Generalised Overlap Measures for Assessment of Pairwise and Groupwise Image Registration and Segmentation

William R. Crum[1], Oscar Camara[1], Daniel Rueckert[2], Kanwal K. Bhatia[2], Mark Jenkinson[3], and Derek L.G. Hill[1]

[1] Centre for Medical Image Computing, Dept. of Medical Physics, University College London, WC1E 6BT, UK
{b.crum, o.camara-rey, Derek.hill}@ucl.ac.uk
[2] Visual Information Processing, Department of Computing, Imperial College London, SW7 2BZ, UK
{dr, bhatia98}@doc.ic.ac.uk
[3] FMRIB Centre, John Radcliffe Hospital, University of Oxford, OX3 9DU, UK
mark@fmrib.ox.ac.uk

Abstract. Effective validation techniques are an essential pre-requisite for segmentation and non-rigid registration techniques to enter clinical use. These algorithms can be evaluated by calculating the overlap of corresponding test and gold-standard regions. Common overlap measures compare pairs of binary labels but it is now common for multiple labels to exist and for fractional (partial volume) labels to be used to describe multiple tissue types contributing to a single voxel. Evaluation studies may involve multiple image pairs. In this paper we use results from fuzzy set theory and fuzzy morphology to extend the definitions of existing overlap measures to accommodate multiple fractional labels. Simple formulas are provided which define single figures of merit to quantify the total overlap for ensembles of pairwise or groupwise label comparisons. A quantitative link between overlap and registration error is established by defining the overlap tolerance. Experiments are performed on publicly available labeled brain data to demonstrate the new measures in a comparison of pairwise and groupwise registration.

1 Introduction

Effective validation techniques are an essential pre-requisite for segmentation and non-rigid registration techniques to enter clinical use. Registration for medical applications seeks a mapping from one image (or set of images) to another such that structural or functional correspondence is achieved i.e. identifiable features or regions are correctly mapped between images. In medical image segmentation, the objective is to identify regions that have some functional or structural significance. If a pre-labeled image can be correctly registered to another image then the labeling problem is solved for that image. Conversely, if a pair of images contains corresponding labeled regions then the registration problem is at least partially solved by constructing a mapping between corresponding labels. Two common scenarios are that automatic image regis-

J. Duncan and G. Gerig (Eds.): MICCAI 2005, LNCS 3749, pp. 99 – 106, 2005.

tration has been performed on the basis of voxel intensity information or that a novel segmentation technique has been applied to an image. The problem with validating these cases at present is the paucity of metrics of quality, especially where the segmentation or registration technique generates fuzzy (i.e. fractional or probabilistic) labels or is evaluated over multiple test images. To date most evaluation has relied on the use of simple measures of regional overlap, defined for single labels, which assume that each voxel is either labeled or not labeled. In this paper we use results from fuzzy set theory and fuzzy morphology to extend existing definitions of overlap to (i) measure overlap of partial volume labels (ii) compute overlap measures for multiple labels defined on multiple image pairs (iii) compute overlap measures for groupwise registration and (iv) establish a link between measures of overlap and estimates of target registration error. Experiments are performed to establish the behavior of the new overlap measures and to compare pairwise and groupwise registration performed on publicly available data.

We consider an existing labeling (E) and a test labeling (T) that may have been obtained by a new segmentation technique or by using the result of image registration to map a label set from one image to another. The most obvious quantitative comparison of regions is by volume [1] however two labelings may have similar volumes but very different shapes, or even locations. The Hausdorff-Chebyshev metric defines the largest difference between two contours or surfaces but can be computationally expensive to compute and is not symmetric between E and T (although it can be made so). The Modified Williams Index has been developed for comparison of multiple expert observers boundaries against computer generated boundaries and is the ratio between the average computer-to-observer agreement and the average inter-observer agreement [2].

For a comparison of voxel-wise binary labelling the number of true and false positives and negatives can be determined and measures of region overlap can be computed. These are generally of the form:

$$O = \frac{N(T \cap E)}{N(T \cup E)}, \text{ or } \frac{N(T \cap E)}{N(E)} \text{ or } \frac{2N(T \cap E)}{N(T) + N(E)} \tag{1}$$

In equation 1 N(E) indicates the number of voxels which belong to the label E etc. In this work we will concentrate on the first (left-most) expression (the Tanimoto coefficient) although a similar development could be made with the other forms. These and other overlap measures are reviewed from the perspective of the so-called "overlapping area matrix" by Beauchemin and Thomson [3]. Measures of correspondence based on information theory have also been proposed [4]. For labels defined in a probabilistic fashion, Gerig [5] suggests a probabilistic overlap and has provided software to compute this and other overlap measures.

The work in this paper extends previous overlap definitions to cope with contemporary applications. The motivation is to develop regional overlap measures that can be intuitively pooled across labels and subjects to provide single figures of merit. Potential applications include assessment of registration of ensembles of subjects, particularly in group-wise ("target-less") applications. We also take the opportunity to formalize a previously suggested link between overlap measures and target registration errors in registration via the definition of the overlap tolerance, τ.

2 Methods

We first redefine the overlap measure for fuzzy labels. In equation 1, the overlap is obtained by summing the number of distinct voxels in the label intersection and union. However, both binary labels which have been interpolated following registration and those which model partial occupancy at a voxel can be characterized by a number [0, 1] at each voxel that defines the fraction of voxel that is labelled. Results from fuzzy set theory for the intersection and union of two fuzzy sets can immediately be applied to rewrite equation 1 to give an overlap, O_F which is a function of the label values at each voxel summed over all voxels, *i*, in the image.

$$O_F = \frac{\sum_{voxels,i} MIN(T_i, E_i)}{\sum_{voxels,i} MAX(T_i, E_i)} \tag{2}$$

Equation 2 computes the overlap of a single pair of fuzzy labels defined on a pair of images. The numerator and denominator can be accumulated across multiple labels and multiple image pairs to compute a single overlap figure, O_{PMF}, which describes the total overlap of a set of fuzzy labels defined on a set of image pairs. The overlap is the ratio of the total fuzzy intersection to the total fuzzy union.

$$O_{PMF} = \frac{\sum_{pairs,k} \beta_k \sum_{labels,l} \alpha_l \sum_{voxels,i} MIN(T_{kli}, E_{kli})}{\sum_{pairs,k} \beta_k \sum_{labels,l} \alpha_l \sum_{voxels,i} MAX(T_{kli}, E_{kli})} \tag{3}$$

In equation 3, α_l is a label-specific weighting factor that affects how much each label contributes to the accumulated overlap and β_k is a pair-specific weighting factor that affects the relative contribution of each image pair to the accumulated overlap. We defer a discussion of the possible values of α_l and β_k to section 2.2. There are now emerging techniques where a single groupwise registration is performed instead of multiple pairwise registrations. A groupwise overlap measure, O_{GMF}, can be constructed by considering permutations of image pairs within the group and applying equation 3 Note that there is no simple relationship between multiple overlaps evaluated separately using equation 2 and the ensemble overlap evaluated using equation 3.

2.1 Overlap and Target Registration Error: Overlap Tolerance

The overlaps described above do not consider the proximity of the non-overlapping label portions that may also provide important information. One previously proposed method of determining this proximity is the tolerance, τ [6]. The standard overlap measures can be considered to be $\tau=0$ since pairs of label voxels have to occupy the same space to be considered overlapping. However, $\tau > 0$ allows labels to be considered overlapping if they lie within τ mm of each other. Therefore as the tolerance increases, the fractional overlap $\rightarrow 1$ as the condition for overlapping voxels is relaxed. Previously this has been described for binary labels and integer values of the tolerance. We now define overlap for fractional labels and non-integer tolerances. Starting from the definition of overlap for fractional labels, O_F, (equation 2) the definition of fuzzy overlap to a tolerance,τ, can be written:

$$O_F(\tau) = \frac{\sum_{voxels,i} MAX\left(MIN\left((D_\tau T)_i, E_i\right), MIN\left(T_i, (D_\tau E)_i\right)\right)}{\sum_{voxels,i} MAX\left(T_i, E_i\right)} \quad (4)$$

In equation 4, D_τ is a fuzzy dilation operator that can be represented as a voxel mask of dilation coefficients centered on each voxel of interest with τ specifying the extent of the operator. In 1D, where the voxel dimension is 1mm for example, $D_0=\{1\}$, $D_1=\{1, 1, 1\}$, $D_2=\{1, 1, 1, 1, 1\}$ etc. When considering fractional tolerances then $D_\gamma=\{\gamma, 1, \gamma\}$, $D_{1+\gamma}=\{\gamma, 1, 1, 1, \gamma\}$ etc where $0 \le \gamma \le 1$. Then the fuzzy dilation applied at a single voxel located at the origin, L(0), in 1D can be written as $L^*(0) = MAX\,(D_\tau(i)L(i))$ where i is in the range $[-k, +k]$ and $k = (\text{int})(1+\tau)$. This definition is consistent with that of [7]: $D_v(\mu)(x) = \sup\{t[v(y-x), \mu(y)], y \in S\}$ where S is the image-space, ν is a structuring element, μ is a fuzzy set and x, y are both elements of S. t is a t-norm which in our case is simply defined as $t(a, b) = ab$.

Note that $O_F(\tau)$ is an increasing function of τ for $\tau \ge 0$.The maximum possible overlap given by equation 4 can be established by assuming that both D and T have fractional labels in the range [0, 1]. Then when $\tau >> 1$ the numerator reduces to $MAX(T_i, E_i)$ and the maximum overlap is therefore 1 as expected. Now consider a pair of misregistered images where every voxel is independently labeled and the same set of labels exists in each image but are not necessarily coincident. Then the overlap of any pair of labels can be computed as described above. For each labeled voxel in the target image the smallest tolerance, τ_1, for which the overlap with its partner in the source image is 1, can be computed. Then the map of tolerances is a map of target registration error. For labels spanning multiple voxels, τ_1 estimates the maximum displacement between non-overlapping voxels belonging to corresponding labels.

2.2 Parameter Choices

In equation 3, weights α and β were introduced to respectively define the relative contribution of labels and subjects to the overlap measures. With $\alpha=1$, all labels are implicitly weighted by their volume. This may not be desirable as smaller labels may represent a greater registration or segmentation challenge. Two alternative choices are to set α either to (i) the inverse mean volume of the current label pair to give all labels equal weighting or (ii) the inverse mean volume squared of the current label pair to weight by the inverse volume. We examine the effect of these different α values below. In the experiments reported in this paper we have set $\beta=1$ but β could be used to weight inversely with the variance of labeling accuracy.

3 Experiments A, B and C

Nine T1-weighted MR-brain images with labels from the Internet Brain Segmentation Repository[1] were used. Each image had ten binary anatomical labels, one for each of the following structures: amygdala, caudate, cerebellum, cortex, hippocampus, lateral ventricle, pallidum, putamen, thalamus and white matter.

[1] http://www.cma.mgh.harvard.edu/ibsr

Experiment A investigates the decrease in overlap measures in response to forced misregistration. Each of the nine images (the targets) was paired with a copy of itself (the sources) and the accumulated overlap measure was computed for each pair. Then a non-rigid misregistration algorithm, which acted to *reduce* an image similarity measure, was applied to each source image and associated labels using tri-linear interpolation. The mean misregistration displacement over each pair was computed together with the new accumulated overlaps. The overlaps were also computed for a single label, S constructed from the union of the ten labels above.

Experiment B examines the relationship between the overlap tolerance and the misregistration error. For each of the misregistered labels in experiment A including the combined label, S, the tolerance was found for which the overlap O_F was 0.99 and compared with the mean misregistration error computed for each label from the applied transformations.

Experiment C compares a group of pairwise registrations and a groupwise registration. The 9 images from experiment A were (i) each registered independently to a tenth image in a pairwise fashion using a B-spline approach [8] and (ii) registered in a groupwise fashion to a common reference frame representing the average shape of the population, also using a B-spline approach [9]. Both techniques overlay a mesh of uniformly spaced control points onto each of the images; deforming the control points deforms the underlying images. The control points are manipulated until the normalised mutual information is maximised. The groupwise technique does not use an explicit anatomical reference; instead an average shape is calculated implicitly by constraining the sum of all the deformations to be equal to zero using a Gradient Projection Method. The cases were compared by assuming that both had been performed in a groupwise fashion and computing the groupwise overlap measure, O_{GMF} by permuting and accumulating all possible pairwise overlaps as in equation 3.

4 Results

Experiment A. Fig. 1 plots the accumulated overlap against the mean applied misregistration for the three different label weightings, α, for each pair in experiment A.

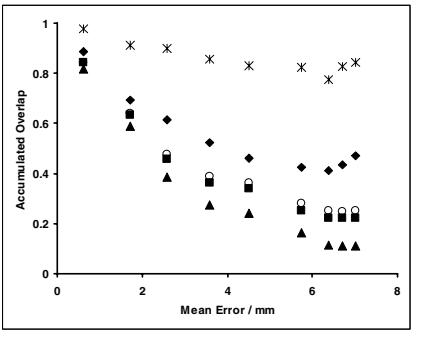

Fig. 1. The results of experiment A. The overlap accumulated over 10 labels plotted against the mean applied misregistration error for 9 image pairs. ♦=Volume-weighting, ■=Equal-weighting, ▲=Inverse Volume-weighting, * = Union Label S and ○ = Simple Average.

Also plotted are the overlap for the single union label S and the average overlap computed for each pair. It can be seen the trend is for the measured overlap to decrease as a function of increasing applied misregistration. There is a distinction between the 3 different weighting schemes with the inverse volume weighted overlaps decreasing fastest. This is to be expected, as the overlap of smaller regions will be more sensitive to misregistration. The average overlap for each pair is nearly coincident with the equally weighted accumulated overlap as expected. The overlap of large structures (as deduced from the volume-weighted and union plots) does not always strictly decrease with increasing misregistration. This result is considered in section 4.

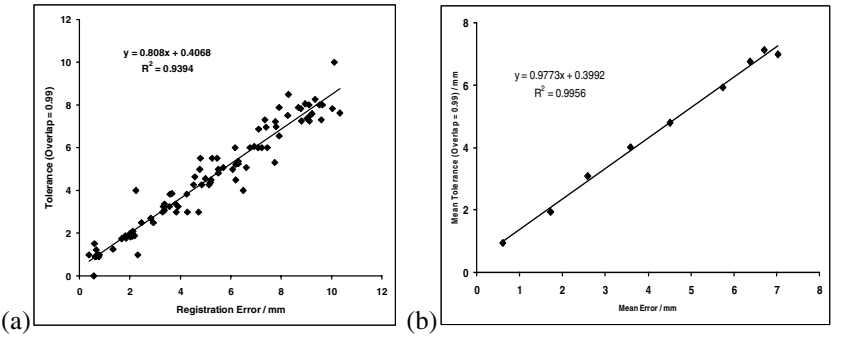

Fig. 2. Results for experiment B. (a) The smallest tolerance that gives an overlap of 0.99 against the applied misregistration error for each of 10 label-pairs on 9 subjects. (b) The results of (a) averaged over each subject.

Experiment B. Fig. 2a plots the tolerance computed iteratively using equation 4 for each label on each pair against the mean applied misregistration for that label. Figure 2b plots the tolerance averaged over all the labels on each pair against the mean applied misregistration calculated over the aggregated label S. In both cases there is a strong linear relationship between the tolerance and the applied misregistration.

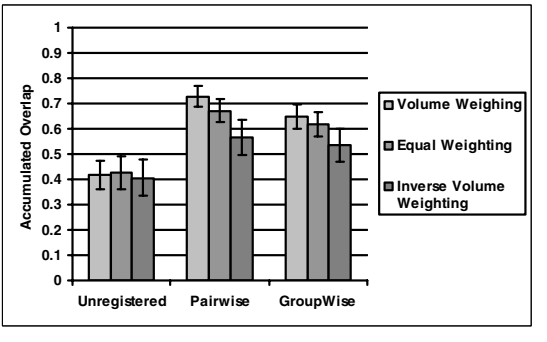

Fig. 3. Results for experiment C. The total accumulated overlap computed in a groupwise fashion for the unregistered, pairwise registered and groupwise registered images. Results are shown for volume-weighting, equal weighting and inverse volume-weighting between labels. Error bars represent the standard deviation of the overlap computed over all subject pairs.

Experiment C. Fig. 3 shows the groupwise overlap, O_{GMF}, for the unregistered, pairwise and groupwise registered images for volume weighting, equal weighting and inverse volume weighting between structures. The overlap consistently ranks the pairwise overlaps above the groupwise overlaps. While the pairwise registration was refined to a control point spacing of 2.5mm, the groupwise was only refined to a spacing of 5mm for computational efficiency. Therefore the scale of non-rigid deformations is restricted in the groupwise registration compared to the pairwise. Also the minimization of the groupwise cost function is more complex and therefore more likely to find local minima; this is an area of continuing research. A more interesting observation is that the inverse volume weighted overlaps are far more similar than the equal and volume weighted overlaps indicating that small structures are being registered more consistently. This is probably because the small, deep brain structures have relatively consistent anatomy whereas the larger structures such as cortex and white matter are known to vary significantly between individuals.

5 Discussion and Conclusion

We have developed overlap measures to allow comparison of multiple fuzzy labels defined on multiple subjects. The specific case of pairwise and groupwise registration has been considered here but these measures could also be applied to related problems of segmentation. We have re-introduced the idea of overlap tolerance and used it to relate registration error to overlap. We have demonstrated a linear relationship between the overlap tolerance and the applied misregistration in one experiment but this cannot be considered a completely general result. The misregistration acted normally to edge features so preferentially displaced high-contrast boundaries of structures rather than rearranging low-contrast features. The tolerance would be an insensitive indicator of registration error if the mis-registrations were occurring within labels; this is a property of labels rather than these overlap measures. Another application for the tolerance might be to initialize registration problems where labels exist on some parts of the image.

The framework presented in this paper allows single overlap measures that encompass multiple labels defined on multiple image pairs to be generated in a natural way. Weighting can be applied to prefer smaller labels and/or to accommodate other prior information about the images. Such evaluation tools are necessary for the clinical adoption of new registration and segmentation techniques.

Acknowledgements

We acknowledge the following support: Medical Images and Signals IRC (EPSRC GR/N14248/01 and MRC D2025/31) (WRC), Modelling, Understanding and Predicting Structural Brain Change (EPSRC GR/S48844/01), (OC) and Integrated Brain Image Modelling (EPSRC GR/S82503/01), (MJ). We thank David Kennedy and the Centre for Morphometric Analysis at MGH for use of the IBSR data.

References

1. Collins, D., Dai, W., Peters, T. and Evans, A. Automatic 3D model-based neuroanatomical segmentation, Human Brain Mapping 3 (1995) 190-208.
2. Chalana, V. and Kim, Y. A methodology for evaluation of boundary detection algorithms on medical images, IEEE Transactions on Medical Imaging, 16(5) (1997) 642-652.
3. Beauchemin, M., Thomson, K.P.B. The evaluation of segmentation results and the overlapping area matrix, International Journal of Remote Sensing 18(18) (1997) 3895-3899.
4. Bello F., Colchester A.C.F., Measuring global and local spatial correspondence using information theory, Proceedings of MICCAI 1998, LNCS 1496: 964-973.
5. Gerig, G., Jomier, M. and Chakos, M. Valmet: A new validation tool for assessing and improving 3D object segmentation, Proceedings of MICCAI 2001 516-528.
6. Crum, W.R., Griffin, L.D., Hill, D.L.G. and Hawkes, D.J., Zen and the Art of Medical Image Registration: Correspondence, Homology and Quality, NeuroImage 20 (2003) 1425-1437.
7. Bloch, I., Fuzzy spatial relationships for image processing and interpretation: a review, Image and Vision Computing 23(2) (2005) 89-110.
8. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O. and Hawkes, D.J. Non-rigid registration using free-form deformations: Application to breast MR images. IEEE Transactions on Medical Imaging, 18(8) (1999) 712-721.
9. Bhatia, K.K., Hajnal, J.V., Puri, B.K., Edwards A.D. and Rueckert, D,. Consistent group-wise non-rigid registration for atlas construction, IEEE Symposium on Biomedical Imaging (ISBI), pp 908-911, Arlington 2004.