Geodesic Information Flows: Spatially-Variant Graphs and Their Application to Segmentation and Fusion

M. Jorge Cardoso*, Marc Modat, Robin Wolz, Andrew Melbourne, David Cash, Daniel Rueckert, and Sebastien Ourselin

***Abstract*—Clinical annotations, such as voxel-wise binary or probabilistic tissue segmentations, structural parcellations, pathological regions-of-interest and anatomical landmarks are key to many clinical studies. However, due to the time consuming nature of manually generating these annotations, they tend to be scarce and limited to small subsets of data. This work explores a novel framework to propagate voxel-wise annotations between morphologically dissimilar images by diffusing and mapping the available examples through intermediate steps. A spatially-variant graph structure connecting morphologically similar subjects is introduced over a database of images, enabling the gradual diffusion of information to all the subjects, even in the presence of large-scale morphological variability. We illustrate the utility of the proposed framework on two example applications: brain parcellation using categorical labels and tissue segmentation using probabilistic features. The application of the proposed method to categorical label fusion showed highly statistically significant improvements when compared to state-of-the-art methodologies. Significant improvements were also observed when applying the proposed framework to probabilistic tissue segmentation of both synthetic and real data, mainly in the presence of large morphological variability.**

Manuscript received January 16, 2015; revised March 23, 2015; accepted March 27, 2015. Date of publication April 14, 2015; date of current version August 28, 2015. The Dementia Research Centre is an Alzheimer's Research Trust Co-ordinating centre and has also received equipment founded by the Alzheimer's Research Trust. SO receives funding from the EPSRC (EP/H046410/1, EP/J020990/1, EP/K005278), the MRC (MR/J01107X/1), the EU-FP7 project VPH-DARE@IT (FP7-ICT-2011-9-601055), the NIHR Biomedical Research Unit (Dementia) at UCL and the National Institute for Health Research University College London Hospitals Biomedical Research Centre (NIHR BRC UCLH/UCL High Impact Initiative—BW.mn.BRC10269). MM is supported by the UCL Leonard Wolfson Experimental Neurology Centre (PR/ylr/18575). MJC receives funding from EPSRC (EP/H046410/1). DC is in part supported through a grant from Brain Research Trust. AM was supported by UK registered charity SPARKS. RW and DR are funded by the 7th Framework Programme by the European Commission (http://cordis.europa.eu/ist/). *Asterisk indicates corresponding author.*

M. J. Cardoso is with the Translational Imaging Group, Centre for Medical Image Computing (CMIC), University College London, WC1E 6BT London, U.K., and also with the Dementia Research Centre (DRC), Institute of Neurology, University College London, WC1N 3AR London, U.K. (e-mail: m.jorge.cardoso@ucl.ac.uk).

M. Modat and S. Ourselin are with the Translational Imaging Group, Centre for Medical Image Computing (CMIC), University College London, WC1E 6BT London, U.K., and also with the Dementia Research Centre (DRC), Institute of Neurology, University College London, WC1N 3AR London, U.K.

A. Melbourne is with the Translational Imaging Group, Centre for Medical Image Computing (CMIC), University College London, WC1E 6BT London, U.K.

D. Cash is with the Dementia Research Centre (DRC), Institute of Neurology, University College London, WC1N 3AR London, U.K.

R. Wolz and D. Rueckert are with the Biomedical Image Analysis (BioMedIA) Group, Imperial College London, WC2R 2LS London, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2015.2418298

***Index Terms*—Information propagation, label fusion, parcelation, tissue segmentation.**

I. Introduction

SINCE the advent of open imaging databases, researchers have struggled with the fact that clinical, structural and anatomical annotations are only available on a small subset of the data. For example, annotations such as voxel-wise labels (characterising structural parcellations or tissue segmentations), landmarks (localising anatomical features) and diagnosis (characterising the patient clinical status) are usually scarce due to the need of human interaction. Ideally, one would like to be able to estimate this information for all subjects in a large database by propagating and extrapolating from a subset of annotated examples. More specifically, this work will focus on the problem of propagating categorical labels and probabilistic segmentations between datasets.

In neuroimage analysis, the most well known example of information propagation and extrapolation is the use of *a priori* probabilistic atlases in the context of tissue segmentation. In segmentation, the observed intensities alone often do not provide sufficient information about the underlying tissue composition. The ill-posed nature of the segmentation problem is the result of several imaging limitations, ranging from reduced signal—(SNR) and contrast-to-noise ratio (CNR) to imaging artefacts (e.g movement, ringing, chemical shift, susceptibility) and intensity non-uniformity (INU). As the spatial localisation of an intensity sample can be informative about its tissue composition, the tissue segmentation problem can be regularised by adding *a priori* information to the model through coordinate mapping and propagation of anatomical priors (see Fig. 1—top). This coordinate mapping can be carried out prior to segmentation [1] or iteratively optimised within the segmentation procedure [2].

The process of generating anatomical priors for healthy or pathological populations starts by manually segmenting a set of subjects, followed by a registration to a mean shape/appearance space, known as a groupwise space. Due to its mathematical properties and computationally efficiency, groupwise averages have been thoroughly used by the medical image community for information propagation and group analysis.

However, groupwise spaces suffer from three main problems: (1) their construction is highly dependent on the choice of image similarity metric and regularisation [3]; (2) the mapping errors to the groupwise space can result in morphological mismatch,

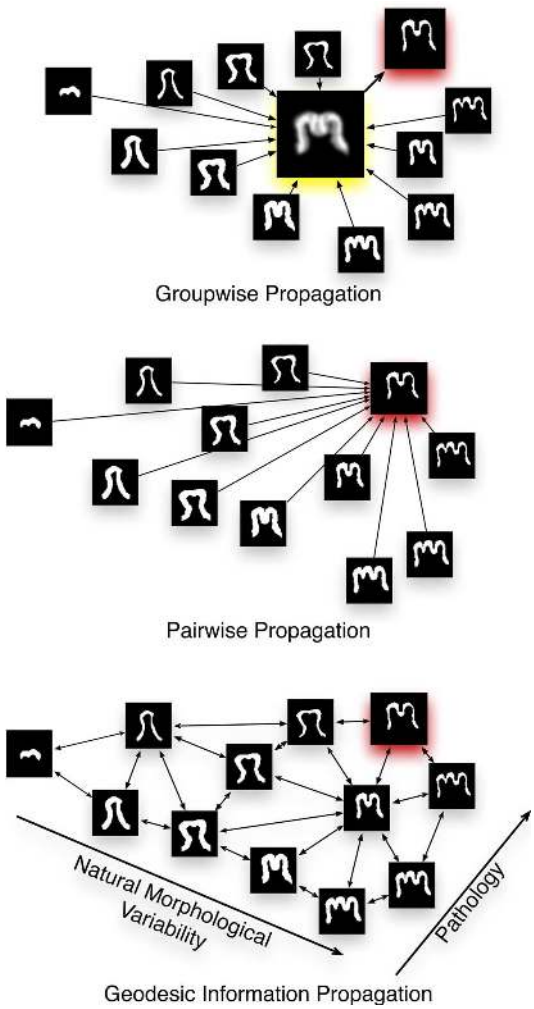

Fig. 1. Information flow from all subjects to a target subject (in red), using a groupwise, pairwise and geodesic aproach. Top) The information from all the subjects flows through the group mean (in yellow) and then to the target subject. Centre) The information is propagated directly from all subjects to the target subject. Bottom) The information flows only between the neighbours of the target subject.

a problem which has generated wide criticism [4], [5]; and (3) groupwise spaces confound the two sources of morphological variation by averaging natural variability (e.g. sulcal patterns, brain shape) and pathological effects (e.g. atrophy). This latter effect is detrimental for the purpose of information propagation.

In order to overcome the problem of mixing normal morphological variability and pathology, some groups [6], [7] have explored the idea of stratifying different pathological subgroups into disease-specific atlases. This process relies on either having *a priori* knowledge about pathological clustering assignments and clinical characteristics (e.g. age, gender) of each subject [6], or by optimising the clustering as part of the model [7]. Even though groupwise atlases become sharper, pathological stratification does not take into account the fact that there are different non-pathology-related morphological subgroups (e.g. different sulcal patterns). Subsequently, there is a need to further stratify the population into local morphological subgroups. In the limit, this population stratification process considers each training subject as independent prior information, which is then used to generate a patient specific prior [8]. Interestingly, this solution can be interpreted as the same problem solved by the multi-atlas segmentation propagation and label fusion community, but instead of generating a final categorical parcellation, one is actually estimating the prior probability of a certain voxel being assigned to a specific class.

Multi-atlas segmentation propagation and fusion uses a pairwise information propagation scheme. Many researchers have shown that propagating structural parcellations from multiple sources by mapping them to new unseen data using pairwise image registration, followed by a label fusion scheme, provides a good estimation of the true underlying parcellation [9], [10] (see Fig. 1—centre). However, this propagation strategy can be problematic in the case of limited and morphologically clustered source of information, e.g. propagating labels from an atlas [11] consisting of 30 young controls to a 90 year old diagnosed with Alzheimer's disease. As these parcellations are defined only on young controls with normal anatomy, it is non trivial to directly map this information to morphologically dissimilar and pathological subjects [12], [13] without introducing large errors. Recently, Wolz *et al.* [14] introduced the LEAP approach (learning embeddings for atlas propagation) for brain segmentation. In the LEAP framework, similarly to the work by Liu *et al.* [15], and by Lafon *et al.* [16], a low dimensional representation of the data is used to find a surrogate measurement of the morphologically similarity between datasets. This morphological similarity can then be used to propagate the segmentation between young subjects and AD subjects via intermediate datasets, greatly increasing the segmentation accuracy. Since the similarity metric used in LEAP is a global metric, the morphological embedding becomes less localised as the size of the structure to be segmented increases, resulting in a decrease in performance. A similar idea, but in the context of geodesic image registration, was introduced by Hamm *et al.* [17] with the GRAM (geodesic registration on anatomical manifolds) method. This method was later expanded to regional manifolds by Ye *et al.* [18]. Two more recent methodologies also provide interesting insights towards this general step-wise propagation idea, one by Jia *et al.* [19], which uses tree-based registration, and another by Wang *et al.* [20], which uses multiple registration paths. This family of step-wise propagation algorithms will become increasingly relevant with the availability of larger unlabeled databases. Ideally, one would like to slowly diffuse any information from the training examples to all the other images in a database in an unbiased manner.

We present a framework, named geodesic information flows (GIF), that propagates information between images using the geodesic path of a spatially-variant graph. This spatially-variant graph represents local patches of an implicit manifold using a heat kernel. GIF is a general formulation that can propogate many different types of information, such as labels, image intensities, or transformation matrices. In this manuscript we present two example applications: The propagation of categorical labels (similarly to multi-atlas segmentation propagation algorithms) and probabilistic tissue segmentations (using patient specific priors). By using a restricted neighbourhood for information propagation the proposed framework is not only more accurate but also less biased than state-of-the-art techniques. This paper is an extension of previous preliminary work [21].

II. Geodesic Information Flows

This section will first introduce the mathematical framework and the spatially-variant undirected graph for geodesic information flow, followed by the morphological similarity metric describing both image intensity similarity and the complexity of the coordinate mapping between images. Further details on the estimation of the geodesic distance and its advantages will then be provided. Finally, after building the local graph embedding, the geodesic information propagation framework will be applied to two well-known types of problems: propagation of categorical labels for image parcellation and propagation of probabilistic tissue priors for image segmentation.

A. Spatially-Variant Graphs

Graphs are ubiquitous in machine learning. They are used for a variety of applications, ranging from classification, image segmentation, dimensionality reduction and information propagation [15], [16].

Graphs can also be used to embed high-dimensional data in a low-dimensional manifold. The work by Wolz *et al.* [14] and by Gerber *et al.* [22] are good examples of graph-based dimensionality reduction strategies applied to brain images. These techniques use a low-dimensional representation of the data to propagate categorical labels between subjects [14], or describe the brain's morphological variability [22]. However, the features used by [14], necessary to project imaging data to a low-dimensionality space (data similarity, distance in the high dimensional space, angle preservation), require equicardinal data samples (i.e. all images should be resampled into a common unbiased discretisation grid, normally a group mean) [14]. Furthermore, due to the complexity and high dimensionality of brain data, the embedded dimensions of the manifold can lack interpretability and usefulness. For example, Gerber *et al.* [22] explored the manifold structure of the space of brain images and concluded that the first dimension of the manifold represents global ventricular expansion due to disease/ageing, while the second dimension is described as "less obvious". This sort of interpretation illustrates that local variations in morphology are hard to capture using a single global manifold. This hypothetical global manifold of brain morphologies would have to capture the local variation in sulcal patterns and subcortical shape between all the brain regions and all the subjects in a population, resulting in a very high-dimensional embedding.

Instead of characterising the morphology of the full brain, one should instead capture the local variation in morphology at separate spatial locations. This can be achieved using a local similarity metric as a measure of distance between mapped anatomical locations in different images [23], [18]. Initial work by Bhatia *et al.* requires a common discretisation space (i.e. resampling to MNI) and has very high computational and memory requirements, making the problem intractable for large datasets. As an example, to store a pairwise distance matrix at every voxel, assuming a set of 120 neighbouring images with average size 200^3, one would need approximately 429 GB of computational memory. Furthermore, the memory requirements will grow with $(N^2) * M$, where N is the number of datasets and M is the number of voxels in the common discretisation space, i.e. an N^2 matrix per voxel M. On the other side, the regional manifold learning method by Ye *et al.* [18] a utilises a common template as a registration target in order to spatially define the regions of interest. However, these regions have to be large ($37 \times 32 \times 32$ voxels) and non-overlapping in order to ensure a smooth and diffeomorphic mapping. Furthermore, while regional methods reduce the computational and memory burden, they are highly dependent on the choice and size of the regions of interest. While one can argue that the size $(N^2) * M$ of the full graph will be greatly reduced with a sparsifying operation (thresholding), it is still not suitable to current memory-limited systems.

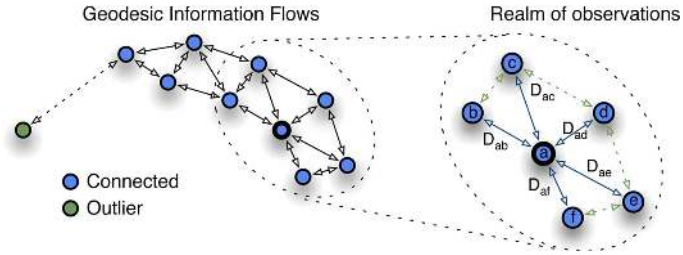

Fig. 2. Left) Implicit manifold with the neighbourhood defined as all the data points within a certain distance. Right) Diagram representing the local graph patch, the observed ($\in \mathcal{F}(a, \vec{v})$) and unobserved ($\notin \mathcal{F}(a, \vec{v})$) connections (in blue and green respectively) and distances from the standpoint of the voxel $\vec{v}$ in image a.

Thus, it is computationally impractical to have an explicit representation of the manifold at the voxel level. Instead of constructing an explicit representation of the manifold, one can assume the existence of an implicit manifold, which is here represented through local graph patches (see Fig. 2), i.e., a per-voxel subgraph that describes the local data morphological neighbourhood}. As this local graph patch is defined in the space of each image independently, it does not require a groupwise mapping between subjects. This obviates, to some degree, the problems related to groupwise matching of anatomical structures, as all registrations are pairwise, and also the problems of discretisation bias, as every image is discretised in the space of every other image. This local graph patch is obtained through pairwise mapping of every image in a database to every other image. While pairwise mapping is an N^2 problem, it can be easily distributed as a pre-processing step as each operation is independent. Furthermore, the amount of memory required to represent a local graph patch is linearly proportional to N, greatly reducing both computational complexity and memory requirements of the graph representation.

B. The Geodesic Information Flow Spatially-Variant Graph

Let a set Y of N images be the full set of observed T1-weighted MRI data with the i-th image of this set denoted by Y_i. Each image Y_i is a vector of size M_i, with the sample at position $\vec{v}$ being a graph vertex denoted by $\mathcal{V}(i, \vec{v})$. Note that M_i is not fixed for all images in Y, as this value is different depending upon field of view and image resolution.

We now define T_{ij} as a coordinate mapping between image Y_i and Y_j, found through pairwise registration. Here, $T_{ij}(\vec{v})$ represents the corresponding real-world location of the vertex $\mathcal{V}(i, \vec{v})$

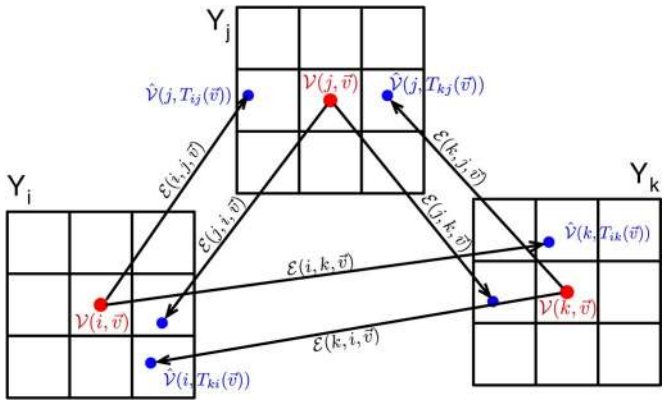

Fig. 3. A graph vertex $V(i, \vec{v})$ (red) in image Y_i is linked to a "virtual"*vertex* $\hat{\mathcal{V}}(j, T_{ij}(\vec{v}))$ (blue) by a graph edge $\mathcal{E}(i, j, \vec{v})$ (black).

in the space of image Y_j. Now, let $\hat{\mathcal{V}}(j, T_{ij}(\vec{v}))$ define a "*virtual*" *leaf vertex* at location $T_{ij}(\vec{v})$. As this *leaf vertex* $\hat{\mathcal{V}}$ is located at a non-integer position, it will sample its value from the underlying discrete grid on Y_j using an appropriate resampling function (e.g. trilinear, nearest neighbour).

One can now define a graph edge $\mathcal{E}(i, j, \vec{v})$ that connects the vertex $\mathcal{V}(i, \vec{v})$ in image Y_i and the corresponding "*virtual*" *vertex* $\hat{\mathcal{V}}(j, T_{ij}(\vec{v}))$ in image Y_j. Each edge will have an associated distance $D(i, j, \vec{v})$ describing the similarity between its two composing vertices. See Fig. 3 for a pictorial example of the graph vertices, "*virtual*" *vertices* and edges.

A subgraph $\mathcal{F}(i, \vec{v})$ is comprised of the vertex $\mathcal{V}(i, \vec{v})$, its corresponding "*virtual*" *vertices* $\{\hat{\mathcal{V}}(j, T_{ij}(\vec{v}))\}$ and their connecting edges $\{\mathcal{E}(i, j, \vec{v})\}, \forall j \neq i$. This subgraph $\mathcal{F}(i, \vec{v})$ connects the location $\vec{v}$ in image i to its mapped location in image $j, \forall j \neq i$, thus describing the local data neighbourhood. Finally, a pruning threshold d_t is introduced over the distances D, i.e. the edges $\mathcal{E}(i, j, \vec{v})$ in the subgraph $\mathcal{F}(i, \vec{v})$ are pruned if $D(i, j, \vec{v}) < d_t$.

An interesting point about this graph construction is that one does not need to explicitly represent the full graph in memory. In order to solve one iteration of the information diffusion problem at a given vertex $\mathcal{V}(i, \vec{v})$, one only needs to keep track of its subgraph $\mathcal{F}(i, \vec{v})$, visually shown in Fig. 2. Note that in the case of very high morphological variability, d_t can be set to 0, thus ensuring that all vertices are connected.

Similarly to diffusion maps, we now introduce a weight $W(i, j, \vec{v})$ characterising the contribution of vertex $\hat{\mathcal{V}}(j, T_{ij}(\vec{v}))$ and edge $\mathcal{E}(i, j, \vec{v})$ to the information of vertex $\mathcal{V}(i, \vec{v})$. The weight $W(i, j, \vec{v})$ is a property of the edge $\mathcal{E}(i, j, \vec{v})$, and characterises the amount of contribution that vertex $\hat{\mathcal{V}}(j, T_{ij}(\vec{v}))$ has to the reconstruction of the vertex $\mathcal{V}(i, \vec{v})$. Here, a heat kernel is used to reconstruct the missing information [24], [25]. This kernel is defined as

$$W(i, j, \vec{v}) = \exp\left(-\frac{(D(i, j, \vec{v}))^2}{\sigma}\right) \quad (1)$$

with σ being a heat kernel temperature that will determine the speed and the distance the information can diffuse. For all experiments in this paper, $\sigma = 1$ and $d_t = 0.1$.

C. The Distance Metric

The heat kernel decay function is based on the assumption that one can calculate a distance that is a surrogate of the morphological similarity between two vertices in the graph. Ideally, this distance should be at least a semi-metric, respecting the coincidence as well as separation axioms and symmetry. In a medical imaging framework, and more specifically in neuroimaging, the local distance between images should take into account both local morphology and local image similarity. To achieve this goal, Gerber *et al.* [22] and Ye *et al.* [18], both propose to use the complexity of the coordinate transformation as a distance metric that informs about the object's morphology. The coordinate transformation maps an image Y_i to an image Y_j by finding the optimal transformation T_{ij} that minimises some cost function. In order for $D(i, j, \vec{v})$ to be a semi-metric, this coordinate transformation has to be symmetric, inverse consistent and diffeomorphic. In our work, we use a symmetric variant of a non-rigid free-form registration algorithm [26]. Under the symmetry and diffeomorphism constraints, the transformation $T_{ij} = T_{ji}^{-1}$ and $T_{ij} \circ T_{ji} = \mathrm{Id}$, with T^{-1} being the inverse of the transformation, $\circ$ being the composition operator and Id the identity transformation. In order to remove the smoothly varying local affine component of the transformation that characterises the global anatomical shape differences, the low frequency component of the transformation is removed using a $20mm$ standard-deviation Gaussian kernel. From the resulting high-frequency version of the transformation, one can then find the displacement field F_{ij} that describes how much (in mm) a voxel $\vec{v}$ in Y_i had to move in order to match the corresponding voxel $T_{ij}(\vec{v})$ in Y_j.

Even though this displacement field will describe the morphological differences between different subjects, we also combine it with an intensity similarity metric in order to assess the local similarity between the images after transformation [27]. This similarity term is necessary to characterise both the local differences in tissue appearance due to pathology (e.g. damaged white matter (WM) in dementia) and also some possible local registration errors. The local similarity between an image Y_i and an image Y_j transformed by T_{ij}, denoted by L_{ij}, can be calculated as the kernel local sum of squared differences (LSSD) between the intensity in these images, using a cubic B-spline as a kernel, i.e. $L_{ij} = B_S * (Y(i, \vec{v}) - Y(j, T_{ij}(\vec{v})))^2$ with B_S being the B-spline kernel and $*$ as the convolution operator. We combine the two semi-metrics together by setting

$$D(i, j, \vec{v}) = \alpha L_{ij}(\vec{v}) + (1 - \alpha) F_{ij}(\vec{v}), \quad (2)$$

with α being a relative weight (here set to 0.5), meaning that both a low displacement and a low LSSD are necessary to obtain a low distance $D(i, j, \vec{v})$ between images. The intensity images Y are z-scored before estimating L, in order to balance the influence of L and F in the metric. The mean and standard-deviation of the observed intensities within the foreground region are used for the z-scoring procedure. The foreground region (i.e. the full head) is obtained through an Otsu threshold.

Note that $F_{ij}(\vec{v})$ can be defined using the differential of either the displacement field or the velocity field of the diffeomorphic registration rather than the proposed heuristic high-pass filtering

methodology. Nonetheless, better results were found with the proposed high-pass filtering methodology.

D. Geodesic Distance Estimation

When propagating the information through the spatially-variant graph, (2) assumes that the distance between the vertices, and thus the quality of the available information is only dependent on their pairwise distance between position $\vec{v}$ in image i and its neighbours $T_{ij}(\vec{v})$. However, one should note that in theory, vertices that are closer to the source of information should have more accurate segmentations, as the extrapolation error is lower. It would thus be ideal if this accuracy metric was also used for the information flow process.

Let K be a set of all manually annotated images in a database, with $K \subset N$. Now, let $G(i,\vec{v})$ be a characteristic of vertex $\mathcal{V}(i,\vec{v})$, describing the amount of information extrapolation. G is defined as the geodesic distance along the graph edges between vertex $\mathcal{V}(i,\vec{v})$ and the closest source of manual labeled information in image Y_j, $\forall j \in K$. Note that by definition, $G(j,\vec{v}) = 0$, $\forall j \in K$, as j is manually annotated. In the geodesic information flow framework, $G(i,\vec{v})$ cannot be directly estimated as one only has access to the subgraph $\mathcal{F}$. However, $G(i,\vec{v})$ can be obtained by iteratively solving at every vertex $\mathcal{V}(i,\vec{v})$ of every image $i \notin K$ by

$$G(i,\vec{v})^t = \underset{j \in D(i,j,\vec{v}) < d_t}{\arg\min} \left(\hat{G}\left(j, T_{ij}(\vec{v})\right)^{(t-1)} + D(i,j,\vec{v}) \right) \tag{3}$$

i.e., the geodesic distance at iteration t and vertex $\mathcal{V}(i,\vec{v})$ is equal to the smallest value of the neighbour's geodesic distance $\hat{G}(j,T_{ij}(\vec{v}))$ at iteration $(t-1)$ plus the pairwise distance $D(i,j,\vec{v})$, for all neighbouring vertices. The value of $\hat{G}(j,T_{ij}(\vec{v}))$ for a "*virtual*" *vertex* $\hat{\mathcal{V}}(j,T_{ij}(\vec{v}))$, required by (3), is obtained through trilinear interpolation of it closest vertices $\mathcal{V}$ in image Y_j. One should also note that for all $i \notin K$, the geodesic distance is initialised to $G(i,\vec{v}) = +\infty$. We will see later in Section III that setting $G(i,\vec{v})$ to $+\infty$ removes the influence of the unsolved node i at $\vec{v}$ in the information propagation step. An example of G is shown in Fig. 6.

Note that this iterative geodesic minimisation algorithm is analogous to the Bellman-Ford algorithm. The main advantage of Bellman-Ford in this context pertains with the fact that one does not need to keep track of the node with the minimum distance value at each iteration, meaning that we can solve the geodesic path search by having access only to $\mathcal{F}$. This minor detail allows for solving the geodesic path problem without storing the full graph or a graph queue in memory. Also, the proposed geodesic distance is an heuristic solution to a problem that could have been solved in a more principled manner by many methods present in the literature. However, these methods commonly require access to the full graph or to the graph laplacian, a structure which is not available in this work due to the memory constraints explained in Section II-A.

III. Application to Label Fusion

The two previous sections have defined the neighbourhood graph and the distance metric. This section will make use of the graph structure to introduce the concept of propagating information between neighbouring vertices of the graph. More specifically, information here refers to the propagation of categorical labels as done in a multi-atlas propagation and fusion.

Let $\mathcal{L}(i,\vec{v})$, defined in the domain of Y_i, represent some annotation or label at vertex $\mathcal{V}(i,\vec{v})$. Under the assumption that only a subset K of the images are initially labeled, $\mathcal{L}(i,\vec{v})$ is only defined $\forall i \in K$, where $K \subset N$. The aim of the information propagation step is to obtain an estimate of $\mathcal{L}(i,\vec{v})$ for $i \notin K$.

As the realm of observations at each spatial location $\mathcal{V}(i,\vec{v})$ is limited by the subgraph $\mathcal{F}(i,\vec{v})$, one can approximate the information at $\mathcal{V}(i,\vec{v})$ by a combination of the information available within the subgraph, using the heat kernel defined in (1). As the degree of the "*virtual*" *vertices* of the subgraph is 1, the reconstruction of the data at $\mathcal{V}(i,\vec{v})$ is equivalent to a normalised weighted sum of the information available within subgraph. Thus, can be obtained by iteratively solving $\mathcal{L}(i,\vec{v})$

$$\mathcal{L}(i,\vec{v})^t = \frac{\sum_j W(i,j,\vec{v}) \hat{\mathcal{L}}\left(j, T_{ij}(\vec{v})\right)^{(t-1)}}{\sum_j W(i,j,\vec{v})}, \tag{4}$$

solved for all $i \notin K$, i.e. for all datasets where the information is not defined. Here, $T_{ij}(\vec{v})$ is the spatially transformed coordinate $\vec{v}$ into the space of image Y_j, mapped using the previously described transformation, and t is the current iteration number. The information flow is thus governed by the heat kernel-derived weights $W(i,j,\vec{v})$. Note that the above equation is similar to most weighted voting and patch based algorithms. Also, if $\hat{G}(j,T_{ij}(\vec{v})) = +\infty$ for all the "*virtual*" *vertices* of $\mathcal{F}(i,\vec{v})$, then $\sum W(i,j,\vec{v}) = 0$, and subsequently $\mathcal{L}(i,\vec{v})$ will not be defined. However, this is not a problem as $G(i,\vec{v})$ (see (3)) will tend to $+\infty$, meaning that image i will always have a weight $W(j,i,\vec{v}) = 0 \forall i \neq j$. Note that (4) is only valid for continuous data and not for categorical labels. The same equation can be reformulated in a weighted label fusion scheme by making $\mathcal{L}(i,\vec{v})$ equal to $p(\mathcal{L}(i,\vec{v}),l)$, representing the probability that location $\vec{v}$ in image i has label l. The value of $\hat{\mathcal{L}}(j,T_{ij}(\vec{v}))$ is a property of the "*virtual*" *vertex* $\hat{\mathcal{V}}(j,T_{ij}(\vec{v}))$, and its value is obtained through interpolation of it closest vertices $\mathcal{V}$. This interpolation process can be either nearest neighbour interpolation for categorical labels or trilinear interpolation for probabilistic labels. Both (4) and (3) are solved iteratively for all $i \notin K$.

In this work, the geodesic distance is taken into account as an estimate of uncertainty due to extrapolation. Thus, one can reformulate (1) as

$$W(i,j,\vec{v}) = \exp\left(-\frac{\left(\hat{G}\left(j, T_{ij}(\vec{v})\right) + D(i,j,\vec{v})\right)^2}{\sigma} \right), \tag{5}$$

The reader should note that under this reformulation, (5) does not represent a pure diffusion process anymore, as it is now dependent on G. Nonetheless, the introduction of G in (5) minimises the length travelled by the propagated label through the graph, which not only reduces extrapolation error, but also propagates information faster (than in (1)).

Using (5), if i is an unsolved or disconnected vertex, then $W(j,i,\vec{v}) = 0$ as e^{-x} will tend to 0 when x tends to ∞. Similarly, a source vertex i will have weight dependent only on $D(i,j,\vec{v})$. As this weighted fusion scheme is analogous to a local weighted voting strategy under geodesic propagation,

GIF will be referred to by the name GIF+LWV in the rest of the paper. Note that the proposed GIF framework is only a graph construct which enables the propagation of information, meaning that the proposed local weighted voting strategy ((4)) represents only one example of its application for label fusion.

In all experiments, we assume that the algorithm has converged when the mean (for all the nodes) change in Geodesic distance between iterations is below 0.01, which normally happens in less than 10 iterations.

IV. Application to Tissue Segmentation

Another interesting application of the GIF framework is the problem of tissue segmentation. This application builds on the work by Van Leemput *et al.* [1]. The tissue segmentation problem is modelled as a *maximum likelihood* (ML) probabilistic model defined as

$$\hat{\Phi} = \arg\max_{\Upsilon} \log \left[\sum_{L} P(Y|L,\Phi) \right]$$

with $\Phi = \{\mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k, c_1, \ldots, c_m\}$ being a vector of model parameters. The observed intensities are modelled as a mixture of K tissue classes with parameters μ and Σ characterising the mean vector and the covariance matrix respectively. As in [1], intensities are assumed to be corrupted by a smoothly varying bias field, modelled using m polynomial basis functions and m basis coefficient.

A. Spatially Variant Prior Over L

The GIF graph over a set of data is introduced into the segmentation framework [1] through a modification of the Markov Random field (MRF) model. The change to MRF model proposed in this work preserved all parameter update equations presented in [1]. We thus refer the reader to the original work by Van Leemput *et al.* for the model optimisation.

As previously defined in Section II, let W and T be a set of model parameters characterising the similarities between subjects in a database and their pairwise coordinate mappings respectively. These parameters are assumed to be given *a priori* and are introduced into Φ, now defined as $\Phi = \{\mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k, c_1, \ldots, c_m, W, T\}$. The MRF energy function ($U_{\mathrm{MRF}_{jk}}$) presented in [1] is modified to incorporate both a spatial constrain, i.e. the segmentation should vary smoothly between neighbouring voxels, and the GIF graph constraints, i.e. two morphologically similar locations in two different images should have similar tissue segmentations. As in [1], we use a mean field approximation and assume independence between the spatial neighbourhood and the graph neighbourhood. Thus, the probability that the hidden label z at location j is of type K is defined as

$$P(z_{jk}|\Phi) = \frac{e^{U_{\mathrm{MRF}_{jk}}}}{\sum_{k'} e^{U_{\mathrm{MRF}_{jk'}}}} \tag{6}$$

where U_{MRF} will contain a term over $\mathcal{N}_{ij}$, i.e. the first-order spatial neighbours of pixel j in image i, and a term over GIF graph, with $\mathcal{S}_j$ representing the first order mapped locations from pixel j in the current image to the corresponding locations on the space of the other images. This MRF has two components, the first enforcing spatial smoothness of the segmentation and the second enforcing smoothness between the different morphologically similar images in the database. This second MRF provides a way for information to flow between subjects in a database and can be seen as the main contribution of this section to the classic probabilistic framework. The U_{MRF} term is defined as

$$U_{\mathrm{MRF}} = U^{\mathcal{N}}_{\mathrm{MRF}_{ijk}} + U^{\mathcal{S}}_{\mathrm{MRF}_{ijk}} \tag{7}$$

The spatial smoothness term is defined, similarly to [13], as

$$U^{\mathcal{N}}_{\mathrm{MRF}_{jk}} = \underbrace{\left[-\beta \sum_{k' \in K} h_{kk'} \left(\sum_{j' \in \mathcal{N}_j} n_{jj'}\, p_{j'k'} \right) \right]}_{\text{spatial neighbourhood}}$$

where, $n_{jj'} = 1/d_{jj'}$, with $d_{jj'}$ being the real-world distance between the centre of voxel j and j' in the image being segmented, β is a scaling term that controls the strength of the neighbourhood constraint and $h_{kk'}$ is neighbourhood energy function defined in [1]. The GIF graph smoothness term is defined as

$$U^{\mathcal{S}}_{\mathrm{MRF}_{jk}} = \underbrace{\left[log \left(\sum_{i \in \mathcal{S}_j} w_{ij} t_{ij}(p_{ik}) \right) \right]}_{\text{population neighbourhood}}$$

where w_{ij} is the morphological similarity weight between the current subject and subject i at location j, i.e. $w_{ij} = W(*, i, j)$ as defined in (5), with $*$ being the current image. Also, t_{ij} is the change of coordinate system between the current subject and subject i in order to sample p_{ik} at location j, i.e. t_{ij} resamples p_{ik} to the space of the current subject.

Note that under the assumption that a database of manually segmented training images is available, if one only wants to segment a single image, then the "population neighbourhood" component of the $U_{\mathrm{MRF}_{jk}}$ energy term will not change at each iteration and (6) can be seen as a common static prior. In this special case, (6) becomes

$$P(z_{jk}|\Phi) = \frac{\pi_{jk} e^{U^{N}_{\mathrm{MRF}_{jk}}}}{\sum_{k'} \pi_{jk'} e^{U^{N}_{\mathrm{MRF}_{jk'}}}}$$

with the π_{jk} term defined as

$$\pi_{jk} = \sum_{i \in \mathcal{S}_j} w_{ij} t_{ij}(p_{ik}).$$

Note that if one assumes that w_{ij} is the same for every image and if the transformation t_i is approximated as the composition of transformations from image i to the groupwise average and then from the groupwise average to the current image, then the proposed formulation becomes the framework proposed by Van Leemput *et al.* [1] and Ashburner *et al.* [2]. One can then see (6) as a generalisation of the classical atlas-based prior probability.

V. Validation: Label Fusion

The data used in this work for the validation of the label fusion application is comprised of four datasets.

- 30 T1-weighted MRI images from young controls with associated structural parcellation of 83 key structures, here denoted as the Hammers dataset [11] (http://www.brain-development.org)

- 90 subjects from the ADNI database. The ADNI database was subdivided into 30 controls, 30 Mild Cognitive Impairment (MCI) and 30 Alzheimer's diseased (AD) patients with associated manual segmentations of the brain, here denoted as the ADNI dataset (http://adni.loni.ucla.edu).
- 35 T1-weighted MRI images from young controls with associated structural parcellation of 143 key structures as provided by Neuromorphometrics for the MICCAI 2012 Grand Challenge on label fusion, here denoted as the Neuromorphometrics dataset (https://masi.vuse.vanderbilt.edu/workshop2012/index.php/Challenge_Details).
- 20 T1-weighted and T2-weighted MRI images from neonatal subjects (5 term subjects and 15 preterm subjects) with associated structural parcellation of 50 key structures, here denoted as the ALBERT dataset [28] (http://www.brain-development.org).

In this and the next section, $\alpha = 0.5$. Optimisation of α will be addressed in future work.

One of the aims of the current work aims is to homogenise databases under the assumption that extra information is only available on a subset of the data. From these sources of information, measuring the information extrapolation accuracy will always be limited by the anatomical and pathological variability within the full dataset and by the range of available segmentations. Furthermore, the most complex sources of information, like the 30 young controls with full brain parcellations, are currently not available in pathological subjects. This makes the validation anecdotal for untested morphologies.

Within the scope of label fusion, the proposed validation will thus have four subsections. The first experiments will access the segmentation accuracy in a leave-one-out (LOO) setup. However, due to the LOO approach, this experiments will only characterise empirically the overall performance of the GIF+LWV propagation strategy against well known fusion strategies. It does not highlight the ability to extrapolate information as the LOO validation strategy makes GIF+LWV analogous to a pairwise LWV approach. The second and third experiments characterise not only the accuracy of information extrapolation by propagating the segmentations from a training dataset to a morphologically different testing dataset, but can also be compared to previously published state of the art methodologies. Here GIF+LWV is compared to a pairwise version of the local weighted voting algorithm (Pair+LWV). This experiment does not attempt to show that the proposed fusion algorithm is better than state-of-the-art methodologies. We are only assessing the impact of the geodesic propagation in comparison to pairwise propagation. The final experiment demonstrates the trivial extension of the proposed distance D to multimodal data. This experiment shows that multimodal data dramatically improves propagation results. This experiment also shows the advantage of using the geodesic distance ((4)) when compared to using only the edge distance ((1)) as proposed in preliminary work.

A. Leave-One-Out Cross Validation on the Hammers Dataset

The accuracy of propagating information through a geodesic path was compared to MAPER [12] using the Hammers dataset. The results for MAPER were kindly provided by the author of [12]. As the amount of parcellations available for validation is limited, a leave-one-out cross validation was performed only on the 30 young controls that have manual brain parcellations. The left-out manual segmentations were then used as the *gold standard* for comparison. One should notice that the limited availability of segmentations restricts the range of morphological variability in the propagation, thus not representing the real performance when segmenting morphologically dissimilar subjects.

TABLE I
MEAN DICE COEFFICIENT FOR A SET OF KEY STRUCTURES, COMPARING THE PROPOSED METHOD (GIF+LWV) WITH MAPER [12] ON CLINICALLY RELEVANT STRUCTURES

Structure	Unilateral Structures		
	GIF+LWV	MAPER	
	Mean Dice	Mean Dice	p-value
All Structures	0.818	0.809	$< 10^{-4}$
Corp. callos.	0.880	0.867	$< 10^{-4}$
Brainstem	0.953	0.938	$< 10^{-4}$
Structure	Left Side		
	Mean Dice	Mean Dice	p-value
Hippocampus	0.844	0.834	0.004
Amygdala	0.826	0.792	$< 10^{-4}$
Cerebellum	0.971	0.966	0.002
Caudate nucl.	0.898	0.892	0.037
Nucleus acc.	0.758	0.683	$< 10^{-4}$
Putamen	0.907	0.892	$< 10^{-4}$
Thalamus	0.921	0.888	$< 10^{-4}$
Pallidum	0.856	0.766	$< 10^{-4}$
Lateral vent.	0.902	0.898	0.001
Structure	Right Side		
	Mean Dice	Mean wDice	p-value
Hippocampus	0.824	0.821	0.212
Amygdala	0.824	0.783	$< 10^{-4}$
Cerebellum	0.973	0.967	$< 10^{-3}$
Caudate nucl.	0.902	0.896	0.027
Nucleus acc.	0.732	0.671	$< 10^{-4}$
Putamen	0.910	0.896	$< 10^{-4}$
Thalamus	0.920	0.885	$< 10^{-4}$
Pallidum	0.855	0.767	$< 10^{-4}$
Lateral vent.	0.912	0.908	0.002

In this paper, the Dice score was used as a measure of accuracy. The mean Dice scores per structure for the LOO cross validation are shown in Table I. Out of 83 structures, 15 structures had a significantly higher Dice score using the GIF+LWV when compared to MAPER, while only two structures (lingual gyrus and superior parietal gyrus) where better segmented in MAPER. Under the LOO setting, the mean Dice score over all structures and all patients for the proposed method (0.8182) was significantly higher ($p < 10^{-4}$) than in MAPER (0.8089) using a two-tailed paired t-test. A parametric t-test was used in this experiment because the pairwise errors were approximately Gaussian.

B. Information Extrapolation Accuracy Using ADNI

In the previous subsection, the accuracy of propagating information through a geodesic path was limited to a morphologically similar set of subjects due to the use of a LOO cross validation strategy. Thus, the previous validation does not capture the GIF+LWV ability to extrapolate information to anatomically disparate subjects. The information extrapolation accuracy is here assessed by using a restricted subset (cognitively

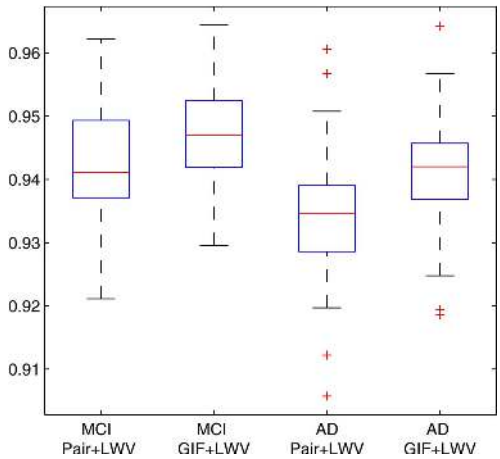

Fig. 4. Dice scores for pairwise (Pair+LWV) and geodesic (GIF+LWV) propagation of the brain mask.

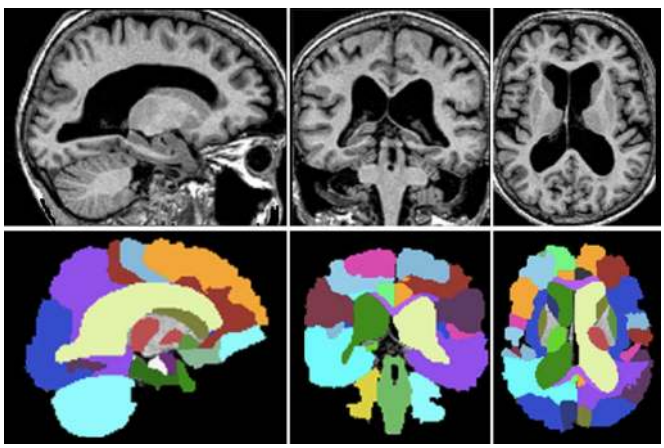

Fig. 5. An example of the propagation of the structural parcellation to an atrophied subject (ID:1049) from the ADNI database. Note the correct ventricle segmentation and the smooth deep grey matter parcellation.

normal elderly control group) of all the manual brain segmentations as training data. This morphologically clustered set of data is then used to segment both the MCI and AD groups, assumed in this work to be morphologically less similar than the subjects within the training population. The manual brain segmentations of the MCI and AD groups were used as *gold standards* for comparisons. The proposed geodesic propagation algorithm is compared to a direct pairwise propagation algorithm, hereafter named Pair+LWV, based on the locally weighted majority voting algorithm with a inverse exponential weight proposed by Yushkevich *et al.* [29]. This algorithm was chosen due to its similarities with the proposed technique, resulting in an experiment that compares mostly the advantages of GIF versus pairwise fusion under the same voting scheme and using the same data pre-processing and pairwise registrations.

The results are presented in Fig. 4 and Fig. 5, with segmentation accuracy measured using Dice score. The mean (std) Dice score for the proposed geodesic method was 0.941(0.008) and 0.949(0.008) for the AD and MCI groups respectively while for the direct method, the mean (std) Dice score was 0.934(0.009) and 0.942(0.008) for the AD an MCI groups respectively. This

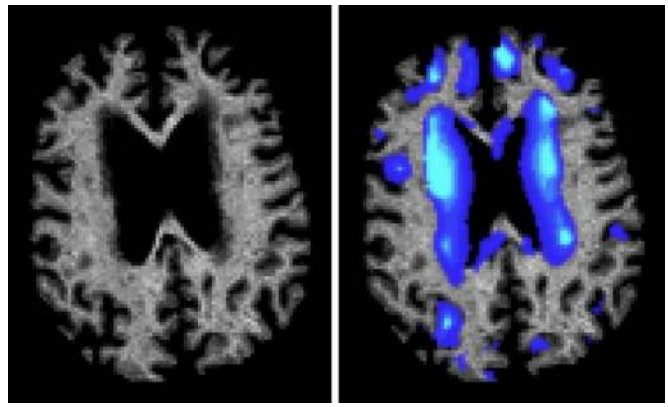

Fig. 6. An example of the geodesic distance G_j for an AD subject when measured from a database of cognitively normal subjects. Areas with high G_j (light blue) correlate with regions known to be associated with AD pathology. The image was skull stripped and low values of G have been set to transparent for visualisation purposes.

represents a statistically significant ($p < 10^{-4}$) increase in segmentation accuracy when using a two-tailed paired t-test for statistical comparison. Note that one should not compare these results with other brain segmentation methods due to the lack of post-processing, the limited size of the training set and the fact that these brain segmentations do not include inter-sulcal CSF.

An interesting outcome of this experiment is presented in Fig. 6, which represents the geodesic distance G_j at convergence for an AD subject when measured from a database of cognitively normal subjects. In other words, the figure presents areas that are morphologically distant from a cognitively normal population. Note that regions with high G_j are associated with AD pathology, i.e. periventricular lesions, sulcal openings.

C. Information Extrapolation Using the MICCAI 2012 Challenge Data

This section validates the GIF methodology on 35 T1-weighted MRI images from the Neuromorphometrics dataset. All subjects were controls with associated structural parcellation of 143 key structures as provided by Neuromorphometrics for the MICCAI 2012 Grand Challenge on label fusion (https://masi.vuse.vanderbilt.edu/workshop2012/index.php/Challenge_Details).

The aims of this experiment are two fold: First, this experiment aims not only to serve as a ground for comparison to other techniques, but also to allow for an unbiased comparison between GIF and the equivalent local weighted fusion algorithm with the same weighting function. This validation is unbiased as it uses exactly the same pre-processing, registration strategies and assessment methods. In this first experiment, 15 subjects are used as training datasets, whilst the other 20 subjects are used as training datasets, as defined in the MICCAI challenge setup. Second, we artificially reduce the size of the training database to 5 subjects and then to only 1 subject, in order to demonstrate the robustness to extreme situations. These 5(1) subjects were selected as the 5(1) youngest females from the initial 15 training subjects in order to select a morphologically clustered subset.

As an extra insight, we show how the accuracy of GIF evolves with each iteration. It is also important to note that GIF with

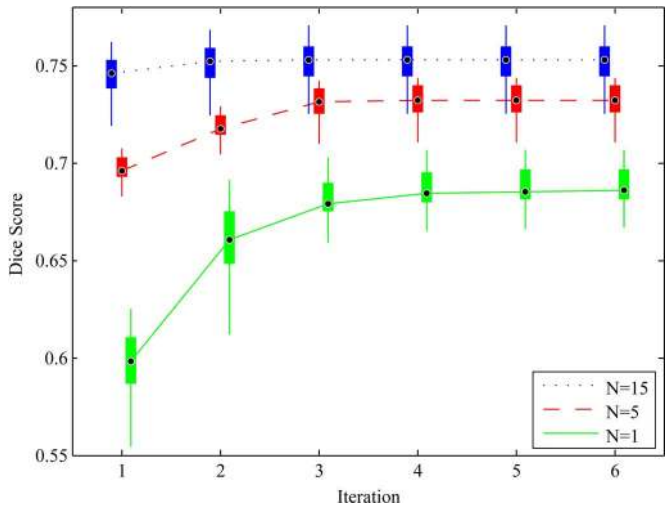

Fig. 7. Average Dice score over all relevant regions and testing subjects for GIF+LWV. Each line represents the same test but starting from different number of training datasets N. Note the improvement in accuracy with each iteration.

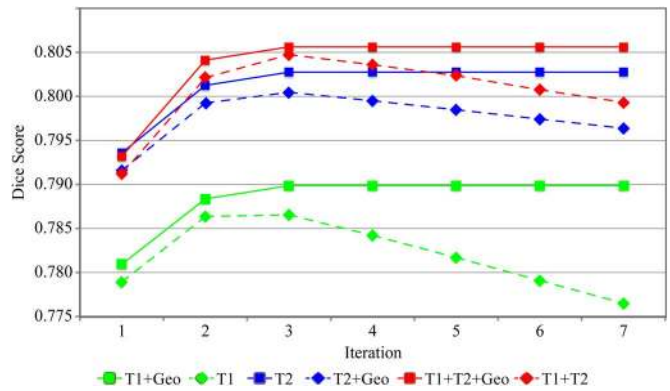

Fig. 8. Average Dice score over all regions and testing subjects for GIF with (full line) and without (dashed line) the geodesic distance using T1 data, T2 data and joint T1/T2 multimodal data. Note the improvement in accuracy with each iteration for the method with the geodesic distance, and the degradation of the results if this is not used.

only one iteration is equivalent to paired weighted label fusion (Pair+LWV). Thus, GIF's results after convergence (here needing 6 iterations) can be directly compared to GIF's results after 1 iteration, providing a comparison between GIF+LWV and Pair+LWV when using exactly the same setup, similarity kernel, pre-processing and implementation.

The average Dice score for all testing subjects and for all relevant regions as defined in the MICCAI challenge website is provided in Fig. 7. GIF after 6 iterations obtained a Dice score of 0.755, in line with the best methods of the MICCAI challenge. More specifically, even with a simpler fusion model, the proposed method would have ranked 4th out of 25. The best method performed 0.01 Dice above GIF, but used a correction strategy that could be used to post-processing of any other algorithm (including GIF). Also, these results should always be carefully compared between methods as each submitted methodology uses a different pre-processing strategy. This pre-processing acts as an accuracy confound.

A more interesting result was the performance of the proposed GIF algorithm on the restricted training dataset. GIF with only 5 training subjects obtained a Dice score of 0.728, an accuracy which is better than 13 other label fusion algorithms from the MICCAI challenge that were using 15 training datasets. Finally, the most extreme experiment (where $\mathrm{N} = 1$) showed that using only one training dataset, the GIF methodology can obtain an average Dice score of 0.681, which represents a gain of 0.084 in Dice score when compared to weighted voting. Note that because $\mathrm{N} = 1$, Pair+LWV becomes equivalent to single atlas propagation. In all three experiments, GIF+LWV performed significantly better ($p < 0.001$) than Pair+LWV using a Wilcoxon Signed-Rank test.

D. The Value of Multimodal Imagin Using the ALBERT Dataset

For the final label fusion validation we use the ALBERT dataset, comprised of 20 T1-weighted and T2-weighted MRI images from neonatal subjects (5 term subjects and 15 preterm subjects) with associated structural parcellation of 50 key structures [28] (http://www.brain-development.org). In this experiment, the 5 term subjects were used as training datasets and the 15 preterm datasets were used as testing datasets. The aims of this experiment are two fold: First we want to show that using the geodesic distance ((4)) is beneficial when compared to using only the edge distance ((1)) both in terms of accuracy and stability; second, we want to show that GIF can be extended to multimodal data in a trivial manner and that this multimodal extension can provide substantial accuracy advantages.

In order to incorporate multimodal information into the algorithm, the local similarity $L_{ij}(\vec{v})$, previously defined as the LSSD between two images, is replaced with the sum of the LSSD between each modality, i.e. $L_{ij}(\vec{v}) = \mathrm{LSSD}_{T1} + \mathrm{LSSD}_{T2}$. The deformation field between image pairs is also estimated using both modalities and the locally normalised cross correlation as a image similarity. All other equations remain the same.

The average Dice score for all testing subjects and for all relevant regions is provided in Fig. 8.

We observed that the performance of the GIF algorithm with the geodesic distance plateaus after 3 iterations, whilst the version of GIF only using pairwise distances starts degrading its accuracy after 3 or 4 iterations. This is caused by label propagation oversmoothing due to the unconstrained heat kernel formulation.

The average accuracy after convergence when starting from only 5 training datasets to the remaining 15 testing was 0.805, which is greatly improved when compared to the results presented in [30] where the author performs a leave-one-out cross validation. This demonstrates that the proposed algorithm can obtain good results even in low-contrast neonatal datasets, mainly when using multimodal data.

VI. Validation: Tissue Segmentation

The validation of the GIF framework for tissue segmentation will be comprised of two sections with three experiments:

1) An experiment on synthetic data with ground truth segmentations using the 20 BrainWeb (www.bic.mni.mcgill.ca/brainweb) datasets in a leave-one-out fashion.

TABLE II
BRAINWEB RESULS (SYNTHETIC DATA). DICE OVERLAP STATISTICS ARE PRESENTED. THE P-VALUES COMPARE EACH METHOD TO BOTH GIF+SEG AND GIF + Seg_{C}, ACCORDING TO THE WILCOXON SIGNED-RANK TEST

	WM			GM		
	Dice Coef. mean (std)	p-value vs. GIF+Seg	p-value vs. GIF+Seg_{C}	Dice Coef. mean(std)	p-value vs. GIF+Seg	p-value vs. GIF+Seg_{C}
GIF+Seg	0.957 (0.007)	-	-	0.944 (0.006)	-	-
GIF+Seg_{C}	0.941(0.013)	-	-	0.935 (0.013)	-	-
GW-EM	0.918 (0.023)	$<10^{-4}$	$<10^{-4}$	0.917 (0.025)	$<10^{-4}$	$<10^{-4}$
SPM	0.931 (0.010)	$<10^{-4}$	$<10^{-4}$	0.925 (0.015)	$<10^{-4}$	$<10^{-4}$
GIF-LWV	0.902 (0.010)	$<10^{-4}$	$<10^{-4}$	0.893 (0.011)	$<10^{-4}$	$<10^{-4}$

2) A leave-one-out cross-validation experiment using the 35 subjects from the Oasis database (www.oasis-brains.org) that have the corresponding silver-standard manual segmentations provided by *Neuromorphometrics, inc*. This dataset includes some highly pathological subjects suffering from ventricular expansion, atrophy, WM hypo-intensities and imaging artefacts.
3) The robustness to discrepant morphologies is assessed by separating the 35 subjects into a training (the 5 youngest females) and test group (the remaining subjects). This validation tests the ability to segment subjects that are highly different from the training population.

A. Synthetic Data With Ground-Truth Segmentations

20 datasets were downloaded from the BrainWeb MR image simulator. Each dataset contained a simulated T1-weighted image and corresponding segmentations of the grey matter (GM), white matter (WM) and Cerebrospinal fluid (CSF). The simulated data was generated using a spoiled FLASH sequence with TR $= 22$ ms, TE $= 9.2$ ms, $\alpha = 30^{\circ}$ and 1-mm isotropic voxel size with simulated 3% noise and 20% INU [31]. A leave-one-out cross-validation strategy was used to validate the segmentation accuracy when compared to the ground-truth segmentation. The tissue segmentations used for MRI simulation are used as *ground truth* for comparison. We validated the proposed tissue segmentation method (abbreviated to GIF+Seg) against a standard EM segmentation based on a groupwise population atlas [1] (abbreviated to GW-EM) as implemented in NiftySeg (niftyseg.sf.net), against SPM12b [2] (abbreviated to SPM) and also against the GIF based locally weighted label fusion algorithm (abbreviated to GIF+LWV) presented in the previous section of this work. All methods are tested using a leave-one-out cross validation strategy, assuming that all the images except the image under analysis are training sets. The GIF method is also tested using test/training jackknifing cross-validation (denoted as GIF + Seg_{C}), where for each testing image, a subset of 10 of the remaining images is randomly selected as training data. This last jackknifing comparison strategy tests, to some degree, the robustness of GIF to a reduced number of training samples.

For both GIF and GIF + Seg_{C}, W_i, T_i are obtained using the procedure described in Section II-D. For the GW-EM method, in order to segment a subject, the remaining 19 subjects are used to create a population prior. This population prior is then registered to the image under study (as illustrated in Fig. 1-left) using a sequence of affine and non-rigid registrations. Finally, for the GIF+LWV method, the weighted majority voting label-fusion technique presented in this work is used. The accuracy of the segmentation was measured using the Dice overlap.

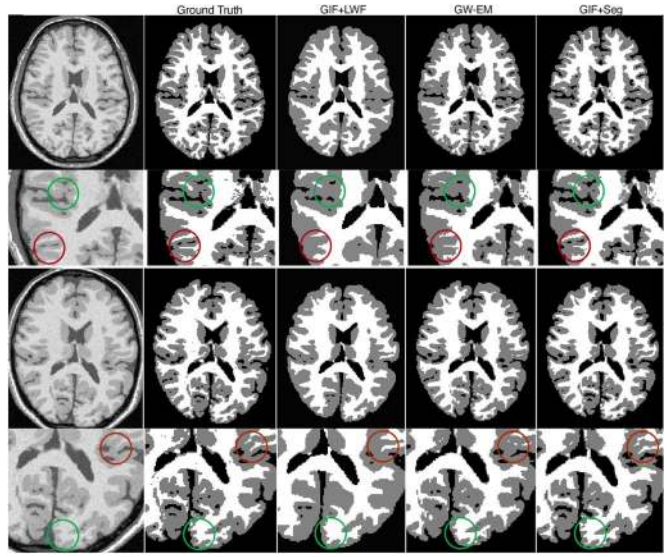

Fig. 9. From left to right: A synthetic T1-weighted image from the Brainweb database and its corresponding ground-truth segmentation, segmentation using label fusion, segmentation using a population prior in a groupwise space and segmentation using GIF+Seg. The first two and last two rows correspond to subjects 04 and 53, respectively. The red and green circles highlight areas with large variations between methods.

Examples of the segmentation results are shown in Fig. 9 the population statistics are shown in Table II. A two-tailed non-parametric Wilcoxon Signed-Rank test was used to assess statistical significance due to the non-Gaussian nature of the pairwise errors. This test was chosen due to the non-Gaussian nature of the Dice coefficient distributions caused by a heavy tail. Using the Wilcoxon Signed-Rank test, the proposed method (GIF+Seg) achieved statistically significantly higher ($p < 10^{-4}$) Dice overlap when compared to the other techniques.

B. Clinical Data With Manual Segmentations

A set of 35 subjects from the OASIS reliability dataset [32] were manually segmented by *Neuromorphometrics, inc*. into 140 different labels. These 140 labels were combined into 8 tissue classes: cortical GM and WM, cerebellar GM and WM, extra-cerebral and ventricular CSF, deep GM structures and pons. It is important to consider that both the deep GM and pons manual segmentations are based on geometrical assumptions and anatomical knowledge and not on an intensity distribution, and thus segmenting these tissues assuming Gaussian intensity

TABLE III
DICE OVERLAP STATISTICS BETWEEN EACH METHOD AND THE SILVER-STANDARD WHEN DOING A LEAVE-ONE-OUT CROSS VALIDATION (TOP) AND USING ONLY 5 SUBJECTS AS TRAINING SAMPLES (BOTTOM). THE HIGHEST MEAN IS IN BOLD. ALL P-VALUES EXCEPT BETWEEN GIF+LWV AND GIF+SEG ON THE DEEP GM (BOTH EXPERIMENTS) AND CEREBELLAR GM (LIMITED DATA EXPERIMENT) REPRESENT STATISTICALLY SIGNIFICANT INCREASE IN ACCURACY FOR GIF. GIF+LWV OUTPERFORMS BOTH GW-EM AND GIF+SEG ON THE DEEP GM AND ON THE CEREBELLAR GM FOR THE SECOND EXPERIMENT

Full Data	Cortical GM			Cortical WM			Cerebellar GM			Cerebellar WM			Deep GM		
	GIF+LWV	GW-EM	GIF+Seg	GIF+LWV	GW-EM	GIF+Seg	GIF+LWV	GW-EM	GIF+Seg	GIF+LWV	GW-EM	GIF+Seg	GIF+LWV	GW-EM	GIF+Seg
Average	0.863	0.912	**0.925**	0.879	0.930	**0.940**	0.924	0.927	**0.933**	0.880	0.905	**0.921**	**0.894**	0.825	0.849
Std	0.018	0.025	0.018	0.011	0.015	0.013	0.019	0.029	0.016	0.007	0.008	0.008	0.009	0.019	0.014
p-value	$<10^{-4}$	$<10^{-4}$	-	$<10^{-4}$	$<10^{-4}$	-	$<10^{-3}$	$<10^{-4}$	-	$<10^{-4}$	$<10^{-4}$	-	$<10^{-4}$	$<10^{-4}$	-
Limited Data	**Cortical GM**			**Cortical WM**			**Cerebellar GM**			**Cerebellar WM**			**Deep GM**		
	GIF+LWV	GW-EM	GIF+Seg	GIF+LWV	GW-EM	GIF+Seg	GIF+LWV	GW-EM	GIF+Seg	GIF+LWV	GW-EM	GIF+Seg	GIF+LWV	GW-EM	GIF+Seg
Average	0.833	0.805	**0.915**	0.848	0.832	**0.936**	**0.916**	0.881	0.912	0.877	0.866	**0.933**	**0.873**	0.789	0.844
Std	0.017	0.043	0.023	0.010	0.039	0.015	0.011	0.018	0.014	0.018	0.033	0.022	0.040	0.083	0.027
p-value	$<10^{-4}$	$<10^{-4}$	-	$<10^{-4}$	$<10^{-4}$	-	$<10^{-4}$	$<10^{-4}$	-	$<10^{-4}$	$<10^{-4}$	-	$<10^{-4}$	$<10^{-4}$	-

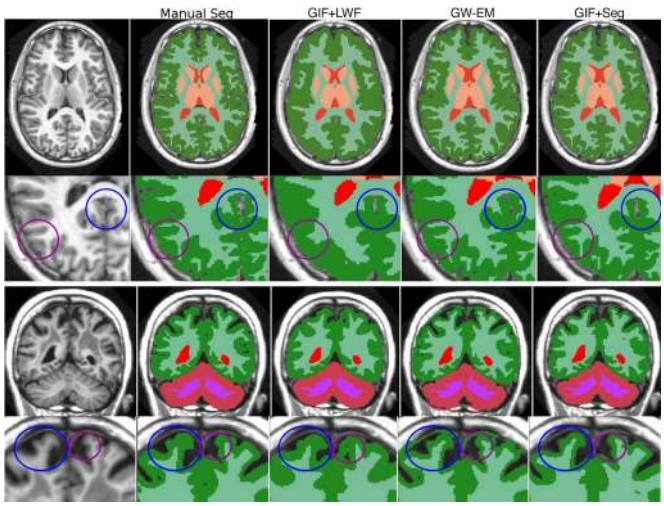

Fig. 10. From left to right: A T1-weighted image from the OASIS database and its corresponding manual segmentation, segmentation using label fusion, using a groupwise population prior and using GIF from the leave-one-out experiment. The first two and last two rows correspond to subjects OAS1_0285 and OAS1_0083, respectively. The purple and blue circles highlight areas with noticeable variations between methods. Note the extent of the periventricular WM damage in subject OAS1_0083.

distributions is not ideal. The validation of segmentation accuracy follows the same algorithmic comparison, leave-one-out methodology, similarity metric and statistical test used in Section VI-A, but with the OASIS data. The SPM12b algorithm was not used for comparison as it is not optimised for such high number of structures.

A second experiment tests the ability to segment morphologically dissimilar subjects. For this purpose, the 5 youngest female subjects in the database (age range $= 18-20$, 100% females) were chosen as training subjects and the remaining 30 subjects (age range $= 20-90$, 63% females) were used as test subjects. This experiment is similar to the jackknifing cross validation strategy from the previous section $(\mathrm{GIF}+\mathrm{Seg_C})$, but with a highly biased training dataset (biased towards a very specific age group and gender), demonstrating GIF's ability to cope with large morphological variability. For the GW-EM method, a new groupwise population prior was created from the 5 training subjects. Both the GIF+Seg and GIF-LWV methods used the 5 subjects as sources of information in the GIF framework. The same similarity metric and statistical tests used in Section VI-A were used for segmentation accuracy estimation.

The population statistics for both experiments are shown in Fig. 3—top, where the proposed method (GIF+Seg) achieved statistically significantly higher ($p < 10^{-4}$) Dice overlap in the cortical and cerebellar GM/WM.

VII. OPEN SOURCE IMPLEMENTATION AND WEB SERVICE

An open-source implementation of GIF will be made available as part of NiftySeg (niftyseg.sf.net). Also, in order to provide a simple and purpose optimised tool for the community, a fully automated GIF-based brain parcelation and tissue segmentation web-service is available at http://cmictig.cs.ucl.ac.uk/softweb/.

VIII. DISCUSSION

This work proposes a framework for information propagation between a population of images. This framework can be exploited for multiple applications, ranging from tissue segmentation and structural parcellation to morphometric analysis and image synthesis. Here, we apply the GIF framework to the problems of multi-atlas label propagation and tissue segmentation and demonstrate improved performance compared to the equivalent pairwise approach, mainly in the presence of morphological differences between the training and testing population.

More specifically, the application to multi-atlas propagation problem showed a small but significant increase in performance when compared to MAPER. It is important to note that GIF was not compared to other more advanced fusion techniques as the proposed geodesic propagation framework is agnostic to the fusion strategy, i.e. GIF can be combined with most fusion techniques by changing (4). Interestingly, Section V-B, which aims at propagating a set of brain masks from control subjects to pathological subjects, demonstrates that geodesic propagation can improve the overall performance when compared to direct pairwise propagation by improving the ability to extrapolate information. Visual inspection (e.g. see Fig. 5) shows good quality results even in the presence of large scale atrophic processes and pathology. Nonetheless, further validation on subjects with larger morphological variability is still necessary, as the current validation is hampered by the limited amount of ground-truth manual parcellations. Combining GIF with other fusion techniques, rather than a simple local weighted voting, should also further improve the accuracy of the results.

Another interesting point pertains to Fig. 6, which represent the value of G at convergence for an AD patient, when measured from a database of cognitively normal subjects. This figure shows that the areas that are morphologically distant from a normal population are located in regions normally

associated with AD pathology. To some degree, Fig. 6 shows visually that the geodesic distance is capturing pathology related features, which could in theory be used for pathological classification in future work.

Section V-C compares the proposed methodology to state of the art algorithms as defined in the MICCAI label fusion challenge. This section also provides insights of the algorithmic performance in situations of very low number of training datasets. Experiments showed that GIF with only 5 training datasets (rather than the 15 datasets defined in the challenge) can provide better results that 13 other methodologies that competed in the challenge using the full (15) training data. A similar, albeit worse, performance can be obtained using GIF with only one single training dataset, demonstrating that GIF performs well even in extreme situations with very limited training data. This characteristic can open new opportunities in labelling very large and time consuming datasets such as 7T brain data, small animal imaging or even microscopy data.

Section V-D extends the validation of label fusion to multimodal data and demonstrates that multimodal information can be exploited to improve the process of label fusion. GIF obtained better results than the state-of-the-art method for this application using only 5 training datasets. This section also shows that the use of the geodesic distance within GIF is paramount for both stability and accuracy of the algorithm, as GIF without the geodesic component starts degrading its accuracy after a few iterations due to label over-smoothing.

The application of the GIF framework to the problem of tissue segmentation is interesting from a more conceptual point of view, as it provides a different way to think about the propagation of *a priori* information between subjects. Again, results show improved performance when compared to both an equivalent technique using groupwise tissue priors, to SPM or to fusion, even when only a subset of the training data is used (see GIF + $\mathrm{Seg_C}$). The results on the OASIS database provide a more captivating view of the advantages of geodesic propagation when compared to pairwise or groupwise propagation due to the large scale morphological differences between the subjects in the OASIS/*Neuromorphometrics, inc*. database. Furthermore, even with gender—and age-group limited training data, the performance of the tissue segmentation does not deteriorate substantially when using GIF, but does so for groupwise or fusion-based approaches. This advantage can have a crucial impact when applied to pathological populations or to the analysis of the developing brain.

This work presents the first step towards a formal, compressive and unified framework for the processing of brain images. The central idea of this paper can also be used in the context of image synthesis [33], atrophy simulation [34] and to stratified voxel based morphometry [35], showing its general applicability. Future work will aim at optimising the multiple parameters in GIF by learning from training datasets and explore different applications of the GIF framework to bias-field correction and outlier detection.

IX. Conclusion

This work presents an algorithm where information is propagated along geodesic paths through a local spatially-variant neighbourhood graph. Application of the geodesic propagation concept to structural parcellation and brain segmentation has demonstrated statistically significant advantages when compared to their pairwise equivalent methods. Overall, the proposed framework can be used to better propagate information from a group of subjects to other morphologically-different subjects in a dataset.

References

[1] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based tissue classification of MR images of the brain," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 897–908, Oct. 1999.

[2] K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, no. 3, 2005.

[3] B. T. T. Yeo, M. R. Sabuncu, R. S. R. Desikan, B. Fischl, and P. Golland, "Effects of registration regularization and atlas sharpness on segmentation accuracy," *Med. Image Anal.*, vol. 12, no. 5, pp. 603–615, Oct. 2008.

[4] F. L. Bookstein, "Voxel-based morphometry should not be used with imperfectly registered images," *NeuroImage*, vol. 14, no. 6, pp. 1454–1462, Dec. 2001.

[5] C. Davatzikos, "Why voxel-based morphometric analysis should be used with great caution when characterizing group differences," *NeuroImage*, vol. 23, no. 1, pp. 17–20, Sep. 2004.

[6] M. R. Sabuncu, S. K. Balci, M. E. Shenton, and P. Golland, "Image-driven population analysis through mixture modeling," *IEEE Trans. Med. Imag.*, vol. 28, no. 9, pp. 1473–1487, Sep. 2009.

[7] A. Ribbens, J. Hermans, F. Maes, D. Vandermeulen, and P. Suetens, "SPARC: Unified framework for automatic segmentation, probabilistic atlas construction, registration and clustering of brain MR images," in *Proc. IEEE Int. Symp. Biomed. Imag., From Nano to Macro*, 2010, pp. 856–859.

[8] A. Ericsson, P. Aljabar, and D. Rueckert, "Construction of a patient-specific atlas of the brain: Application to normal aging," in *Proc. IEEE Int. Symp. Biomed. Imag., From Nano to Macro*, 2008, pp. 480–483.

[9] T. Rohlfing, D. B. Russakoff, and C. R. Maurer, Jr., "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 983–994, Aug. 2004.

[10] S. K. Warfield, K. H. Zou, and W. M. Wells, III, "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.

[11] A. Hammers *et al.*, "Statistical neuroanatomy of the human inferior frontal gyrus and probabilistic atlas in a standard stereotaxic space," *Human Brain Map.*, vol. 28, no. 1, pp. 34–48, Jan. 2007.

[12] R. A. Heckemann *et al.*, Alzheimer's Disease Neuroimag. Initiative, "Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation," *NeuroImage*, vol. 51, no. 1, pp. 221–227, May 2010.

[13] M. J. Cardoso *et al.*, "LoAd: A locally adaptive cortical segmentation algorithm," *NeuroImage*, vol. 56, no. 3, pp. 1386–1397, Jun. 2011.

[14] R. Wolz, P. Aljabar, J. V. Hajnal, A. Hammers, and D. Rueckert, Alzheimer's Disease Neuroimag. Initiative, "LEAP: Learning embeddings for atlas propagation," *NeuroImage*, vol. 49, no. 2, pp. 1316–1325, Jan. 2010.

[15] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2368–2382, Dec. 2011.

[16] S. Lafon, Y. Keller, and R. R. Coifman, "Data fusion and multicue data matching by diffusion maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1784–1797, Nov. 2006.

[17] J. Hamm, D. H. Ye, R. Verma, and C. Davatzikos, "GRAM: A framework for geodesic registration on anatomical manifolds," *Med. Image Anal.*, vol. 14, no. 5, pp. 633–642, Oct. 2010.

[18] D. H. Ye, J. Hamm, D. Kwon, C. Davatzikos, and K. M. Pohl, "Regional manifold learning for deformable registration of brain MR images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*. New York: Springer, 2012, vol. 7512, LNCS, pp. 131–138.

[19] H. Jia, P.-T. Yap, and D. Shen, "Iterative multi-atlas-based multi-image segmentation with tree-based registration," *NeuroImage*, vol. 59, no. 1, pp. 422–430, Jan. 2012.

[20] H. Wang, A. Pouch, M. Takabe, B. Jackson, J. Gorman, R. Gorman, and P. A. Yushkevich, "Multi-atlas segmentation with robust label transfer and label fusion," in *Information Processing in Medical Imaging*. Berlin, Germany: Springer, Jan. 2013, pp. 548–559.

[21] M. J. Cardoso, R. Wolz, M. Modat, N. C. Fox, D. Rueckert, and S. Ourselin, "Geodesic information flows," in *Medical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer, 2012, pp. 262–270.

[22] S. Gerber, T. Tasdizen, S. Joshi, and R. Whitaker, "On the manifold structure of the space of brain images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*. New York: Springer, vol. 5761, LNCS, pp. 305 –312.

[23] K. K. Bhatia, A. Rao, A. N. Price, R. Wolz, J. Hajnal, and D. Rueckert, "Hierarchical manifold learning," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*. New York: Springer, vol. 7510, LNCS, pp. 512–519.

[24] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," in *Proc. 19th Int. Conf. Mach. Learn.*, Jan. 2002, pp. 315–322.

[25] R. R. Coifman *et al.*, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *PNAS*, vol. 102, no. 21, pp. 7426–7431, May 2005.

[26] M. Modat *et al.*, "Fast free-form deformation using graphics processing units," *Comput. Methods Programs Biomed.*, vol. 98, no. 3, pp. 278–284, Jun. 2010.

[27] R. Souvenir and R. Pless, "Image distance functions for manifold learning," *Image Vis. Comput.*, vol. 25, no. 3, pp. 365–373, Mar. 2007.

[28] I. S. Gousias *et al.*, "Magnetic resonance imaging of the newborn brain: Manual segmentation of labelled atlases in term-born and preterm infants," *NeuroImage*, vol. 62, no. 3, pp. 1499–1509, Sep. 2012.

[29] P. A. Yushkevich *et al.*, "Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI," *NeuroImage*, vol. 53, no. 4, pp. 1208–1224, Dec. 2010.

[30] I. S. Gousias *et al.*, "Magnetic resonance imaging of the newborn brain: Automatic segmentation of brain images into 50 anatomical regions," *PLoS ONE*, vol. 8, no. 4, 2013.

[31] B. Aubert-Broche, M. Griffin, G. B. Pike, A. C. Evans, and D. L. Collins, "Twenty new digital brain phantoms for creation of validation image data bases," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1410–1416, Nov. 2006.

[32] D. S. Marcus *et al.*, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, demented older adults," *J. Cognitive Neurosci.*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007.

[33] N. Burgos *et al.*, "Attenuation correction synthesis for hybrid PET-MR scanners: Application to brain studies," *IEEE Trans. Med. Imag.*, vol. 33, no. 12, pp. 2332–2341, Dec. 2014.

[34] M. Modat, I. J. A. Simpson, M. J. Cardoso, D. M. Cash, N. Toussaint, N. C. Fox, and S. Ourselin, "Simulating neurodegeneration through longitudinal population analysis of structural and diffusion weighted MRI data," in *Medical Image Computing and Computer-Assisted Intervention*. New York: Springer, Jan. 2014, pp. 57–64.

[35] M. J. Cardoso, M. Modat, I. Simpson, and S. Ourselin, "Stratified voxel-based morphometry (sVBM)," *Math. Foundat. Comput. Anatomy*, p. 49, 2013.