A Generic Method for Evaluating Appearance Models

Anonymous CVPR submission

Paper ID 249

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 the 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. The approach is validated using Active Appearance Models (AAMs) of face and brain images, and shows that these measures both degrade 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.

Generative models of appearance are generally statistical in nature, and derived from training sets of images. AAMs use models that are linear in 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 [8, 1]. 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 also 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) 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 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 [14, 11]. An alternative approach refines an initial estimate of correspondence so as to code the training set of images as efficiently as possible [5]. We 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 a method of evaluation to compare models built using non-rigid registration [14, 11] and the minimum description length groupwise registration approach of Twining et al.

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 it's 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. We discuss the choice of an appropriate distance metric in section 3.3.

3.1. Generalisation

Generalisation of a model defines its ability to generalise to or represent well images of the modeled class both seen (in the training set) and unseen (not in the training set). A model that comprehensively captures the variation in the modeled class of object should be close, i.e. exhibit low distance, to all the images from that class). In practice this means that all the training examples used to construct the model should be close to model distribution sampled by the model-generated synthetic examples. Given the framework defined for evaluation of specificity above, i.e. a large set of synthetic example images sampled from the model $\{I_j : j = 1..m\}$ and a measure of the distance between images $|\cdot|$, Generalisation G of a model and the standard error on its measurement σ_G can be defined as follows:

$$G = \frac{1}{n} \sum_{i=1}^{n} min_j \, |I_i - I_j|, \tag{3}$$

$$\sigma_{\mathbf{G}} = \frac{SD(min_j \, |I_i - I_j|)}{\sqrt{n-1}}, \tag{4}$$

i.e. it is the average distance from each training image to its nearest neighbour in the image set generated by the model. Once again, good models exhibit low values of Generalisation indicating that the modelled class is well-represented by the model.

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

3.2. Specificity

Specificity of an appearance model defines its ability to generate realistic, new examples of the modelled class. A model that correctly describes the variation within an object class should be able to produce new examples of the class that would appear realistic compared to the original training set used to create the model. Conversely, a degraded model would be unable to articulate the main modes of object appearance and would only produce new examples disparate from the original training set. This definition is used to practically measure Specificity. Specifically, given $\{I_j : j = 1..m\}$ as a large set of synthetic example images sampled from the model and having the same distribution, Specificity S is defined as the average distance between each of the synthetic examples and it closest neighbour in the original training set:

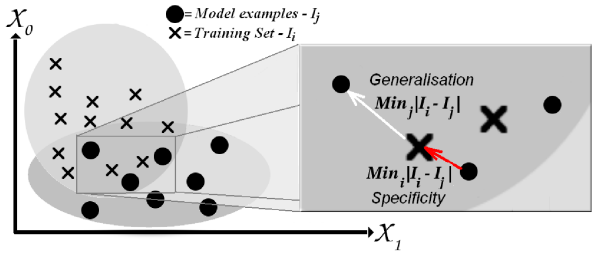

Figure 2. Hyperspace representation of the model (metric) evaluation approach

$$S = \frac{1}{m} \sum_{j=1}^{m} min_i |I_i - I_j|. \quad (5)$$

where I_i is the îth training image, $|\cdot|$ describes the distance between two images and SD is the standard deviation. Equivalently, the standard error in the measurement σ_S is thus:

$$\sigma_{\mathbf{S}} = \frac{SD(min_j |I_i - I_j|)}{\sqrt{m-1}}. \quad (6)$$

Generally, for a good model the Specificity is low as the images generated by the model are similar (low distance) to original training examples. Conversely, as a model degrades the generated examples get further away from the training examples increasing the distance and consequently Specificity.

3.3. Measuring Distances Between Images

The most straightforward way to measure the distance between images is to evaluate the 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. In the context of model and image registration evaluation considered here, such direct measurement results in a distance that increases rapidly even for quite small image misalignments. Robustness can be added to the distance evaluation by considering a 'shuffle difference', 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 corresponding pixels, the mean of minimum absolute difference between each pixel in one image and pixels in a *shuffle neighbourhood* around the corresponding pixel in the other is used. This approach is less sensitive to small misalignments, and provides a more robust image distance evaluation. Furthermore, the sensitivity to misalignment is directly determined by the size and type of the shuffle neighbourhood. One obvious choice is a square box around the corresponding pixel. A more even treatment of the local region is provided by a shuffle disc, of radius r, that only considers pixels located within r of the central pixel. Examples of shuffle distance evaluation with varying r between two brain examples, the original image and missaligned version, are shown in Figure 3. The effect of the shuffle neighbourhood radius on the distance misalignment sensitivity is obvious as the distance perceivably decreases in areas of small missalignment and becomes less noisy as we go from $r = 0$ to $r = 3.7$ (roughly equivalent to a 7x7 square window).

4. Experimental Evaluation

The proposed model evaluation approach is demonstrated in two stages. In the first instance, a set of validation

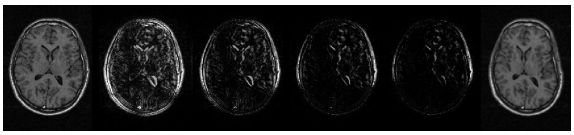

Figure 3. Shuffle distance evaluation: left - original image , right - warped image, centre from left: distance with r = 0 (abs diff), 1.5, 2.9 and 3.7

experiments are performed where the behaviour of the metrics is observed for a deliberate and controlled degradation of set of appearance models. The approach is then practically demonstrated on 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 intended to investigate the effect of varying shuffle radius. Experiments were performed using two very different data sets. The first consisted of corresponding 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.

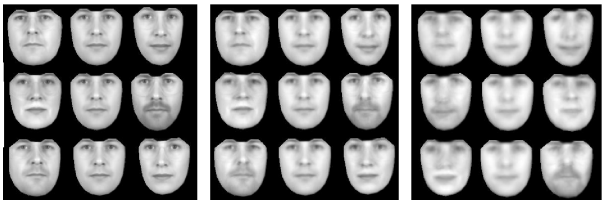

Figure 4. Model constructed from ground-truth annotation: left, and models constructed with increasingly degraded registration: centre and right (variation of $\pm 2.5\sigma_0$) in first three modes

The first 3 modes of variation of the the face model built using the ground-truth correspondence is shown in Figure 4(left). Keeping the shape vectors defined by the landmark locations fixed, smooth pseudo-random spatial warps, based on bi-harmonic Clamped Plate Splines (CPS) were then applied to the training images. The warps comprised 25 knot-points and the extent of these warps was carefully studied. 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 4(centre and right) shows the models obtained using progressively degraded training data.

Models degraded using a range of mean pixel displacement (from the correct registration) were evaluated using the method described in section 3, using Euclidean distance (r = 0) and three different values of shuffle radius r = 1.5, 2.9 and 3.7. In each case, $m = 1000$ images were synthesised using the first 10 modes of the model, and Specificity and Generalisation were estimated.

Results are shown for the brain data in Figure 5. The results for the face data are similar, as shown in 6, but they are based on a single instantiation rather than 10, which makes the curves worse beyond the range of a 3 pixel mean displacement. 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 metrics, even for fairly small registration perturbations (eg the model of 4(center)). The steepness of the curve in the results for Euclidean distance already suggests that the use of shuffle distance gives better results.

4.2. Comparison with Ground Truth?

Model and registration? Overlap?

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 (ground truth was not required for this experiment). 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] we argued that a 'groupwise' approach that took proper account of the whole group might be expected to perform better. We built three models, one using the pairwise approach, and two variants of our groupwise approach. The results, including different numbers of modes in the models, are shown in Figure 7 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 we are able to detect increasing model degradation reliably. The results ob-

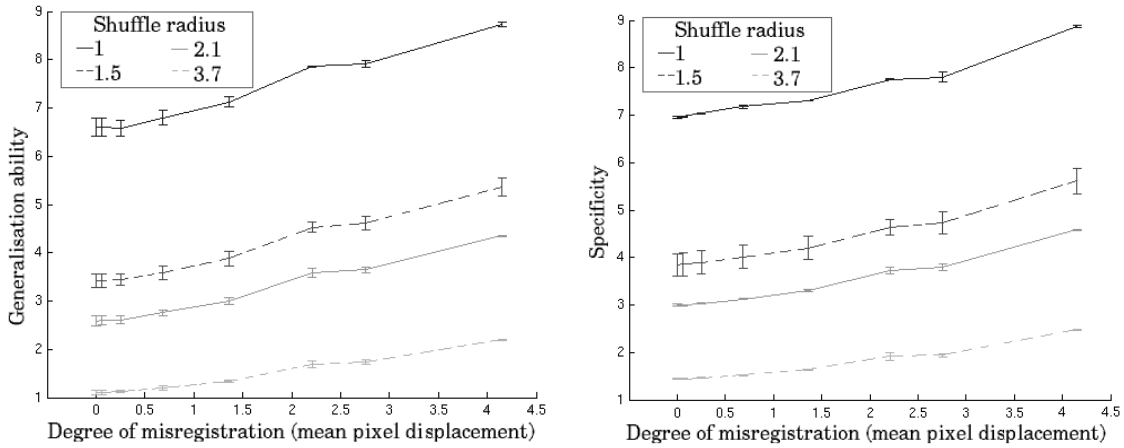

Figure 5. Specificity and Generalisation of degraded brain models.

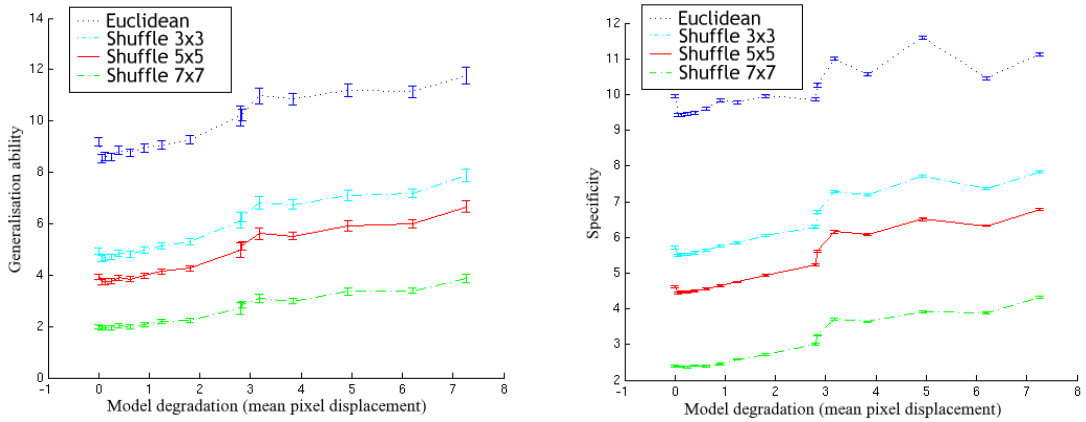

Figure 6. Specificity and Generalisation (with error bars) of degraded faces models.

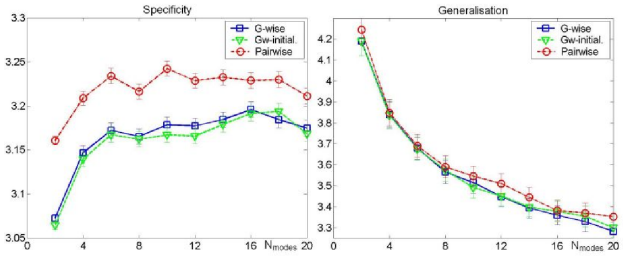

Figure 7. Specificity and Generalisation of the three automatic model construction approaches

tained 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.

Models and registration? Overlap?

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*, Lecture Notes in Computer Science 1613:322-333, 1999.

[8] T. F. Cootes, P.Kittipanya-ngam, Comparing Variations on the Active Appearance Model Algorithm. In *Proceedings of BMVC 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 Imaging*, 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. To be presented in *Information Processing in Medical Imaging*, 2005.

1 1 1 1 2 1, 2 1 1 1, 2, 4 1, 2, 4 3 2, 4