Evaluating Brain Registration using Models of Appearance

Abstract

Appearance models are an applicable approach to the analysis of anatomical variability. They are able to distinguish between groups, e.g. normal and diseased, as a model encapsulates the properties of a group from which it was derived. The construction of such models is closely-related to the task of registration and it requires one-to-one correspondence, which registration is able to obtain. We developed a framework which evaluates both appearance models and registration, based on the statistics of large sets of images. The framework is capable of distinguishing between good models of the brain and worse ones. Furthermore, it provides a method of validating the models and evaluating registration. It does so by measuring how well a model and its (potentially registered) data fit together. Two measures are defined which reflect on the quality of a model. The first of these – specificity – approximates the level to which data generated by the model fits data from which the model was constructed. The complementary measure – generalisation – is able to quantify 'distance' between data from which the model was constructed and model-generated data. Results show that as models degrade in quality, their specificity and generalisation ability rise, as expected. The algorithms are used to compare models of the brains, which were built *automatically* by independent registration approaches. This greatly helps in identifying better model construction algorithms, which are analogous to registration algorithms. The algorithm is purely data-driven and requires no manual annotation.

1. Introduction

One powerful method for the modelling of anatomy was introduced by Edwards et al. [4] and it is known as appearance models – a natural successor to shape models [3]. This method's prerequisite is a large enough set of data, which is representative of a population and ideally spans its full variability. Appearance models are able to *learn* what characterises inter-subject or intra-subject changes and determine the prominence of the main characteristics. Hence, it is able to identify changes and derive a model that encapsulates change – all in a data-driven manner.

Non-rigid image registration is ubiquitously used as the basis for analysis of medical images. The results of registration can be used for structural analysis, atlas matching, and analysis of change. Methods for obtaining registration are are well-established and quite uniform in nature. A goal is achieved by warping pairs of images so that they appear more similar. The similarity leads to overlap, which allows corresponding structures to be identified. This problem is complementary to that of modelling groups of images. Statistical models of a group of images need dense correspondence to be defined across the group; non-rigid registration provides exactly that.

Ever since the emergence of appearance models, attempts have been made to reproduce and improve it. To name a few such efforts, Stegmann [5] built

4-dimensional cardiac appearance models and Reuckert et al. [14] derived statistical deformation models from several registrations of the brain. Models have been built in a variety of ways and what is yet lacked is the ability to compare them. It becomes clear from experience that attempts to distinguish between them by eyesight are hopeless. More recently, appearance models were built automatically using piece-wise affine registration [16]. Evaluation of models in this particular case enabled evaluation of registration algorithms.

The approach we present is based on the observation that one of the things one can do with a registered set of image is build a statistical model. So, our proposal is that you can measure the quality of registration in terms of statistical models quality.

The idea of evaluating models was successfully exampled. Davies et al. [2] explored the evaluation of *shape* models and ultimately developed a robust framework for the task. This paper outlines a principled approach to the evaluation of *appearance* models, which is a challenging task since their complexity is far higher. The approach is shown to be reliable in evaluation of brain models[1] and, more importantly, it is then used to *evaluate registration algorithms*, from which appearance can be been derived. The way evaluation has been done so far is by deliberately warping the same image and observing whether the registration algorithm helped in recovering the same answer. The other alternative was to make use of ground truth, though in this the method we present, no such knowledge is needed. Evaluation of registration only requires the registered data and the entire process is automatic.

2. Background

The paper unifies ideas from models of appearance and image registration. It proposes a measure that does not only characterise model quality, but is also capable of reasoning about registration. These two strands – modelling and registration – inherently work towards achieving the same goal. Registration of a group of images leads to correspondence, which then forms a model of shape and intensity. The model can be used inversely to argue about the quality of its seminal registration. The important concepts are explained in the remainder of this section.

2.1. 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 over a decade ago. The principles were later extended to sample the variation in pixel intensities (also commonly referred to as textures) and create a model of full variation (see Fig. 1). That model 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 the principles actually lies.

The integrity of models breaks down if correspondences, annotated in the form of spatial landmarks, are inappropriately identified. Furthermore, the annotation

[1] Examples from non-medical domains are beyond the remit of this paper, though they have been successful as well.

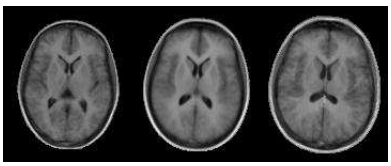

Fig. 1. The first mode of a brain appearance model. ± 2.5 standard deviations are shown.

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.

2.2. 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 must be forced. 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) [19, 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 the entire groups of images should be accounted for when an optimal (correct) solution is sought.

2.3. Model Evaluation

Shape models were previously evaluated using the two measures named specificity and generalisation ability (generalisability in short). Model complexity measures were initially investigated by Kotcheff [18] and further use was made of them in the work of Davies where shape models were optimised.

The idea behind this reverts back to fundamentals where models of visual forms in fact describe clouds in a high-dimensional space (visualised in Fig. 2). A model

is essentially a descriptor of a volume in space – a mean point with knowledge about its extent of variation in the different directions.

If each model is in fact a simple cloud in space, models can be compared by measuring the overlap of clouds analytically. Models are evaluated by comparing them with respect to the training set, which are interchangeably just a set of points in that space. The models are best represented by a large collection of syntheses which are derived from them.

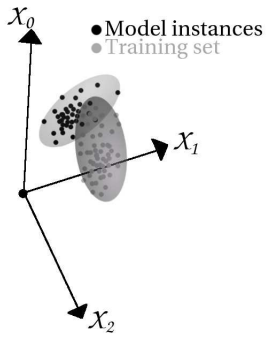

Fig. 2. Training set and model synthesis in hyperspace

3. Methodology

In order to measure the quality of a model, a measure was devised which reflects on its occupancy in a high-dimensional space. The method measures how well one cloud of points fits another. One cloud comprises the training set of the model whereas the other accommodates many model syntheses. A similar approach was used to evaluate shape models[2], thereby building optimal models of shape automatically.

The measures are based on distances between examples in two groups of instances. The first of these measures, namely specificity, is intended to discover how well a model describes its training set. The other – generalisation ability, indicates how closely the training set fits within the model. To represent the model, many synthetic instances need to be generated from it. The safe assumption is that a large enough number of instances is capable of approximating the model's behaviour.

The shuffle transform is a robust filter that can measure difference between images that have localised discrepancies. Shuffling was used to derive images which are reminiscent of the originals, but highlighted regions of inconsistency.

Shuffle distance is defined to be the minimum value for every pixel with respect to a group of pixels in its vicinity in a second image (see Fig. 4).

To calculate 'distance' between two images, the me an intensity of the resulting shuffle distance image should be taken. This simple idea has proven to be fast, as well as powerful.

Let $\{I_a(X_0) : a = 1, ...\eta\}$ be a large image set which has been generated from the model and has the same distribution as the model. The distance between two images is described by $|\cdot|$ and this allows us to define:

[2] The performance peaked when a full minimum description length framework was utilised.

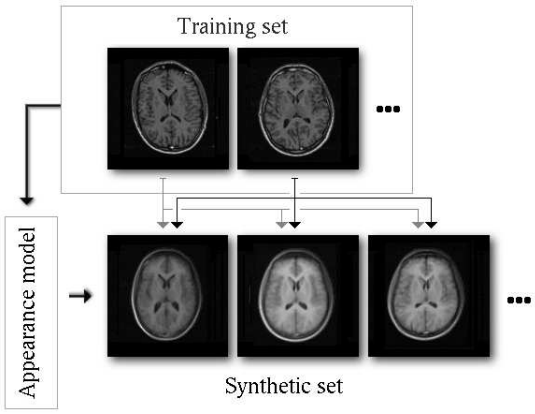

Fig. 3. The model evaluation framework. Each image in the training set is compared against all model-generated syntheses

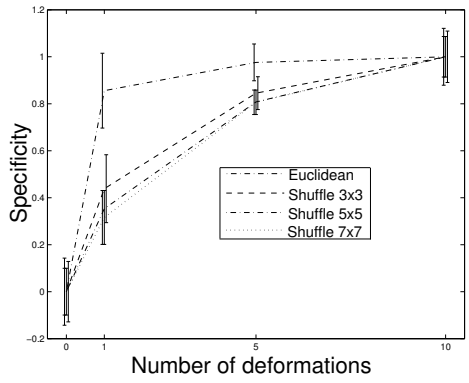

Fig. XXX. Varying number of warps are applied to the training set of a model whose specificity is then evaluated using Euclidean distance and shuffle distance with varying window sizes.

$$\text{Generalisation ability: } G = \frac{1}{N} \sum_{i=1}^{N} min_{(w.r.t.\ a)} \left(|I_{\tau_i}(X_0) - I_a(X_0)|\right),$$

$$\text{Specificity: } S = \frac{1}{\eta} \sum_{i=1}^{\eta} min_{(w.r.t.\ i)} \left(|I_{\tau_i}(X_0) - I_a(X_0)|\right).$$

Algorithmically, a method can then be devised for evaluating models. It is based on generalisation ability and specificity and it consists of the following steps:

- ➤ Generate a set of model syntheses (I_{syn})
- ➤ For all images in the training set (I_{model}):
  - ✦ Pre-process the image if necessary, e.g. resize, crop
  - ✦ For all image generated from the model:
    - ◆ Pre-process the image if necessary
    - ◆ Calculate the shuffle distance between I_{syn} and I_{model} and record it in an appropriate location of a matrix of size $I_{syn} \times I_{model}$
- ➤ Using the equations above, derive specificity and generalisation ability from the matrix

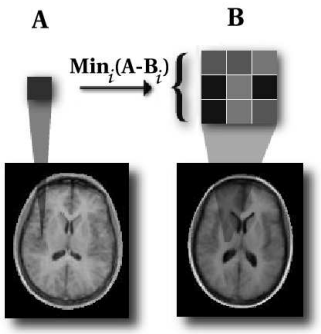

Fig. 4. Illustrating the essence of a shuffle distance image

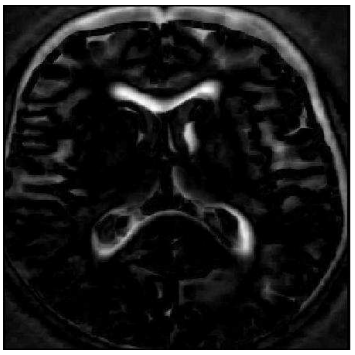

Fig. 5. An example of the shuffle distance image when applied to two brains

4. Experiments and Results

4.1. Validity of the Method

To satisfy ourselves that the method behaves properly, we demonstrated that a hand-annotated model gets assigned fixed values for specificity and generalisation. Noise was then applied to the annotation, the model re-built and as a result of that noise, values of specificity and generalisability measures were negatively affected. Noise that was applied to the mark-up resulted a steady rise in these value (indicating exacerbation) so even less trivial experiments were embarked upon.

To established even more confidence in the evaluation criterion, subtle changes were made to the *images* rather than the mark-up. To do so, images were transformed locally while the landmark points upon them remained unchanged. This means that correspondences will be badly affected and lose their meaning. By applying a different number of warps at each stage (while keeping older warps in place), landmark points should *usually* be located at worse positions. Hence, the model representing all images under consideration is expected to be aggravated. Shown below in Fig. 6 (also confer Fig. 1 where the corresponding correct model is included) are models created from images that were subjected to a varying number of small, localised clamped-plate spline warps. It can be seen that the model becomes more fuzzy and less realistic as more warps are applied.

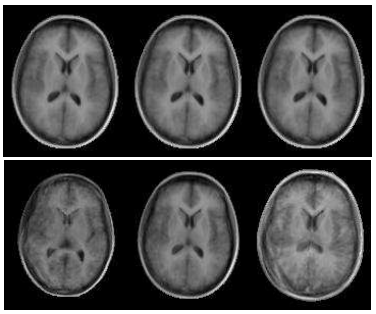

Fig. 6. The first mode of an appearance model of the brain whose training set was subjected to diffeomorphic warps – 15 warps for the model at the top and 30 for the one at the bottom. ± 2.5 standard deviations are shown.

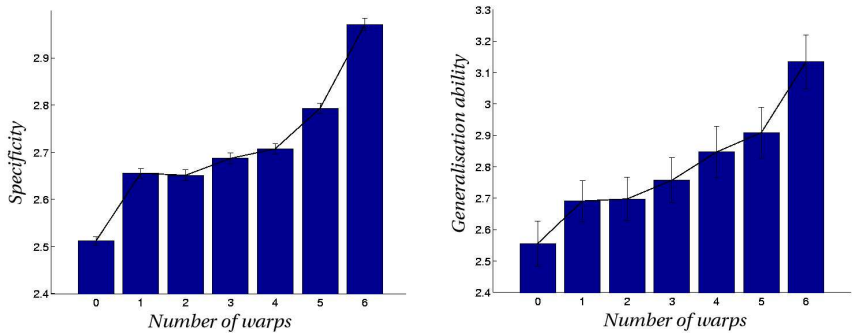

Fig. 7. On the left: The effect of image perturbation on specificity; On the right: The effect of image perturbation on generalisation ability.

4.2. Evaluating Registration

The method was used to learn about registration algorithms that often lack benchmark tools. Crum et al. [17] have devised validation methods based on overlap measures of markup. These can measure the quality of registration, however they require ground-truth annotation.

Fig. 8 shows an appearance model which was built automatically using group-wise registration. Group-wise registration account for the entire set when registering, whereas pair-wise register with respect to a single reference. The initialisation algorithm is an information-theoretic group-wise algorithm that is described in [16].

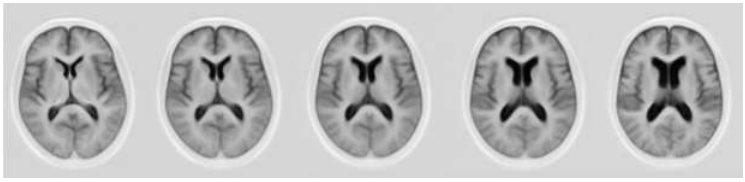

Fig. 8. Appearance model which was built automatically by group-wise registration. First mode is shown, ± 2.5 standard deviations.

It can be inferred from the results (Fig. 9) that group-wise registration outperforms pair-wise registration. This means that a group-wise registration leads to better model and is hence *better representative* of the group.

It was ensured that the comparison is not dependent upon the process where syntheses are extracted from the model. Since a finite number of modes is selected in synthesis, the range of 2-20 modes was investigated and the figure shows the average. Nonetheless, the properties of the bar charts did remain consistent throughout.

5. Discussion

Models of appearance can be evaluated in a simple principled way. Results presented throughout this paper illustrated the method's applicability to the case of brain model evaluation. Although results have not been shown for other data types, it is known to be a generic method that requires no tweaking and remains robust, regardless of various parameters. By fitting and comparing data sets in a

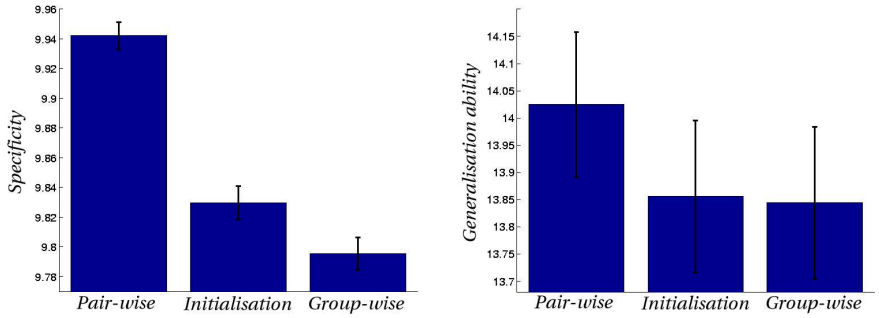

Fig. 9. Registration evaluation which compares 3 different registration algorithms. Specificity is shown on the left and generalisation ability on the right. Values are the mean over a wide range of modes in the model.

suitable manner, good models can be identified and evaluated quantitatively. For the task, one is better off using robust measures such as the shuffle transform, which is immune to small local discrepancies.

The method was used to evaluate models built by registration and it is therefore valuable as a benchmark tool. Furthermore, a provision emerges for validation of registration, assuming that optimal registrations lead to immaculate models.

Acknowledgements. The project is funded by the EPSRC and forms a part of the Medical Image and Signal (MIAS) Interdisciplinary Research Collaboration. The brain data was made available thanks to the Wellcome Trust and was generously contributed by Paul Bromiley.

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.

[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 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] 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.

[17] W. R. Crum, R. I. Scahill, and N.C. Fox. Automated hippocampal segmentation by regional fluid registration of serial MRI: Validation and application in Alzheimer's disease. *NeuroImage* 13:847-855, 2001.

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

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