models built using manual annotation to establish correspondence. In section 4.3 we use our method of evaluation to compare models built using non-rigid registration [11, 14] and the minimum description length groupwise registration approach of Twining et al [16].

3 Appearance Model Evaluation

Our approach to model evaluation is based on measuring, directly, key properties of the model. This approach is based on the work of Davies et al [6], who defined specificity and generalisation ability for shape models. To be effective, a model needs the ability to generate a broad range of examples of the class of images that have been modelled. We refer to this as *Generalisation* ability. Although this property is necessary , it is not sufficient. We also require that the model can only generate examples that are consistent with the class of images modelled. We refer to this as *Specificity*. We define both of these measures by comparing the distribution of training images and the distribution of images generated using the model. An overview of the approach is given in Figure 2. Any image can be considered as a point in a high-dimensional space (defined by its intensity values). The training set forms a cloud of points in such a space. If we sample from the model, we generate a second cloud of points in this space. For an ideal model, the two clouds are coincident. We define *Generalisation* and *Specificity* in terms of the distance from each training image to the nearest model-generated image, and the distance from each model-generated image to the nearest training image respectively. We discuss the choice of an appropriate distance metric in section 3.3.

3.1 Generalisation

The Generalisation ability of a generative appearance model measures the extent to which it is able to represent images of the modelled class both seen (in the training set) and unseen (not in the training set). A model that comprehensively captures the variation in the

modelled class should generate a distribution of images that overlaps the training distribution as completely as possible. This means that, if we generate a large set of synthetic images, $\{I_\alpha : \alpha = 1,\ldots m\}$, from the model, each image in the training set should be close to a synthetic image. Given a measure, $|\cdot|$, of the distance between images, we define the Generalisation G of a model and its standard error, σ_G, as follows:

$$G = \frac{1}{n}\sum_{i=1}^{n} \mathbf{min}_\alpha |I_i - I_\alpha|, \tag{3}$$

$$\sigma_\mathbf{G} = \frac{SD(\mathbf{min}_\alpha |I_i - I_\alpha|)}{\sqrt{n-1}}, \tag{4}$$

where I_i is the i^{th} training image, $\min_\alpha$ is the minimum over α (the set of *synthetic* images), and SD is standard deviation. That is, Generalisation is the average distance from each training image to its nearest neighbour in the synthetic image set. A good model exhibits a low value of Generalisation, indicating that the modelled class is well-represented by the model.

3.2 Specificity

The Specificity of a generative appearance model measures the extent to which images generated by the model are similar to those in the training set. A specific model should generate a distribution of images that overlaps the training distribution as completely as possible. If we take a synthetic image set such as that defined previously, $\{I_\alpha : \alpha = 1,\ldots m\}$, each synthetic image should be close to an image in the training set. We define the Specificity, S, and its standard error, σ_S, as follows:

$$S = \frac{1}{m}\sum_{\alpha=1}^{m} \mathbf{min}_i |I_i - I_\alpha|, \tag{5}$$

$$\sigma_\mathbf{S} = \frac{SD(\mathbf{min}_i |I_i - I_\alpha|)}{\sqrt{m-1}}. \tag{6}$$

That is, Specificity is the average distance from each synthetic image to the nearest training image. A good model exhibits a low value of Specificity, indicating that it generates synthetic images, all of which are similar to those in the training set.

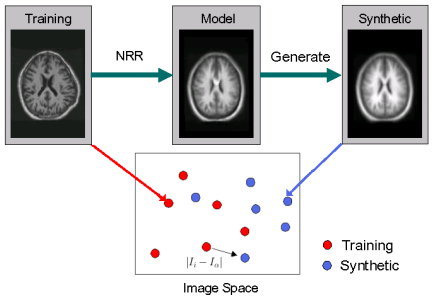

Figure 2: A simplified representation of the model evaluation approach.

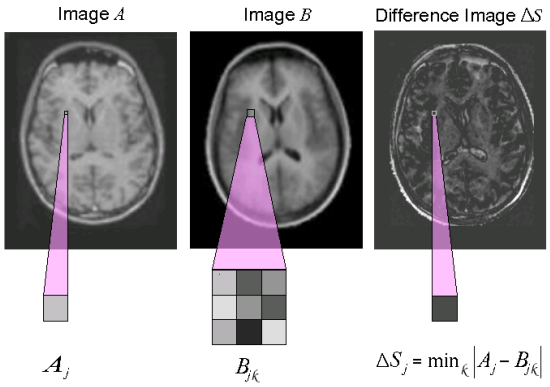

Figure 3: Calculating the shuffle difference image

3.3 Measuring Distances Between Images

The most straightforward way to measure the distance between images is to evaluate the mean absolute difference between them, or alternatively treat them as vectors by concatenating pixel/voxel values and take the Euclidean distance. Although this has the merit of simplicity, it does not provide a very robust distance measurement because it is very sensitive to small image misalignments. Robustness can be enhanced by considering a 'shuffle distance', inspired by the 'shuffle transform' [15]. The idea is to seek correspondence with a wider area around each pixel. Instead of taking the mean absolute difference between exactly corresponding pixels, we take each pixel in one image in turn, and compute the *minimum* absolute difference between it and pixels in a *shuffle neighbourhood* of the exactly corresponding pixel in the other image to produce a shuffle difference image $\triangle S$ (see Figure 3). The shuffle distance is given by $\sum_j \triangle S_j$ where $\triangle S_j$ are the elements of $\triangle S$. This approach is less sensitive to small misalignments, and provides a more robust measure of image distance. The sensitivity to misalignment is determined directly by the size and shape of the shuffle neighbourhood. One obvious choice is a square box around the corresponding pixel, but this is inherently anisotropic. Instead, we consider a shuffle disc, of radius r, which contains all pixels within a distance r of the central pixel.

Figure 4 shows examples of shuffle distance between an original image and a mis-

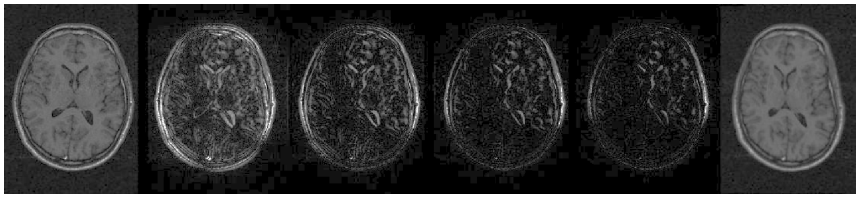

Figure 4: Shuffle distance calculation: **Left:** original image, **Right:** warped image, **Centre, from left to right:** shuffle difference images for $r = 1$ (abs. diff.), 1.5, 2.1 & 3.7 respectively.

aligned version evaluation, for varying values of the radius r. The effect of the shuffle neighbourhood radius on the sensitivity to misalignment is obvious as the contribution to distance perceivably decreases in areas of limited misalignment, as we go from $r = 1$ to $r = 3.7$ (roughly equivalent to a 7×7 square window).

4 Experimental Evaluation

We demonstrate the proposed approach to model evaluation in two stages. Firstly, a set of validation experiments are performed in which the behaviour of specificity and generalisation ability are observed for a deliberate and controlled degradation of a set of appearance models. The approach is then applied to the problem of choosing an optimal non-rigid registration algorithm for automatic construction of appearance models.

4.1 Validation

The purpose of the validation experiment was to establish if our measures of Specificity and Generalisation were able to detect a known model degradation. We also wished to investigate the effect of varying shuffle radius. Experiments were performed using two very different data sets. The first consisted of equivalent 2D mid-brain T1-weighted slices obtained from 3D MR scans of 36 subjects. In each of the images, a fixed number (167) of landmark points were positioned manually on key anatomical structures (cortical surface, ventricles, caudate nucleus and lentiform nucleus), and used to establish a ground-truth dense correspondence over the entire set of images, using locally affine interpolation. The second consisted of 68 frontal face images with blacked out backgrounds (to avoid biasing the distance measurements), with ground truth correspondence defined using 68 landmark points positioned consistently on the facial features in each image.

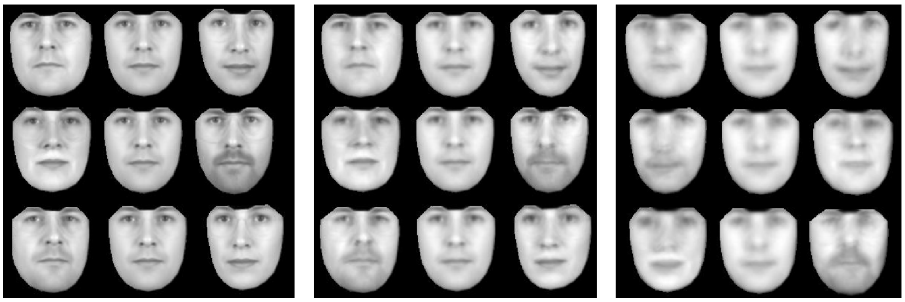

Figure 5: **Left:** Model constructed from ground-truth annotation. **Centre and right:** models constructed with increasingly degraded registration. Variation of $\pm 2.5\sigma_0$ about the mean in first three modes.

The first 3 modes of variation of the the face model built using the ground-truth correspondence are shown in Figure 5 (left). Keeping the shape vectors defined by the landmark locations fixed, smooth pseudo-random spatial warps, based on biharmonic Clamped Plate Splines (CPS) were then applied to the training images. The warps were controlled by sets of 25 randomly placed knot-points, each displaced in a random direction by a distance drawn from a Gaussian distribution. The relationship between the

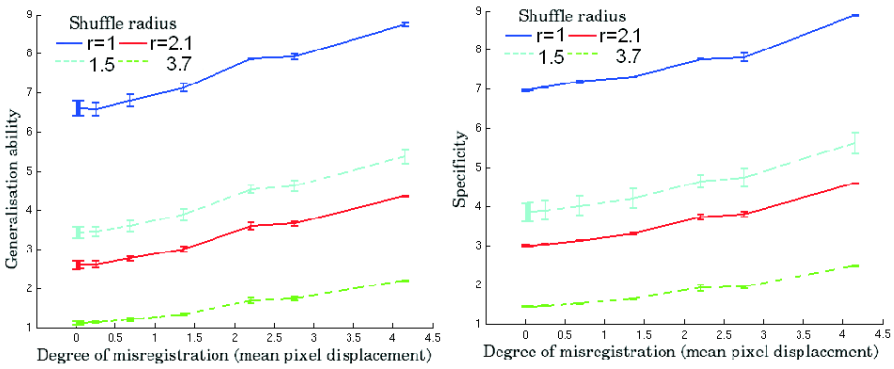

Figure 6: Specificity and Generalisation of degraded brain models.

mean of the displacement distribution and the mean pixel displacement for the whole image was carefully calibrated. This allowed a controlled misregistration to be introduced by changing the parameters of the displacement distribution.

By increasing the warp magnitude, successively increasing mis-registration was achieved. The mis-registered training images were used to construct degraded versions of the original model. Figure 5 (centre and right) shows the models obtained using progressively degraded training data. Models degraded using a range of values of the mean pixel displacement (from the correct registration) were evaluated using the method described in section 3. The image distances used were Euclidean distance (r = 1) and three different values of shuffle radius $r = 1.5$, 2.1 and 3.7. In each case, $m = 1000$ images were synthesised using the first 10 modes of the model, and Specificity and Generalisation were then estimated.

Results for the brain data are shown in Figure 6. Each point represents the average of 10 random instantiations of the perturbing warps. The results for the face data are similar, and shown in 7, but they are based on a single instantiation of each warp, which results in more noisy data. As expected, Specificity and Generalisation both degrade (increase in value) as the mis-registration is progressively increased. In most cases there is a monotonic relationship between Specificity/Generalisation and model degradation, but this is not the case when Euclidean distance is used. Note that there is a measurable difference in both measures, even for fairly small perturbations to the initial registration (Figure 5 (center)).

4.2 Sensitivity

It is useful to compare the performance of different measures of model quality. For a given measure, the level of model degradation that can be detected in the validation experiments described above depends on both the change in the value of the measure as a function of model degradation and the uncertainty in the value. To quantify this, we define the *sensitivity* D of a measure as follows.

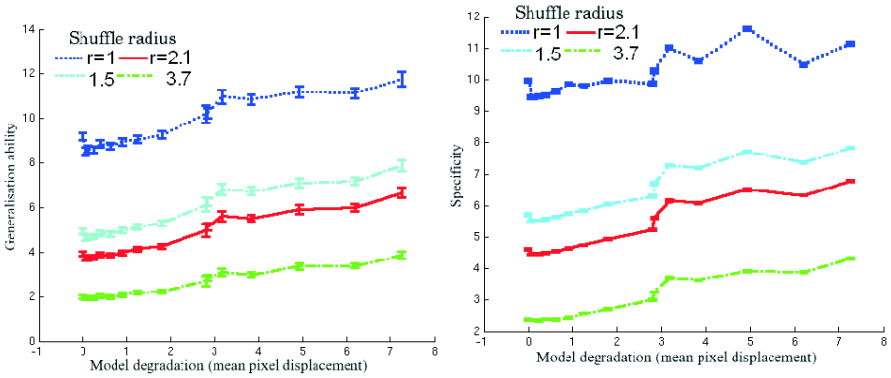

Figure 7: Specificity and Generalisation (with error bars) of degraded face models.

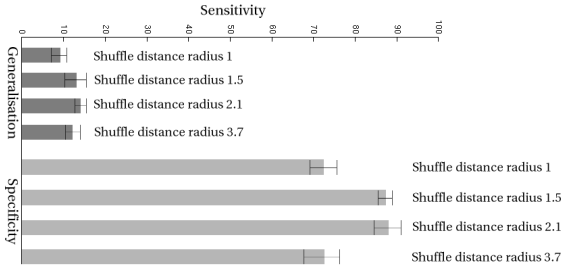

Figure 8: The sensitivity of Specificity and Generalisation

$$D(m,d) = \frac{1}{\bar{\sigma}} \left(\frac{m(d) - m(0)}{d} \right), \tag{7}$$

where $m(d)$ is the value of the measure for some degree of degradation d, $\overline{\sigma}$ is the mean error in the estimate of m over the range. $D(m,d)$ is reciprocal of the change in d required for $m(d)$ to change by one noise standard error, which indicates the lower limit of change in quality d which can be detected by the measure. Sensitivities for the specificity and generalisation for different values of shuffle radius are shown in Figure 8. These results demonstrate that specificity is a more sensitive measure of model quality than generalisation, and that the use of shuffle distance improves the sensitivities of both measures over those obtained using Euclidean distance.

4.3 Application to Model Evaluation

We used our new method to evaluate three different models built using an enlarged set of the brain data containing 104 affine aligned images. It has been shown previously [14, 11] that an appearance model can be built by registering each image in a set (pairwise) to a reference image. In [16] it was argued that a 'groupwise' approach which took proper account of the whole set of images might be expected to perform better. We built three

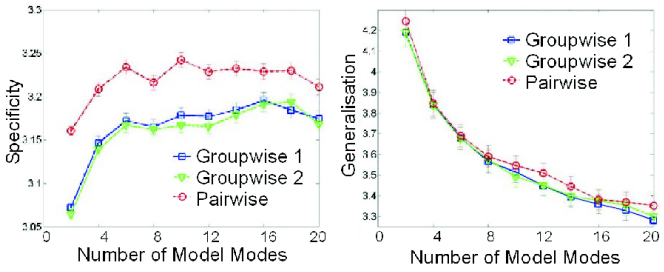

Figure 9: Specificity and Generalisation of the three automatic model construction approaches.

models, one using the pairwise approach, and two variants of our groupwise approach. The results, including the effect of including different numbers of modes in the models, are shown in Figure 9 and demonstrate a clear advantage in terms of both Specificity and Generalisation for both groupwise methods over the pairwise approach. It was not possible to discriminate between the two groupwise methods.

5 Summary and Conclusions

We have introduced an objective method of assessing appearance models, that depends only on the model to be tested and the training data from which it was generated. Validation experiments, based on perturbing correspondences obtained using ground truth, show that, using specificity in particular, we are able to detect small changes in model quality (due to sub-pixel displacements) reliably over a wide range of misregistration values. The results obtained for different sizes of shuffle neighbourhood show that the use of shuffle distance rather than Euclidean distance ensures monotonicity and increases the sensitivity of the method. We have also shown that the approach is capable of detecting statistically significant differences between models based on different approaches to automated model building. We believe that this work makes a valuable contribution, by providing an objective basis for comparing different methods of constructing generative models of appearance.

References

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Proceedings of European Conference on Computer Vision*, 2:484-498, 1998.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:681-685, 2001.

[3] Y. Li, S. Gong, and H. Liddel. Constructing facial identity surfaces in a nonlinear discriminating space. In *Proceedings of Computer Vision and Pattern Recognition*, pages 258-263, 2001.

[4] S. Romdhani, S. Gong, and A. Psarrou. A multi-view nonlinear active shape model using kernel PCA. In *Proceedings of the British Machine Vision Conference*, pages 483-492, 1999.

[5] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:1380-1384, 2004.

[6] R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton, and C. J. Taylor. A minimum description length approach to statistical shape modeling. *IEEE Transactions on Medical Imaging*, 21(5):525-537, 2002.

[7] T. F. Cootes, C. Beeston, G. J. Edwards, and C. J. Taylor. A unified framework for atlas matching using active appearance models. In *Proceedings of Information Processing in Medical Imaging (IPMI)*, 1613:322-333, 1999.

[8] T. F. Cootes, P.Kittipanya-ngam. Comparing variations on the Active Appearance Model algorithm. In *Proceedings of the British Machine Vision Conference 2002*, Vol 2:837-846.

[9] G. J. Edwards, T. F. Cootes, and C. J. Taylor. Face recognition using active appearance models. In *Proceedings of European Conference on Computer Vision*, 2:581-595, 1998.

[10] M. B. Stegmann, B. K. Ersboll, and R. Larsen. FAME - a flexible appearance modeling environment. *IEEE Transactions on Medical Imaging*, 22(10):1319-1331, 2003.

[11] D. Rueckert, A. F. Frangi, and J. A. Schnabel. Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration. *IEEE Transactions on Medical Imaging*, 22(8)1014-1025, 2003.

[12] T. F. Cootes, S. Marsland, C. J. Twining, K. Smith, and C. J. Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. In *European Conference on Computer Vision*, 2034:316-27, 2004.

[13] W. R. Crum, T. Hartkens, and D. L. G. Hill. Non-rigid image registration: theory and practice. *British Journal of Radiology*, 77:140-153, 2004.

[14] A. F. Frangi, D. Rueckert, J. A. Schnabel, and W. J. Niessen. Automatic construction of multiple-object three-dimensional statistical shape models: application to cardiac modelling. *IEEE Transactions on Medical Imagin*g, 21:1151-66, 2002.

[15] K. N. Kutulakos. Approximate N-view stereo. In *Proceedings of European Conference on Computer Vision*, 1:67-83, 2000.

[16] C. J. Twining, T. F. Cootes, S. Marsland, S. V. Petrovic, R. S. Schestowitz, and C. J. Taylor. A unified information-theoretic approach to groupwise non-rigid registration and model building. In *Proceedings of Information Processing in Medical Imaging (IPMI),* 3565:1-14, 2005.