Multi-Atlas Segmentation Using Partially Annotated Data: Methods and Annotation Strategies

Lisa Margret Koch, Martin Rajchl, Wenjia Bai, Christian Frederik Baumgartner, Tong Tong, Jonathan Passerat-Palmbach, Paul Aljabar, and Daniel Rueckert, *Fellow, IEEE*

Abstract—Multi-atlas segmentation is a widely used tool in medical image analysis, providing robust and accurate results by learning from annotated atlas datasets. However, the availability of fully annotated atlas images for training is limited due to the time required for the labelling task. Segmentation methods requiring only a proportion of each atlas image to be labelled could therefore reduce the workload on expert raters tasked with annotating atlas images. To address this issue, we first re-examine the labelling problem common in many existing approaches and formulate its solution in terms of a Markov Random Field energy minimisation problem on a graph connecting atlases and the target image. This provides a unifying framework for multi-atlas segmentation. We then show how modifications in the graph configuration of the proposed framework enable the use of partially annotated atlas images and investigate different partial annotation strategies. The proposed method was evaluated on two Magnetic Resonance Imaging (MRI) datasets for hippocampal and cardiac segmentation. Experiments were performed aimed at (1) recreating existing segmentation techniques with the proposed framework and (2) demonstrating the potential of employing sparsely annotated atlas data for multi-atlas segmentation.

Index Terms—Multi-atlas segmentation, partial annotations, Markov Random Field, unifying framework, continuous max-flow, annotation strategies

1 Introduction

In recent years, major efforts have been undertaken towards building large medical image databases such as ADNI [1]. Segmenting anatomical structures in these images is often necessary to better understand physiological and pathological processes through quantitative analysis. As the wealth of data increases, manually annotating the images becomes prohibitive, especially for large 3D or 4D image datasets. Automated segmentation approaches may face challenges in large databases due to large variability in shape and appearance of the structures of interest, the presence of pathologies, or different imaging protocols used to acquire the images. In particular, it becomes increasingly desirable to develop robust and accurate segmentation techniques that rely on minimal manual input or weak supervision.

Multi-atlas segmentation [2], [3], [4] has proven to be a successful and robust tool and is widely used in the medical imaging community [5]. The approach generally relies on label propagation from multiple atlases (i.e., fully annotated training images) to a target image. Using multiple atlases offers the important advantage of capturing anatomical variability. Ideally, the atlases should match the population to be segmented [6]. However, suitable atlases are not always available for large image databases, especially if the images in the database exhibit large variabilities, e.g., due to the presence of disease or aging processes. This motivates the use of training data obtained with different annotation strategies, where atlas images are only *partially* annotated, drastically reducing the labelling effort per image and therefore allowing expert raters to (partially) annotate more training images in the same time. To employ partially annotated atlas data while building on the success of multi-atlas segmentation (MAS), we propose a generalisation of the labelling problem in existing MAS methods. In the following paragraphs, we review relevant work in the field before identifying the main contributions of this paper.

Many MAS techniques use non-linear registration to warp segmentations from multiple suitable atlases to a target image [2], [3], [4], [7], [8], [9]. The target segmentation can be formed by fusion of the propagated labels, for example by applying a majority vote rule [2], [8] or another combination strategy such as a weighted average based on global or local similarity measures between the target and atlas images [7], [10]. In [9], a probabilistic framework was presented where the above-mentioned vote rules are expressed with a generative label fusion model. This was extended in [10] to incorporate non-local label fusion and registration uncertainty, and

- *L.M. Koch, M. Rajchl, W. Bai, C.F. Baumgartner, T. Tong, J. Passerat-Palmbach, and D. Rueckert are with the Biomedical Image Analysis Group, Imperial College London, London SW7 2AZ, United Kingdom. E-mail: {l.koch, m.rajchl, w.bai, c.baumgartner, t.tong11, j.passerat-palmbach, d.rueckert}@imperial.ac.uk.*
- *P. Aljabar is with the Division of Imaging Sciences & Biomedical Engineering, King's College London, London WC2R 2LS, United Kingdom. E-mail: paul.aljabar@kcl.ac.uk.*

Manuscript received 27 Oct. 2015; revised 2 Dec. 2016; accepted 22 May 2017. Date of publication 21 Aug. 2017; date of current version 12 June 2018. (Corresponding author: Lisa Margret Koch.)
Recommended for acceptance by T. Cootes.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2017.2711020

in [11] to allow the use of atlases annotated with different labelling protocols. Other combination strategies include STAPLE [12], where label fusion weights are estimated with an expectation-maximisation algorithm, or Joint Label Fusion [13], where correlations among atlases are taken into account. To account for high local anatomical variability between images, and to relax the requirement for accurate registration, patch-based segmentation [14], [15] has been introduced. Using this approach, the label fusion step employs a non-local weighted average of voxel labels in a small neighbourhood of the atlases, with weights based on the similarities of patches centred on the compared voxels. Considerable improvements in segmentation accuracy can be achieved by using the label propagation results as prior probabilities in subsequent refinement steps, combining them with regularisation terms and an intensity model in a Markov Random Field (MRF) formulation [16], [17], [18], [19], [20]. This was first suggested by [16] in combination with graph-cuts [21], whereas [17] proposed an expectation-maximisation approach, which was also adopted in [18] and [19].

All of the above methods rely on the availability of a fully annotated atlas dataset with the aim to segment an individual target image. It has been shown that, in general, segmentation accuracy decreases when fewer [2] or less similar [8] atlases are used. However, segmentation methods requiring fewer atlases (i.e., training data) while preserving accuracy are highly desirable, as they could reduce the workload of raters who manually annotate these atlases. Recently, a number of methods have been proposed for iterative label propagation, which allow labels from a small set of annotated atlas images to be propagated to similar images or image regions in the test population [6], [22], [23], [24], [25]. These methods avoid error-prone registration between dissimilar images by only propagating information between similar images which are easy to register. They therefore exploit the unlabelled test population in a semi-supervised learning setup and thus reduce the amount of labelled atlas data necessary to achieve accurate segmentation results.

Other strategies to reduce the manual workload that have been proposed in the computer vision and medical imaging community employ weak supervision. This includes annotations in the form of bounding boxes around an object instead of pixel-wise labelling, such as proposed in GrabCut [26] and recently extended to 3D bounding boxes in [27], scribbles that only annotate part of an image (e.g., [28]), or image tags which only describe which class is present in an image (e.g., [29]). [30] give a good summary of the various forms of weak supervision and propose a unified framework for segmentation in computer vision datasets. In the context of MAS, [31] proposed a modification of the STAPLE algorithm [12] that can deal with missing annotations in the atlases.

A frequently used method to efficiently solve the labelling problem is to express it as an MRF energy function [32] and minimise it using min-cut/max-flow techniques [21], [28], [33], [34]. The MRF is normally defined by a graph constructed on a regular grid that represents the target image. However, some applications formulate an MRF energy function on graphs connecting *multiple* images. Recently, [35] applied graph-cuts for co-segmentation of pairs of PET and CT images by minimising an MRF energy function which penalises tumour segmentation differences between a PET and CT image of the same subject. [36] used an extension of continuous max-flow [33] for simultaneous prostate segmentation in multiple 2D slices while penalising segmentation differences between slices. Continuous max-flow (CMF) solves the continuous counterpart to the discrete min-cut/max-flow problem [33] and it can be computed using a reliable, inherently parallelisable multiplier-based algorithm with guaranteed convergence. This makes it suitable for the optimisation of large labelling problems.

1.1 Our Contribution

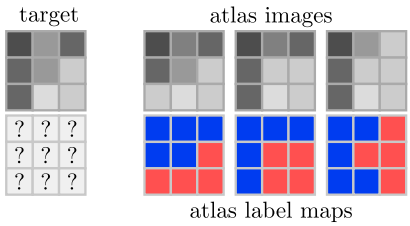
In this paper, we propose methods and annotation strategies which enable the use of partially annotated data for MAS, with the main goal of reducing the required manual labelling effort. As a first contribution, we propose a unifying framework for MAS using a novel graphical representation of the labelling problem. In Section 2 we demonstrate how label fusion, spatial regularisation, and data models can be expressed simultaneously using this representation. We then show in Section 3 how the framework can be used to go beyond the abilities of existing MAS techniques: The proposed flexible graph structure allows a relaxation of the annotation requirements in atlas images. This means that our framework naturally allows the use of atlases that were only partially annotated, resulting in a reduced manual labelling effort for expert raters. We examine different partial annotation strategies and investigate modifications in the graph configuration to optimally exploit partially annotated atlas data in the segmentation process. To optimise the arising MRF energy function, we provide an efficient optimisation scheme based on continuous max-flow [33], [34] in Section 4. Experiments on hippocampal (Sections 5.1 and 5.2) and cardiac segmentation (Section 5.3) highlight the performance of the proposed framework and shed light on some of the possibilities it offers for employing partial annotations such as missing slices or scribbles. A preliminary version of this work was presented in [37]. In comparison, this paper includes a more comprehensive description of our work and its context, the exploration of additional partial annotation strategies, an extension of the proposed method to multi-label segmentation, and improved and more comprehensive experimental evaluation on two datasets.

2 Unified Framework for Multi-Atlas Segmentation

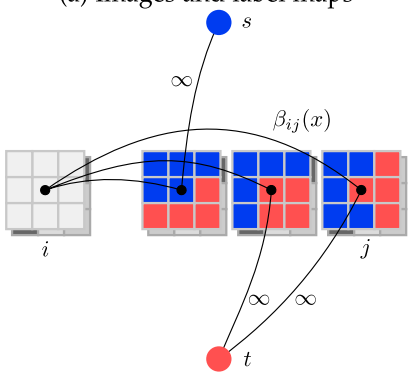
In this section, we first revisit the labelling problem in existing MAS methods [2], [7], [8], [16], [17] and reformulate it as an MRF energy optimisation problem defined on a graph comprising multiple images (i.e., the target and atlases). In particular, we show how the proposed graphical approach can incorporate label fusion (Section 2.1), spatial regularisation (Section 2.2), as well as a data term and missing atlas labels (Section 2.3). It is important to note that this unifying framework also provides a flexible way to employ partially annotated data and leverage unlabelled data as later introduced in Section 3, which is not possible with the existing MAS techniques it can express, and is inspired by.

2.1 Label Fusion

For MAS using R images, all atlas images $j \in \{1, \dots, R\}$ are registered to the target image i. For convenience we assume



(a) Images and label maps



(b) Graphical representation of label fusion

Fig. 1. (a) Toy dataset with an unlabelled target image on the left, atlas images and corresponding manual annotations (blue and red depict different labels) on the right. (b) In MAS, each voxel x in target image i is labelled by label propagation from atlases $j \in \{1, \ldots, R\}$ with fusion weights $\beta_{ij}(x)$. This can also be interpreted as an MRF optimisation problem, where atlas voxels are connected to the terminal nodes with infinitely weighted edges and inter-image edges $\beta_{ij}(x)$ encode label fusion.

$i = R + 1$. The label maps l_j associated with the atlas images j are then propagated to the target. Fig. 1a shows an example atlas set with corresponding label maps, and an unlabelled target image. Each voxel $x \in \Omega$ in the target image i is labelled using some combination strategy, e.g., a weighted average of atlas labels $l_j(x)$

$$l_i(x) = \arg\max_L \sum_{j=1}^{R} \beta_{ij}(x)\delta(l_j(x) = L). \tag{1}$$

Here $\delta(\cdot)$ is an indicator function. The weights $\beta_{ij}(x)$ can be uniform (which is equivalent to the majority vote rule as used in [2], [3], [8]) or based on global or local similarity measures between images i and j as in [7], [9], [10].

As an alternative perspective, we can use a graphical representation to model the relationship of shared information between the atlases and the target using an MRF [32]. According to the above labelling scenario, this graph connects each voxel x in the target image i to the corresponding voxels in the atlases j with an edge weighted by $\beta_{ij}(x)$. The manual annotations in the atlases can be encoded by the unary potential function

$$V(l_j(x)) = \begin{cases} 0 & l_j(x) = G_j(x), \\ \infty & \text{otherwise}, \end{cases} \tag{2}$$

where $G_j(x)$ is the ground truth label given by the expert rater, assigning infinite cost to the hypothetical scenario of assigning a different label to the atlas voxel. Fig. 1b visualises this configuration and in Section 2.3, these terminal graph connections are discussed in more detail. To find a labelling on the graph, we can formulate a pairwise potential function that penalises conflicting labels in voxels connected by a high weight $\beta_{ij}(x)$, e.g.,

$$V(l_i(x), l_j(x)) = \beta_{ij}(x)\delta(l_j(x) \neq l_i(x)). \tag{3}$$

This assigns a high penalty when the target and atlas labels differ and the atlas is considered similar to the target i, as defined by the similarity measure $\beta_{ij}(x)$. In the case of a majority vote, the weights are uniform, e.g., $\beta_{ij}(x) = 1$. The cost for labelling an individual voxel x in image i can then be calculated as follows:

$$E_{\text{propagation}}(l_i(x)) = \sum_{j=1}^{R} V(l_i(x), l_j(x)) \tag{4}$$

$$= \sum_{j=1}^{R} \beta_{ij}(x) - \sum_{j=1}^{R} \beta_{ij}(x)\delta(l_j(x) = l_i(x)). \tag{5}$$

As we assume the graphical model encodes Markov properties, voxels in the target image are conditionally independent given the atlas images since spatially neighbouring voxels in the target image are not connected in the graph (in contrast to the setting for regularisation in many vision problems [32]). Since the atlas labels are fixed and assumed to be independent of each other (a common assumption in MAS), it follows that the target voxels are statistically independent, and the optimal label can be found by minimising $E_{\text{propagation}}(l_i(x))$ independently for all voxels

$$l_i(x) = \arg\min_L \quad E_{\text{propagation}}(l_i(x) = L) \tag{6}$$

$$= \arg\max_L \quad \sum_{j=1}^{R} \beta_{ij}(x)\delta(l_j(x) = L). \tag{7}$$

This leads to the same result as the vote rule in Eq. (1), demonstrating that MAS can be expressed in terms of a graph optimisation problem. Patch-based segmentation (PBS [14], [15]) can also be expressed in this framework. In this case we use a slightly different graph structure as the label fusion step in PBS takes into account multiple voxels in a neighbourhood of x in each atlas instead of just one voxel at location x. By denoting the patch-based label fusion weights as $\beta_{ij}(x, y), y \in \mathcal{N}_x$ to reflect the non-local nature of these methods, a labelling can be found for this scenario as well. Here, multiple patches in the atlases are used at locations y in a neighbourhood $\mathcal{N}_x$ around location x. This scenario is visualised in Fig. 2. A similar configuration could be used to express registration uncertainty as presented in [10], where labels from multiple "candidate locations" in each atlas were fused using weights based on registration uncertainty. While the proposed formulation holds for these non-local techniques, the graph structure becomes more complex. In the scope of this paper, we limit ourselves to graphs on regular grids where voxels in different images are only connected if they are at corresponding locations.

It is important to note that the graphical model presented so far is an ineffective way to encode label fusion. The ∞-weighted terminal connections as introduced in Eq. (2) are never cut. Therefore all atlas voxels could be collapsed with the terminal nodes they are associated with, and label fusion could be encoded by unary potentials on the target image only. However, the proposed novel perspective on

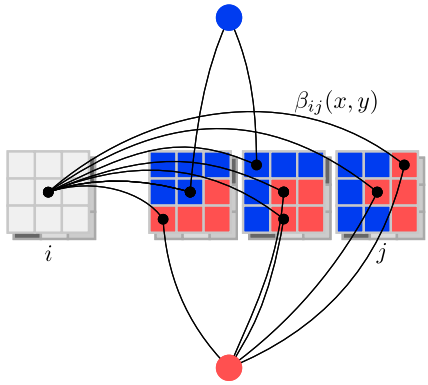


Fig. 2. Graph configuration representing patch-based segmentation. $\beta_{ij}(x,y)$ is determined by a patch similarity measure between a patch centred around voxel x in image i and voxel y in image j. Not all connections are drawn for better visibility and to reflect that in practice, dissimilar patches are omitted in the label fusion [14].

label fusion has two advantages: (1) it readily allows the integration of additional components and therefore provides a unifying reformulation for existing multi-atlas segmentation methods, and (2) the graphical approach extends to segmentation using partially annotated atlases (Section 3), where the proposed model does not permit the trivial reduction to unary potentials mentioned above.

2.2 Spatial Regularisation

In the previous section, we proposed assigning pairwise potentials between target and atlas voxels for label propagation. In addition, we can incorporate spatial regularisation with pairwise potentials between adjacent voxels x, y within an image i

$$V(l_i(x), l_i(y)) = \alpha_i(x,y)\delta(l_i(x) \neq l_i(y)). \tag{8}$$

This simple modification of the graph structure is shown in Fig. 3a. Regularisation enforces spatial consistency by penalising different label assignment in adjacent voxels. If the regularisation weights $\alpha_i(x,y)$ are based on intensity gradients, consistent labels can be enforced in adjacent labels that are similar in appearance, while allowing different labels across intensity boundaries. A graph configuration as shown in Fig. 3a models the scenario where regularisation is used to refine label fusion results, e.g., as in [16], [38].

2.3 Data Term and Missing Labels

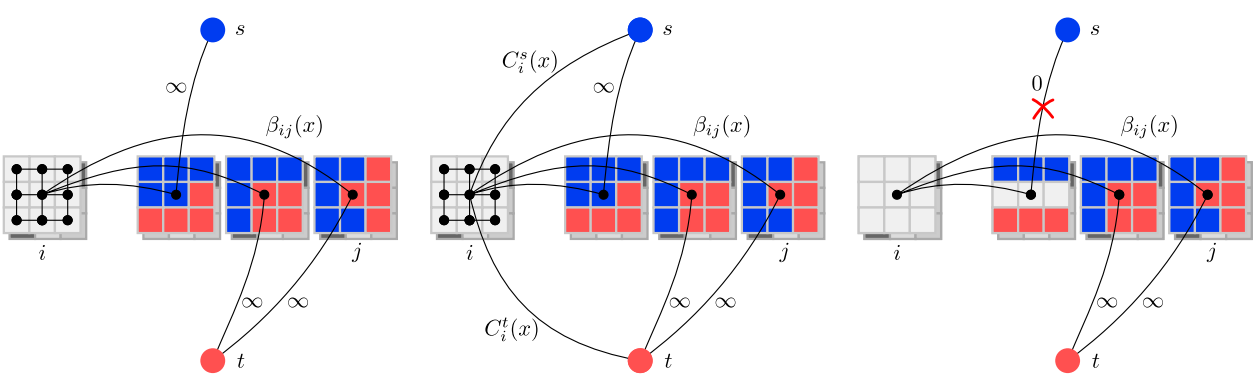
In Eq. (2) we showed how manual annotations can be encoded as unary potentials which are often referred to as a data term [28], [32]. The ground truth nature of these annotations is reflected in the graph structure by infinitely weighted terminal connections for each atlas voxel according to the manual label given. As can be seen in Figs. 1b or 3a, the voxels in the target image are not connected to the terminals as they are assumed to be unlabelled and no prior knowledge is available for them. It is important to note that a data term could be specified for the target image as well using prior probabilities, intensity models of the data, or a combination of both. This is a common technique when using MRFs in vision problems [16], [17], [32], [39] and can be incorporated by extending the graph structure as visualised in Fig. 3b. However, in the scope of this work, such unary potentials on the target image were not investigated. Furthermore, missing labels can be easily accounted for by removing terminal connections (i.e., unary potentials) for voxels where annotations are not available, as shown in Fig. 3c. The important implications of this property will be discussed in detail in Section 3 in conjunction with partially annotated atlas data.

2.4 Summary

We propose to interpret both the target image and the set of atlas images as a single graph structure (in which each voxel is a node) encoding Markov properties. On this graph we can use unary potentials to define the data term E_{data} to encode manual annotations or other prior knowledge, or to reflect missing labels. We then showed how pairwise potentials can be used to encode label fusion through *inter-image* connections and to build a propagation energy term $E_{\text{propagation}}$. Another pairwise potential term $E_{\text{regularisation}}$ encodes spatial regularisation through *intra-image* edges. The propagation, data, and regularisation terms can be combined to a comprehensive labelling energy function defined for the whole graph

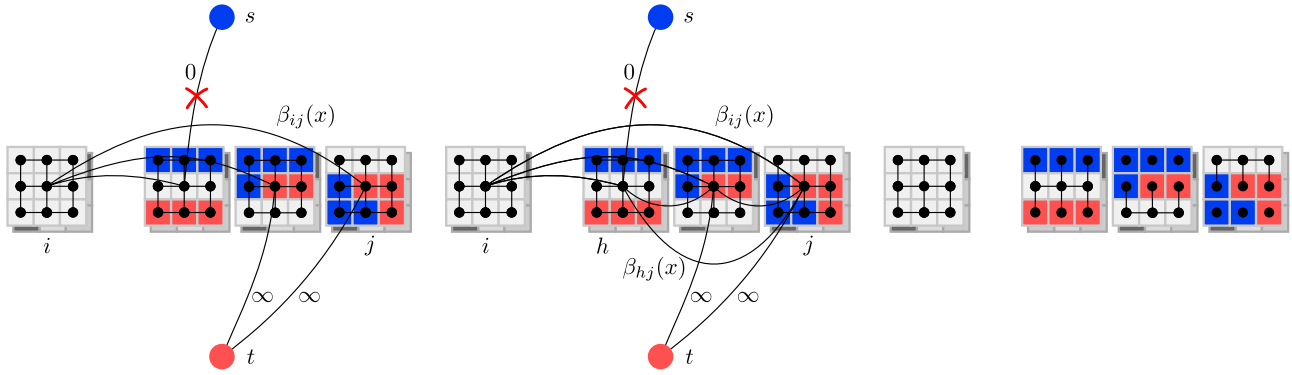
$$E(l) = E_{\text{data}}(l) + E_{\text{regularisation}}(l) + E_{\text{propagation}}(l). \tag{9}$$

As mentioned in the introduction, many existing multi-atlas segmentation methods (e.g., [16], [18]) use an MRF formulation to improve label propagation results with the benefits



(a) Multi-atlas segmentation with regularisation (b) Additional data term (c) Missing labels

Fig. 3. Different graph configurations representing (a) MAS with spatial regularisation in the target image, (b) an additional data term in the target image, i.e., encoding intensity models for the data, (c) MAS with missing atlas labels. Missing labels are reflected in the graph structure by missing terminal connections.



(a) Graph configuration 1 (CONF1) (b) Graph configuration 2 (CONF2) (c) Sparse regularisation term

Fig. 4. Graph configurations for employing partially annotated atlas data. Voxels with missing labels (white) are disconnected from terminal nodes. In contrast to Fig. 3c, spatial regularisation is enabled in *all* images. (a) Voxels at each location x in the target image are connected to voxels in atlases j. (b) Additionally, atlas voxels are connected to voxels in other atlases. (c) Shows a possible graph sparsification by removing obsolete edges between labelled atlas voxels.

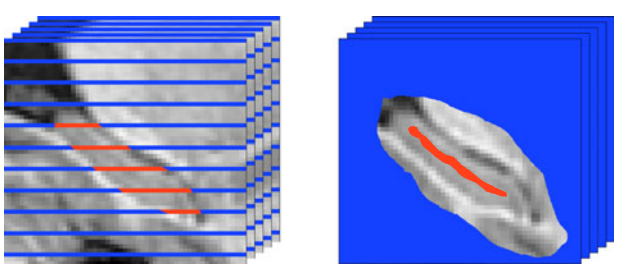
of regularisation and intensity data models. However, these approaches use probabilistic label propagation results as prior probabilities (i.e., unary potentials) in a *subsequent* refinement step, therefore adding the MRF optimisation as a separate post-processing step. The above comprehensive formulation treats label propagation as part of the optimisation process, and unifies all the components within a single framework. Furthermore, as we show in Section 3, the flexibility of the proposed graph structure lends itself naturally to exploit partially annotated data.

3 Partial Annotation Strategies

Manually annotating medical images is very time consuming, placing a major burden on clinical experts tasked with labelling large datasets. Using the proposed unified framework, it is possible to perform segmentation using *partially* annotated atlas data, going beyond the scope of existing multi-atlas segmentation techniques. We showed in Section 2.3 how our graphical model can easily accommodate missing labels through missing terminal connections in the graph structure. By applying our framework to any of the existing approaches discussed throughout Section 2, this would lead to a segmentation that is inferred from the *available* labels only, ignoring missing information.

Additionally, spatial consistency in the atlas images can be exploited to employ unlabelled atlas data as well. As neighbouring voxels are expected to share the same label, particularly if the voxels exhibit similar intensity patterns, we propose to use spatial regularisation within the atlas images as a form of *intra-image* label propagation. This way, labels may be shared between similar regions with labelled and unlabelled voxels in the atlases and propagated to the target image. This modification in the graph structure leads to a configuration as shown in Fig. 4a. Another possible configuration combines this with an additional inter-atlas propagation scheme which allows atlases to share information as well (shown in Fig. 4b). This serves to facilitate the propagation, especially when manual labels are very scarce at some locations x. The reader may note that this spatial regularisation scheme contains obsolete edges between labelled atlas voxels. These could be removed to improve computational efficiency as shown in Fig. 4c.

With this framework, it becomes interesting to pursue strategies which aim to efficiently build partially annotated datasets which may then be used as training data for segmentation tasks. In the remainder of this section, we propose two partial annotation strategies, which are evaluated in the Experiments Sections 5.2 and 5.3.



(a) Partial annotation strategy A: Slicewise (b) Partial annotation strategy B: Scribbles

Fig. 5. Illustration of partial annotation strategies: (a) a volumetric image with partial slice-by-slice annotation and (b) the same image with scribbles placed on each slice. Red and blue depict foreground and background, respectively, and voxels in grey remained unlabelled.

3.1 Strategy A: Slicewise Annotation

Medical volumetric images are often manually annotated slice-by-slice. Therefore reducing the proportion of annotated slices while retaining robust and accurate segmentation is an important goal. To simulate partially annotated atlases, only annotations from a proportion of evenly spaced 2D slices are used, and the remaining labels are set to be "missing". As an example, Fig. 5a shows a cross-section of a 3D image where every fifth slice is annotated. It is important to note that in the selected slices, the structures of interest are delineated in detail, i.e., all voxels in that slice are labelled.

3.2 Strategy B: Scribbles

Scribbles are often used to annotate images in the context of interactive segmentation [26], [28]. This strategy typically involves placing brush strokes (i.e., "scribbles") on parts of the image considered within the structure of interest, or within the background. As scribbles do not delineate the structure boundary, this only requires a very short user interaction and could potentially require less expertise. These properties make "scribbling" an attractive annotation

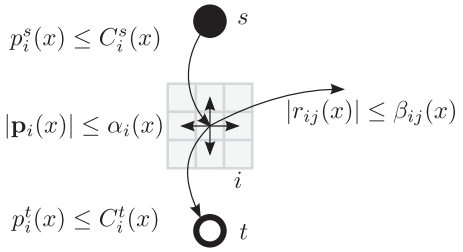


Fig. 6. Flow constraints $\beta_{ij}(x), C_i^{s,t}(x), \alpha_i(x)$ for label propagation, data term and spatial regularisation, and corresponding inter-image flows $r_{ij}(x)$, source and sink flows $p_i^{s,t}(x)$ and spatial flows $\mathbf{p}_i(x)$ at location x in image i.

strategy if it can be shown their use leads to competitive segmentation results. Fig. 5b shows an example image with scribbles for both the structure of interest (i.e., the hippocampus) and the background. We propose to annotate the training dataset by efficiently placing scribbles covering large areas (without delineating boundaries), as this can be done efficiently and is expected to make the segmentation task easier than very sparse, small scribbles.

4 Optimisation Using Continuous Max-Flow

The MRF energy function proposed in Eq. (9) consists of unary and pairwise terms. The pairwise terms encoding propagation (Eq. (3)) and regularisation (Eq. (8)) are both chosen as a Potts model. The MRF energy is therefore *submodular* for binary labelling and *metric* for multiple labels. Such discrete pairwise MRFs are graph-representable in the binary case and their global minimum can be found in polynomial time using min-cut/max-flow approaches [39]. In the multi-label case with metric costs, approximate solutions within a known factor of the global minimum can be found [21]. Discrete graph-cuts [21] on regular voxel grids may suffer from metrication artifacts, in particular when spatial regularisation is only enforced in a 6-neighbourhood. This can be addressed by increasing the number of edges in the graph to allow more isotropic regularisation, thus increasing the computational burden. Increasing the neighbourhood size may therefore be problematic for large graphs between multiple images. Recently, [33] proposed a max-flow algorithm in the continuous 2D or 3D domain (i.e., an image) which avoids this metrication bias and is inherently parallelisable in contrast to many discrete graph-based methods [33]. As the proposed energy function needs to be optimised for a large graph consisting of voxels in all images and their interactions, this approach was adopted and extended for graphs between multiple images.

Analogous to discrete max-flow approaches, the energy function on the graph can be optimised by maximising a source flow p^s through the network, subject to flow conservation and capacity constraints on the edges. In the original continuous max-flow algorithm [33], spatial flows $\mathbf{p} = [p_x, p_y, p_z]^T$ exist between adjacent voxels in the image domain Ω (for regularisation) and source and sink flows $p^{s,t}$ between voxels and terminal nodes. The optimisation is performed with a variational approach by introducing a Lagrange multiplier $u(x)$ to incorporate the constraints [33]. It has been shown that the resulting $u(x)$ corresponds to the globally optimal labelling [33] in the binary case.

We present a generalisation of CMF from a single image to an arbitrary configuration of interconnected images to account for any user-defined choice of inter-image relationships $\beta_{ij}(x)$. Fig. 6 shows the capacity constraints and introduces the notation for inter-image flows $r_{ij}(x)$ (for label propagation), spatial flows $\mathbf{p}_i(x)$ (for regularisation) and terminal flows $p_i^{s,t}(x)$ (for the data term). The regularisation constraints $\alpha(x)$ determine the smoothness of the result. To enforce greater smoothness in homogeneous image regions than along intensity boundaries, $\alpha(x)$ can be defined based on the image gradient $\nabla I(x)$

$$\alpha(x) = a \exp\left(-\frac{\|\nabla I(x)\|^2}{2\sigma_1^2}\right), \tag{10}$$

with parameters a and σ_1. This measure is the continuous equivalent of the regularisation term used in in [16], one of the pioneering works combining regularisation and MAS. To satisfy flow conservation, the sum of all in- and outgoing flows $\rho_i(x)$ at each node must be zero, i.e., $\forall i, x \in \Omega$

$$\rho_i(x) = \text{div } \mathbf{p}_i(x) - p_i^s(x) + p_i^t(x) + \sum_{j=1, j\neq i}^{R} r_{ij}(x) = 0, \tag{11}$$

where $r_{ij}(x) = -r_{ji}(x)$ and R is the number of images in the graph. We propose to adapt the definitions of the discrete gradient and divergence operators to account for anisotropic voxel dimensions $[s_x, s_y, s_z]$, which are often found in medical images

$$\nabla \mathbf{p} = \left[\frac{\delta_x \mathbf{p}}{s_x}, \frac{\delta_y \mathbf{p}}{s_y}, \frac{\delta_z \mathbf{p}}{s_z}\right]^T \tag{12}$$

$$\text{div } \mathbf{p} = \nabla \cdot \mathbf{p}. \tag{13}$$

Here, $\delta_x, \delta_y, \delta_z$ are the intensity differences between neighbouring voxels in different orientations, respectively. Using the augmented Lagrangian method [40], the following augmented Lagrangian function can be defined

$$L(u, p^s, p^t, \mathbf{p}, r) = \sum_{i=1}^{R}\left(\int_\Omega p_i^s dx + \int_\Omega u_i \rho_i dx - \frac{c}{2}\|\rho_i\|^2\right). \tag{14}$$

Eq. (14) can be maximised iteratively by optimising each variable $u, p^s, p^t, \mathbf{p}, r$ separately. The novel component compared to [33], [36] is the use of inter-image flows $r_{ij}(x)$ between any pair of images i, j. We therefore show in particular the optimisation step at iteration k for $r_{ij}(x)$, while fixing all other variables

$$r_{ij}^{k+1} = \underset{|r_{ij}| \leq \beta_{ij}}{\arg\max}\ L(u, p^s, p^t, \mathbf{p}, r). \tag{15}$$

This leads to

$$r_{ij}^{k+1} = \begin{cases} -\beta_{ij}, & \frac{1}{2}(J_j^k - J_i^k) \leq -\beta_{ij}, \\ \frac{1}{2}(J_j^k - J_i^k), & |\frac{1}{2}(J_j^k - J_i^k)| \leq \beta_{ij}, \\ \beta_{ij} & \text{otherwise.} \end{cases} \tag{16}$$

where

$$J_i^k = (\text{div } \mathbf{p}_i - p_i^s + p_i^t)^k + \sum_{l=1, l\neq i, j}^{R} r_{il}^k - \frac{u_i^k}{c}, \tag{17}$$

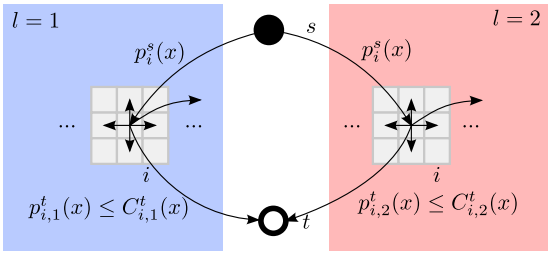


Fig. 7. Schematic showing graph configuration for multi-label CMF using the Potts Model. The graph (in this figure only one image i is shown) is replicated for each label l. The data term is encoded in the sink constraints for every label.

A more detailed derivation of this result is given in the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2017.2711020. After convergence, a segmentation can be found by discretising the resulting solution for u, e.g., by thresholding at 50 percent.

CMF has been extended to multi-label segmentation problems in [34] using a Potts model approach. To optimise for multiple labels, the graph structure is duplicated for every label. The data term is encoded in the sink constraints of each "sub-graph" while the source connections remain unconstrained. The same changes can be applied to the the graph in our framework, as shown in Fig. 7, and Eq. (14) can be adapted accordingly

$$L(u, p^s, p^t, \mathbf{p}, r) = \sum_{i=1}^{R} \left(\int_\Omega p_i^s dx + \sum_{l=1}^{L} \int_\Omega u_{i,l} \rho_{i,l} dx - \frac{c}{2} \sum_{l=1}^{L} \|\rho_{i,l}\|^2 \right). \tag{18}$$

Here, $u_{i,l}$ is the labelling function for label $l \in 1, .., L$ in image i and $\rho_{i,l}$ is the new flow conservation constraint

$$\rho_{i,l}(x) = \text{div } \mathbf{p}_{i,l}(x) - p_i^s(x) + p_{i,l}^t(x) + \sum_{j=1, j \neq i}^{R} r_{ij,l}(x) = 0. \tag{19}$$

5 Experiments and Results

In the previous sections, we proposed a unified MAS framework which can naturally accommodate partially annotated atlas data. We showed how the proposed graphical representation can implement a number of existing techniques through changes in the graph configuration. In the following experiments, we first employ our framework to perform hippocampal segmentation using three existing techniques (Section 5.1). We then investigate how the framework can be used—with further modifications of the graph structure—to employ partially annotated atlases for segmentation. This is done using both the slicewise partial annotation strategy (Section 5.2) and scribbles (Section 5.3).

The experiments were carried out on two datasets: (1) brain MR images from the ADNI database for hippocampal segmentation (a binary segmentation problem) and (2) cardiac MR images for segmentation of the right and left ventricular cavities and the left ventricle myocardium (i.e., segmentation with multiple labels).

5.1 Evaluation of Proposed Framework for MAS

To explore the proposed unifying framework, a number of different configurations were compared which correspond to existing segmentation techniques. To acquire a labelling on a target image, selected atlas images were aligned with the target image using non-rigid registration [41] and a graph was constructed using each of the chosen configurations. The optimisation proposed in Section 4 was performed to achieve a segmentation result.

First, we studied segmentation using the majority vote label fusion step (MAS-MV) [2], [3], [8]. For this, we assume a graph structure as shown in Fig. 1b and label propagation weights were set to $\beta_{ij}(x) = 1$. We compared MAS-MV to locally weighted label fusion (MAS-LW) as explored in [7], [9], [10]. Propagation weights $\beta_{ij}(x)$ were based on a local similarity measure between the target and the atlases

$$\beta_{ij}(x) = K \cdot \exp\left(-\frac{(P_i(x) - P_j(x))^2}{2\pi\sigma_2^2 \cdot |P|} \right), \tag{20}$$

where $P(x)$ is a patch centred around voxel x and $|P|$ is the patch size. K does not influence the label fusion result and was set to 1. By modifying the graph configuration to additionally incorporate intra-image edges in the target image, we added a regularisation term as described in Section 2.2 and shown in Fig. 3a. This configuration (further referred to as MASr-LW) implements simultaneous label fusion and regularisation similar to [16], [17]. It is important to note that these approaches incorporated an additional probability term based on intensity models of the data. However, in preliminary experiments, we achieved better results without this term.

5.1.1 Data and Experiment Setup

The proposed method was applied to 202 images from the ADNI database [1] for which reference segmentations of the hippocampus were made available through ADNI. All images were affinely aligned to the MNI152 template space with a voxel spacing of $1\ \text{mm}^3$ and intensity-normalised [42]. The data were split randomly into two equally sized sets, one for parameter tuning and one for evaluation. Optimal parameters were chosen for locally weighted label fusion (i.e., the propagaton term) and for spatial regularisation. The tuning procedure is described in Section 5.4.1. The terminal connections encoding the data term simply consisted of infinite weights in voxels where manual annotations were available, and zero weight in unlabelled voxels.

5.1.2 Results

For evaluation, a 10-fold cross-validation was performed within the evaluation set. In each fold, every test subject was segmented by selecting the R most similar images from the *remaining folds* as atlases and transforming them to the target space using nonrigid registration [41]. Similarity was assessed with normalised mutual information. This was repeated for $R = \{5, 10, 15, 20\}$ to measure the influence of the number of atlases on segmentation accuracy. Fig. 8 shows the mean Dice coefficients of the pooled results. Segmentation results generally increased with the number of atlases used. Majority vote (MAS-MV) was more robust

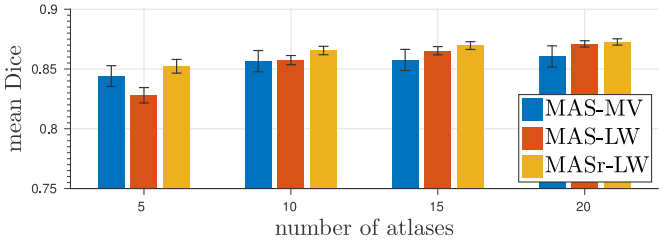


Fig. 8. Mean Dice coefficients for MAS-MV, MAS-LW and MASr-LW using $R = \{5, 10, 15, 20\}$ atlases. The error bars depict the standard error.

than locally weighted fusion (MAS-LW) when using 5 or 10 atlases, but for larger atlas sets, MAS-LW achieved better results. With additional spatial regularisation, MASr-LW consistently outperformed both MAS-LW and MAS-MV.

5.2 Evaluation of Partial Annotation Strategy A: Slicewise (PA-SW)

This experiment investigates the performance of our framework on the same segmentation task when using atlases which were partially labelled slice-by-slice as proposed in Section 3.1. This strategy provides exact delineations of structure boundaries in the annotated slices, which is desirable due to poor contrast between the hippocampus and neighbouring tissue. As proposed in Section 3, we examined two graph configurations using different propagation schemes. In the first configuration (PA-SW-CONF1, Fig. 4a), the regularisation term included spatial regularisation in *all* images (i.e., target and atlases). The propagation term allowed label propagation from the atlases to the target. In addition, in the second configuration (PA-SW-CONF2, Fig. 4b), label propagation *between atlases* was allowed by expanding the propagation term with inter-atlas connections. To demonstrate how these approaches benefit from unlabelled data in the atlas set, they were compared to a third configuration, PA-SW-baseline, using the same partial annotations, but no spatial regularisation within the atlases. In this configuration, inter-image edges connected target voxels to the atlas voxels, and spatial regularisation in the target was performed as in MASr-LW in the previous section.

5.2.1 Data and Experiment Setup

The same data was used as in the previous experiment (Section 5.1). To simulate partially annotated atlas data, manual labels of a proportion q of evenly distributed slices in 20 atlases were used for segmentation of the target image. To determine which slices were used, for each atlas a different (random) offset was added to the determined slice positions. The partial annotations were then transformed to the target space using nonrigid registration [41]. The data term was built by establishing terminal connections at labelled voxels, while leaving unlabelled voxels unconnected, as explained in Section 2.3. The proportion of labelled atlas slices ranged from $q = 1$ (i.e., fully labelled) to $q = 0.1$ (i.e., every 10th slice) to investigate how strongly the atlas label maps could be sub-sampled while achieving robust segmentation results. The parameters for the propagation term were chosen as in the previous experiment and optimal choices for the regularisation coefficients a, σ_1 were obtained through parameter tuning as described in Section 5.4.2.

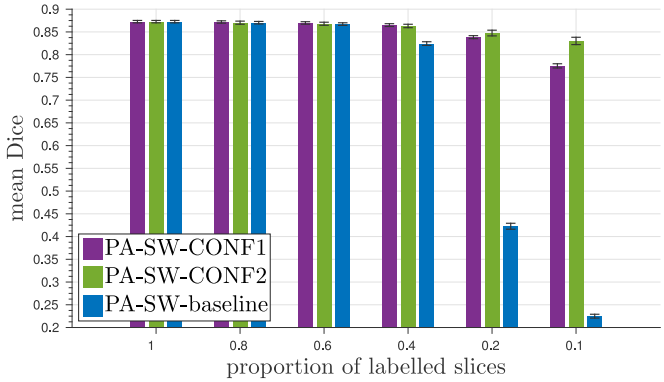


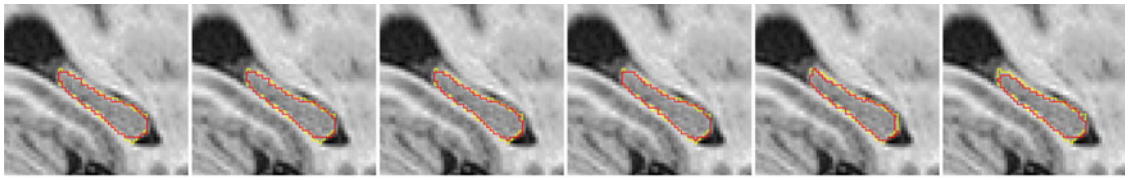
Fig. 9. Mean Dice coefficients for slicewise partial annotation for different proportions q of labelled atlas slices. PA-SW-CONF1, PA-SW-CONF2, PA-SW-baseline describe graph configurations and the error bars depict the standard error.

5.2.2 Results

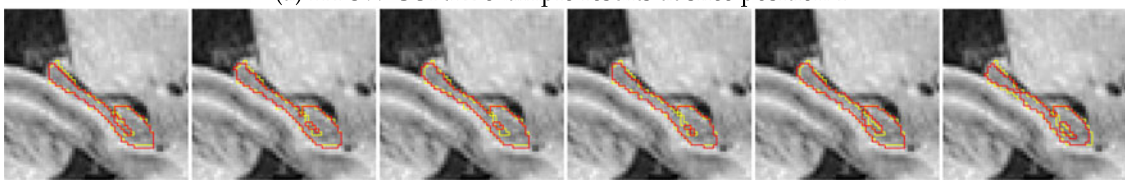
Results on the evaluation set were obtained using the same 10-fold cross-validation as described in Section 5.1.2. Fig. 9 shows the mean Dice coefficients pooled from all folds for all tested proportions of labelled slices q. For $q = 1$ (i.e., the group on the left), all atlas slices were labelled. In this case, the proposed graph configurations PA-SW-CONF1 and PA-SW-CONF2 are equivalent to MAS with regularisation refinement (MASr-LW). It can be seen that reducing the proportion of labelled atlas slices to $q = 0.4$ still yields comparable results for both tested configurations. When using fewer labelled slices, the performance decays rapidly for PA-SW-CONF1. For the second configuration CONF2, accuracy decreases as well, but more steadily. However, it is important to remember that the performance trade-off for, e.g., $q = 0.1$ stems from one tenth of the labelling effort. In contrast to this, results for PA-SW-baseline (where unlabelled atlas data was ignored, shown in blue) deteriorated rapidly when decreasing the annotation rate. Fig. 10 shows example segmentation results for one subject at two different slice positions (top and bottom rows) for decreasing values of q (left to right). For the slice in Fig. 10a, even using only every tenth atlas slice (i.e., $q = 0.1$ on the very right) did not influence the segmentation result. The slice in Fig. 10b was more challenging to segment due to the complex shape of the hippocampus. There, reducing the proportion of labelled atlas slices lead to failure in detecting the folding of the structure. Incorporating constraints preventing holes in the segmentation could potentially help reduce this effect.

5.3 Evaluation of Partial Annotation Strategy B: Scribbles (PA-SC)

Finally, we examined the performance of our framework when using data annotated with scribbles as proposed in Section 3.2. This experiment was carried out on *cardiac* MR data (Section 5.3.1) to demonstrate the applicability of our method to different (and multi-label) segmentation tasks. Scribbles are a suitable annotation strategy for this type of data due to the small number of slices (therefore it was feasible to annotate each slice) and good image contrast. In a first group of experiments, we investigated the scenario when the scribbles were available only on the atlases. This



(a) PA-SW-CONF2 example results at slice position 1



(b) PA-SW-CONF2 example results at slice position 2

Fig. 10. An example segmentation for PA-SW-CONF2 is shown in red and the ground truth segmentation in yellow. The same subject is shown at different slice positions in (a) and (b). From left to right, the proportion of labelled atlas slices q was $1, 0.8, 0.6, 0.4, 0.2, 0.1$.

partial annotation scenario will be referred to as PA-SC-A and was compared against MASr-LW with fully annotated atlases as a gold standard. We used the graph configuration CONF1 (as shown in Fig. 4a) since manual labels were available in roughly the same locations in all images (as opposed to the slicewise annotation strategy where entire slices remained unlabelled). Therefore, the complex propagation scheme CONF2 was not deemed necessary. In the second group of experiments, we examined scenarios which involve placing scribbles on a target image before automated segmentation, closely related to [28]. In the simplest configuration, scribbles were placed solely on the target image (PA-SC-T) [28], and no atlases were used. We then investigated if, in addition, a "scribbled" atlas database would improve these results (PA-SC-A+T). Here, scribbles were available both in the atlas database and the target image. Lastly, we used fully annotated atlases in combination with a scribbled target image (PA-SC-AF+T) to obtain a target segmentation with the proposed framework.

5.3.1 Data and Experiment Setup

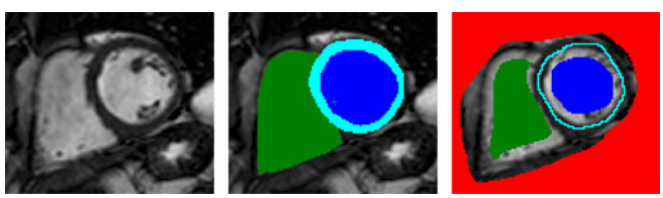
These experiments were performed for multi-label cardiac segmentation. The proposed method was tested on a short-axis cardiac MR dataset of 28 subjects in the end-diastole phase. The data were acquired on a 1.5T Philips Achieva system (Best, The Netherlands) using a 32-channel coil and the balanced-steady state free precession (b-SSFP) sequence. Images in the left ventricular short-axis plane were acquired using the following parameters: 320×320 mm field-of-view; 3.0 ms repetition time; 1.5 ms echo time; 50 ms shot duration; 8 mm section thickness with a 2 mm gap. The reconstructed MR images are of dimension $288 \times 288 \times 12$, with voxel spacing $1.23 \times 1.23 \times 10$ mm. The LV cavity, LV myocardium, and the RV cavity were manually annotated by two experienced imaging scientists. Ten subjects were labelled by one observer, whereas the other 18 were labelled by the second observer. The annotation time for a complete image was approximately 30 min. In addition, all images were partially annotated by a third observer. For this purpose, scribbles were placed on every slice for all structures (including the background). The task was set such that the observer should rapidly label large areas while not delineating the structure boundaries. This allowed the annotation time to be reduced to a mean time of 3.9 ± 0.6 min, i.e., a speedup of a factor > 7 compared to a full annotation. All manual annotations were done using ITK-SNAP [43].

The propagation weights β_{ij} for label fusion were chosen as in [10], where the same cardiac dataset was used. There, an exponential kernel was proposed based on the sum of squared distances between two patches centred around corresponding voxels in the target and atlas image. The optimal kernel width was found to be 50 and the patch size $3 \times 3 \times 1$ voxels. Suitable parameters for spatial regularisation a, σ_1 were found in a tuning step as described in Section 5.4.3.

5.3.2 Results

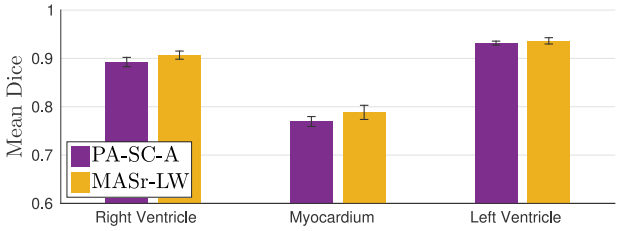
The proposed configurations were evaluated using each image not used during parameter tuning as a target image. The remaining images were used as atlases, respectively. For each target subject, the 15 most similar remaining images were used as atlases as in [10] (measured with normalised mutual information).

Fig. 12a shows mean Dice coefficients for the first group of experiments, where scribbles were placed on the atlases, and completely unlabelled target images were segmented using the proposed framework. It can be seen that using scribbled atlases (PA-SC-A) yielded results comparable to MASr-LW (where fully annotated atlases were used) for the right and left ventricle. For the myocardium, using scribbled atlases could not match the accuracy achieved when using fully annotated atlases. Fig. 13 shows example segmentation results for one subject. It can be seen that the results of PA-SC-A and MASr-LW are similar. However, since there is no boundary delineation in the scribbled atlases, the resulting segmentation results for PA-SC-A were more intensity

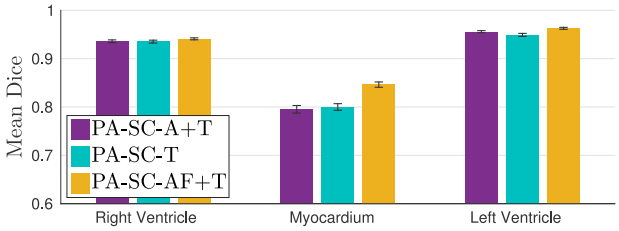


(a) Example image (b) Full annotation (c) Scribbles

Fig. 11. Example cardiac data: (a) shows an image of the heart and (b) shows the complete annotation of the left ventricular cavity (blue), the left ventricular myocardium (cyan) and the right ventricular cavity (green). (c) shows scribbles placed on the same image using ITK-SNAP [43].



(a) Results for experiments using scribbled atlases and MASr-LW



(b) Results for experiments using scribbled targets

Fig. 12. Mean Dice coefficients for experiments employing scribbles. (a) compares the performance of configurations using scribbled atlas data to fully annotated atlas data and in (b), results are shown for all configurations where the target itself contains scribbles as well.

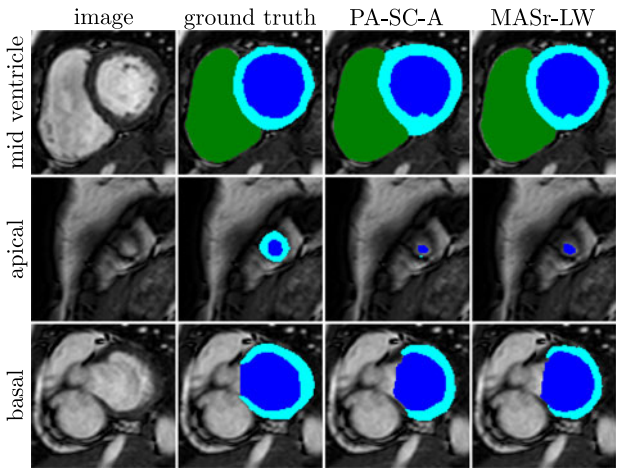


Fig. 13. Visual results for a mid-ventricular (top), apical (middle) and basal slice (bottom) for one subject. From left to right: The example image, ground truth segmentation and segmentation obtained with PA-SC-A and MASr-LW.

driven as can be seen for example in the myocardium in the mid-ventricular view.

The results for the second group of experiments are shown in Fig. 12b. Here, the target images to be segmented contained scribbles. In the simplest configuration PA-SC-T, a target segmentation is obtained from the scribbled target image only. Adding the scribbled atlases (PA-SC-A+T) yielded results very similar to PA-SC-T. However, placing scribbles in a target image to aid segmentation using *fully annotated* atlases (PA-SC-AF+T) yielded considerable improvements over both PA-SC-T (as seen in Fig. 12b) and MASr-LW (as seen in Fig. 12a. Visual results for these experiments are shown in Fig. 14 for the same subject as above. It can be seen that all three methods containing target scribbles were able to detect the myocardium in the apical slice, which was not possible using only atlas information (as seen in the middle row in Fig. 13). Furthermore, it can be seen that the segmentation obtained with fully annotated

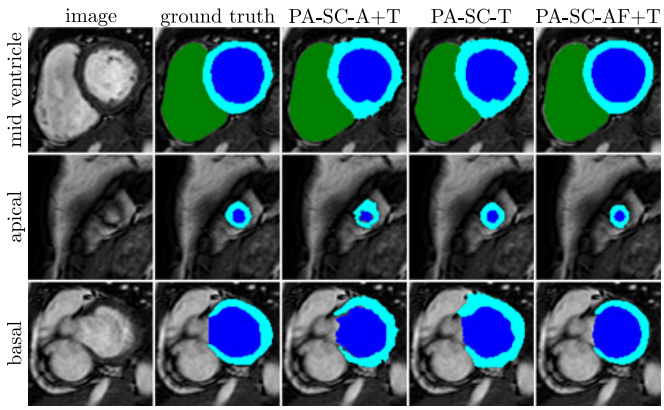


Fig. 14. Visual results for a mid-ventricular (top), apical (middle) and basal slice (bottom) for one subject. From left to right: the example image, ground truth segmentation and segmentation obtained with PA-SC-A+T, PA-SC-T, and PA-SC-AF+T.

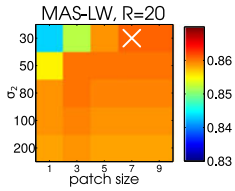


Fig. 15. Mean Dice coefficients for a grid search of the parameter choices for MAS-LW on $R = 20$ atlases. The white cross marks the optimal parameter choice.

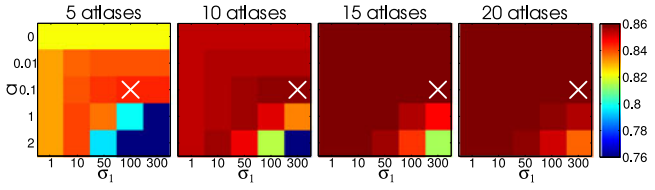


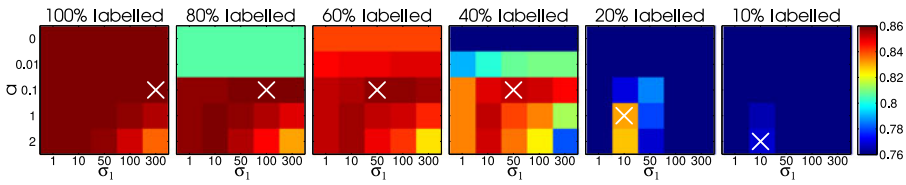
Fig. 16. Mean Dice coefficients for a grid search of the parameter choices for MASr-LW using $R = \{5, 10, 15, 20\}$ atlases (left to right). The white cross marks the optimal parameter choice for each experiment.

atlases and a scribbled target image (PA-SC-AF+T) is visually very similar to the ground truth segmentation, which is also reflected in the high Dice scores reported in Fig. 12b.

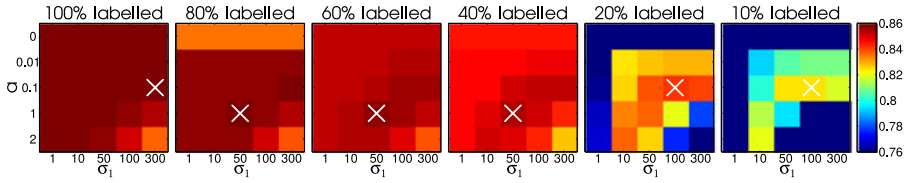
5.4 Analysis of Parameter Sensitivity

5.4.1 *Multi-Atlas Segmentation*

In this section, we describe the parameter selection procedure for the experiments performed in Section 5.1. First, we determined parameter values $\{\sigma_2, |P|\}$ for MAS-LW as introduced in Eq. (20). To do this, 10 target subjects were randomly drawn from the parameter tuning data. For each target image, the 20 most similar images in the remaining tuning images were used as atlases as recommended in [8] and the segmentation experiments were performed for a parameter range of $|P| = \{1, 3, 5, 7, 9\}$ and $\sigma_2 = \{30, 50, 80, 100, 200\}$. The parameter set yielding the highest mean Dice coefficient were used for evaluation and subsequent tuning of the regularisation coefficients a, σ_1 for MASr-LW. These parameters were tuned for $R = \{5, 10, 15, 20\}$ atlases, as we expected the number of atlases to have an influence on the optimal regularisation coefficients. The explored parameter range was $a = \{0, 0.01, 0.1, 2\}$ and $\sigma_1 = \{1, 10, 50, 100, 300\}$. Figs. 15 and 16 show the results of parameter tuning.



(a) Parameter tuning for PA-SW-CONF1



(b) Parameter tuning for PA-SW-CONF2

Fig. 17. Mean Dice coefficients for a grid search of the parameter choices using a proportion of $q = \{1, 0.8, 0.6, 0.4, 0.2, 0.1\}$ labelled slices in the atlases (left to right). The white cross marks the optimal parameter choice for each q. The colours encode the Dice coefficient (see colorbar on the right). The top (a) and bottom (b) rows show results for CONF1 and CONF2, respectively.

5.4.2 *Slicewise (SW) Partial Annotation Strategy*

For the experiments using slicewise partial annotations (Section 5.2), the regularisation parameters a, σ_1 were tuned on the same dataset as above. The parameters were tuned separately for both examined graph configurations CONF1 and CONF2. Fig. 17 shows optimal parameter choices for both PA-SW-CONF1 (Fig. 17a) and PA-SW-CONF2 (Fig. 17b) when using different proportions q of annotated atlas slices. The parameters with the highest mean Dice score for each configuration and each q were used during the evaluation.

5.4.3 *Scribbles (SC) Partial Annotation Strategy*

Here, parameter selection is discussed for the final experiment (Section 5.3) where scribbles are used for cardiac segmentation. To find parameter settings for regularisation, 10 random subjects were selected as target images. For each target subject, the 15 most similar images from the remaining population were used as atlases as in [10]. The parameter space was explored on the selected target subjects and the best performing set was used for the remaining population. The spatial regularisation parameters a, σ_1 were explored in a range of $\{0, 0.001, 0.01, 0.1, 1\}$ and $\{1, 10, 50, 100, 300\}$, respectively. Fig. 18 shows the tuning results for all experiment configurations, with optimal parameter choices marked with a white cross.

6 Discussion

In the experiments section, we first demonstrated how our framework can be used to express state-of-the-art techniques through modifications in the graphical representation of the labelling problem (Section 5.1). In particular, label fusion using the majority vote rule [2], [8] and locally weighted vote rule [7], [9], [10] were compared against locally weighted label fusion with added regularisation for spatial coherence. As expected, using more atlases generally improved segmentation accuracy [2]. The parameters for locally weighted label fusion were tuned using 20 atlases, which may explain the drop in performance of MAS-LW compared to MAS-MV when using fewer (i.e., 5 or 10) atlases. More elaborate parameter tuning should remove this effect as locally weighted fusion has been shown to outperform majority vote in similar settings [9]. Regularisation in the target image (MASr-LW) performed consistently better than MAS-LW. However, improvements became smaller for larger datasets where label fusion from many atlases caused inherent smoothness, yielding decreased benefit from additional spatial regularisation.

By re-interpreting label fusion as a pairwise component on an MRF energy function, it is possible to go beyond the scope of existing applications for multi-atlas segmentation. An important point is that the modular graph structure, where pairwise terms can be used for label propagation (between images) or spatial regularisation (within images) and where a unary term can be used to encode manual annotations, allows a relaxation of the annotation requirements for atlases. Therefore, the proposed framework can employ partially annotated images and represent unlabelled voxels simply by removing terminal links in the graph structure. Furthermore, the label propagation and regularisation schemes can be configured in different ways to facilitate information propagation in the graph. In Section 5.2, two configurations were used for hippocampal segmentation using partially labelled atlases where only a proportion of slices in each image were annotated. With both configurations, it was possible to achieve robust results when using as little as 40 percent of the annotations. Using the configuration where labels were propagated between atlases as well as to the target image (PA-SW-CONF2), it was possible to reduce the amount of labelled slices even further while still obtaining mean Dice coefficients of 0.83 ± 0.08 for $q = 0.1$. In that case for example, only every tenth slice was labelled in the atlases. Depending on the application, this performance trade-off could be acceptable, and this would mean that partially annotated atlas databases could be built in 10 percent of the time required to create a

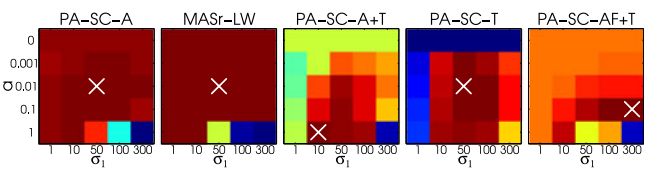


Fig. 18. Results of parameter tuning for experiments using scribbles. The color encodes a measure of combined segmentation accuracy in all structures of interest.

fully labelled dataset. When allowing propagation only between each atlas and the target image (PA-SW-CONF1), the performance decayed as the proportion of labelled atlas slices was reduced. This can be explained by the increased distance between labelled slices, making it more difficult for intra-image regularisation to interpolate labels. In contrast to CONF2, in CONF1 each voxel in the atlases is connected only to its spatial neighbours and the target image. Therefore, there may be large distances (on the graph) between unlabelled and labelled nodes. CONF2 addresses this problem by facilitating propagation between atlases as well, therefore reducing the distances of unlabelled nodes to nodes with strong data terms. A comparison of the above configurations to a configuration where unlabelled atlas data was ignored (PA-SW-baseline) revealed that exploiting unlabelled data as in PA-SW-CONF1 and PA-SW-CONF2 was crucial for achieving robust results when decreasing the annotation rate.

In the slicewise annotation strategy discussed above, the selected slices were completely annotated with detailed delineations of structures of interest. In contrast, scribbles were proposed as an alternative partial annotation strategy in Section 3.2, with the aim to save time by not requiring the observer to delineate the structure boundaries. We chose to design the task such that the scribbled areas were as large as possible without sacrificing speed on annotating details (as shown in Fig. 11c). Placing smaller scribbles could further increase speed, but likely at the expense of segmentation accuracy. The results presented in Fig. 12a show that using scribbled atlases yielded comparable performance to MASr-LW, albeit with slightly worse accuracy in the myocardium. The final set of experiments assumed the infrastructure for placing manual scribbles is available at segmentation time, as for example in interactive segmentation [28]. Results (Fig. 12b) showed that in this case, the additional help of scribbled atlases did not greatly influence segmentation results, indicating that scribbles in the target directly are sufficient for obtaining an accurate segmentation with the proposed framework. However, it can be seen that in combination with a scribbled target image, a fully annotated atlas set can improve segmentation results considerably in the myocardium, which is the most challenging structure to segment accurately.

6.1 Limitations and Future Work

The proposed method involves two computationally expensive steps: (1) the pairwise non-rigid registration between the atlases and each target image (approx. 2-10 min per registration step), and (2) the MRF energy minimisation step (approx. 5-10 min per segmentation task). To increase computational efficiency, an extension to the proposed framework could move from a voxel-wise representation of the images to a supervoxel representation. This change in the graphical representation could enhance the scalability of the proposed method to larger databases.

The formulation proposed in this paper assumes that the atlases are a good representation of the anatomy of the target image. This was achieved by atlas selection based on global image similarity as commonly used in multi-atlas segmentation [8]. To account for remaining anatomical variability in the selected atlases, we used *local* similarity measures for label fusion. However, when scaling the proposed method to larger databases of dissimilar images, the aforementioned assumption may no longer hold, and sparse connections between similar images only could ensure accurate label propagation, as well as alleviate computational burden due to registration.

In the scope of this paper, the data term was used exclusively to encode manual annotations. However, as briefly described in Section 2.3, the data term could also incorporate conditional label probabilities based on the observed intensities. These intensity models could be learned from the annotated data similar to [26] and applied to unlabelled regions in all images. This could make it feasible to further reduce the annotation rate while maintaining robust segmentation results. Furthermore, it would be of great interest to extend the data term to incorporate weak annotations such as bounding boxes or image tags.

7 Conclusion

In this paper, we proposed a unifying formulation for label propagation and regularisation based on a novel graphical representation of the labelling problem which is flexible and easily extendable. Small modifications in its configuration allow the use of partially annotated atlas data for segmentation. Experiments on two datasets demonstrated the usefulness of the proposed framework for segmentation using different partial annotation strategies. Pursuing these annotation strategies can save time and make annotating large databases feasible, while leading to robust segmentation results when combined with existing concepts in multi-atlas segmentation.

Acknowledgments

The authors thank Dr. Declan P. O'Regan from MRC Clinical Sciences Centre, Hammersmith Hospital, Imperial College London, for providing the cardiac MR data and the Alzheimer's Disease Neuroimaging Initiative for providing the brain MR data used in this manuscript. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007 2013) under grant agreement no. 601055, VPH-DARE@IT.

References

[1] C. R. Jack, et al., "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Resonance Imag.*, vol. 27, no. 4, pp. 685–691, 2008.

[2] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, 2006.

[3] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer, "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains," *NeuroImage*, vol. 21, no. 4, pp. 1428–1442, 2004.

[4] A. Klein, B. Mensh, S. Ghosh, J. Tourville, and J. Hirsch, "Mindboggle: Automated brain labeling with multiple atlases," *BMC Med. Imag.*, vol. 5, 2005, Art. no. 7.

[5] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, 2015.

[6] R. Wolz, P. Aljabar, J. V. Hajnal, A. Hammers, and D. Rueckert, "LEAP: Learning embeddings for atlas propagation," *NeuroImage*, vol. 49, no. 2, pp. 1316–1325, 2010.

[7] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solórzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1266–1277, Aug. 2009.

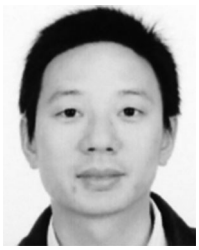
[8] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," *NeuroImage*, vol. 46, no. 3, pp. 726–738, 2009.
[9] M. R. Sabuncu, B. T. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," *IEEE Trans. Med. Imag.*, vol. 29, no. 10, pp. 1714–1729, Oct. 2010.
[10] W. Bai, et al., "A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: Application to cardiac MR images," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1302–1315, Jul. 2013.
[11] J. E. Iglesias, et al., "An algorithm for optimal fusion of atlases with different labeling protocols," *NeuroImage*, vol. 106, pp. 451–463, 2015.
[12] 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.
[13] H. Wang, J. Suh, S. Das, J. Pluta, C. Craige, and P. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 611–623, Mar. 2013.
[14] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, "Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation," *NeuroImage*, vol. 54, no. 2, pp. 940–954, 2011.
[15] F. Rousseau, "A supervised patch-based approach for human brain labeling," *IEEE Trans. Med. Imag.*, vol. 30, no. 10, pp. 1852–1862, Oct. 2011.
[16] F. van der Lijn, T. den Heijer, M. Breteler, and W. J. Niessen, "Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts," *NeuroImage*, vol. 43, no. 4, pp. 708–720, 2008.
[17] J. M. Lötjönen, et al., "Fast and robust multi-atlas segmentation of brain magnetic resonance images," *NeuroImage*, vol. 49, no. 3, pp. 2352–2365, 2010.
[18] A. Makropoulos, et al., "Automatic whole brain MRI segmentation of the developing neonatal brain," *IEEE Trans. Med. Imag.*, vol. 33, no. 9, pp. 1818–1831, Sep. 2014.
[19] C. Ledig, et al., "Robust whole-brain segmentation: Application to traumatic brain injury," *Med. Image Anal.*, vol. 21, no. 1, pp. 40–58, 2015.
[20] M. Rajchl, et al., "Hierarchical max-flow segmentation framework for multi-atlas segmentation with Kohonen self-organizing map based Gaussian mixture modeling," *Med. Image Anal.*, vol. 27, pp. 45–56, 2016.
[21] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[22] L. M. Koch, et al., "Graph-based label propagation in fetal brain MR images," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2014, pp. 9–16.
[23] D. Kuettel, M. Guillaumin, and V. Ferrari, "Segmentation propagation in ImageNet," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 459–473.
[24] M. Rubinstein, C. Liu, and W. T. Freeman, "Annotation propagation in large image databases via dense image correspondence," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 85–99.
[25] M. J. Cardoso, et al., "Geodesic information flows: Spatially-variant graphs and their application to segmentation and fusion," *IEEE Trans. Med. Imag.*, vol. 34, no. 9, pp. 1976–1988, Sep. 2015.
[26] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.
[27] L.-C. Chen, S. Fidler, A. L. Yuille, and R. Urtasun, "Beat the MTurkers: Automatic image labeling from weak 3D supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3198–3205.
[28] Y. Boykov and M. Jolly, "Interactive organ segmentation using graph cuts," in *Proc. 3rd Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2000, pp. 276–286.
[29] J. Xu, A. G. Schwing, and R. Urtasun, "Tell me what you see and i will show you where it is," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3190–3197.
[30] J. Xu, A. G. Schwing, and R. Urtasun, "Learning to segment under various forms of weak supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3781–3790.
[31] B. A. Landman, A. Asman, A. Scoggins, J. Bogovic, F. Xing, and J. Prince, "Robust statistical fusion of image labels," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 512–522, Feb. 2012.
[32] S. Li, "Markov random field models in computer vision," in *Proc. Eur. Conf. Comput. Vis.*, 1994, pp. 361–370.
[33] J. Yuan, E. Bae, and X. Tai, "A study on continuous max-flow and min-cut approaches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2217–2224.
[34] J. Yuan, E. Bae, X. Tai, and Y. Boykov, "A continuous max-flow approach to potts model," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 379–392.
[35] D. Han, J. Bayouth, Q. Song, and A. Taurani, "Globally optimal tumor segmentation in PET-CT images: A graph-based co-segmentation method," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2011, pp. 245–256.
[36] W. Qiu, J. Yuan, E. Ukwatta, Y. Sun, M. Rajchl, and A. Fenster, "Prostate segmentation: An efficient convex optimization approach with axial symmetry using 3D TRUS and MR images," *IEEE Trans. Med. Imag.*, vol. 33, no. 4, pp. 947–960, Apr. 2014.
[37] L. M. Koch, M. Rajchl, T. Tong, J. Passerat-Palmbach, P. Aljabar, and D. Rueckert, "Multi-atlas segmentation as a graph labelling problem: Application to partially annotated atlas data," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2015, pp. 221–232.
[38] R. Wolz, C. Chu, and K. Misawa, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE Trans. Med. Imag.*, vol. 32, no. 9, pp. 1723–1730, Sep. 2013.
[39] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.
[40] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.
[41] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.
[42] L. G. Nyúl and J. K. Udupa, "On standardizing the MR image intensity scale," *Magn. Resonance Med.*, vol. 42, no. 6, pp. 1072–1781, 1999.
[43] P. A. Yushkevich et al., "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, 2006.

Lisa Margret Koch studied electrical and biomedical engineering at ETH Zurich, Switzerland, where she received the MSc degree, in 2011. After one year working as a embedded software and hardware engineer, she joined the Biomedical Image Analysis Group, Imperial College London, in 2012. She is now working toward the PhD degree, which revolves around graphical data representation and graphical models for image segmentation.

Martin Rajchl studied biomedical engineering, Vienna and received the PhD degree from the Robarts Research Institute, Western University, Canada, in 2014. He is now a research associate in the Department of Computing, Imperial College London working on computer vision and machine learning methods for large-scale analysis of medical imaging data. His research interests include graphical optimisation, machine learning, and high-performance computing.

Wenjia Bai received the BEng and MEng degrees on automation from Tsinghua University, in 2004 and 2007, respectively, and the DPhil degree on engineering science from Oxford University, in 2011. His thesis was on respiratory motion correction for PET imaging. Since 2011, he has been working with Imperial College London as a research associate on medical image analysis. His research interests include machine learning on medical imaging and large-scale medical image analysis.

Christian Frederik Baumgartner received the the MSc degree in biomedical engineering from the Federal Technical Institute (ETH), Zurich, Switzerland. He received the PhD degree from the Division of Imaging Sciences, King's College London. He is currently working as a research associate with Imperial College London, BioMedIA Group, where he is specialising on machine learning in medical image analysis.



Tong Tong received the bachelor's and master's degrees in biomedical engineering from the Beijing Institute of Technology and University of Science and Technology of China, respectively, and the PhD degree in computing from Imperial College London, in 2015. He is now a research fellow with Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital / Harvard Medical School, working on brain image analysis.

Jonathan Passerat-Palmbach received the PhD degree from Blaise Pascal University, Clermont-Ferrand, France. He is currently a research associate with Imperial College London, in the BioMedIA Group. His work is focused on high performance computing tools and software engineering applied to medical imaging and neuroinformatics.

Paul Aljabar received the PhD degree in computational image analysis and studied mathematics at King's College London and worked as a high school teacher before re-training. He has developed frameworks to provide normative estimates for biological structures, image-based classification or methods to cluster imaging datasets to support subsequent analysis. His research interests include machine-learning for image analysis tasks such as segmentation and registration.

Daniel Rueckert received the diploma in computer science (equiv to MSc degre) from the Technical University, Berlin and the PhD degree in computer science from Imperial College London. He is a professor of visual information processing in the Department of Computing, Imperial College London, where he leads the Biomedical Image Analysis Group. His research include image registration, image segmentation and the application of machine learning in medical imaging. He is a fellow of the Royal Academy of Engineering, the IEEE, and the MICCAI Society.