A Framework for Evaluation of Appearance Models - DRAFT

Abstract

Models of appearance are powerful tools for capturing data variability and they are capable of synthesising data. Such models have been shown to posses rich 'knowledge' of what data within a set comprises and the way such data can be decomposed and hence simplified. A framework was developed which is able to evaluate appearance model. It is able to tell apart models of varying quality, thereby promoting better algorithms for construction of appearance models. The method also allows the validation of models. By measuring 'distances' between images, it quantifies the proximity between a model and its data. To measure distance between images, a shuffle transform is used, which is robust. Two separate measures reflect on the quality of an entire model, given a large matrix of distances. Specificity measures how well data generated by the model fits data from which the model was constructed, whereas generalisation compares data from which the model was constructed and data generated by the model. The methods were shown to work well when applied to face data and MR brain data. In both cases, progressively perturbed models were correctly analysed by our measures. The framework was used to compare models of the brains, which were built automatically. Models which are known to be superior were merited by the framework.

1 Introduction

When approaching the problem of image interpretation, one has several paradigms for solving it. One such paradigm is the modelling of objects and the use of models to learn something about an object. Along this simple approach, shape models [3] were developed and they assisted in analysing *contours* of objects in an image. The natural extension to shape models was one which encapsulated intensity information, as well as contours. The work of Edwards et al. [4] brought about models which were capable of synthesising photo-realistic images.

Following the success of this approach, several groups have built appearance models using different methods and obtained results of differing quality. Stegmann [5], for example, reproduced algorithms for appearance model construction and extended them to account for 4-D models that include the dimension of time. Rueckert et al. [14] have taken this approach and embedded it in registration algorithms so that statistical deformation models are built subsequent to registration.

Davies et al. [2] have adopted a method for the evaluation of shape models, i.e. the derivation of values for a given shape descriptor. This was done by a minimum description length [16] approach, which relates to the simplicity of a model. Ever since, the method allowed evaluation and comparison between shape models – possibly built by different algorithms – to be compared. Furthermore, it enabled the formation of an information-theoretic objective function. Such as objective function, when treated an an optimisation problem, allowed optimal shape models to be constructed.

While methods of evaluation are available for shape models and, therefore, quantitative comparison is possible, none is available for appearance models. Since appearance

models are far more complex and far heavier than shapes (as they include texture information), a method has been thought for evaluating them and arguing about their validity.

This paper outlines a successful method for the evaluation of appearance models. The method is shown to be well-behaved and its applicability to faces and brains is illustrated. Furthermore, it is used to compare different methods of model construction, all of which do so without need for manual mark-up of the data. Finally, it is shown that the method is able to *correctly* distinguish between models that appear identical to the naked eye.

2 Background

0.1. Active Appearance Models

The task of image analysis, especially in the bio-medical domain, must take into consideration the variation in shape and appearance of objects. The invariant presumption is that corresponding objects in all images are of one particular class so we can typify the contents of the image by training an entity that captures inter-subject variation as well as atrophies.

Statistical analysis of shapes [3] which obtains a model of deformation goes back a decade ago. The principles were later extended to sample the variation in pixel intensities (also commonly referred to as textures) to create a model of full variation that is able to synthesise full appearances [4] and their successful application to medical data has been frequently demonstrated [5]. The correlations between shape and intensity are learned using Principal Component Analysis [6] where much of the power of these principles lies.

0.2. The Correspondence Problem

The integrity of models breaks down if correspondences, annotated in the form of spatial landmarks, are inappropriately identified. Furthermore, the annotation process involves a preliminary segmentation process which highlights parts of the data where landmarks can and should be placed. Although this has become a solved problem in statistical modeling of shape, it is yet difficult to select good landmarks in images which strive to retain full appearances rather than contours or surfaces solely. Several attempts have been made to resolve the issue [7, 8, 9], but none was optimal or even quite satisfactory. Alignment has become the means by which this crucial limitation can be solved and the foundations of image registration assist in establishing this alignment.

Figure x: Shuffle distance... First mode of appearance model, ± 2.5 standard deviations.

0.3. Image Registration

In the medical domain, one of the more fundamental problems is the requirement for the setting of images in a state which makes them appear collectively similar [10]. This greatly simplifies the analysis of a group of images which bear common information, as in the case of brain slices fusion or comparison of patient data, either acquired using different modalities or collected at different time instances.

The problem is trivial if the difference is a rigid one – a difference due to rotation, scale and translation. More realistically, the problem is far more complex and images are inconsistent (primarily in the case of inter-subject registration) so affine and non-rigid transformations are required. In the case of non-rigid registration, transformation is merely unbounded. However, to avoid corruption and distortion of constituent finer parts of the image, limitations to their freedom and certain conditions must be met. Clamped-plate splines (CPS), which are based on Green's function, have proven to be a useful family of warps, allowing for highly flexible manipulation of images. Their attributes are reminiscent of those developed by Lötjönen and Mäkelä [11].

To drive transformation in the right direction and attain convergence, minimisation of the difference perceived in the images must be pursued. To measure discrepancies, or contrariwise, the similarity between two images, mean of squared differences (MSD) or mutual information (MI) [12] are traditionally used as metrics although new techniques are perpetually introduced [13].

Overall, the process of registration comprises the transformation of images followed by similarity measures, where transformations are chosen to iteratively maximise that similarity. Conventionally, a reference is selected in the process [14], but our contention is that this need not be the case if an optimal solution is sought. The technique according to which the registration problem will be solved is entirely described by the objective function as Section 4 illustrates.

3 Experiments

3.1 Shuffle Distance

Change illustrative figures that are the same so that they appear distinct.

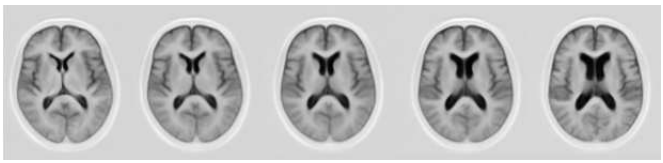

Figure x:

Similar to MICCAI

Cannot tell the difference of the 15 models (neither can you?, send 16 pictures). Algorithm can. Added CPS from 0 to 15. Existing CPS stayed in place, maybe try more in future and see plots trend.

Figure x: First mode of 5 cps

3.2 Model Evaluation

In order to measure the quality...

Figure x: First mode of 10 cps

Figure x: First mode of 15 cps

3.3 Normalisation

$$S_{Total} = \frac{S}{S_0},\ G_{Total} = \frac{G}{G_0} \tag{1}$$

$$S_{\sigma_{Total}} = \sqrt{\frac{S_0^2\sigma_s^2 + S^2\sigma_{s0}^2}{S_{s0}^4}} \tag{2}$$

$$G_{\sigma_{Total}} = \sqrt{\frac{G_0^2\sigma_g^2 + G^2\sigma_{g0}^2}{G_{g0}^4}} \tag{3}$$

where:

- S: Specificity using training set
- G: Generalisation using training set
- S_0: Specificity using pseudo training set
- G_0: Generalisation using pseudo training set
- σ_s: standard error of S
- σ_g: standard error of G
- σ_{s0}: standard error of S_0
- σ_{g0}: standard error of G_0

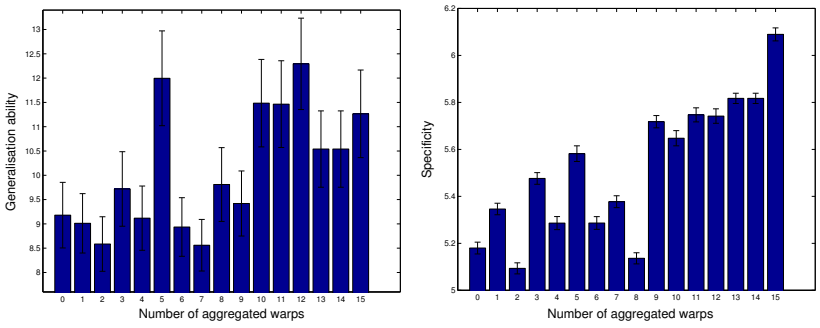

Figure x:

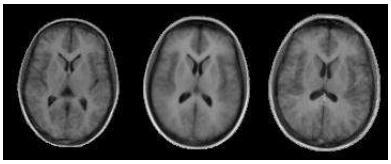

Figure x: Model built from correctly-annotated images

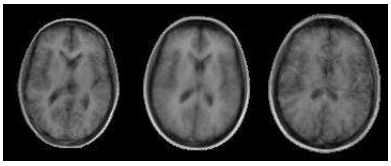

Figure x: Model built from correctly-annotated images with each image affected by 3 warps

4 Results

Experiments... then results...

Face landmark points have been perturbed by noise. The magnitude of noise is defined by a standard distribution and the value sigma of that distribution varies.

To put the algorithm in a challenging position, images were warped while landmark points remained the same. This obtained fuzzier models, but models that appear merely identical. To control the amount of noise, a progressively-increasing number of clamped-plate splines (REF) was applied to all images. All previous warped remained unchanged since the previous passes for stability, but since warps can improve models as well as degrading them, monotonous were not expected. The trend, however, was expected to show the movement of the images around the points resulted in worse models, as one would expect.

In the case of the brains, an even greater amount of warps was applied to see the effect in a larger scale. The data used was MR-..... brains obtained from.... aligned affinely and sliced....

5 Summary and Conclusions

The evaluation of appearance models becomes practical through the use of large sets of data. Fitting the many possible cross-pairings of data and model instance leads to measures which are robust independently of data properties. Results have been shown for brain data as well as a challenging set of face data and graphs have always appeared encouraging. The evaluation method relies upon a distance measure between a pair of images and measures such as shuffle distance appear most suitable to handle the task.

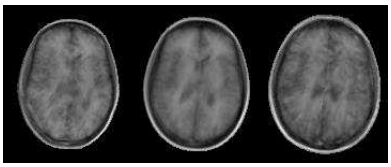

Figure x: Model built from correctly-annotated images with each image affected by 6 warps

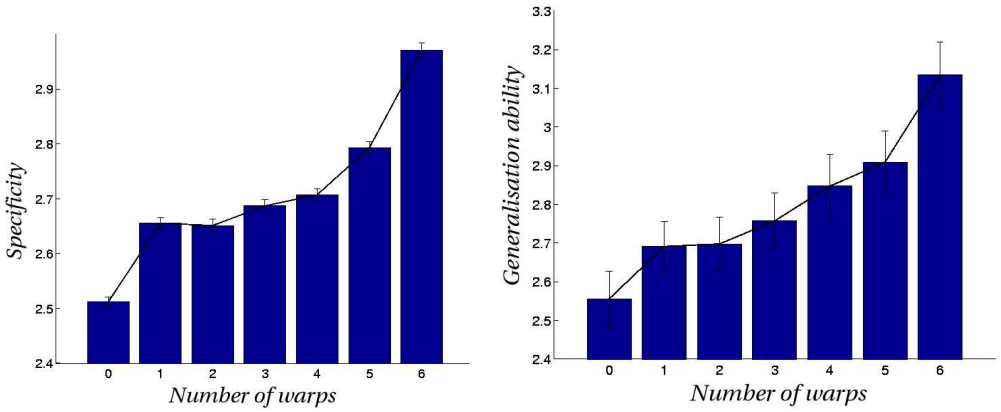

Figure x: Generalisability and Specificity of models with an increasing amount of deformation

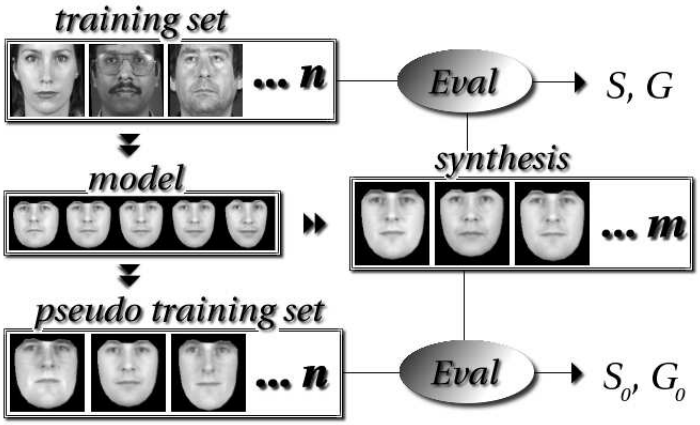

Figure x: The framework of normalisation. Two sets of training data, one which is real and one which is synthesised are compared to model-generated examples.

The method able to compare similar models and distinguish between them successfully. This was shown to be the case when models were corrupted intentionally, but also in cases where models and their quality were poorly understood. In such circumstances, the method was able to provide answers and be used for benchmarking. It appears to suggest that registration in a group-wise manner results in better models of appearance. This opens the door to a framework which validates registration. The observation which motivates it is that correct registration identifies the correspondence perfectly, and therefore builds optimal models.

References

[1] S. Marsland, C. J. Twining, and C. J. Taylor. Groupwise non-rigid registration using polyharmonic clamped-plate splines. In *proceedings of MICCAI 2003*, pages 771-779, Montreal, Canada, 2003.

[2] 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.

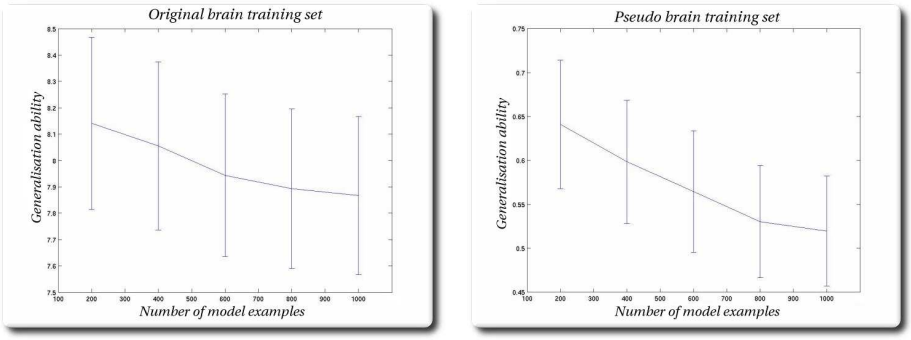

Figure x: Comparison between the Generalisation ability derived from an original data set and a pseudo training set, as function of the number of model examples.

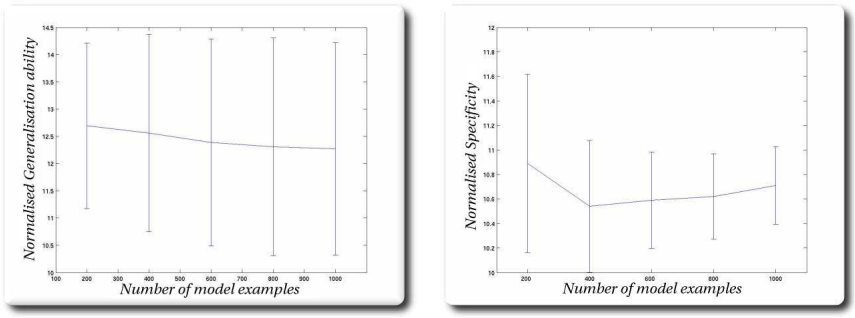

Figure x: Normalised Generalisation ability and Specificity as a function of the number of model examples.

[3] 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.

[4] 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.

[5] 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.

[6] I. T. Joliffe. Principal component analysis. In *Springer Series in Statistics*, Springer, New York, 1986.

[7] A. D. Brett and C. J. Taylor. A method of automated landmark generation for automated 3D PDM construction. *Image and Vision Computing*, 18(9):739-748, 2000.

[8] K. N. Walker, T. F. Cootes, and C. J. Taylor. Automatically building appearance models from image sequences using salient features. *Image and Vision Computing*, 20(6):435-440, 2002.

[9] A. Hill, C. J. Taylor, and A. D. Brett. A framework for automatic landmark identification using a new method of nonrigid correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3):241-251, 2000.

[10] J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes. Medical image registration. Boca Raton, Fla. ; London: CRC Press, 2001.

[11] J. Lötjönen and T. Mäkelä. Elastic matching using a deformation sphere. In *Proceedings of MICCAI 2001*, pages 541-548, 2001.

[12] C. Studholme, D. L. G. Hill, and D. J. Hawkes. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32(1):71-86, 1999.

[13] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22(8):986 - 1004, 2003.

[14] D. Rueckert, A. F. Frangi, and J. A. Schnabel. Automatic construction of 3-D statistical

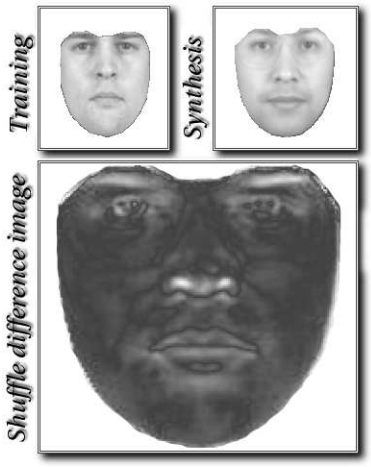

Figure x: Example of shuffle distance...

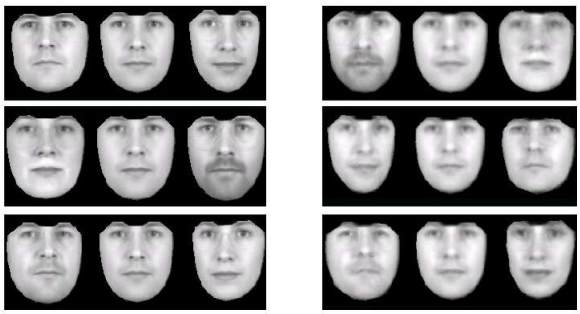

Figure x: The 3 Principal modes of variation of a correct face model (left) and the corresponding model whose landmarks have been perturbed by Gaussian noise of 3 standard deviations.

deformation models of the brain using nonrigid registration. *IEEE Transactions on Medical Imaging*, 22(8)1014-1025, 2003.

[15] S. K. Warfield, J. Rexilius, P. S. Huppi, T. E. Inder, E. G. Miller, W. M. Wells, III, G. P. Zientara, F. A. Jolesz, and R. Kikinis. An entropy measure to assess nonrigid registration algorithms for statistical atlas construction. In *Proceedings of MICCAI 2001*, pages 266-274,2001.

[16] J. R. Rissanen. Stochastic complexity in statistical inquiry. In *World Scientific Series in Computer Science*, Singapore, 1989.

[17] A. C. W. Kotcheff and C. J. Taylor. Automatic construction of eigenshape models by genetic algorithm. In *Information Processing in Medical Imaging*, 1997.

[18] P. Viola and W. M. Wells. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24:137-154, 1997.

[19] M. Rabbani and R. Joshi. An overview of the JPEG 2000 still image compression standard. *Signal Processing: Image Communication*, 17:3-48.

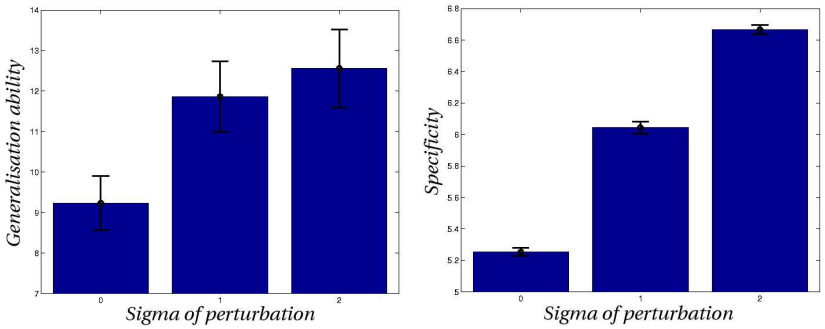

Figure x:

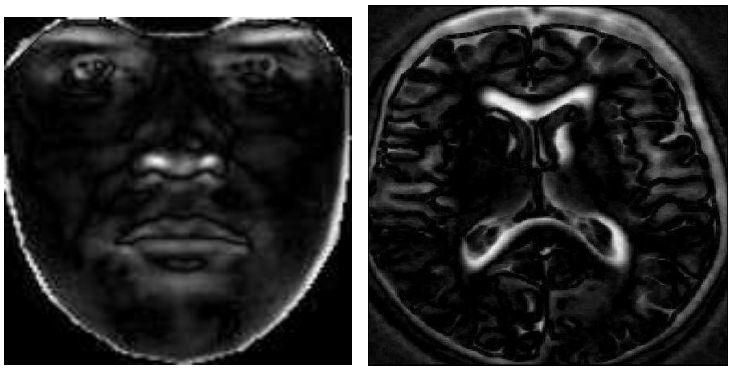

Figure x: Two face images and the resulting shuffle difference image with neighbourhood of size 3×3 shuffle Figure x: Shuffle difference image of the brain 3×3.

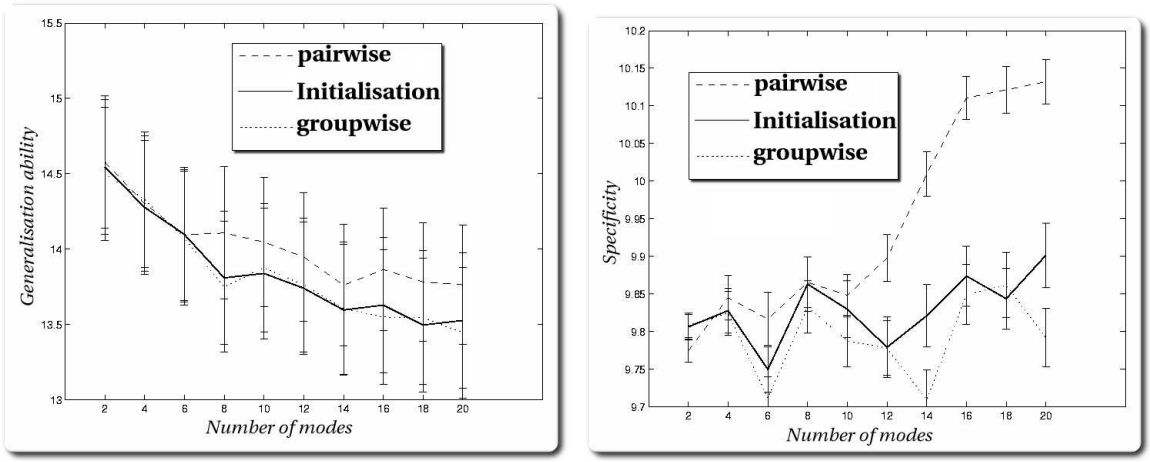

Figure x: