Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis

William R. Crum*, Oscar Camara, and Derek L. G. Hill, *Member, IEEE*

***Abstract*—Measures of overlap of labelled regions of images, such as the Dice and Tanimoto coefficients, have been extensively used to evaluate image registration and segmentation algorithms. Modern studies can include multiple labels defined on multiple images yet most evaluation schemes report one overlap per labelled region, simply averaged over multiple images. In this paper, common overlap measures are generalized to measure the total overlap of ensembles of labels defined on multiple test images and account for fractional labels using fuzzy set theory. This framework allows a single "figure-of-merit" to be reported which summarises the results of a complex experiment by image pair, by label or overall. A complementary measure of error, the overlap distance, is defined which captures the spatial extent of the nonoverlapping part and is related to the Hausdorff distance computed on grey level images. The generalized overlap measures are validated on synthetic images for which the overlap can be computed analytically and used as similarity measures in nonrigid registration of three-dimensional magnetic resonance imaging (MRI) brain images. Finally, a pragmatic segmentation ground truth is constructed by registering a magnetic resonance atlas brain to 20 individual scans, and used with the overlap measures to evaluate publicly available brain segmentation algorithms.**

***Index Terms*—Fuzzy sets, Hausdorff distance, morphological operations, registration, segmentation, validation.**

I. Introduction

OBJECTIVE evaluation of image analysis methods is both vital and generally difficult in medical applications. Two of the most common and important classes of image analysis algorithm with medical applications are image registration and image segmentation. Registration refers to a set of techniques with the shared goal of establishing anatomical or functional correspondence between images acquired at different times, of different subjects, or with different modalities [1], [2]. Segmentation aims to decompose an image into a number of labelled regions maximising a measure of homogeneity within labels and a measure of heterogeneity between labels. Classically, these homogeneity and heterogeneity measures were based solely on the intensity characteristics of the voxels in the image [3]. However, in medical applications, image acquisition can affect local intensity characteristics, important biological structures may be composed of more than one tissue type, and boundaries between different tissue classes within single voxels result in intensities that are not characteristic of either tissue. For example, the hippocampus is a structure of considerable importance in volumetric and morphometric studies of ageing and dementia [4], but it cannot be separated from surrounding structures in magnetic resonance (MR) images by intensity alone because it is composed of grey and white matter and suffers partial volume effects on its boundaries. Therefore, prior knowledge and modelling of image acquisition and/or variation in appearance under imaging is often necessary to obtain biologically meaningful delineations.

The goal of most current work in medical image segmentation is to reduce or remove the need for manual intervention. Evaluation is generally difficult as, although it is possible to image phantom objects with known "tissue" properties, in applications of interest the underlying tissue classification is unknown. Haralick [5] notes, "*Measuring how well an algorithm does on perfect data is not interesting. Performance characterisation has to do with establishing the correspondence of the random variations and imperfections which the algorithm produces on the output data caused by the random variations and the imperfections on the input data*." Therefore, expert manual segmentation of real images (i.e., by delineating the external contour of a region) is regarded as a practical gold standard against which new segmentation algorithms can be compared. Another approach to evaluation is to use synthetic data where images are constructed from a model where the true structure boundaries are known. Synthetic data rarely captures the rich variety of structure and artifact seen in real images but is often used, especially in early testing. One of the most used synthetic data sets is the MNI BrainWeb digital phantom [6], constructed from multiple scans of a single individual, which does allow some imaging artifact to be simulated. Where tissue is subject to mechanical deformation, applications can be validated using biomechanical modelling approaches (e.g., [7]).

Image registration can also be evaluated using a segmentation approach. Corresponding regions are segmented on two images that are then registered using an image-driven approach (i.e., without knowledge of the segmentation). The segmentations before and after registration can be compared to assess how well the registration brings them into alignment.[1] Therefore, both registration and segmentation can be evaluated quantitatively if suitable measures of region agreement are available. There have been many attempts to quantify region agreement (or disagreement) including the Hausdorff distance [8], the Modified Williams Index [9] and measures based on information theory

Manuscript received April 27, 2006; revised June 20, 2006. This work was supported in part by the Medical Images and Signals IRC (EPSRC GR/N14248/01 and UK Medical Research Council Grant D2025/31) and in part by the Modelling, Understanding and Predicting Structural Brain Change Project (EPSRC GR/S48844/01). *Asterisk indicates corresponding author.*

*W. R. Crum is with the Center for Medical Image Computing, University College London, London WC1E 6BT, U.K. (e-mail: b.crum@ucl.ac.uk).

O. Camara and D. L. G. Hill are with the Center for Medical Image Computing, University College London, London WC1E 6BT, U.K. (e-mail: o.camara-rey@ucl.ac.uk; derek.hill@ucl.ac.uk).

Digital Object Identifier 10.1109/TMI.2006.880587

[1]The segmentation must be of high quality to avoid the propagation of segmentation errors into the registration evaluation.

[10]; some of these quantities can be computed using the Valmet software described in [11]. The two most common measures of region overlap are the Dice Similarity Coefficient (DSC) [12] and the Tannimoto Coefficient (TC) [13] (alternatively known as the Jaccard Similarity [14]) which although often treated separately are related by $\mathrm{DSC} = 2\mathrm{TC}/(\mathrm{TC}+1)$ [15] so that they are equal at the extrema $\{0, 1\}$ and $\mathrm{DSC} > \mathrm{TC}$ between those limits.

This work is motivated by three observations. The first is that it is increasingly common for automated segmentation algorithms to provide a so-called "partial volume" estimate i.e., rather than each voxel being labelled as belonging or not belonging to a region there is a notion of partial belonging. Where a binary voxel labelling is characterized by the values $\{0, 1\}$ a partial volume labelling takes values in the continuous range $[0, 1]$ at each voxel. Examples include the use of multi-spectral analysis [16] and statistical models of partial volume estimation [17]–[22]. This situation also occurs in registration when a label is interpolated as it is transformed from one image space to another. The overlap measures commonly used do not explicitly account for this fractional labelling. Some researchers have tried to account for fractional labels by applying thresholds to the continuous distribution of label membership values. For instance, Shattuck *et al.* [15] compared a partial volume brain classification with a binary one using the TC by simply labelling each voxel with the pure tissue class having the largest fractional membership. Ashburner and Friston [23] take a similar thresholding approach. Zou *et al.* [24] threshold the fractional membership and integrate the DSC over all possible thresholds and Anbeek *et al.* [21] take the maximum value of DSC over all possible thresholds. However, a more powerful approach is to accept that labels that reflect the proportion of tissue in a voxel are inherently nonbinary and apply an evaluation framework that exploits this property. The second observation is that evaluation studies are becoming larger and more complicated. For segmentation algorithms, this means evaluation on many different data sets, possibly acquired at different times, different sites or under different conditions. For registration algorithms, this means evaluation on multiple image-pairs, again possibly acquired under different conditions and also using multiple different labels to assess correspondence of many different important structures or tissue classes. Recent progress in registration has made evaluation even harder with the emergence of groupwise/target-less schemes (e.g., [25]) where multiple registrations are performed simultaneously either to a known reference space or to one which is determined dynamically from the images as the registration proceeds. In all of these cases, it would be desirable to compute a single "figure of merit," based on overlaps, which describes the overall effectiveness of the segmentation or registration algorithm. Such a figure of merit should accumulate results from multiple subjects and labels and cope with fractional labels in a natural way. The third observation is that overlap measures do not indicate the scale of the region mismatch, only the proportion of the region match. A measure of the scale of mismatch complementary to the overlap measure would provide a more complete summary of an evaluation study.

This paper expands on the initial work in [26] to present a generalized framework for overlaps that achieves these goals by defining partial-volume, multilabel overlap measures with an associated error measure. The new overlap measures are first validated on synthetic data which models partial volume effects and for which an analytic result for the overlap is known. The potential of the overlap measure to drive nonrigid registration in conjunction with image data is tested, as it is thought that labels can help resolve ambiguities in image registration. Finally, they are applied to evaluate three of the most widely used brain tissue segmentation algorithms, those found in SPM2 [27], [28], SPM5 [23], and FAST [22] using a segmentation ground truth constructed from real data.

II. Theory

For two overlapping regions, A and B, the TC is defined as the ratio of the number of voxels in their intersection to that in the union and the DSC is defined as the ratio of the number in the intersection to the mean label volume. In (1), $N()$ indicates the number of voxels in the enclosed set

$$\mathrm{TC} = \frac{N(A \cap B)}{N(A \cup B)} \quad \mathrm{DSC} = \frac{2N(A \cap B)}{N(A) + N(B)}. \tag{1}$$

In this section, we first develop fuzzy overlap measures that can summarize the overlap of multiple labels, then discuss different ways to weight the contribution of different labels, and show how a distance parameter can be used to measure the scale of labelling error. We work exclusively with the TC but the development could equally be applied to the DSC and the two are intimately related, as discussed in Section I.

A. Multiple Fractional Labels

First, the TC is redefined for fractional labels. To do this, we use established results from fuzzy set theory for the intersection and union of fuzzy sets (e.g., [29]). The amount of labels A and B at a voxel i, is written as $A_i, B_i \in [0,1]$. Then, the fuzzy intersection is the amount of label in common at each voxel and is therefore equal to $\mathrm{MIN}(A_i, B_i)$. Similarly the fuzzy union is the total label at each voxel (counting the shared component only once), and is, therefore, equal to $A_i + B_i - \mathrm{MIN}(A_i, B_i) = \mathrm{MAX}(A_i, B_i)$. Therefore, (1) can be rewritten to give the simple fuzzy overlap TC_F of a pair of fuzzy labels

$$\mathrm{TC}_F = \frac{\sum_{\mathrm{voxels},i} \mathrm{MIN}(A_i, B_i)}{\sum_{\mathrm{voxels},i} \mathrm{MAX}(A_i, B_i)}. \tag{2}$$

The numerator and denominator of (2) can both be accumulated across multiple labels to compute a single overlap figure, TC_{MF}, which describes the total overlap of a set of fuzzy labels defined on a single image pair. This overlap is the ratio of the total fuzzy intersection to the total fuzzy union of all labels

$$TC_{\mathrm{MF}} = \frac{\sum_{\mathrm{labels},l} \alpha_l \sum_{\mathrm{voxels},i} \mathrm{MIN}(A_{li}, B_{li})}{\sum_{\mathrm{labels},l} \alpha_l \sum_{\mathrm{voxels},i} \mathrm{MAX}(A_{li}, B_{li})}. \tag{3}$$

In (3), α_l is a label-specific weighting factor that affects how much each label contributes to the overlap accumulated over all

labels. We defer a discussion of the possible values of α_l to Section II-B. Note that this accumulation preserves the individual contributions of the labels to the total overlap and is not the same as calculating the overlap of a single large label equal to the union of the smaller ones. A further accumulation over all pairs of images gives $\mathrm{TC_{PMF}}$, describing the total fuzzy overlap of a set of labels defined over an ensemble of image pairs

$$\mathrm{TC_{PMF}} = \frac{\sum_{\text{pairs},k} \beta_k \sum_{\text{labels},l} \alpha_l \sum_{\text{voxels},i} \mathrm{MIN}(A_{kli}, B_{kli})}{\sum_{\text{pairs,k}} \beta_k \sum_{\text{labels},l} \alpha_l \sum_{\text{voxels},i} \mathrm{MAX}(A_{kli}, B_{kli})}. \quad (4)$$

In (4), β_k is a pair-specific weighting factor that affects the relative contribution of each image pair to the overlap accumulated over all labels and pairs of images. We again defer a discussion of the possible values of β_{k} to Section II-B. Equation (4) allows each pairwise comparison to contribute according to the weighting factors. The overlaps defined in (2)–(4) ($\mathrm{TC_F}$, $\mathrm{TC_{MF}}$, $\mathrm{TC_{PMF}}$) will be referred to collectively as generalized Tanimoto coefficients (GTC) in the remainder of the paper. We will also make use of label-specific overlaps accumulated over all voxels and subjects denoted $\mathrm{TC_{PF}}$ and obtained by evaluating (4) with a fixed label index, l.

An alternative formulation for some applications (e.g., where each pairwise comparison is in the same reference space) would be to compute the ratio of the joint intersection, (JI) and the joint union (JU) at each voxel over all n image pairs directly, where $JI = \mathrm{MIN}(A_{1li}, A_{2li}, \ldots, A_{nli})$ and $JU = \mathrm{MAX}(A_{1li}, A_{2li}, \ldots, A_{nli})$. This would result in a very conservative estimate of the overlap as the smallest label intensity will dominate the JI and the largest label intensity will dominate the JU at each voxel. Therefore, we will not consider this formulation further in this paper.

From a registration perspective, (4) corresponds to an experiment where multiple pairwise registrations are performed. For targetless registration applications, a reference image is defined dynamically and pairwise comparisons are not appropriate. Therefore we also propose a groupwise overlap measure $\mathrm{TC_{GMF}}$ constructed by considering all overlaps of pairwise permutations of n images ($= n(n-1)/2$) and applying (4). The number of permutations rapidly becomes very large with increasing n and the direct computation of (4) is slow. Groupwise evaluation can be made more practicable by considering all permutations of a label voxel i in n images together. Define the label intensity of this voxel in the n images as $A_1, A_2, \ldots, A_n$–1, An and sort by label intensity (so that $A_1 > A_2 > \cdots > A_n$). Then expanding the MAX() terms in (4) gives the groupwise union at the voxel as $U_{GW} = (n-1)A_1 + (n-2)A_2 + \ldots + 2A_{n-2} + A_{n-1}$ and expanding the MIN() terms gives the groupwise intersection as $I_{GW} = (n-1)A_n + (n-2)A_{n-1} + \ldots + 2A_3 + A_2$ Therefore, for the groupwise case, the intersection and union are weighted sums of the individual voxel label contributions. Assuming that sorting is an $n.\log(n)$ process [30], then for n images with L labels and N voxels per image, direct evaluation of (4) carries a computational cost $\sim L.N.n(n-1)/2$ compared with $\sim L.N.n.\log(n)$ by applying the sorting approach outlined above. For $n = 10$ direct evaluation requires a factor $\sim$5 more operations than the sorting approach and for $n = 100$ direct evaluation requires a factor $\sim$25 more operations.

TABLE I
WEIGHTINGS USED BETWEEN DIFFERENT LABELS

α	1	$1/V_l$	$(1/V_l)^2$	$\overline{\lvert\nabla A_i\rvert}, A_i > 0$
weighting	volume (implicit)	equal	inverse volume	complexity

B. Label Weightings and Error Estimation

In (3) and (4), weights α and β, respectively, define the relative contribution of labels (l) and pairs (k) to the GTC. With $\alpha = 1$, labels are implicitly weighted by their volume so large labels contribute most to the overlap. This may not be desirable as the overlap of smaller labels representing a greater registration or segmentation challenge will not strongly affect the overall overlap. Some alternative choices of α are detailed in Table I and allow all labels to be contribute equally to the overall overlap or to contribute with a volume or complexity dependent weighting. By direct substitution in (3), it can be shown that setting α equal to the inverse union of each label pair gives a GTC equal to the mean of the individual TCs computed for each label pair. This confirms that simple averaging of TCs is a special case of the GTC. The complexity measure used here (mean absolute label intensity gradient across nonzero label voxels) will be high for labels with large surface areas and/or voxels with large intensity gradients as might be found in well-defined labels of complicated structures. Low complexity values will be associated with labels with small surface areas and/or voxels with small intensity gradients as might be found in labels of approximately spherical structures, or labels with many partial-volume voxels. The second parameterβ is used to weight images (or combinations of images) in different ways. It can be used to down-weight labels defined on poor quality images or, where labels have been obtained from different sources such as two manual segmentors, weight them according to reproducibility and accuracy. In the experiments reported in this paper, we have set $\beta = 1$. When the overlap is accumulated over multiple labels on a pair of images it becomes a measure of the agreement of two partitionings of that image space. When the overlap is accumulated over multiple labels on multiple images, it can be considered as a measure of the agreement of two partitionings of a "super-space" composed of the individual image spaces. Different choices of α and β change the relative importance of various pairings and labelings within the space.

There are two standard overlap results reported in this paper: 1) the overlap for a specific label accumulated over multiple image pairs, $\mathrm{TC_{PF}}$ and 2) the overlap accumulated over a set of labels and image pairs $\mathrm{TC_{PMF}}$ as in (4). For case 1) we assume that each pairwise label overlap is an independent measurement drawn from a normal distribution. Therefore, the associated variance can be computed in the standard way. For case 2), we assume that each overlap accumulated over all labels in each pair of images is an independent measurement drawn from a normal distribution. Again the associated variance can be computed.

C. Tolerance and the Scale of Nonoverlap

Simple label overlaps do not provide any information about the nonoverlapping label portions such as the scale of mismatch. Pichon *et al.* [31] noted some previously proposed estimates of distance error, which are functions of the distance of each missegmented voxel to the ground truth, and then introduced a more symmetric error distance and defined some statistical quantities related to the misclassification probability and the mean and worst-case distance error. Although shown to be related, at least in part, to the DSC these measures are not a natural extension of overlap measures and have not been defined for fractional labels. Another previously proposed method of learning about the error scale directly through the overlap is to apply a spatial tolerance τ to the label intersection [32]. The standard GTC is, by definition, $\tau = 0$ since pairs of label voxels have to occupy the same space to be considered overlapping. However, $\tau > 0$ allows label voxels to be considered overlapping if they lie within τ millimeters of each other. In the overlap measures considered here, the relative overlap is defined by the label intersection (numerator) which is then divided by the union (Tanimoto) or mean volume (Dice) to give the normalized overlap. Therefore relaxing the criterion for overlap can be implemented, in both cases, by relaxing the criterion for label intersection. This can be incorporated into the existing framework by applying a morphological dilation operator, D_τ, to each label in turn. The fuzzy dilation operator is a generalization of the familiar binary morphological operator. D_τ is represented as a voxel mask of dilation coefficients centered on the voxel of interest with τ representing the extent of the operator. In one-dimension (1-D), where the voxel dimension is 1 mm 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 $0 \leq \gamma \leq 1$. Then the fuzzy dilation applied at a single voxel in 1-D is $(D_\tau A)_i = \arg\max_j (D_\tau(j) A_{i-j})$ where $j = \{-k, \ldots, 0, \ldots, +k\}$ and $k = (\mathrm{int})(\tau + 0.5)$; this is consistent with the definition of [33]. A graphical example of fuzzy dilation is shown in Table II where dilation kernels for $0 \leq \gamma \leq 1$ are applied to an example 1-D array of image pixels. Starting from the definition of overlap for fractional labels TC_F (2) we define fuzzy overlap to a noninteger tolerance τ in a symmetrical way, as shown in

$$\mathrm{TC}_F(\tau) = \frac{\sum_{\mathrm{voxels},i} MAX\left(\mathrm{MIN}\left((D_\tau A)_i, B_i\right), \mathrm{MIN}\left(A_i, (D_\tau B)_i\right)\right)}{\sum_{\mathrm{voxels},i} \mathrm{MAX}(A_i, B_i)}. \quad (5)$$

TABLE II
SCHEMATIC OF THE EFFECT OF FUZZY DILATION IN 1-D

τ	Dilation Vector	Dilation Kernel	Dilated Image
0.00	D = [0, 1, 0]		
0.25	D = [¼, 1, ¼]		
0.50	D = [½, 1, ½]		
0.75	D= [¾, 1, ¾]		
1.0	D = [1, 1, 1]		

In (5), it can be seen that the normalizing denominator (union term) is unchanged but the numerator (intersection term) takes the maximum value at each voxel of the intersection of A with dilated B and B with dilated A. The maximum possible overlap given by (5) can be established by assuming that A_i and B_i are fractional labels in the range [0, 1] with both having at least one label voxel equal to 1. Then, when $\tau \gg 1$ the summand in the numerator reduces to $\mathrm{MAX}(A_i, B_i)$ since $(D_\tau A)_i = (D_\tau B)_i = 1$ for large τ) and the maximum overlap is 1 as expected. Note that $TCF(\tau)$ is an increasing function of τ for $\tau \geq 0$. In cases where the maximum label voxel is <1 but identical for A and B the above argument holds. For cases where A and B have different maxima then both can be upper thresholded at MIN(MAX(A), MAX(B)) and the argument above will hold but τ may be underestimated.

The smallest value of τ for which the overlap is 1 can be used as a measure of the scale of the nonoverlapping region. We define this as the overlap distance (OD) where $\mathrm{OD} = \tau : \inf_\tau \{TC_F(\tau) = 1\}$. This quantity is related to a classical measure of distance between sets, the Hausdorff distance (HD). For overlapping binary sets, the HD can be obtained by generating, for each set in turn, the set of distances from each point in one set to the nearest point in the other set, and then taking the largest of all these distances [8]. There have been a number of attempts to define the HD for fuzzy-sets or for grey-level images; see [34] for a review. In [35], the HD is defined in terms of dilation operators as $\mathrm{HD} = \mathrm{MAX}(L(U, V), L(V, U))$ where U and V are nonempty compact overlapping sets and $L(U, V)$ returns the smallest value of a dilation parameter λ for U such that $D_\lambda U \supseteq V$. This is a more conservative definition than we have used in (5). In our nomenclature, the definition of an overlap incorporating the HD as a tolerance parameter according to [35] is shown in (6) at the bottom of the page.

$$\mathrm{HD} : \inf \left\{ \mathrm{TC}_F(\mathrm{HD}) = \frac{\mathrm{MAX}\left(\sum_{\mathrm{voxels,i}} \mathrm{MIN}\left((\mathrm{D_{HD}A})_\mathrm{i}, \mathrm{B_i}\right), \sum_{\mathrm{voxels,i}} \mathrm{MIN}\left(\mathrm{A_i}, (\mathrm{D_{HD}B})_\mathrm{i}\right) \right)}{\sum_{\mathrm{voxels,i}} \mathrm{MAX}(\mathrm{A_i}, \mathrm{B_i})} = 1 \right\} \quad (6)$$

Fig. 1. Basic petal object with zero offset and subtraction of a pair of petal objects with increasing angular offset with respect to each other.

In (6), the intersection is computed for all voxels in each image being dilated in turn and then the largest of the resulting intersections is chosen. In our definition of the OD, the maximum intersection at each voxel in turn is chosen. Therefore, the OD is less sensitive to outliers and smaller than the HD since the maximum overlap will be achieved for $\mathrm{OD} < \mathrm{HD}$. This is easily seen by comparing (5) and (6) and noting that for two real sets, a_i and $b_i \geq 0$

$$\sum_i \mathrm{MAX}(a_i, b_i) \geq \mathrm{MAX}\left(\sum_i a_i, \sum_i b_i\right). \tag{7}$$

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 (voxels) can be computed as described above. For each labeled voxel in the target image, the OD can be computed. Then the map of OD for all voxels is a map of target registration error. For labels spanning multiple voxels, the OD gives an estimate of the residual displacement between corresponding labels.

III. Experiments

Experiments were performed to validate the theory and implementation of the GTC, use it as a similarity measure to register sets of labels and to evaluate three publicly available brain tissue classification algorithms.

A. Validation of the GTC Using Synthetic "Petal" Data

To validate the theory and software implementation of the fuzzy overlap calculation, we used a synthetic test object designed to be simple enough that an analytic expression for the overlap could be obtained, but to also feature a nontrivial border which would exhibit partial volume effects in an image of the object. A "petal" object meets these criteria and can be described in two-dimensional (2-D) polar coordinates (r, θ) by (8) where r_0 and a are constant lengths $(r_0 \geq a)$, n is the number of wavelengths (petals) and δ is an angular offset

$$r(\theta) = r_0 + a\sin(n\theta + \delta). \tag{8}$$

Fig. 1 (left panel) shows the object described by (8) with $r = 64$ mm, $a = 16$ mm, $\delta = 0$ and $n = 6$. By varying δ, a set of pairs of images with varying overlap can be created. The other panels in Fig. 1 show subtraction images of pairs of petal object of increasingly different angular offset δ. We evaluated the overlap between such pairs of images in two ways 1) analytically using (8) and 2) by measurement from images of the objects. For details of the analytic computation and the generation of partial volume images of these shapes, see the Appendix. For the same pairs of shapes, we computed the HD numerically by exhaustive sampling of points on the external contour of both shapes, and compared this with the OD computed from the corresponding image pairs using (5).

B. Application of the GTC to Image and Label Registration

The use of the GTC to act as a similarity measure for nonrigid registration was evaluated using 9 of the 18 publicly available three-dimensional (3-D) T1-weighted labelled MR-brain images from the Internet Brain Segmentation Repository.[2] The images used (numbers 02, 04, 06, 07, 08, 10, 11, 12, and 16) were qualitatively assessed to have the best contrast and least artifact to minimise associated labelling errors. Each image had $256 \times 256 \times 128$ voxels of either $0.9375 \times 0.9375 \times 1.5$ mm or $1.0 \times 1.0 \times 1.5$ mm and 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. A tenth image (09) with the same labels was chosen as a reference. Each image was registered to the reference using a fluid registration algorithm [7] with two resolution levels (half and full resolution). The registration was performed in three different ways: 1) the grey-level image data was registered by maximizing the intensity cross correlation; 2) images containing the label-sets defined above were registered by maximizing the GTC computed over all labels; 3) grey-level image and label image pairs were registered by maximizing the sum of the intensity cross correlation and total overlap. To do this, the GTC was implemented as a voxel-wise force to drive the fluid registration. For the specific case here, where the target image labels are binary, each fluid force component at voxels with nonzero labels in the target image, is proportional to the difference in label intensities between source voxels neighboring the corresponding target voxel i.e., a centered difference scheme that strives to match the intensity of corresponding source and target label voxels.

C. Application of the GTC to Segmentation Evaluation

The final experiment applies the GTC to evaluate three widely used brain tissue segmentation algorithms SPM2,[3] SPM5[4] and FAST.[5] SPM2 employs a modified mixture model cluster analysis technique [27] including a correction for image intensity nonuniformity [28]. SPM5 again uses Gaussian mixture model techniques but treats tissue segmentation as one part of an integrated spatial normalization process that includes estimation of nonuniformity correction and warping parameters. FAST uses a hidden Markov Random Field model to explicitly incorporate spatial information into the process [22]. A data set with known tissue class segmentation was constructed for the evaluation. Twenty 3-D T1-weighted volumetric MR brain scans of 10 cognitively normal subjects imaged twice were obtained from the Dementia Research Centre, Institute of Neurology, London, U.K. The scans were acquired on a 1.5T General Electric Signa scanner with acquisition parameters: TR/TE $= 35/5$ ms and a

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

[3]http://www.fil.ion.ucl.ac.uk/spm/software/spm2.

[4]http://www.fil.ion.ucl.ac.uk/spm.

[5]http://www.fmrib.ox.ac.uk/analysis/research/fast.

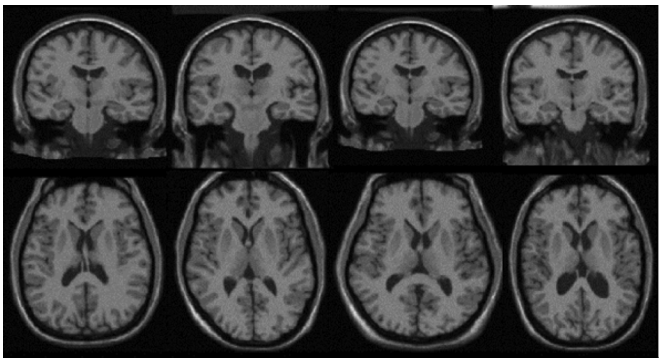

Fig. 2. Examples slices from four of the test images used for the segmentation evaluation. Coronal and axial slices were selected visually to show the same anatomy. Images were obtained by warping the MNI BrainWeb noise-free T1-weighted image onto normal control images.

35° flip angle. Each reconstructed scan consisted of 124 1.5 mm slices with 256 × 256 voxels of size 0.9375 × 0.9375 mm in-plane. The MNI BrainWeb T1-weighted brain [6] was then fluidly registered at half resolution to each brain in turn [7]. The resulting warp-fields were used to transform both the MNI BrainWeb T1-weighted noise-free brain image, and its associated partial volume labels for grey-matter, white-matter and CSF into the space of each subject in turn using trilinear interpolation. The 20 transformed BrainWeb images with added Rician noise (2% of maximum intensity on two independent Gaussian channels) formed the test-set for the segmentation evaluation.

Fig. 2 shows example axial and coronal slices from four of the transformed images showing the range of appearance despite all images being derived from the same reference scan. To check the consistency of the registration, the CSF, grey matter and white matter volumes were computed from the warped tissue maps for each image and compared for time-point 1 and time-point 2 in each case under the assumption that as the subjects were normal controls, no volume change should have occurred. The mean (sd) of the absolute difference in tissue volume between time points as a percentage of the mean tissue volume was CSF: 2.0% (2.2%), grey matter: 1.0% (1.1%), white matter 0.7% (0.9%). These figures should be compared with the percent standard deviations in each tissue volume for the ten time-point one scans which were: CSF: 10.4%; grey matter: 7.1%; white matter: 7.1%. The results indicate that tissue class warping is consistent to 1% by volume in grey-matter and white matter with approximately twice that variation in CSF. The transformed BrainWeb tissue labels were used as the gold-standard segmentation results. This experiment does not depend on perfect inter-subject brain registration since the evaluation is done purely on transformed BrainWeb images.

To perform the evaluation, each segmentation algorithm was applied to the 20 test images in turn to produce partial volume estimates of the three tissue classes. Each algorithm was applied "out of the box" in that all default parameter settings were used. In the case of FAST, the Brain Extraction Tool (BET) was used to presegment brain from nonbrain as this is the recommended use. One case was subsequently removed from the study as BET failed to achieve a good brain extraction.

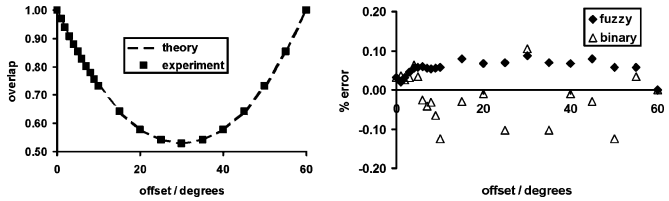

Fig. 3. Left graph compares the analytical result for the overlap of a pair of petal objects with varying relative angular offset with that measured from generated images incorporating partial volume effects. Graph shows the percentage fractional overlap error for the fuzzy petal labels and where each fuzzy petal label voxel has been thresholded at the 50% level.

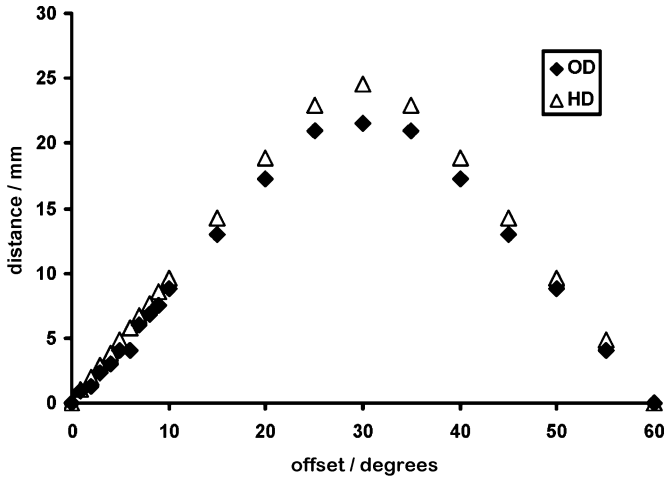

Fig. 4. Comparison of the HD computed for pairs of the petal object with different relative angular offsets with the OD computed from the corresponding image pairs. Theory predicts that the $\mathrm{OD} \leq \mathrm{HD}$. See text for further details.

IV. Results

A. Validation of the GTC Using Synthetic "Petal" Data

Fig. 3(a) shows excellent agreement between the overlap computed analytically and that measured experimentally from the partial-volume images for a range of offset values. This confirms that the software implementation of the overlap measure is correct and that the overlap of structures with ill-defined borders can be recovered using partial volume assumptions. The percentage fractional error of the experimental result compared with the theoretical result is shown in Fig. 3(b) for two cases: 1) with the fuzzy label voxels described above and 2) where each fuzzy label voxel was thresholded at the 50% intensity level to create a binary label voxel equivalent. The fuzzy case has mean (standard deviation) error 0.06% (0.02%) whereas the binary case has mean (s.d.) −0.16% (0.06%). In this simple experiment the error as a fraction of measured overlap is low in both cases but the fuzzy measurement slightly overestimates the computed overlap suggesting an effect of the discretization of the petal shape. Fig. 4 shows the relationship between the OD and the HD, again as a function of the angular offset. As predicted $\mathrm{HD} > \mathrm{OD}$ with the difference becoming larger for larger angular offsets and smaller overlaps.

B. Application of the GTC to Image and Label Registration

Fig. 5 shows the relationship between the intensity cross correlation (voxel similarity) and the GTC (label similarity)

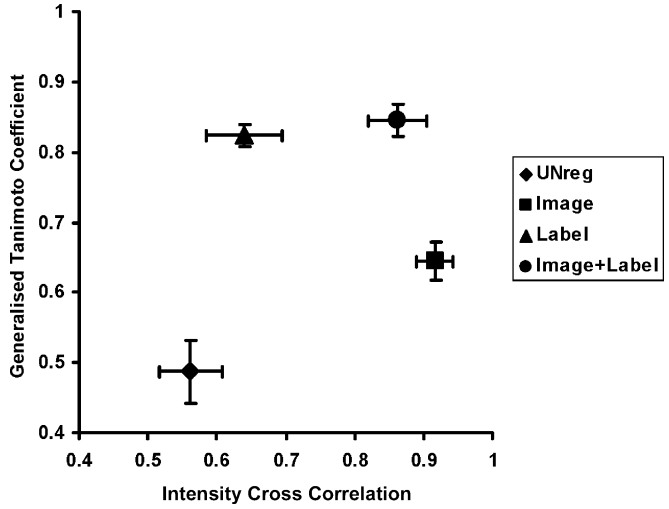

Fig. 5. Relationship between the mean image similarity measure and the GTC computed over all labels and subjects for the unregistered, image-registered, label-registered and image+label-registered cases. Image similarity was computed over the entire image volume in each case.

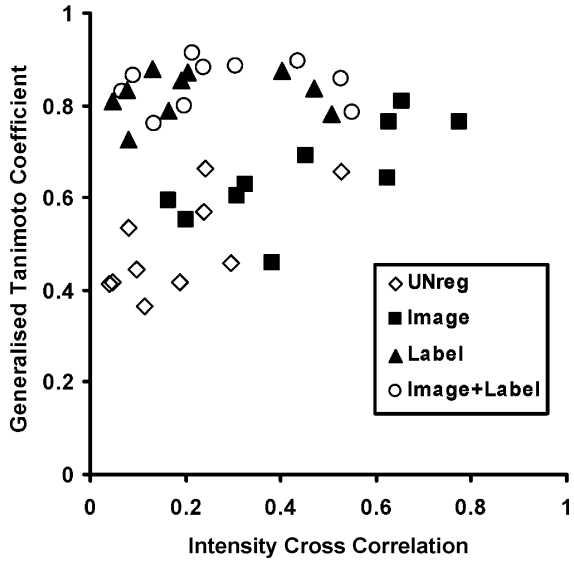

Fig. 6. Breakdown of the results in Fig. 5 for the individual labels. Image similarity was computed over the voxels comprising each label in the target space.

for the unregistered, image-registered, label-registered, and image+label-registered experiments. As expected, the unregistered data has the lowest image similarity and label overlap. The image-registered data has the largest image-similarity and also improves the label overlap. The label-registered data has a significantly larger label overlap and an improved image similarity over the unregistered data. The key result is that the image+label registered data has an image similarity larger than the label-registered case—which we would expect—and a label overlap larger than the label registration case—which we might not expect. Therefore, the image and label information used together for registration is resulting in high image similarity and high label overlap representing a compromise between image and label features. The labels contain a relatively sparse set of information for registration compared with the original image data raising the possibility of local minima during the optimization procedure. The results here suggest that the additional voxel intensity information used in the image+label registration experiment has a beneficial effect in avoiding such minima. Fig. 6 shows the breakdown of the results of Fig. 5 for each label. The significance of the difference of label overlaps between each pair of registration experiments was assessed using the paired, two-tailed student t-test and found to be significant $(p < 0.001)$ in all cases.

C. Application of the GTC to Segmentation Evaluation

We applied the three segmentation procedures to the test images to produce fuzzy tissue maps for grey matter, white matter, and CSF. These were compared against the transformed MNI tissue maps. Fig. 7 shows the GTC over all subjects and labels for the three methods compared with the known segmentations. Results are presented for the four different label weightings discussed in Section II-B. The performance of FAST and SPM5 are comparable for the equal-weighting case with SPM5 significantly better for the inverse volume weighting case and FAST significantly better for the volume weighting case. SPM2 has significantly poorer performance than the other two methods.

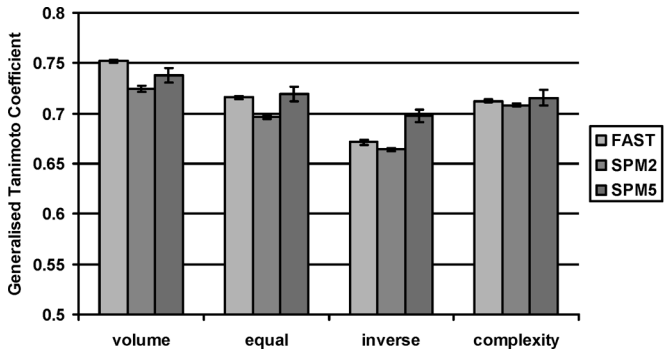

Fig. 7. The GTC computed over all labels and subjects for the segmentation evaluation experiment. Groupings on the x axis correspond to different label weightings in the overall overlap. See text for full details.

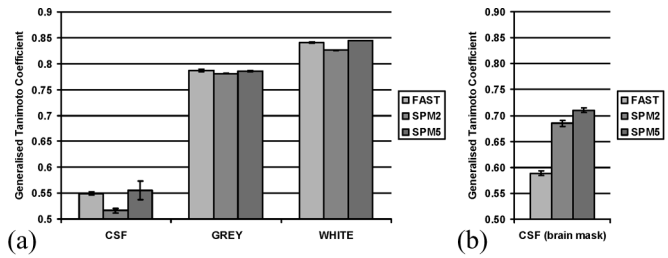

Fig. 8. (a) GTC computed over all subjects for each label in the segmentation evaluation experiment. Across subjects, each label was weighted equally. (b) GTC for CSF after application of a brain mask.

These results can be broken down further by examining the overlaps for each tissue class [Fig. 8(a)]; again the results were comparable for SPM5 and FAST for each tissue class with the overall results for CSF worse than GM and WM. Visual examination of the CSF tissue labels showed that both SPM (both versions) and FAST are overestimating the cortical CSF compared with the gold standard; this tissue class is the most challenging to classify due to the poorly defined boundary in MR between cortical CSF and the meninges. To assess the effect of misclassification outside the brain the GTCs were recomputed

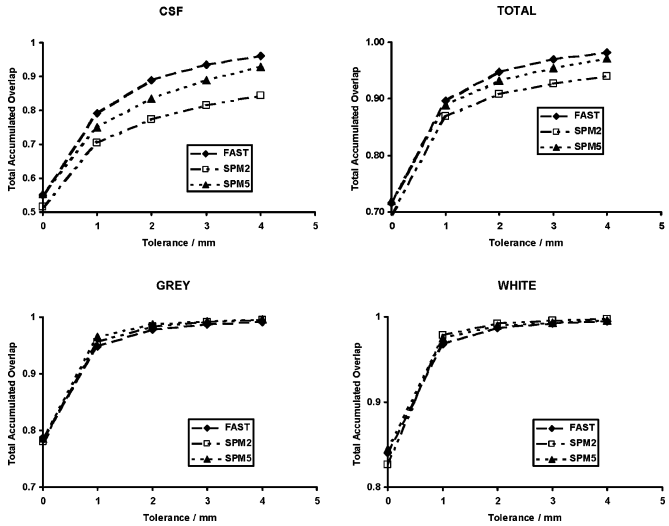

Fig. 9. Relationship between overlap and overlap tolerance for grey matter, white matter and CSF and overall for the three segmentation algorithms. Standard errors are not shown for clarity but are $\sim$0.01 for $\tau = 0.0$ falling to $\sim$0.001 for $\tau = 5.0$.

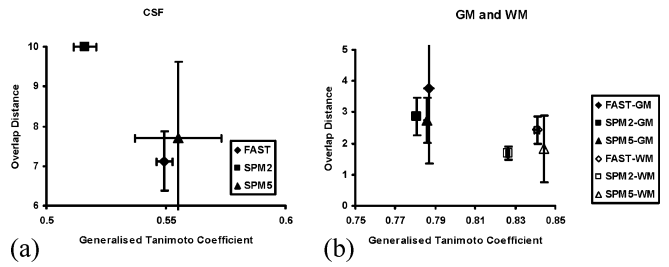

Fig. 10. Relationship between the generalized Tanimoto Coefficient and the OD for CSF, GM, and WM for three segmentation methods. Note that the OD was limited to a maximum of 10.0 mm in the calculation; in the case of the CSF calculation for SPM2 all ODs were limited resulting in an apparent standard deviation of 0.0.

after the application of a brain mask derived from the known GM, WM, and CSF maps for each subject. The results for GM and WM were unchanged but there was significant improvement in the results for CSF in all three cases and SPM5 improved the most [Fig. 8(b)]. These results confirm that all three algorithms tended to misclassify (or over-classify) extra-cortical voxels as CSF. Fig. 9 shows how the unmasked overlap varies with the tolerance overall and for each tissue class. For grey matter and white matter in all cases, a large increase in overlap is achieved for a tolerance of 1 mm indicating that the tissue boundary determined by these algorithms is spatially close to the gold-standard. For tolerances $>$0.0 mm, FAST is consistently superior for CSF segmentation and this effect is large enough to result in a similar trend in the total overlap. Fig. 10 shows the relationship between GTC and OD for each tissue class in each segmentation. In terms of OD, the average performance of the three techniques is similar, giving OD $\sim$2.0 mm for white matter and $\sim$3.0 mm for grey matter. The OD is larger and exhibits more variation for CSF consistent with the overlap results in Figs. 7 and 8.

V. Discussion and Conclusion

There is a need for better evaluation techniques for image segmentation and registration algorithms. One of the biggest problems in medical image analysis remains the lack of gold standards for many segmentation applications. Time-consuming manual segmentation with its inherent variability remains necessary but is often limited by resource and expertise. As recently as 2005, a comparative study of brain segmentation methods used expert manual segmentation of just two slices from just one normal MR brain study as a ground truth [36].

One alternative strategy is to use the segmentation estimates themselves to produce a plausible ground truth. Warfield *et al.* have recently introduced the STAPLE (Simultaneous Truth and Performance Level Estimation) algorithm for validation of image segmentation that estimates the "true" segmentation from a set of input segmentation estimates [37]. The ground truth estimate will depend strongly on the input data (since STAPLE does not perform segmentation using image data) and the assumptions inherent in STAPLE (see later). An objective ground-truth that can be generated without excessive manual intervention remains a desirable goal. Our approach in this paper was to use a standard template for brain anatomy but to warp it to resemble volunteer brains. This is clearly only a partial solution, not least due to the assumption that the effects of interpolation are consistent between the grey-level image and the tissue maps. The other limitation is that there is nonrigid variation in the gross neuroanatomy but essentially the same cortical structure in our test images. It is possible in principle to use the transformed tissue templates to simulate new more realistic MR volumes but this does not avoid the fact that there would still be little variation in the geometry of the cortex. More sophisticated methods of simulating variation in neuroanatomy [38] may result in improved evaluation data becoming available.

The results of the segmentation evaluation (Section IV-C) provide some evidence for convergence in the performance of contemporary brain tissue classification algorithms; it is worth noting that we did not try to optimize the performance of any of the algorithms by preprocessing the data or by using non-default parameter settings. SPM5 was previously evaluated in [23] (Table I) where the DSC was computed for grey matter and white matter after thresholding a fuzzy segmentation of the T1-weighted BrainWeb brain to give Dice overlaps of 0.93 (grey matter) and 0.96 (white). Converting our GTC to a GDSC (generalized Dice Similarity Coefficient) gives GDSC overlaps of 0.88 (grey matter) and 0.92 (white matter), which are lower by approximately 5% but have the same trend. Of future interest is how the increasingly subtle differences in performance between automated brain classification algorithms affect clinical research studies.

To be useful in large-scale studies, measures of segmentation quality must be amenable to statistical analysis. The relationship of the DSC to the Kappa statistic [39] is often mentioned in the literature on segmentation evaluation but rarely explicitly exploited. Similarly it is also often reported that a value for the $\mathrm{DSC} > 0.7$ indicates "excellent agreement" between data under consideration but the relevance of this to segmentation evaluation is questionable and as noted in [39], "*most of its merit lies in the fact that it provides a value that can be used to compare the similarities between measurement pairs.*" Zou *et al.* [24] suggest a fuller statistical framework for the DSC.

One key step in their formulation is to apply the logit transform $(\mathrm{logit}(x) = x/(1-x))$ to transform the domain of the DSC from [0, 1] to $[-\infty, +\infty]$. We note in passing that effectively this means a different overlap measure is being analyzed, defined as the ratio of the (Dice) overlap to the nonoverlap. A similar analysis could be applied to the measures in this work; to date we have compared total overlaps using associated variances in a very simple way.

Ultimately, definitive validation requires a number of approaches that will be dictated by the application at hand. The STAPLE approach is potentially very powerful, especially for comparison of trained manual segmentors, but does come with assumptions. From [37], "*Implicit in this model is the notion that experts have been trained to interpret the images in a similar way, the segmentation decisions may differ due to random or systematic rater differences* ...". It is not clear how far these assumptions will apply to automated segmentation techniques. Of particular relevance to this work is the assumption of STAPLE that the ground truth at each voxel is a single binary label. In addition the ability of STAPLE to establish a ground-truth from a number of segmentation estimates should improve given more segmentation estimates but also weaken given fewer. As the estimate of the ground truth becomes worse, so does the ability to compare raters. Simple overlap measures like the ones described in this paper remain of value for these reasons. Indeed in [40], where a comparative study of STAPLE and the Williams index for brain tissue classification was performed, it was concluded that "*When no ground truth is required, we recommend the use of Williams index as it is easy and fast to compute*." Pairwise generalized overlap measures are also relatively fast to compute requiring minutes on contemporary desk-top processors for the entire sets of images and labels used in this paper. However, at present the computation of the OD is costly because fuzzy dilation is relatively slow to compute (several minutes per label per pair) and an optimization over several fuzzy dilations is required. Measures related to the HD are typically difficult to compute efficiently, however our implementation of this calculation is currently far from optimal so there is potential for significant improvement.

In addition to validation applications, the generalized overlaps can be used to drive registration using label information. Registration of labels has previously been reported by D'Agostino *et al.* [41] who used the Kullback–Leibler distance between an ideal joint tissue class probability distribution and one evaluated during fluid registration as a similarity measure. Camara *et al.* [42] used the root mean square difference of voxel label intensities to register thoracic/abdominal structures in positron emission tomography (PET) and computerized tomography (CT) images. Frangi *et al.* [43] used the label consistency measure, which is essentially a TC, and a Kappa statistic measure that is related to the DSC, for Free Form Deformation registration of labelled MR cardiac images and found no significant difference between them. Even earlier, Christensen *et al.* [44] used segmentation of key structures in pelvic CT scans to drive fluid registration of serial CT scans of subjects undergoing intracavitary brachytherapy. The GTC combines the advantages of the previous work, specifically, that multiple fractional labels can be registered together by maximizing a global similarity measure which incorporates partial volume labelling effects. Using labels simplifies the correspondence problem and can mask confounding image information, where there are large differences between images. We also reported one way of combining image and label information—by summing image and label voxel forces in fluid registration—that in our experiment slightly improved the overall label match. In general, the purpose of combining label and image information is to make the registration robust away from the labels and guarantee known correspondence in the vicinity of the labels. The registration transformation could be initialized using the difference in position of centres of mass of corresponding labels but we have not attempted this here.

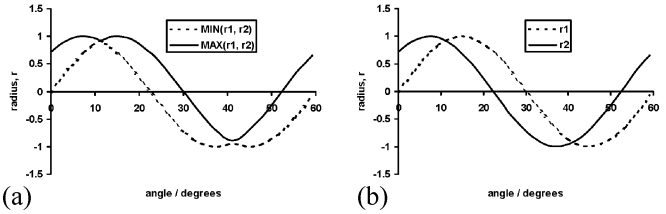

Fig. 11. Radial profiles of the petal object over the angular range spanning one petal (n = 6) (a) the radius of two petal objects with an angular offset (b) the MIN and MAX of the profiles in (a) corresponding with the intersection and union object.

In future work, we hope to explore the relationship between the OD and the GTC more fully and develop a statistical framework for the analysis of overlap computed over different subsets of the images and labels under consideration. We plan to explore the potential for a small number of labels to resolve problems in contemporary nonrigid registration problems. In summary, we have presented a flexible framework for computing generalized label overlaps which offers a natural way to summarize the results of complex registration and segmentation studies.

APPENDIX

To compute the overlap analytically, we consider a pair of images described by (8), one with $\delta = 0$ and the other with $\delta > 0$ so that $r_1 = r_0 + a\sin(n\theta)$ and $r_2 = r_0 + a\sin(n\theta + \delta)$ (Fig. 11). Then the areas of intersection I and union U of the overlapping pair are given by (9)

$$I = \frac{1}{2}\int_0^{2\pi} r_I^2(\theta)d\theta \quad U = \frac{1}{2}\int_0^{2\pi} r_U^2(\theta)d\theta. \tag{9}$$

In (9), $r_I = \mathrm{MIN}(r_1, r_2)$ corresponds to the outline of the intersection of the objects and $r_U = \mathrm{MAX}(r_1, r_2)$ corresponds to the outline of the union of the objects. To compute these integrals, we observe that (Fig. 11), the external contours of the pair of images defined above intersect at angles given by $n\theta = \arctan(\sin\delta/(1-\cos\delta))$ (i.e., when $\sin(n\theta) = \sin(n\theta + \delta)$) and that there are two such crossings θ_1 and θ_2 in each wavelength. Therefore, the integrals in (9) can be split into three segments $[0, \theta_1]$, $[\theta_1, \theta_2]$, and $[\theta_2, 2\pi/\mathrm{n}]$ where MIN($r1, r2$) or MAX($r1, r2$) in the expressions for r_I and r_U consistently returns r_1 or r_2 for the whole of the segment. Expressions for the

area of intersection and union are then obtained by expanding the squared terms in (9) and using elementary methods to evaluate each segment separately. For the validation experiments, we generated voxelated partial volume image pairs representing the same shapes. First an empty image was defined and then, all voxels outside the maximum extent of the shape were set to zero and all voxels within r_0 of the centre of the shape to 255. Then each voxel in the "petal" region was set to zero and subdivided into a grid of $50^2 = 2500$ points. Each point on the grid was identified as being inside or outside the shape by computing its 2-D radial polar coordinates and using (9). Then the intensity was set to 255.number_points_inside/total_number_of_points to represent partial volume effects at each voxel.

ACKNOWLEDGMENT

The authors would like to thank Prof. N. Fox of the Dementia Research Centre, Institute of Neurology, University College London, for the MR brain data used to construct the segmentation gold standard. This paper benefited and stemmed from discussions within the Integrated Brain Image Modelling project (EPSRC GR/S82503/01). The authors would also like to thank two anonymous reviewers for constructive comments that have improved this paper.

REFERENCES

[1] B. Zitova and J. Flusser, "Image registration methods: A survey," *Image Vision Comput.*, vol. 21, no. 11, pp. 977–1000, 2003.

[2] W. R. Crum, T. Hartkens, and D. L. G. Hill, "Non-rigid image registration: Theory and practice," *Br. J. Radiol.*, vol. 77, pp. S140–S153, 2004.

[3] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognit.*, vol. 29, no. 8, pp. 1335–1346, 1996.

[4] N. C. Fox, E. K. Warrington, P. A. Freeborough, P. Hartikainen, A. M. Kennedy, J. M. Stevens, and M. N. Rossor, "Presymptomatic hippocampal atrophy in Alzheimer's disease—A longitudinal MRI study," *Brain*, vol. 119, no. 6, pp. 2001–2007, 1996.

[5] R. M. Haralick, "Performance characterization in computer vision," *CVGIP: Image Understanding*, vol. 60, no. 2, pp. 245–249, 1994.

[6] R. K.-S. Kwan, A. C. Evans, and G. B. Pike, "MRI simulation-based evaluation of image-processing and classification methods," *IEEE Trans. Med. Imag.*, vol. 18, no. 11, pp. 1085–1097, Nov. 1999.

[7] W. R. Crum, C. Tanner, and D. J. Hawkes, "Anisotropic multi-scale fluid registration: Evaluation in magnetic resonance breast imaging," *Phys. Med. Biol.*, vol. 50, pp. 5153–5174, 2005.

[8] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.

[9] V. Chalana and Y. Kim, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Trans. Med. Imag.*, vol. 16, no. 5, pp. 642–652, May 1997.

[10] F. Bello and A. C. F. Colchester, "Measuring global and local spatial correspondence using information theory," in *Proc. MICCAI*, 1998, vol. 1496, pp. 964–973.

[11] G. Gerig, M. Jomier, and M. Chakos, "Valmet: A new validation tool for assessing and improving 3-D object segmentation," in *Proc. MICCAI*, 2001, pp. 516–528.

[12] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297–302, 1945.

[13] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[14] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bulletin de la Societe Vaudoise des Sciences Naturelles*, vol. 44, pp. 223–270, 1908.

[15] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy, "Magnetic resonance image tissue classification using a partial volume model," *NeuroImage*, vol. 13, pp. 856–876, 2001.

[16] G. Harris, N. C. Andreasen, T. Cizadlo, J. M. Bailey, H. J. Bockholt, and V. A. Magnotta *et al.*, "Improving tissue classification in MRI: A three-dimensional multispectral discriminant analysis method with automated training class selection," *J. Comput. Assist. Tomogr.*, vol. 23, no. 1, pp. 144–154, 1999.

[17] J. Tohka, A. Zijdenbos, and A. Evans, "Fast and robust parameter estimation for statistical partial volume models in brain MRI," *NeuroImage*, vol. 23, no. 1, pp. 84–97, 2004.

[18] M. A. G. Ballester, A. P. Zisserman, and M. Brady, "Estimation of the partial volume effect in MRI," *Med. Image Anal.*, vol. 6, no. 4, pp. 389–405, 2002.

[19] T. J. Grabowski, R. J. Frank, N. R. Szumski, C. K. Brown, and H. Damasio, "Validation of partial tissue segmentation of single-channel magnetic resonance images of the brain," *NeuroImage*, vol. 12, no. 6, pp. 640–656, 2000.

[20] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "A unifying framework for partial volume segmentation of brain MR images," *IEEE Trans. Med. Imag.*, vol. 22, no. 1, pp. 105–119, Jan. 2003.

[21] P. Anbeek, K. L. Vincken, G. S. van Bochove, M. J. P. van Osch, and J. van der Grond, "Probabilistic segmentation of brain tissue in MR imaging," *NeuroImage*, vol. 27, pp. 795–804, 2005.

[22] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, Jan. 2001.

[23] J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, pp. 839–851, 2005.

[24] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. C. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index," *Academic Radiol.*, vol. 11, no. 2, pp. 178–189, 2004.

[25] C. Studholme and V. Cardenas, "A template free approach to volumetric spatial normalization of brain anatomy," *Pattern Recognit. Lett.*, vol. 25, no. 10, pp. 1191–1202, 2004.

[26] W. R. Crum, O. Camara, D. Rueckert, K. K. Bhatia, M. Jenkinson, and D. L. G. Hill, "generalized overlap measures for assessment of pairwise and groupwise image registration and segmentation," in *Proc. MICCAI*, 2005, vol. 3749, pp. 99–106.

[27] J. Ashburner and K. J. Friston, "Multimodal image coregistration and partitioning—A unified framework," *NeuroImage*, vol. 6, pp. 209–217, 1997.

[28] ——, "Voxel-based morphometry—The methods," *NeuroImage*, vol. 11, pp. 805–821, 2000.

[29] D. Dubois and H. Prade, *Fundamentals of Fuzzy Sets*. New York: Kluwer, 2000.

[30] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C++: The Art of Scientific Computing*. New York: Cambridge Univ. Press, 2002.

[31] E. Pichon, A. Tannenbaum, and R. Kikinis, "A statistically based flow for image segmentation," *Med. Image Anal.*, vol. 8, no. 3, pp. 267–274, 2004.

[32] W. R. Crum, L. D. Griffin, D. L. G. Hill, and D. J. Hawkes, "Zen and the art of medical image registration: Correspondence, homology and quality," *NeuroImage*, vol. 20, pp. 1425–1437, 2003.

[33] I. Bloch, "Fuzzy spatial relationships for image processing and interpretation: A review," *Image Vision Comput.*, vol. 23, no. 2, pp. 89–110, 2005.

[34] I. Bloch and H. Maitre, "Fuzzy mathematical morphologies—A comparative-study," *Pattern Recognit.*, vol. 28, no. 9, pp. 1341–1387, 1995.

[35] D. Dubois and H. Prade, "On distances between fuzzy points and their use for plausible reasoning," in *Proc. Int. Conf. Syst. Man Cybernetics*, 1983, pp. 300–303.

[36] M. Bach Cuadra, L. Cammoun, T. Butz, O. Cuisenaire, and J.-P. Thiran, "Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images," *IEEE Trans. Med. Imag.*, vol. 24, no. 12, pp. 1548–1565, Dec. 2005.

[37] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

[38] Z. Xue, D. Shen, B. Karacali, and C. Davatzikos, "Statistical representation and simulation of high-dimensional deformations: Application to synthesizing brain deformations," in *Proc. Med. Image Comput. Computer-Assisted Intervention*, 2005, vol. 3749, pp. 500–508.

[39] A. P. Zijdbenos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, "Morphometric analysis of white matter lesions in MR images: methods and validation," *IEEE Trans. Med. Imag.*, vol. 13, no. 4, pp. 716–724, Apr. 1994.

[40] M. Martin-Fernandez, S. Bouix, L. Ungar, R. W. McCarley, and M. E. Shenton, "Two methods for validating brain tissue classifiers," in *Proc. MICCAI*, 2005, vol. 3749, pp. 515–522.

[41] E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens, "An information theoretic approach for non-rigid image registration using voxel class probabilities," in *Proc. Med. Image Computing Computer Assisted Intervention*, 2003, vol. 2879, pp. 812–820.

[42] O. Camara, G. Delso, and I. Bloch, "Free form deformations guided by gradient vector flow: A surface registration method in thoracic and abdominal PET-CT applications," in *Proc. Workshop Biomed. Registration*, 2003, vol. 2717, LNCS, pp. 224–233.

[43] 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 Trans. Med. Imag.*, vol. 21, no. 9, pp. 1151–1166, Sep. 2002.

[44] G. E. Christensen, B. Carlson, K. S. C. Chao, P. Yin, P. W. Grigsby, and K. Nguyen *et al.*, "Image-based dose planning of intracavitary brachytherapy: Registration of serial-imaging studies using deformable anatomic templates," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 51, no. 1, pp. 227–243, 2001.