Automatic segmentation of the striatum and globus pallidus using MIST: Multimodal Image Segmentation Tool

Visser, E.; Keuken, M.C.; Douaud, G.; Gaura, V.; Bachoud-Levi, A.C.; Remy, P.; Forstmann, B.U.; Jenkinson, M.

Published in:
NeuroImage

DOI:
10.1016/j.neuroimage.2015.10.013

Link to publication

Creative Commons License (see https://creativecommons.org/use-remix/cc-licenses):
CC BY

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Automatic segmentation of the striatum and globus pallidus using MIST:
Multimodal Image Segmentation Tool

Eelke Visser a,⁎, Max C. Keuken b, Gwenaëlle Douaud a, Veronique Gaura c,d, Anne-Catherine Bachoud-Levi e,f,g,h, Philippe Remy c,i, Birte U. Forstmann b, Mark Jenkinson a

ARTICLE INFO

Article history:
Received 20 February 2015
Accepted 6 October 2015
Available online 19 October 2015

Keywords:
Segmentation
Multimodal
Striatum
Globus pallidus
Brain
Huntington

ABSTRACT

Accurate segmentation of the subcortical structures is frequently required in neuroimaging studies. Most existing methods use only a T1-weighted MRI volume to segment all supported structures and usually rely on a database of training data. We propose a new method that can use multiple image modalities simultaneously and a single reference segmentation for initialisation, without the need for a manually labelled training set. The method models intensity profiles in multiple images around the boundaries of the structure after nonlinear registration. It is trained using a set of unlabelled training data, which may be the same images that are to be segmented, and it can automatically infer the location of the physical boundary using user-specified priors. We show that the method produces high-quality segmentations of the striatum, which is clearly visible on T1-weighted scans, and the globus pallidus, which has poor contrast on such scans. The method compares favourably to existing methods, showing greater overlap with manual segmentations and better consistency.

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Introduction

The subcortical structures of the brain are a set of nuclei with unique anatomical shapes and, due to their histological composition, specific MR contrast properties. Automatic segmentation of these structures is often desired, as manually segmenting them is a process that is both laborious and prone to labelling inconsistencies and errors. Unlike cortical segmentation, which is a nearly universal process for all of the cortex, subcortical segmentation requires methods that are tailored to these structures specifically, due to both their unique anatomies and contrasts. These methods produce an estimate of the extent of the structures, either as voxel labels or in the form of a mesh describing the outer surface of the shape.

Numerous methods have been proposed to automatically segment the subcortical structures and a number of these are used routinely in imaging studies. The input data typically consist of a single T1-weighted magnetic resonance imaging (MRI) volume. Such an image is often a standard part of the acquisition protocol in neuroimaging studies and serves as the anatomical reference in the analysis. The T1-weighted contrast is most suitable for structures such as the putamen and the caudate nucleus, which are clearly visible on such scans. Other structures, like the globus pallidus and the red nucleus are barely visible or invisible. In such cases, a segmentation method that only uses a T1-weighted volume needs to rely more on anatomical priors.

Iosifescu et al. (1997), Khan et al. (2005) and Lin et al. (2010) describe methods that segment the subcortical structures by non-linearly registering an atlas of the subcortical anatomy to the new scan. This idea is extended by Gouttard et al. (2007) through the use of a probabilistic atlas. Yan et al. (2013) describe an adaptive atlas and Van Leemput (2009) applies Bayesian inference to optimise atlas and registration parameters during atlas generation. Multiple manual
segmentations can be individually registered to a new scan and combined into a final segmentation by building a probabilistic atlas (Svarer et al., 2005) or using a decision fusion scheme such as using the modal label (Heckemann et al., 2006), support vector machine-based selection (Hao et al., 2014), graph-based selection (Wang et al., 2014), joint label fusion (Wang et al., 2013; Yusheevich et al., 2015) or fusion of the likelihoods under the different atlases (Tang et al., 2013). J. Morra et al. (2008a) and J.H. Morra et al. (2008b) use a classifier trained on non-local image features using an extension to AdaBoost (Freund and Schapire, 1997). A method using a single template and explicit modelling of an hierarchy of structures is described by Wu and Chung (2009).

Fischl et al. (2002, 2004) and Han and Fischl (2007) take a different approach in the method that is used in the FreeSurfer software package (http://surfer.nmr.mgh.harvard.edu/). A probabilistic atlas, which is derived from a number of manual segmentations, is still used, but in this case it is used as a prior on a Markov random field (MRF). The spatial constraints imposed by the atlas allow the MRF to segment the image into the large number of classes needed to segment the subcortical structures. This idea is combined with a multi-atlas approach by van der Lijn et al. (2008) and Wolz and Aljabar (2009). A mesh-based spatial prior can also be used (Puonti et al., 2013). Khan et al. (2008) propose to use the segmentation output from the method by Fischl et al. (2002) to drive registration of a secondary manually labelled template.

The information fusion approach described by Barra and Boire (2001) segments structures based on a model of expert knowledge. General descriptions of experts of segmentation criteria are converted into image-based pieces of information, from which a fusion process finds the final segmentation. Fuzzy neural networks have also been proposed for segmentation (Jian et al., 2002). An approach based on discriminative dictionary learning was proposed by Benkarim et al. (2014).

Finally, a number of methods segment structures by fitting a mesh to the imaging data. The method described by Belitz and Rohr (2006) uses an active surface model (ASM), which is based on image-based edge detection. Asl and Soltanian-Zadeh (2008) also use a shape model, but optimise with respect to the image entropy within structures instead. Liu et al. (2007) propose a method that incorporates a probabilistic atlas and Uzunbas et al. (2010) and Cerrolaza et al. (2012) take advantage of the spatial relationships between structures. First (Patenaude et al., 2011) uses a Bayesian appearance model that links the intensities around a deformable shape to the spatial configuration of the shape. The intensities and possible shape variations are derived from a set of manually labelled training segmentations.

All of the methods described above segment subcortical structures on the basis of only a $T_1$-weighted scan. In the case of a poorly visible structure like the globus pallidus, a different contrast, like a $T_2$-weighted scan, can be more useful. Magnotta et al. (1999) propose a method that uses multiple contrasts to segment the brain tissue into grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF). An artificial neural network is then trained using the tissue-type maps and a set of manual labels to label the subcortical structures. Multi-modal extensions have also been proposed to atlas-based methods (Xiao et al., 2012) and random-field-based segmentation (Marrakchi-Kecem et al., 2010, related to Poupon et al., 1998). Kim et al. (2014) propose a method that combines explicit edge detection on multiple modalities with a statistical shape model that is trained using multiple manual segmentations.

Many of the methods described above need to be trained using manually labelled data. The required amount of training data can be either a single segmentation for deterministic atlas-based methods or a larger set of labelled scans for the classifier and probabilistic atlas-based methods. The degree to which methods are constrained to the properties of these training segmentations varies. Fully atlas-based methods will not capture details at a higher resolution than that of the deformation field, while the random-field-based approach in FreeSurfer should allow it to detect smaller variations more flexibly. FIRST parameterises shapes using a limited number of modes of variation that are derived from the training segmentations.

If segmentation is constrained by the variations that were seen in the training data, this may have undesirable consequences when applying the method to populations that are different from the training set. An example would be patient groups, where anatomy might be significantly different from the normal population. Segmentation results in such a group would be biased towards the appearance of the structure in the training data. The same holds for the less typical cases within the normal population; these might not be within the variability observed in the training group. In the case of pathology, there might also be unique variations in single subjects.

Because of these potential differences compared to the training group, it is a desirable property of a segmentation method to require only a limited number of manual segmentations, as this makes retraining the method for a new population more tractable. The segmentation method could then be re-trained to the images that are to be segmented, without requiring a full set of manually labelled training segmentations. This would remove much of the bias towards the original training group, but would not solve the more idiosyncratic cases. To address this problem, a more flexible approach with weaker constraints is required. A further potential disadvantage of using manually labelled training segmentations is that it introduces some subjectivity in the segmentation process. We would like a new method to be more data-driven and less reliant on manual segmentations. When a method can be more easily retrained, it can also be adapted if the image contrast in a study is fundamentally different from that in the training data. This might be the case if acquisition is optimised for the segmentation of a particular structure, for example the globus pallidus, which is hard to segment on a $T_1$-weighted image but much clearer on a $T_2$-weighted one.

In this paper, we describe a multimodal segmentation method that can simultaneously use information in multiple volumes with different contrasts. This has two potential advantages. First, if not all boundaries of a structure are visible in each contrast, it allows the method to combine the complementary information. Secondly, in cases of low contrast, the availability of multiple images can help segmentation as it effectively increases the contrast-to-noise ratio. The method detects edges in images based on their intensity profiles, allowing it to capture shape variations that did not occur in the training set. We use a generative model for the intensity profiles perpendicular to a deformable mesh, somewhat like Patenaude et al. (2011), but unlike that method it can learn in an unsupervised fashion from unlabelled training data. The user input is limited to the reference mesh, which may be derived from a single manual segmentation or from an existing atlas, and a set of priors describing the general form of contrast changes at the boundaries of the structure. We describe an automatic procedure for setting these parameters for the striatum and globus pallidus, to enable retraining on different datasets and populations. To use the method with new contrasts and structures, the user can specify the priors by hand using the same principles described in the paper. We will refer to the method as 'multimodal image segmentation tool (MIST)'.

**Methods**

We aim to integrate the information in multiple imaging modalities to produce accurate segmentations of subcortical structures. As a starting point, we take a reference mesh that roughly corresponds to the structure that we are interested in and map this to a subject’s data using nonlinear registration. The reference mesh may, for example, be derived from an atlas. The method should then deform the mesh to optimally align with the physical boundaries of the structure of interest. The intensity profiles along the normal vectors at the vertices of the reference mesh are used to estimate the adjustments that are needed to make the reference mesh align with the anatomy.
An important question is how the method will be able to relate observed image intensities to the actual structural boundaries. If we had used manually labelled training segmentations, this would be trivial, as the physical boundary would be defined as the point where the person that labelled the image drew the boundary. However, as we do not have such manual segmentations, we need to devise an alternative method. The approach that we take is based on the simultaneous alignment of the profiles observed in an unlabelled training group, which is combined with a prior that specifies, in general terms, what intensity changes we expect near the boundary. Such a prior resolves the ambiguity in the connection between an edge in the observed intensity profiles and the physical boundary.

The framework that we will use is a probabilistic one, where a generative model describes how between-subject anatomical variability gives rise to the observed intensities. We will use a Bayesian approach where we find estimates of the model parameters in a training phase. The trained model is then used to obtain an approximate maximum a posteriori (MAP) estimate of the deformations needed to segment a given image.

In the next sections, an overview of the generative model will be given first, followed by a discussion of how we obtain estimates of its parameters and how the priors need to be set. More practical issues, such as prerequisite image registration, are discussed in the final part of this section. We will assume for now that the reference mesh is roughly aligned with the images that are to be segmented.

**Generative model**

The model that we use can be seen as consisting of two parts: A part that we will call the *shape model*, which describes how the displacements of different vertices vary and covary, and a second part, the *intensity model*, which is independent for each vertex and which describes how the intensities arise given a displacement.

**Shape model**

We assume that the displacements along the vertex normals are sampled from a multivariate normal (MVN) distribution with mean $\mu^s$ and precision $\Lambda^s$:

$$p(\delta^s | \mu^s, \Lambda^s) = N_{y}(\delta^s | \mu^s, \Lambda^s).$$

In this equation, $\delta^s$ is the $N$-dimensional vector consisting of the displacements of all $N$ vertices along the normals.

**Priors for the shape model**

The role of the shape model is to learn the differences between the mean location of the structural boundary in the training data and the reference mesh. It also constrains the segmentation by controlling the covariance between vertices. We use a Normal–Wishart distribution as the prior for the mean and precision of the shape model:

$$N_{w}(\mu^s, \Lambda^s) = N_k(\mu^s | \mu_0^s, \Lambda^s_0)W_{k}(\Lambda^s | \alpha^s, \beta^s).$$

The prior mean $\mu_0^s$ is set to zero and the elements of the covariance (inverse precision) prior are specified as a Gaussian function of the surface distances between the vertices:

$$\Sigma_{ij} = h^2 \exp \left( -\frac{d_{ij}}{w} \right),$$

where $d_{ij}$ is the number of edges between vertices $i$ and $j$, $w$ is the width parameter and $h^2$ is the expected on-diagonal value of the covariance matrix; we will refer to $h$ as the height parameter. The width parameter controls how strong we expect the correlation between neighbouring vertices to be. The parameters $\mu_0^s$ and $\alpha^s$ control the width of the prior distributions. More details about all of these parameters are given in Appendix A.

**Intensity model**

In order to be able to later fit the mesh to imaging data, we need to define the relationship between the two. In this section we shall only be concerned with a single vertex on the mesh, as the form of the model is identical and independent for all vertices. We start with a simple model where the intensities sampled at equidistant points along the normals of the anatomical structure are represented by an MVN distribution:

$$p(y | \mu, \Lambda) = N_{k}(y | \mu, \Lambda),$$

where $y$ is the $k$-dimensional vector of the intensities that make up the profile and $\mu$ and $\Lambda$ are the mean and precision. The profile $y$ is centred around the physical boundary of the structure.

To be able to estimate displacements, the profile that is considered in the image is shorter than the reference profile $\mu$. The observed profile $y$ has length $k$ and is centred around the reference shape. We assume that this is sampled from the distribution described by Eq. (4). As the observed profile is shorter than $\mu$, however, we need to marginalise over the unobserved points. For an MVN distribution, this is as simple as dropping the corresponding dimensions. The number of dimensions that need to be dropped on either end of the profile is determined by the integer displacement $\hat{\delta}$ at this vertex:

$$p(y | \hat{\delta}) = \sum_{r=1}^{\Delta} p(y | \hat{\delta}, \mu, \Lambda)r(r) = N_{k}(y | \mu^s, \Lambda^s),$$

with the integer displacement $\hat{\delta}$ defined with respect to the maximum negative displacement. The maximum integer displacement is $\Delta = k - k'$. In practice, we will use the same value for $\Delta$ and $k'$, which means that, for example, an observed profile of 20 points would have maximum displacements of 10 points inwards and 10 points outwards. The length of the reference profile would be $k = 40$ in this case. We will use the superscript $\hat{\delta}$ to denote taking the subvector or submatrix corresponding to an integer displacement $\hat{\delta}$.

It is useful to note at this point that in the training stage, due to the lack of manually labelled training data, the displacement $\hat{\delta}$ will be unknown. As we will see later, this ambiguity can be resolved by using informative priors.

The anatomy that is observed along the normal at a given vertex may not be the same for all subjects. Examples of this might be vessels, whose exact location varies with respect to the vertices, or the exact points where neighbouring structures start to be connected to the structure of interest. In order to model such variability, we extend the simple model in Eq. (5) by replacing the single MVN by a mixture model with $N_r$ components:

$$p(y | \hat{\delta}, \mu_0, \Lambda_0, \ldots, \mu_{N_r}, \Lambda_{N_r}, \theta) = \sum_{r=1}^{N_r} p(y | \hat{\delta}, \mu_r, \Lambda_r)\text{Cat}(r | \theta) = \sum_{r=1}^{N_r} N_{k}(y | \mu_r^s, \Lambda_r^s)\text{Cat}(r | \theta),$$

where $\text{Cat}(r | \theta)$ is the categorical distribution and where $\mu_r$ and $\Lambda_r$ are the mean and precision matrix for component $r$.

The final extension that we need to make to the profile model is the extension to multiple modalities. This is relatively straightforward:

$$p(y_0, \ldots, y_{N_r} | \hat{\delta}, M, L, \theta) = \prod_{m=1}^{r_{N}} \sum_{r=1}^{N_r} N_{k}(y_m | \mu_{mr}^s, \Lambda_{mr}^s)\text{Cat}(r | \theta_m).$$
where \( N_m \) is the number of modalities and where we have used the shorthands \( M, L \) and \( \theta \) to denote all mean vectors \( \mu_m \), all mean precision matrices \( \Lambda_m \) and all mixing parameter vectors \( \theta_m \), respectively.

**Parameterisation of the covariance matrix**

The precision matrices in the intensity model are parameterised through the inverse of the covariance matrix:

\[
(\Lambda_m)^{-1} = \Sigma_m = DG_m G, \tag{10}
\]

where \( D_m \) is the diagonal matrix that parameterises the covariance matrix and \( G \) is a symmetric matrix that determines the smoothness. When fitting the model, \( G \) is a constant parameter, while \( D_m \) needs to be determined from the data. Both \( \Sigma \) and the precision matrix \( \Lambda = (\Sigma)^{-1} \) need to be computed frequently and this can be done quickly as \( G^{-1} \) is constant and needs to be computed only once and \( D_m \) being diagonal, is trivial to invert. The elements of the matrix \( G \) are:

\[
G_{ij} = e^{-\alpha^2 \sigma^2}, \tag{11}
\]

where \( \sigma^2 \) is the smoothness parameter of the intensity model.

**Priors for the intensity model**

The priors that we will set on \( \mu_m \) and \( \Lambda_m \) are crucial for the method to be able to detect edges rather than to rely on manual segmentations to build the reference profiles. Without prior information, the relationship between the physical boundary and the mean profile intensity is undefined. By setting a prior that specifies what we expect an edge to look like, this ambiguity is resolved. We use a Normal–Wishart prior for \( p(\mu_m, \Lambda_m) \):

\[
\text{NW}_k(\mu_m, \Lambda_m) = \text{NW}_k(\mu_m, \Lambda_m | n_0 \Lambda_m) \text{NW}(\Lambda_m | \alpha^0, \beta^0), \tag{12}
\]

where \( \text{NW} \) denotes a Wishart distribution. The hyperparameters that we need to set are the parameters \( \mu_m^0 \), \( \Lambda_m^0 \), \( \alpha^0 \) and \( \beta^0 \). The parameters \( \mu_m^0 \) and \( \Lambda_m^0 \) control the mean and variance of the distribution on \( \mu_m \) and the parameters \( \alpha^0 \) and \( \beta^0 \) control the mean and variance of \( \Lambda_m \).

In general, when setting up the method to segment a structure, we will have some general idea about what the intensity profiles perpendicular to the structural boundaries look like. We would like to specify these ideas once for the structure (i.e. we want to use the same priors for all vertices) and therefore the specification needs to be relatively general. The model can then learn more specific features of the profile and the probabilities of the mixture components appearing at a specific vertex. We will parameterise the prior probabilities by selecting one of the functions in Table 1.

We have parameterised the covariance matrices through Eq. (10) and we will use a similar structure for the covariance hyperparameter \( \beta^0 \) of the Normal–Wishart distribution. The Wishart distribution includes covariance matrices that cannot be generated through Eq. (10) and the parameterisation of \( (\Lambda_m)^{-1} \) effectively acts as an additional constraint to enforce the desired off-diagonal structure. While in principle different values can be specified for different points along the profile, there is little advantage to doing so and we use a scalar value \( \beta_m \) for all of the coefficients:

\[
\beta_m^0 = \beta_m G G. \tag{13}
\]

For the mixing parameters \( \theta_m \) we use a Dirichlet prior \( \text{Dir}(\theta_m | \alpha) \). All components of the hyperparameter \( \alpha \) are set to the same value. This value determines a preference towards either equal mixing or selection of a single component.

**Full model**

Combining Eqs. (1) and (9) yields the joint distribution for the full generative model:

\[
p(Y_1, ..., Y_N; \delta(M, L, \theta_1, ..., \theta_N, \mu_r, \Lambda_r)) = N_N(\delta | \mu_r, \Lambda_r) \prod_{r=1}^{N_r} \prod_{m=1}^{N_m} \sum_{s=1}^{N_s} N_k \left( y_{im} | \mu_{mr}, \Lambda_{mr} \right) \text{Cat}(r | \theta_m), \tag{14}
\]

where \( N \) is the number of vertices and where \( Y \) denotes all vectors \( y_{im} \). The subscripts \( \delta \) refer to the integer displacements corresponding to the components of \( \delta \). Eq. (14) will only be evaluated for these integer values.

Eq. (14) shows that the individual profile mixture models at different vertices are thus linked through the shape distribution \( N_N(\delta | \mu_r, \Lambda_r) \). The model parameters will be learned in the training stage, which will be described below. The training stage makes use of the priors introduced above to find posterior estimates of the parameters. Due to the complexity of the model, in practice we will train the profile and shape parts of the model sequentially. For the shape model, we will use the analytical conjugate process and this part of the model is fully Bayesian. Maximum a posteriori (MAP) parameter estimates are used in the intensity model.

**Model estimation**

The intensity models at different vertices are in principle linked through the shape model. Estimating the full model is computationally intensive, however, given the large number of vertices. For this reason, we estimate the intensity models for all vertices independently and use point estimates for their parameters when estimating the shape model.

**Estimating \( \mu_m, \Lambda_m \) and \( \theta_m \) from training data**

The parameters \( \mu_m, \Lambda_m \) and \( \theta_m \) are unknown and we want to learn these from a set of training data. From Eq. (9), it is easily seen that, for a single vertex, the combined probability for the full set of intensity profiles \( z_{im} \) in the training data is:

\[
p(z_1, ..., z_S; \delta_1, ..., \delta_S; M, L, \theta) = \prod_{i=1}^{S_m} \prod_{m=1}^{N_m} \sum_{r=1}^{N_r} N_k \left( z_{im} | \mu_{mr}, \Lambda_{mr} \right) \text{Cat}(r | \theta_m), \tag{15}
\]

where \( \delta_i \) is the displacement for subject \( s \), \( N_i \) is the number of training subjects and the vertex index \( i \) has been omitted. Applying Bayes’ rule yields

\[
p(M, L, \theta, \delta_1, ..., \delta_S | z_1, ..., z_S) = \prod_{i=1}^{S_m} \prod_{m=1}^{N_m} \sum_{r=1}^{N_r} N_k \left( z_{im} | \mu_{mr}, \Lambda_{mr} \right) \text{Cat}(r | \theta_m) \tag{16}
\]

The priors \( N(\delta_i | \mu_i, \Lambda_i) \) replace the shape model. This replacement is necessary to be able to train the profile part of the model separately for each vertex, but it is an approximation to the full model. The

<table>
<thead>
<tr>
<th>Description</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat</td>
<td>( f_{\text{flat}}(a) = a )</td>
</tr>
<tr>
<td>Step</td>
<td>( f_{\text{step}}(a, b) = \begin{cases} a &amp; \text{if } x &lt; 0 \ b &amp; \text{if } x \geq 0 \end{cases} )</td>
</tr>
<tr>
<td>Exponential</td>
<td>( f_{\text{exp}}(v, a, b) = \begin{cases} a &amp; \text{if } x &lt; 0 \ (a + (b-a)e^v) &amp; \text{if } x \geq 0 \end{cases} )</td>
</tr>
</tbody>
</table>
parameter $\mu_i$ is set to zero, i.e. the distribution is centred around the reference shape.

When fitting the full model to new data, we will use point estimates for the parameters $\mu_{mr}, \Lambda_{mr}$ and $\theta_m$. These are the MAP estimates in the single-vertex isolated intensity model. To simplify the problem of finding these, it is useful to first marginalise out $\delta_1, \ldots, \delta_5$:

$$p(M, L, \theta | z_1, \ldots, z_N) \propto \prod_{m=1}^{N_m} \prod_{r=1}^{N_r} p(\mu_{mr}, \Lambda_{mr}) \prod_{r=1}^{N_r} \sum_{\delta_r=0}^{\delta_{r \max}} \sum_{r=1}^{N_r} N_{r} \left( z_{mr} | \mu_{mr}, \Lambda_{mr} \right) \text{Cat}(\theta_m),$$

(17)

where the sum is over the integer values of $\delta_r$. The MAP estimates of the parameters $\mu_{mr}, \Lambda_{mr}$ and $\theta_m$ are the values where this function is at its global maximum. For $\Lambda_{mr}$ the parameterisation that was described above is used.

To perform the optimisation, the method of moving asymptotes (MMA, Svanberg, 2002) is used, as implemented in the NLopt optimisation library (Johnson, http://ab-initio.mit.edu/nlopt). This algorithm was chosen because it converged more quickly in practice than the alternatives that were considered. Optimisation is performed using the logarithm of the probability given by Eq. (17). The logarithm of the probability and its derivatives, which are needed for MMA, are given in the Supplementary material.

Training the shape model

As we train the intensity models in isolation, we can take the most likely displacements for all subjects and use these to train the shape part of the model. We use the standard training process using conjugate priors; the details are given in the Supplementary material.

Estimating the approximate full model

With the trained intensity models and the shape model we can obtain an estimate of the vector of displacements $\delta$ for a new set of data given the training data. It is not difficult to see that the conditional probability of the continuous displacements $\delta$ is proportional to Eq. (14):

$$p(\delta | Y_1, \ldots, Y_N, J, \delta_0, \ldots, \delta_N) = \frac{p(Y_1, \ldots, Y_N, \delta, J, \delta_0, \ldots, \delta_N)}{\int p(Y_1, \ldots, Y_N, \delta, J, \delta_0, \ldots, \delta_N) d\delta_1 \ldots d\delta_N} \propto p(Y_1, \ldots, Y_N, \delta | J) p(\delta | \delta_0, \ldots, \delta_N),$$

(18)

where $J = [M_I, L_I, \theta_1, \ldots, M_{17}, L_{17}, \theta_1]$ denotes the parameters of the intensity models at all vertices. As mentioned before, this depends on $M_I, L_I$ and $\theta_1$ explicitly, as we use MAP point estimates from the training stage when fitting to new data. The final segmentation is given by the vector $\delta$ for which this distribution has its maximum. The details of the optimisation procedure are given in the Supplementary material.

Parameters

The most important parameters to be set up before training the model are the shapes and intensities of the prior mean profiles. As will be seen below, the shapes, i.e. the selection of a function from Table 1 along with the specification of any shape parameters, depends mostly on anatomy, whereas good choices for the intensities can be derived from the training data themselves in an automated fashion.

While there are a significant number of parameters to be specified in setting up the priors, the meaning of these parameters is relatively transparent. The problem’s geometry is illustrated by Fig. 1, where the light grey shape represents the structure to be segmented. For purposes of specifying the prior, it is assumed that image intensity $I_m$ inside the structure is close to homogeneous. In this example, the structure of interest has two neighbouring structures with intensities $I_I$ and $I_2$. These intensities are highly acquisition-dependent, but sufficiently good estimates can be found by automatically determining the intensities inside atlas-derived volumetric masks in the training dataset.

In addition to specifying the intensities, suitable functions need to be selected from Table 1. These need to have some resemblance to the profiles actually encountered in the training data, as their role is to determine what point on the intensity profile corresponds to the physical boundary. In cases where the outside intensity is homogeneous in a region that extends beyond the profile length considered (e.g. $I_2$), a step function is a good choice, but in other cases, for example where there is a thin sheet of white matter, an exponential prior may be more appropriate as this forces the boundary to be adjacent to the sheet; a step function would leave the location within the relevant half ambiguous.

In practice, we use a Python script to automatically set up the hyperparameters in a rule-based fashion. This script allows a user of the method to automatically retrain it for any dataset that has similar modalities to those used in this paper. The script uses the median intensities in a number of predefined regions in the user’s images to automatically set the intensity parameters of the priors. A full description of how the priors and other parameters are set up is given in Appendix A. An automated procedure to set the parameters in order to retrain the method on a new dataset is also explained there.

In order to be able to directly compare uni-modal and multi-modal segmentation accuracy with the same method, we also trained the model on only the $T_1$-weighted data with the same parameters that were used for the $T_1$-weighted modality in the multimodal case.

Registration and intensity normalisation

The approach we take in segmenting a structure is to first register the brain to a standard space template volumetrically. The first aim of this step is to remove the pose of the head and any global scaling, which is achieved by the affine registration step. A nonlinear registration step is also run to remove as much between-subject anatomical variability as possible. Note that this is merely to reduce the size of the variations that need to be explained by the shape model of the segmentation method; the final segmentation output will still be in the undistorted subject-native coordinates.

Registration is performed using the FLIRT and FNIRT tools in FSL (Jenkinson et al., 2002; Andersson et al., 2010). The $T_1$-weighted volume is registered to the 2 mm isotropic resolution version of the MNI152 template. Where needed, the other modalities are then registered to the $T_1$ weighted volume using mutual information or, in the case of the diffusion data, using boundary-based registration of the fractional anisotropy (FA) image (Greve and Fischl, 2009). In the 7 T dataset, which will be described below, all modalities were registered to the FLASH volumes instead.

Before fitting the model, the intensity of all volumes for each modality is scaled automatically to have the same mean intensity across subjects in a bounding box around the structure to be segmented. For structures like the caudate nucleus, the amount of cerebrospinal fluid (CSF) inside such a box can vary substantially from subject to subject.
In many contrasts, CSF exhibits either very high or very low intensities and as this might interfere with normalisation, CSF voxels were not included in the calculation of the mean for any of the structures. FAST (Zhang et al., 2001) was used to create the CSF mask. Normalisation was not used for the FA volumes (see below) and for the QSM volumes (see below) an additive normalisation procedure was used instead in order to deal with negative values correctly.

Mesh generation

The reference meshes for our segmentation method were generated from the probabilistic Harvard–Oxford subcortical atlas (supplied with FSL) using the marching cubes algorithm as implemented in VTK (http://www.vtk.org/). The 2 mm isotropic resolution version of the atlas was used, meaning that points on the meshes are approximately 2 mm apart. For the putamen and globus pallidus, the threshold was set at 50%. For the caudate nucleus and nucleus accumbens simple thresholding did not produce a suitable mesh, as the tail of the caudate has relatively low atlas probabilities due to anatomical variation. To compensate for this, voxels were upweighted based on their Euclidean distance from the highest-probability region of the structure. A regularisation and smoothing step implemented in FIRST (Patenaud et al., 2011) was applied to all meshes to produce more regular triangles and smoother surfaces.

We chose to use a single mesh representing the caudate nucleus and nucleus accumbens joined together, because there is little to no visible structure dividing the two in the images we are using. The merged structure was created by taking the sum of the probabilistic maps for the two structures before generating the mesh.

Datasets

The three datasets described below were used to evaluate the method.

Human Connectome Project (HCP80) dataset

The first is the 80 unrelated subjects subset of the Human Connectome Project (HCP80) dataset. All subjects are healthy adults. The following scans were acquired on a Siemens Skyra 3 T system with a 32-channel head coil (Van Essen et al., 2012, 2013):

- The average of two 0.7 mm isotropic resolution $T_1$-weighted MP-RAGE acquisitions with a $320 \times 320 \times 256$ imaging matrix, repetition time (TR) = 2400 ms, echo time (TE) = 2.14 ms, inversion time (TI) = 1000 ms, flip angle = 8° and acceleration factor 2. The acquisition time for each scan was 7:40.
- The average of two 0.7 mm isotropic resolution $T_2$-weighted 3D-SPACE acquisitions with a $320 \times 320 \times 256$ imaging matrix, TR = 3200 ms, TE = 565 ms, variable flip angle and acceleration factor 2. The acquisition time for each scan was 8:24.
- An extensive diffusion-weighted imaging (DWI) acquisition with a $125 \times 125$ mm isotropic resolution, multiband echo-planar imaging (EPI) with a $168 \times 168$ imaging matrix, 111 slices, TR = 5520 ms, TE = 89.5 ms, 6/8 phase partial Fourier and multiband factor 3 (Moeller et al., 2010; Feinberg et al., 2010; Setsompop et al., 2012; Sotiropoulos et al., 2013). All volumes were acquired twice with left–right and right–left phase encoding polarities. The acquisition time was approximately one hour. Only the 90 directions with diffusion weighting $b = 1000 \text{ s/mm}^2$ and 18 volumes with $b = 0$ were used for calculating the FA images.

All of these images were corrected for gradient nonlinearity induced distortions by the HCP pipeline (Glasser et al., 2013). The pipeline additionally corrected the intensities of the $T_1$- and $T_2$-weighted images for $B_0$ inhomogeneities. Geometric distortions due to $B_0$ inhomogeneity and eddy currents were corrected by the pipeline using EDDY and TOPUP, making use of the additional images with reversed phase encoding (Andersson et al., 2003, 2012) Fractional anisotropy (FA) images were calculated using FSL’s DTIFIT.

The HCP80 dataset contains 77 usable subjects, 10 of which were used for experimentation and 10 were manually labelled for the quantitative evaluation of our method. This left a training group of 57 subjects. All results shown in this paper were obtained by training the model on this training group and segmenting the subjects in the labelled group.

7 T dataset

The second set of structural data were the young subjects of the publicly available dataset described by Forstmann et al. (2014). These data were acquired using a 7 T Siemens Magnetom MRI using a 24-channel head array Nova coil (NOVA Medical Inc., Wilmington MA):

- A $T_1$-weighted MP2RAGE slab (Marques et al., 2010) which consisted of 128 slices with an acquisition time of 9:07 min (TR = 5000 ms; TE = 3.71 ms; TI$_1$/TI$_2$ = 900/2750 ms; flip angle = 5°/3°; bandwidth = 240 Hz/Px; voxel size = 0.6 mm isotropic).
- A $T_1$-weighted whole brain MP2RAGE acquisition which had 240 sagittal slices with an acquisition time of 10:57 min (repetition time (TR) = 5000 ms; echo time (TE) = 2.45 ms; inversion times TI$_1$/TI$_2$ = 900/2750 ms; flip angle = 5°/3°; bandwidth = 250 Hz/Px; voxel size = 0.7 mm isotropic).
- A $T_2$-weighted 3D FLASH acquisition. The FLASH slab (Haase et al., 1986) consisted of 128 slices with an acquisition time of 17:18 min (TR = 41 ms and three different echo times (TE): 11.22/20.39/29.57 ms; flip angle = 14°; bandwidth = 160 Hz/Px; voxel size = 0.5 mm isotropic).

The imaging slab for the limited field-of-view acquisition was 64 mm thick for the FLASH protocol and 76.8 mm for the MP2RAGE protocol and positioned to capture the subcortical structures. Both slab sequences consisted of axial slices tilted –23° to the true axial plane in scanner coordinates. The QSM volume was calculated using the phase information of the FLASH MRI sequence and the method proposed by Schweser et al. (2013).

The 7 T dataset consisted of 29 usable subjects; one subject was not used because registration of the whole-brain MP2RAGE volume to the FLASH slab failed. Fourteen of these subjects were male and 15 were female. Their mean age was 23.8 years with a standard deviation of 2.3 years. For quantitative evaluation, a ’leave 5 out’ scheme was used, where per block of 5 or 4 subjects the method was trained on the other 24 or 25 subjects.

Clinical HD dataset

To investigate how well the method performs on data acquired using more typical protocols and in the presence of pathology, it was also applied to the dataset from a Huntington’s disease (HD) study. HD is an informative test case as the disease is associated with severe atrophy of subcortical grey matter. These data and the manual segmentations are described in detail in the original papers (Douaud et al., 2006, 2009) and acquired as part of the MIG-HD (Multicentric Intracerebral Grafting in HD) project.

The dataset consists of ten patients and six control subjects from the original study. The number of control subjects included was limited to six, as this was the only consistent subset of the control data that had both identical acquisition protocols for the structural images and for which diffusion data were available. For each subject the following scans were obtained on a General Electric 1.5 T Signa system with a birdcage head coil with 40 mT/m maximum gradient strength:

- A $T_1$-weighted scan using a 3D inversion recovery fast spoiled gradient recalled (IR-FSPGR) acquisition with 0.9375 mm ×
0.9375 mm × 1.2 mm resolution, matrix size 256 × 256 × 124, TI = 620 ms, TE = 2 ms, TE = 20.3 ms, flip angle = 10° and bandwidth = 7.81 kHz.

• Diffusion-weighted echo-planar imaging (EPI) with matrix size 128 × 128, reconstructed as 256 × 256 to yield 0.9375 × 0.9375 in-plane resolution with 2 mm slice thickness, 60 slices, TE = 71 ms, TR = 2500 ms, flip angle = 90°, bandwidth = 125 kHz, 41 diffusion directions, $b = 700$ s/mm$^2$ and one volume with $b = 0$.

Manual segmentations

7 T dataset

The manual segmentations in the 7 T data were described in Keuken et al. (2014). The putamen, caudate nucleus and nucleus accumbens, along with the islands of grey matter between the putamen and the caudate nucleus and nucleus accumbens, were segmented as a single structure. Because all segmentation methods evaluated in this paper segment the different parts of the striatum separately and none segment the grey matter islands, these islands were removed from the segmentation masks. The internal and external parts of the globus pallidus were labelled using the QSM volumes. As there is a gap between the parts, voxels for which the sum of the Euclidean distances to the internal and the to the external part was smaller than 2.5 mm were added to the combined mask. This effectively closes the gap between the two parts.

HCP80 dataset

Manual segmentation of the striatum in the HCP80 dataset was performed by a single rater using the $T_1$-weighted volumes and the same guidelines as used in Keuken et al. (2014). The globus pallidus was not segmented as this dataset did not include QSM images.

HD dataset

Manual segmentation of the putamen, caudate nucleus and ventral striatum in the HD dataset was described in Douaud et al. (2006). The scans were resliced to have the anterior commissure (AC) and posterior commissure (PC) in the same axial plane. This plane was used as the lower boundary of the globus pallidus in the manual segmentations. In the comparisons between methods for the HD dataset, the automatic segmentations were also cut off in the AC–PC plane to be consistent with the manual segmentation protocol.

Methods comparison

The segmentations produced by MIST, FIRST (Patenaude et al., 2011) and FreeSurfer (Fischl et al., 2002) were quantitatively compared with manual segmentations in all three datasets (HCP80, 7 T and HD). To compare the methods, three different comparisons were performed. The first of these compares the Dice scores between different methods.$^1$ The second comparison looks at the mean distance between the meshes.

$^1$ Dice’s coefficient is defined as $\frac{2NA}{NB + NA}$, where $NA$ and $NB$ are the number of voxels in masks $A$ and $B$, and $NA\cap B$ is the number of voxels in the overlapping region (Dice, 1945).
produced by the different methods. Finally, the third comparison compares the correlation coefficients between the automatically segmented mask volumes and manual segmentations. The general procedure used for this comparison is described here, as well as a number of processing steps that are specific to one or two of the datasets.

The mesh-based output from MIST was converted to voxel-based masks in order to be able to compare to the manual segmentation. This was achieved by testing for all voxels' vertices whether they were inside the mesh. A voxel was considered to be inside the mesh if at least one of its vertices was inside. The vtkSelectEnclosedPoints filter in VTK was used for this procedure.

In order to be able to compute the average distances between meshes, the manual segmentations and FreeSurfer segmentations were converted to meshes using the vtkMarchingCubes filter. Where meshes needed to be combined for the comparison, such as for the caudate and accumbens meshes from FIRST, only points that are at least 2 mm from the other mesh were used in the distance calculation to ensure that no internal boundaries were present.

In the HCP80 and 7 T datasets, results were compared to those obtained using FIRST with and without its boundary correction post-processing step. This step uses a voxel-based mixture model to reassess which voxels on the boundary of a structure should and which should not be part of the structure. In the 7 T dataset, the segmentations also needed to be registered to the scans on which the manual segmentations were performed. The partial FOV MP2RAGE scans for striatum and the QSM volume for the globus pallidus. All segmentation methods subdivide the striatum into multiple structures; the masks for the putamen, caudate nucleus and nucleus accumbens were combined into a single striatum mask.

In the HCP80 dataset, the FreeSurfer segmentations that are provided with the pre-processed structural datasets by the Human Connectome Project were used. FIRST was run with its default options on the $T_1$-weighted volume. The segmentations produced by multi-modal segmentation were also output in the coordinates of this volume.

In the 7 T dataset, FreeSurfer was run on the brain-extracted whole-brain MP2RAGE volume as whole brain coverage is a requirement of its processing pipeline. The segmentation masks were linearly upsampled to match the full resolution MP2RAGE volume and registered to both the partial FOV MP2RAGE and FLASH volumes as the manual segmentations were performed on these images. FIRST was run on both the whole-brain and partial FOV MP2RAGE volumes and the resulting segmentations were registered with the partial FOV MP2RAGE and FLASH images. In the segmentations using MIST, the partial FOV MP2RAGE volumes were used.

**Results**

*Automatic edge alignment*

An important question is whether the profile model, that is fitted for each vertex separately, is successful at finding a mean profile that...
captures the local intensity features and that is properly aligned with the physical boundary of the structure. Fig. 2 shows the unaligned profiles that were sampled for the different training subjects at a single vertex on the inferior boundary of the putamen. It also displays the prior and posterior means and the training profiles after they have been aligned to the posterior mean (i.e. at their most probable displacement $\delta_0$). It shows that, while the unaligned profiles are not closely aligned with each other, the aligned ones overlap much better; this is particularly clear for the FA profiles. In addition, the step in intensity is now located at the centre of the reference profile, which means that it is aligned with the prior that defines the physical boundary.

**Segmentation results**

The full set of segmented structures in an example subject is shown in Fig. 3. In this figure there are some clear examples of areas where multiple contrasts can complement each other. One example is the globus pallidus, which is essentially invisible on the $T_1$-weighted image. Another example is the lateral boundary of the putamen, which, although there is contrast in the $T_1$- and $T_2$-weighted images, is difficult to see or detect due to the resolution of the volume. In the FA image, however, the boundary is much clearer. Further examples of segmentations in different subjects in the HCP80 dataset can be found in Figs. S1 (putamen), S2 (caudate nucleus and nucleus accumbens) and S3 (globus pallidus) in the Supplementary material. A comparison with the meshes produced by FIRST is given in Fig. S4.

The advantage of multimodal segmentation for segmenting the globus pallidus can also be seen in the 7 T dataset. Fig. 4 shows the results obtained using the full multimodal model, as described in Table A.2, and a model using only the $T_1$-weighted volume. The lateral boundary, which is clearly visible on the MP2RAGE scan, is accurately segmented in both cases, whereas the other boundaries improve with...

---

**Comparison between multimodal and T1-based segmentation**

<table>
<thead>
<tr>
<th>Subject 1</th>
<th>Subject 2</th>
<th>Subject 3</th>
<th>Subject 4</th>
<th>Subject 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Axial</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QSM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coronal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QSM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 4.** Segmentation results for the globus pallidus in the 7 T dataset using all three modalities (red) and $T_1$-weighted only (green). Top three rows: axial slices, bottom three rows: coronal slices. The QSM volume was not used for segmenting the caudate nucleus and nucleus accumbens.
the inclusion of the other modalities. Examples of all three structures that are segmented are given in Fig. S5 in the Supplementary material.

Voxel-based overlap with other methods

To quantitatively assess the performance of MIST, the results were compared with manually labelled scans. The Dice overlap scores for segmentation of the striatum in the HCP80 dataset using MIST, FIRST and FreeSurfer are shown in Fig. 5. The scores show that our method produces accurate segmentations in the HCP80 dataset with overlaps that are larger than those obtained with FIRST and FreeSurfer. Multimodal segmentation using MIST produces significantly more accurate results than segmentation using only the $T_1$-weighted volume. The picture for the mesh-based distances is fairly similar, although the difference between MIST and FIRST is not significant in this case. Mesh distance is more sensitive than the Dice score to differences between segmentation in areas such as the tail of the caudate, a thin and elongated part of the structure.

In the comparison of volume correlations it is remarkable that boundary-corrected FIRST and FreeSurfer fare better than in the overlap- and distance-based comparisons. The advantage of comparing methods in terms of these correlations is that they show whether a method produces larger segmentations in subjects where a structure is larger, without being sensitive to the exact handling of the boundaries. In this case, the correlations indicate that FIRST and FreeSurfer capture the variability in striatal volume between subjects, despite the imperfections in boundary placement suggested by the Dice scores.

Examples of the voxel-based masks (as used to calculate the Dice scores) generated by the different methods are shown in Fig. 6.

Methods comparison in HCP80 dataset

Fig. 5. Dice overlap (first column) with manual segmentations and mean mesh distance in mm (second column) of segmentations produced by different methods in HCP 80 dataset (10 subjects). Correlation coefficients between manual and automatic mask volumes are shown in the third column. MIST: multimodal segmentation, $T_1$: MIST with $T_1$-weighted images only, Ft−/+: FIRST without and with boundary correction, FS: FreeSurfer. Data points that are outside the box by more than 1.5 times the interquartile range are treated as outliers. A significant difference in performance between a method and MIST is denoted by an asterisk ($p \leq 0.05$, Wilcoxon signed rank test for the boxplots and Williams’s test for the correlation coefficients. Computed using R, http://www.r-project.org/).

Fig. 6. Example masks for the striatum using different methods in an example subject (499566) from the HCP80 dataset. Green: manual labelling, red: automatic segmentation, yellow: overlap. BC denotes boundary correction.
the sense that the difference between the lowest and highest scores is smaller. In the case of the globus pallidus, the improvement over the existing methods is even clearer than for the striatum. The correlation plots indicate that while FIRST and FreeSurfer show reasonable overlap with the manual segmentations, they do not accurately capture the differences in volume between subjects of the globus pallidus (in the left hemisphere only, in the case of FreeSurfer). For the globus pallidus, multimodal segmentation using MIST produces more accurate segmentations than segmentation using only the $T_1$-weighted volume, whereas in the case of the striatum, which has clear contrast on a $T_1$-weighted scan, MIST performs slightly better in the $T_1$-only case.

**Methods comparison in 7T dataset**

Fig. 7. Dice overlap (first column) with manual segmentations and mean mesh distance in mm (second column) of segmentations produced by different methods in the 7 T dataset (29 subjects). Correlation coefficients between manual and automatic mask volumes are shown in the third column. MIST: multimodal segmentation, $T_1$: MIST with $T_1$-weighted images only, $FtS−/+$: FIRST without and with boundary correction on limited FOV MP2RAGE data, $FtW−/+$: FIRST without and with boundary correction on whole brain MP2RAGE data. Data points that are outside the box by more than 1.5 times the interquartile range are treated as outliers. A significant difference in performance between a method and MIST is denoted by an asterisk ($p \leq 0.05$).
Example segmentations in a patient in the HD dataset are shown in Fig. 8. This patient has significant atrophy of the subcortical GM. Despite this, MIST is successful in segmenting the different structures.

Fig. 9 shows the metrics of the segmentation performance of the different methods in the HD dataset. For the globus pallidus, MIST and FIRST perform comparably in terms of overlap and mesh distances and MIST performs slightly better than FreeSurfer (though statistically significant). In the comparison of volume correlations, MIST and FreeSurfer are comparable and FIRST shows significantly lower correlations.

For the striatum, Dice scores for MIST and FIRST are again higher (and mesh distances lower) than for FreeSurfer. The volumes of the FIRST segmentations of the left striatum correlate more weakly with the manual labellings than the volumes obtained with the other methods. This appears to be caused by two subjects in which FIRST performed poorly when segmenting this structure (see the Dice scores). This did not occur for the right striatum and this may explain the higher correlation in this case.

Example segmentations using MIST in clinical HD dataset

![Example segmentations using MIST in clinical HD dataset](image)

**Fig. 8.** Example patient in the HD dataset showing segmentations of putamen (red), globus pallidus (green) and caudate + nucleus accumbens (blue) on axial (top row) and coronal (bottom row) slices.

**General observations in the comparison between methods**

The three measures of performance (Dice overlap, mean mesh distance and volume correlation) are sensitive to different aspects of the segmentations produced by the automatic methods, as illustrated above. MIST performs well in terms of all three measures in all datasets, while both FIRST and FreeSurfer are less consistent in this respect.

The methods comparison shows that FIRST achieves reasonable Dice scores for the globus pallidus in the 7 T dataset (above 0.7, Fig. 7), although the low volume correlations indicate that it is not successful in capturing between-subject variability. A possible explanation is that in the absence of image contrast, the segmentations are almost exclusively driven by prior knowledge. In this situation, there could still be substantial overlap of the segmentations with the globus pallidus due to the similarity of its appearance between subjects, but the variations themselves would not be captured by the segmentations. This would result in the low volume correlations. FIRST performs significantly better in the HD dataset, where contrast for the globus pallidus is much better in the T1-weighted volume.

The degree of asymmetry between left and right hemisphere results is remarkable. Volume correlations for the right striatum in the HCP80
dataset are significantly lower when using MIST with just the T1-weighted data compared to the multimodal case, while they appear similar for the left striatum (Fig. 5). As the testing dataset is relatively small (10 subjects), it is difficult to know whether this represents a real asymmetry in performance. A similar situation exists for the difference in the striatal volume correlation for FIRST in the HD dataset, which is significantly lower than for MIST in the left hemisphere only (Fig. 9). FIRST seems to perform poorly in terms of Dice scores in two subjects in the left striatum and it is hard to tell if the fact that no such outliers occur in the right hemisphere is caused by an actual asymmetry in the segmentation process or simply due to chance. A more convincing example of asymmetry is the performance of FreeSurfer in the globus pallidus in the 7 T dataset, which is substantially better in the right hemisphere (Fig. 7). We do not see an obvious explanation for this asymmetry in FreeSurfer’s performance.

**Importance of the size of the training set**

To investigate how sensitive segmentation performance is to the size of the training set, the striatum models for the HCP80 data were trained using a number of reduced-size subsets of the full training set. The results from this comparison are shown in Fig. 10. As expected,
Segmentations become less accurate with smaller training datasets, although performance in terms of these summary measures is still very reasonable even with the smallest training set. The reason for this behaviour is illustrated by Fig. 11. In this figure it can be seen that segmentation is successful for large parts of the structure even with very little training data by mostly relying on the intensity priors. In regions where anatomy is more complex, though, such as near the external capsule, learning the intensity profiles from the training data has a clear benefit. Another area where segmentation improves with more training area is the anterior end of the putamen. This may be caused by the islands of grey matter that are present between the putamen and caudate nucleus and whose appearance differs significantly between subjects and scans.

Contribution of non-linear registration

To investigate how large the relative contributions of non-linear registration and the subsequent segmentation steps are, Fig. 12 compares the result of the full segmentation procedure in the HCPR0 dataset to the results obtained by applying non-linear registration to the reference shape without performing segmentation, i.e. setting all the displacements to zero. Non-linear registration produces a reasonable approximation to the structure, but the full segmentation procedure is required to obtain a high quality delineation.

Discussion

Segmentation results

The results in the previous section show that MIST can produce high-quality segmentations of the striatum and globus pallidus. The successful segmentation of the globus pallidus, which has poor contrast in a T1-weighted scan, illustrates the advantage of multi-modal segmentation. Although both the T2-weighted and FA contrasts provide more information on the boundaries of the structure, neither of these shows all aspects. The combination of the three images, however, allows the

Fig. 10. Dice overlap (first column) with manual segmentations and mean mesh distance in mm (second column) of segmentations produced after training on subsets of different sizes in the HCPR0 dataset. Correlation coefficients between manual and automatic mask volumes are shown in the third column. Subsets of the full training set of 57 subjects were used to investigate how segmentation performance changes with smaller numbers of training subjects. A significant difference in performance between a subset and the full training set is denoted by an asterisk ($p \leq 0.05$).

Fig. 11. Segmentation of the putamen in an example subject in the HCPR0 dataset for different numbers of training subjects. Red outline: automatic segmentation, blue overlay: manual labelling.
method to successfully segment the entire structure. This puts it at an advantage over FIRST and FreeSurfer, which only use a \( T_1 \)-weighted scan, and there is indeed an improvement in segmentation accuracy of the globus pallidus over these methods. MIST also produced good results in the 1.5 T HD dataset, confirming that the method works well with more typical clinical acquisitions as well. Contrast in the results in the 1.5 T HD dataset, contrast in the globus pallidus over these methods. MIST also produced good overlap and mesh distances. This confirms that the method can handle more significant differences in anatomy than those that are typically present in healthy subjects.

Of the segmentations produced by the different methods that were evaluated, those produced by MIST showed the strongest volume correlations with the manual segmentations. This is an important result, as these correlations are probably more descriptive of the relevant properties of a method than the overlap and distance measures. Overlap and distance are very sensitive to issues of subjectivity, such as the preference of either a person doing manual segmentation or an automated method to include more or fewer boundary voxels. While such differences are important in quantifying the exact volume of a structure, they are not very relevant at all when the volumes are compared between groups and the question is just whether the volumes differ. A second scenario where correlation is the more relevant measure is when the numbers are to be used as predictor for some other measure, for example when using the volumes as a regressor in a general linear model.

MIST produced more consistent results than the other methods, especially in the HCP80 and 7 T datasets. It is difficult to make a definite statement about the cause of the difference, as the methodology underlying the three methods differs substantially. The larger number of images available to the multi-modal method is a clear advantage and this should make it more resilient to image noise. MIST was trained on images from the same population and with the same acquisition protocols as the images to be segmented. This is also likely to be advantageous, as this means that the trained model is more appropriate for the images that are segmented. Finally, considering that the largest improvements in segmentation quality and consistency were obtained in the HCP80 and 7 T dataset, it seems likely that the fact that the new method can more easily take advantage of the high quality data also contributes to better consistency.

MIST also performed well when used on only a \( T_1 \)-weighted image, in particular in the 7 T dataset. For the striatum, which is clearly visible on a \( T_1 \)-weighted scan, this may not be completely unexpected. Nevertheless, it confirms that the intensity model can reliably identify edges in the training data and that it is at least as sensitive to image information as the methods to which we compared. Fig. 4 shows that for the globus pallidus, the additional contrast is advantageous in producing good-quality segmentations. In Fig. 7, MIST performs slightly better on segmentation of the striatum when using only the \( T_1 \)-weighted data. The differences in the masks produced by multimodal and \( T_1 \)-only segmentation are rather subtle and it is difficult to conclusively determine what causes this difference. The effect of partial volume averaging and data smoothness will be slightly different in the multimodal case and due to the fact that there is more intensity data, the relative weighting of the shape prior will be somewhat lower. It is likely that the difference is caused by a combination of these effects. Both multimodal and \( T_1 \)-only MIST show better consistency than the other methods. It should also be kept in mind that the manual segmentations were done on the \( T_1 \)-weighted data, meaning that they have not taken the additional information in other modalities into account.

The DWI protocol used in the HCP dataset has a total acquisition time of about an hour, but it is useful to note that for purposes of subcortical segmentation, high angular resolution and multiple \( b \)-values are not required and acquisition could be much quicker, as long as the spatial resolution is relatively high. It should also be noted that while the
acquisition protocol for the $T_1$-weighted scan in the HCP data has been optimised heavily for cortical contrast, subcortical contrast is significantly poorer.

Parameter setup

Using the script that is described in Appendix A, MIST can be retrained fully automatically on any dataset that contains one or more of the contrasts that were used in this paper or similar contrasts. If the dataset contains fundamentally different contrasts, the user can add priors for these modalities to the generated setup. The fact that the method can be easily retrained eliminates bias that might otherwise arise when the model is applied to a population that is different from the one it was trained on. This could be an important advantage in patient studies, where brain anatomy could be substantially different from the healthy population. If desired, a completely unbiased result could be obtained by creating a reference mesh from the study average image; this would require a manual labelling of only that image. Note that, while in this paper we used separate training and evaluation groups, in applications there would be no problem with training on a balanced and representative subgroup of the study subjects.

A suitable specification of the priors is essential in order to train the model to achieve good quality segmentations. This specification should reflect our a priori belief about the shape and amplitude of an intensity profile, which is determined by two factors: Anatomical variability and image properties. The first of these two, the anatomy and the anatomical differences between subjects, are what dictates the choice of function to be used as a prior (see Table 1). In the case where we expect only a single image edge to be present in the measured profiles, specifying the prior profile would be simple, as a simple step function (i.e. $f_{step}$ in Table 1) with the intensity inside the structure as the first parameter and the outside intensity as the second parameter represents this situation accurately. In reality, however, there are likely to be additional image features that we are not interested in, such as edges farther away from the boundary of the structure of interest. The advantage of the exponential shape, which we have mostly used, is that it contains one sharp edge, which helps alignment, but not any other edges that might cause the method to unintentionally align to other structures. The advantage over the step function, which also contains only one edge, is that the amplitude returns to the value for the inside (left) section of the profile. This is beneficial in assigning profiles to a mixture component, as it reduces the influence of intensity differences that are not related to the edge feature.

The second factor in specifying the priors is finding an initial estimate of the image intensities. These intensities can vary significantly between acquisitions and scanners and we have used the automated setup script described in Appendix A to find these for the different datasets presented in this paper. The script uses rules based on the intensities inside different structures to find suitable values to be used in the prior specifications. This should automate the retraining process for the striatum and globus pallidus in datasets that have the types of contrasts as used here. Fundamentally different contrasts (i.e. other than $T_1$, $T_2$-weighted, FA or QSM) that show different aspects of structures will require the user to manually specify the shapes and intensities of the prior profiles for those contrasts. To segment a completely new structure (not the striatum or globus pallidus), the user will need to provide a reference mesh and specify the profile priors manually (or extend the Python setup script).

General considerations

By considering displacements along the shape normals only, we were able to formulate a model that is relatively straightforward and in which the intensity models factorise if the shape model is not taken into account. As mentioned in the introduction, this type of deformation was chosen because most of the intensity information that is relevant to find the location of the structural boundary is along the normals. The results we obtained suggest that with an appropriately constructed reference shape, this parameterisation is flexible enough to capture the anatomical variability that is present.

The meshes that we have used are relatively regular in terms of point spacing and triangle areas. We have chosen to do this, as we want the model to be able to capture small and unique variations like vessels with the same resolution anywhere along the surface. It is conceivable that in areas of high curvature a higher mesh resolution would be advantageous, but for the structures that we segment in this paper we did not see any resolution-related problems with the current meshes. Using lower resolutions for parts of the mesh would compromise our aim of being able to capture small variations anywhere along the mesh and instead impose a degree of regularity.

The primary purpose of the shape model is to regularise segmentations, as the displacements that are determined for all vertices independently can be noisy in areas of weaker contrast. To overcome this problem, the shape model learns the mean displacement at each vertex and the covariance between vertices. This regularises the segmentations by only generating large displacements with respect to the mean learned from the training data if the images provide strong enough evidence in terms of edge contrast. It also favours segmentations that respect the covariance between displacements observed in the training data and the smoothness specified by the covariance prior.

For further statistical analysis, there are two forms of output that can be used. The first option is to use the vertex-wise displacements. These are generated in MNI coordinates, but as the non-linear transformation from native to MNI coordinates already includes some of the shape variability, the best option appears to be to undo the non-linear part of the transformation, but not the affine part. This means that global scaling factors, which primarily represent head size, are not included in the displacements, but that all local deformations are included. Standard multivariate analysis tools can then be used for inference. A second option is to use the voxel-based output for further analysis. This can be a good approach to get only a scalar volume for each structure, or to get a mask that can be used to define a region of interest (ROI) for further analysis.

We chose to treat the caudate nucleus and the nucleus accumbens as a single structure. There is virtually no contrast between these structures in the images that were used here. Because of this, a subdivision would not be driven by the data, but rather it would be based solely on some form of prior knowledge. In the case of a shape-based analysis using the vertex-wise displacements, this may not be a problem, as different patterns of variation within the merged structure can still be represented. In the case of a voxel-based further analysis, we propose to use a subdivision based on the MNI coordinates if this is required.

The runtime of the training stage of the method depends on many factors, the most important of which are the size of the training dataset and the number of modalities. Evaluation time of the cost function scales linearly in both factors, although total runtime might deviate from this linear behaviour due to the non-linear optimisation. Training the method on the structures and datasets presented in this paper took between an hour and half a day of processor time on recent PC hardware. In practice, we have parallelised the training stage, which means the wall time spent is much lower. Segmenting a new image after training takes under a minute on a single processor.

In this paper, we have described a flexible new method for subcortical segmentation. It can successfully combine information from multiple modalities and can be retrained for datasets with different characteristics. We have applied the method to the striatum and globus pallidus, but the framework is more general and can be applied to other structures. Retraining is automatic for the structures and contrast types considered in this paper. The method can be applied to new structures and contrasts by the user through the specification of suitable priors and a reference mesh in the case of new structures. The implementation of the method that was used to produce the results in this paper will be...
made publicly available through inclusion in a future release of FSL (http://fsl.fmrib.ox.ac.uk) and will be called MIST.

**Acknowledgments**

Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH institutes and centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

We wish to thank the Wellcome Trust for funding through the Strategic Award for “Integrated multimodal brain imaging for neuroscience research and clinical practice” (grant number 098369/Z/12/Z).

This work is supported by the Medical Research Council UK (MRC) MR/K006673/1 (to G.D.).

The Huntington’s dataset was acquired as part of the MIG-HD program and was supported by the Délegation Régionale à la Recherche Clinique (PHRC P001106), the Association Française contre les Myopathies and the Établissement Français des Greffes.

This work was supported by an ERC starter grant (B.U.F., SPEED 313481).

The plots in this paper were generated using matplotlib (Hunter, 2007).

**Appendix A. Parameter setup**

In order to make it as straightforward as possible for users of our method to retrain the models on new datasets, we use a script that automatically sets all the parameters by combining prior anatomical knowledge with properties of the training images. This script was also used to set the parameters for all datasets presented in this paper.

Two pieces of prior anatomical knowledge are required by the model: A reference mesh, the generation of which was described in the main text, and the specification of the profile priors. While the intensities of the priors depend strongly on the acquisition, their shapes and the mediants is the value filled in into the templated defined by Table A.2. The same is true for the different T₁-weighted scans.

The prescriptions for the profile priors that were used for all datasets in this paper are displayed in Table A.2. The shapes of the functions have been specified manually through considering the anatomy of the structures’ boundaries and some experimentation.

For the T₁ and T₂ modalities in the putamen, priors are specified for two cases: The case where the neighbouring structure at a vertex is either the globus pallidus or WM and the case where there is a vessel instead. For the first case, two exponentially decaying priors are specified to better capture the range of profiles that may be observed. For example, on the lateral side of the putamen there is the external capsule, a thin sheet of WM, while the globus pallidus on the medial side is a much thicker structure. The diameter of vessels is typically small, which is why the prior that is specified to handle them decays quickly. We have used a single outside intensity for the globus pallidus and

<table>
<thead>
<tr>
<th>Table A.2</th>
<th>Table A.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modality</td>
<td>Prior standard deviation of $f$</td>
</tr>
<tr>
<td>$T_1$</td>
<td>$0.1$</td>
</tr>
<tr>
<td>$T_2$</td>
<td>$0.1$</td>
</tr>
<tr>
<td>FA</td>
<td>$0.1$</td>
</tr>
<tr>
<td>QSM</td>
<td>$0.02$</td>
</tr>
</tbody>
</table>

The intensities are found automatically by the script from the user’s dataset by extracting the intensities inside regions of interest (ROIs). The ROI masks are derived from the Harvard–Oxford subcortical atlas that is provided with FSL. The median of all values with above 75% probability in the atlas is extracted from each subject and the mean of these medians is the value filled in into the templated defined by Table A.2. It should be noted that, as the intensity setup is derived from the image, each component in the table can be used with a range of similar contrasts. In this paper, for example, the $T_2$- and $T_2^*$-weighted images in the different datasets have quite different appearances, yet they can all use the $T_2$ setup from the table. The same is true for the different $T_1$-weighted scans.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Modality</th>
<th>Type</th>
<th>Inside</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>Putamen</td>
<td>$T_1$</td>
<td>Exponential (1 and 4 mm)</td>
<td>Self</td>
<td>Average of GP and WM</td>
</tr>
<tr>
<td></td>
<td>$T_2$</td>
<td>Exponential (1 mm)</td>
<td>Self</td>
<td>Average of GP and WM</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>Exponential (4 mm)</td>
<td>0.1</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exponential (1 mm)</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>QSM</td>
<td>Exponential (2 mm)</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Flat</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Globus pallidus</td>
<td>$T_1$</td>
<td>Exponential (3 mm)</td>
<td>Self</td>
<td>Putamen</td>
</tr>
<tr>
<td></td>
<td>$T_2$</td>
<td>Exponential (3 mm)</td>
<td>Self</td>
<td>WM (\times 3)</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>Step</td>
<td>0.2</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>QSM</td>
<td>Step</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Caudate and accumbens</td>
<td>$T_1$</td>
<td>Step and exp. (1 and 4 mm)</td>
<td>Self</td>
<td>Ventricles (\times 1.5)</td>
</tr>
<tr>
<td></td>
<td>$T_2$</td>
<td>Step and exp. (1 mm)</td>
<td>Self</td>
<td>WM (\times 0.67)</td>
</tr>
</tbody>
</table>

In this section we describe the template used by the automatic setup script to generate a full set of parameters that should apply to any dataset that contains similar modalities to the ones described here, or a subset of them.
WM as the intensity changes with respect to the putamen are the same for both (i.e. higher on T1-weighted and lower on T2-weighted volumes).

For the FA and QSM modalities, wide and narrow priors are specified for the same reasons as in the T1 and T2 modalities. As vessels are not as clear on these modalities, no vessel prior is included for these modalities. The QSM prior with a positive edge is included for the boundary with the globus pallidus, whereas the negative edge should handle boundaries with WM. The intensities for the FA and QSM modalities are not derived from the images as the former has a fixed range and the latter is quantitative.2

2 Although the QSM contrast is quantitative, the QSM data used in this paper use arbitrary units and as such do not have physical units.

The parameters w and h control the prior variance and covariance between vertices. They can be used to control the smoothness of the final segmentation. It is important to note however, that this smoothness concerns the replacements with respect to the mean shape and as we train all vertices in isolation, there is no guarantee that this mean will be smooth. The parameters are useful to regularise the model in the case of noisy data, however. The final segmentation result is not very sensitive to changes of w and h around their present values. Low values of w and high values of h reduce the regularisation effect, while low values of h constrain the mesh to the mean. Very high values of w may interfere with non-linear optimisation as this causes the parameters to be highly collinear. As the meshes used for the different structures in this paper have similar numbers of vertices per unit of surface, we expect similar smoothness and the same values for w and h are used for all of them. Weighting of the priors relative to the training data is controlled by the parameters n0, α, and n0. We have set these close to their minimum values as there is no obvious advantage to increasing these weights. One reason for setting higher values might be if stronger regularisation using w and h is desired. Finally, the parameter α determines whether the intensity model will prefer mixtures where one component explains all observations versus mixtures with more equal mixing. The current setting gives a slight preference to the first case, although in practice the effect of this parameter appears limited.

Appendix B. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2015.10.013.

References
