Brain Extraction from Normal and Pathological Images: A Joint PCA/Image-Reconstruction Approach

Xu Han[a,*], Roland Kwitt[b], Stephen Aylward[c], Spyridon Bakas[d], Bjoern Menze[e], Alexander Asturias[f], Paul Vespa[g], John Van Horn[f], Marc Niethammer[a]

[a]*Department of Computer Science, University of North Carolina at Chapel Hill, USA*
[b]*Department of Computer Science, University of Salzburg, Austria*
[c]*Kitware Inc., USA*
[d]*Center for Biomedical Image Computing and Analytics, Perelman School of Medicine, University of Pennsylvania, USA*
[e]*Department of Computer Science, Technical University of Munich, Germany*
[f]*Institute of Neuroimaging and Informatics, University of Southern California, USA*
[g]*David Geffen School of Medicine, UCLA Medical Center, USA*

Abstract

Brain extraction from 3D medical images is a common pre-processing step. A variety of approaches exist, but they are frequently only designed to perform brain extraction from images *without* strong pathologies. Extracting the brain from images exhibiting strong pathologies, for example, the presence of a brain tumor or of a traumatic brain injury (TBI), is challenging. In such cases, tissue appearance may substantially deviate from normal tissue appearance and hence violates algorithmic assumptions for standard approaches to brain extraction; consequently, the brain may not be correctly extracted.

This paper proposes a brain extraction approach which can explicitly account for pathologies by jointly modeling normal tissue appearance and pathologies. Specifically, our model uses a three-part image decomposition: (1) normal tissue appearance is captured by principal component analysis (PCA), (2) pathologies are captured via a total variation term, and (3) the skull and surrounding tissue is captured by a sparsity term. Due to its convexity, the resulting decomposition model allows for efficient optimization. Decomposition and image registration steps are alternated to allow statistical modeling of normal tissue appearance in a fixed atlas coordinate system. As a beneficial side effect, the decomposition model allows for the identification of potentially pathological areas and the reconstruction of a quasi-normal image in atlas space.

We demonstrate the effectiveness of our approach on four datasets: the publicly available IBSR and LPBA40 datasets which show normal image appearance, the BRATS dataset containing images with brain tumors, and a dataset containing clinical TBI images. We compare the performance with other popular brain extraction models: ROBEX, BEaST, MASS, BET, BSE and a recently proposed deep learning approach. Our model performs better than these competing approaches on all four datasets. Specifically, our model achieves the best median (97.11) and mean (96.88) Dice scores over all datasets. The two best performing competitors, ROBEX and MASS, achieve scores of 96.23/95.62 and 96.67/94.25 respectively. Hence, our approach is an effective method for high quality brain extraction for a wide variety of images.

Keywords: Brain Extraction, Image Registration, PCA, Total-Variation, Pathology

1. Introduction

Brain extraction[1] from volumetric magnetic resonance (MR) or computed tomography images [1] is a common pre-processing step in neuroimaging as it allows to spatially focus further analyses on the areas of interest. The most straightforward approach to brain extraction is by manual expert delineation. Unfortunately, such expert segmentations are time consuming and very labor intensive and therefore not suitable for large-scale imaging studies. Moreover, brain extraction is complicated by differences in image acquisitions and the presence of tumors and other pathologies that add to inter-expert segmentation variations.

Many methods have been proposed to replace manual delineation by automatic brain extraction. In this paper, we focus on and compare with the following six widely-used or recently published brain extraction methods, which cover a wide range of existing approaches:

- *Brain Extraction Tool (BET):* BET [2] is part of FMRIB Software Library (FSL) [3, 4] and is a widely

*Corresponding author
Email address: xhs400@cs.unc.edu (Xu Han)

[1]We avoid the commonly used term skull stripping. We are typically interested in removing more than the skull from an image and are instead interested only in retaining the parts of an image corresponding to the brain.

used method for brain extraction. BET first finds a rough threshold based on the image intensity histogram, which is then used to estimate the center-of-gravity (COG) of the brain. Subsequently, BET extracts the brain boundary via a surface evolution approach, starting from a sphere centered at the estimated COG.

- *Brain Surface Extraction (BSE):* BSE [5] is part of BrainSuite [6, 7]. BSE uses a sequence of low-level operations to isolate and classify brain tissue within T1-weighted MR images. Specifically, BSE uses a combination of diffusion filtering, edge detection and morphological operations to segment the brain. BrainSuite provides a user interface which allows for human interaction. Hence better performance may be obtained by interactive use of BSE. However, our objective was to test algorithm behavior for a fixed setting across a number of different datasets.

- *Robust Learning-based Brain Extraction System (ROBEX):* ROBEX [8, 9] is another widely used method which uses a random forest classifier as the discriminative model to detect the boundary between the brain and surrounding tissue. It then uses an active shape model to obtain a plausible result. While a modification of ROBEX for images with brain tumors has been proposed [10], its implementation is not available. Hence we use the standard ROBEX implementation for all our tests.

- *Deep Brain Extraction:* We additionally compare against a recently proposed deep learning approach for brain extraction [11, 12] which uses a 3D convolutional neural network (CNN) trained on normal images and images with mild pathologies. Specifically, it is trained on the IBSR v2.0[2] [13], LPBA40 [14, 15] and OASIS [16, 17] datasets. We use this model as is without additional fine-tuning for other datasets.

- *Brain Extraction Based on non-local Segmentation Technique (BEaST):* BEaST [18, 19] is another recently proposed method, which is inspired by patch-based segmentation. In particular, it identifies brain patches by assessing candidate patches based on their sum-of-squared-difference (SSD) distance to known brain patches. BEaST allows using different image libraries to guide the brain extraction.

- *Multi-Atlas Skull Stripping (MASS):* MASS [20], uses multi-atlas registration and label fusion for brain extraction. It has shown excellent performance on normal (IBSR, LPBA40) and close to normal (OASIS) image datasets. One of its main disadvantages is its runtime. An advantage of MASS, responsible for its performance and robustness, is that one can easily make use of dataset-specific brain templates. However, this requires obtaining such brain masks via costly manual segmentation. For a fair comparison to all other methods, and to test the performance of a given algorithm across a wide variety of datasets, we select 15 anonymized templates for MASS's multi-atlas registration. These templates were obtained from various studies and are provided along with the MASS software package [21], as well as through CBICA's Image Processing Portal [22].

In addition to these methods, many other approaches have been proposed. For example, Segonne et al. [23] proposed a hybrid approach which combines watershed segmentation with a deformable surface model. Watershed segmentation is used to obtain an initial estimate of the brain region which is then refined via a surface evolution process. 3dSkullStrip is part of the AFNI (Analysis of Functional Neuro Images) package [24, 25]. It is a modified version of BET. In contrast to BET, it uses image data inside and outside the brain during the surface evolution to avoid segmenting the eyes and the ventricles.

Even though all these brain extraction methods exist and are regularly used, a number of challenges for automatic brain extraction remain:

- Many methods show varying performances on different datasets due to differences in image acquisition (e.g., slightly different sequences or differing voxel sizes). Hence, a method which can reliably extract the brain from images acquired with a variety of different imaging protocols would be desirable.

- Most methods only work for images which appear normal or show very minor pathologies. Strong pathologies, however, may induce strong brain deformations or strong localized changes in image appearance, which can impact brain extraction. For example, for methods based on registration, the accuracy of brain extraction will depend on the accuracy of the registration, which can be severely affected in the presence of pathologies. Hence, a brain extraction method which works reliably even in the presence of pathologies (such as brain tumors or traumatic brain injuries) would be desirable.

Inspired by the low-rank + sparse (LRS) image registration framework proposed by Liu et al. [26] and our prior work on image registration in the presence of pathologies [27], we propose a brain extraction approach which can tolerate image pathologies (by explicitly modeling them) while retaining excellent brain extraction performance in the absence of pathologies.

The contributions of our work are as follows:

- *(Robust) brain extraction:* Our method can reliably extract the brain from a wide variety of images. We

[2]This is a different dataset than the IBSR dataset that we use in this paper.

achieve state-of-the-art results on images with normal appearance, slight, and strong pathologies. Hence our method is a generic brain extraction approach.

- *Pathology identification:* Our method captures pathologies via a total variation term in the decomposition model.
- *Quasi-normal estimation:* Our model allows the reconstruction of a quasi-normal image, which has the appearance of a corresponding pathology-free or pathology-reduced image. This quasi-normal image also allows for accurate registrations to, e.g., a normal atlas.
- *Extensive validation:* We extensively validate our approach on four different datasets, two of which exhibit strong pathologies. We demonstrate that our method achieves state-of-the-art results on all these datasets using a *single* fixed parameter setting.
- *Open source:* Our approach is available as open-source software.

The remainder of the paper is organized as follows. Section 2 introduces the datasets that we use and discusses our proposed model, including the pre-processing, the decomposition and registration, and the post-processing procedures. Section 3 presents experimental results on 3D MRI datasets demonstrating that our method consistently performs better than BET, BSE, ROBEX, BEaST, MASS and the deep learning approach for all four datasets. Section 4 concludes the paper with a discussion and an outlook on possible future work.

2. Materials and Methods

2.1. Datasets

We use the ICBM 152 non-linear atlas (2009a) [28] as our normal control atlas. ICBM 152 is a 1x1x1 mm template with 197×233×189 voxels, obtained from T1-weighted MRIs. Importantly, it also includes the brain mask. As the ICBM 152 atlas image itself contains the skull, we can obtain a *brain-only* atlas simply by applying the provided brain mask.

We use five different datasets for our experiments. Specifically, we use one (OASIS, see below) of the datasets to build our PCA model and the remaining four to test our brain extraction approach.

OASIS. We use images from the Open Access Series of Imaging Studies (OASIS) [16, 17] to build the PCA model for our brain extraction approach. The OASIS cross-sectional MRI dataset consists of 416 sagittal T1-weighted MRI scans from subjects between 18 and 96 years of age. In this data corpus, 100 of the subjects over 60 years old have been diagnosed with very mild to mild Alzheimer's disease (AD). The original scans were obtained with in-plane resolution 1×1 mm (256×256), slice thickness = 1.25 mm and slice number = 128. For each subject, a gain-field corrected atlas-registered image and its corresponding masked image in which all non-brain voxels have been assigned an intensity of zero are available. Each image is resampled to $1 \times 1 \times 1$ mm isotropic voxels and is of size $176 \times 208 \times 176$.

We *evaluate* our approach on four datasets, which all provide brain masks. Although in our study, we focus on T1-weighted images only, our model can be applied to other modalities as long as the PCA model is also built from data acquired by the same modality. The datasets we use for validation are described below.

IBSR. The Internet Brain Segmentation Repository (IBSR) [29] contains MR images from 20 healthy subjects of age 29.1±4.8 years including their manual brain segmentations, provided by the Center for Morphometric Analysis at Massachusetts General Hospital. All coronal 3D T1-weighted spoiled gradient echo MRI scans were acquired using two different MR systems: ten scans (4 males and 6 females) were performed on a 1.5T Siemens Magnetom MR system (with in-plane resolution of 1×1 mm and slice thickness of 3.1 mm); another ten scans (6 males and 4 females) were acquired from a 1.5T General Electric Signa MR system (with in-plane resolution of 1×1 mm and slice thickness of 3 mm).

LPBA40. The LONI Probabilistic Brain Atlas (LPBA40) dataset of the Laboratory of Neuro Imaging (LONI) [14, 15] consists of 40 normal human brain volumes. LPBA40 contains images of 20 males and 20 females of age 29.20 ± 6.30 years. Coronal T1-weighted images with slice thickness 1.5 mm were acquired using a 1.5T GE system. Images for 38 of the subjects have in-plane resolution of 0.86×0.86 mm; the images for the remaining two subjects have a resolution of 0.78×0.78 mm. A manually segmented brain mask is available for each image.

BRATS: We use twenty T1-weighted image volumes of low and high grade glioma patients from the Brain Tumor Segmentation (BRATS 2016) dataset [30] that include cases with large tumors, deformations, or resection cavities. We do not use the BRATS images available as part of the BRATS challenge as these have already been pre-processed (i.e., brain-extracted and co-registered). Instead, we obtain a subset of twenty of the originally acquired images. The BRATS dataset is challenging as the images were acquired with different clinical protocols and various different scanners from multiple ($n = 19$) institutions [31]. Our subset of twenty images is from six different institutions. Furthermore, the BRATS images have comparatively low resolution and some of them contain as few as 25 axial slices (with slice thickness as large as 7mm). The in-plane resolutions vary from 0.47×0.47 mm to 0.94×0.94 mm with image grid sizes between 256×256 and 512×512 pixels. We manually segment the brain in these images to obtain an accurate brain mask for validation.

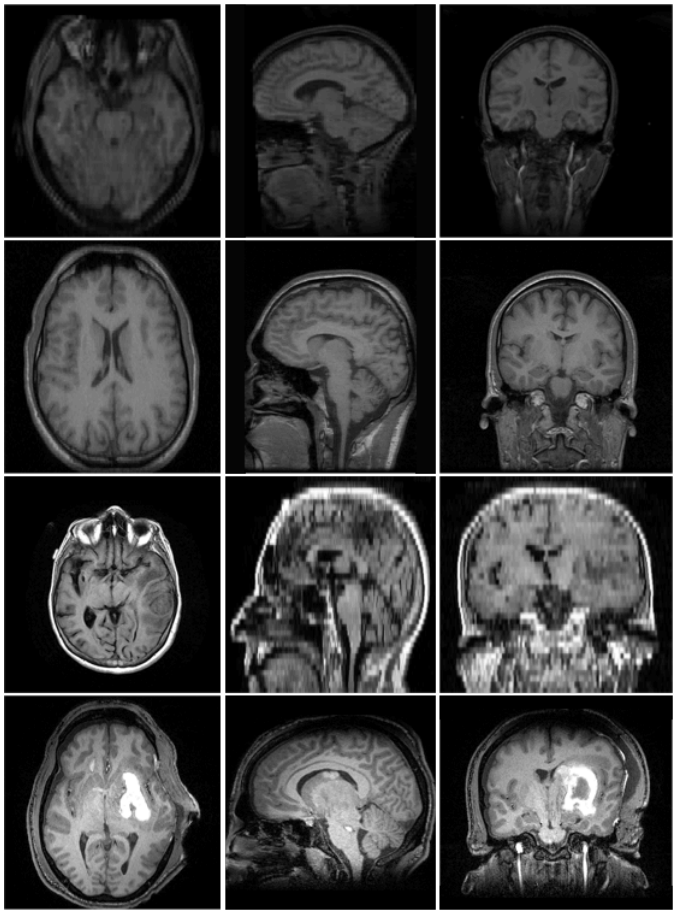


Figure 1: Illustration of image appearance variability on a selection of images from each (evaluation) database. From *top* to *bottom*: IBSR, LPBA40, BRATS and TBI.

TBI. Finally, we use our own Traumatic Brain Injury (TBI) dataset which contains 8 TBI images as well as manual brain segmentations. These are standard MPRAGE [32] T1-weighted images with no contrast enhancement. They have been resampled to $1\times1\times1$ mm isotropic voxel size with image size between $192\times228\times170$ and $256 \times 256 \times 176$. Segmentations are available for healthy brain, hemorrhage, edema and necrosis. To generate the brain masks, we always use the union of healthy tissue and necrosis. We also include hemorrhage and edema if they are contained within healthy brain tissue.

Fig. 1 shows example images from each dataset to illustrate image variability. IBSR and LPBA40 contain images from normal subjects and include large portions of the neck; BRATS has very low out-of-plane resolution; and the TBI dataset contains large pathologies and abnormal skulls.

2.2. Dataset processing

2.2.1. PCA model

We randomly pick 100 images and their brain masks to build our PCA model of the brain. Specifically, we register the brain-masked images to the brain-masked ICBM atlas using a B-spline registration. We use `NiftyReg` [33] to perform the B-spline registration with local normalized cross-correlation (LNCC) as similarity measure. To normalize image intensities, we apply an affine transform to the image intensities of the warped images so that the 1st percentile is mapped to 0.01 and 99th percentile is mapped to 0.99 and then clamp the image intensities to be within $[0, 1]$. We then perform PCA on the now registered and normalized images and retain the top 50 PCA modes, which preserve 63% of the variance, for our statistical appearance model. This is similar to an active appearance model [34].

2.2.2. IBSR refined segmentation

For IBSR, segmentations of the brain images into white matter, gray matter and cerebrospinal fluid (CSF) are provided. While, in principle, the union of the segmentations of white matter, gray matter and CSF should represent the desired brain mask, this is not exactly the case (see Fig. 2). To alleviate this issue for each segmentation, we use morphological closing to fill in remaining gaps and holes inside the brain mask and, in particular, to disconnect the background inside the brain mask from the surrounding image background. The structuring element for closing is a voxel and its 18 neighborhood[3]. We then find the connected component for the background and consider its complement the brain mask. Fig. 2 shows the pre-processing result after these refinement steps, compared to the original IBSR segmentation (i.e., the union of white matter, gray matter, and the CSF).

(a) (b) (c)

Figure 2: Example coronal slice of (a) an IBSR MR brain image, (b) the corresponding original IBSR brain segmentation (i.e., union of white matter, gray matter and CSF) and (c) the refined brain segmentation result.

2.3. Review of related models

As mentioned previously, brain extraction is challenging because it requires the identification of all non-brain tissue which can be highly variable (cf. Fig. 1). Our brain extraction approach is based on image alignment to an atlas space where a brain mask is available. However, this requires a reliable registration approach which can tolerate variable image appearance as well as pathologies (i.e.,

[3]The 18-voxel connectivity is also used for other morphological operations in this manuscript.

brain tumors, traumatic brain injuries, or general head injuries resulting in skull deformations and fractures). In both cases, no one-to-one mapping between image and atlas space may be available and a direct application of standard image similarity measures for image registration may be inappropriate.

A variety of approaches have been proposed to address the registration of pathological images. For example, cost function masking [35] and geometric metamorphosis [36] exclude the pathological regions when measuring image similarities. However, these approaches require prior segmentations of the pathologies, which can be non-trivial and/or labor intensive. A conceptually different approach is to learn the normal image appearance from population data and to estimate a quasi-normal image from a pathological image. Then, the quasi-normal image can be used for registration [37]. The low-rank + sparse (LRS) image registration framework, proposed by Liu et al. [26], follows this idea by iteratively registering the low-rank components from the input images to the atlas and then recomputes the low-rank components. After convergence, the image is well-aligned with the atlas.

Our proposed brain extraction model builds upon our previous PCA-based approach for pathological image registration [27] which, in turn, builds upon and removes many shortcomings of the low-rank + sparse approach of Liu et al. [26]. We therefore briefly review the low-rank + sparse technique in Sec. 2.3.1 and the PCA approach for pathological image registration in Sec. 2.3.2. We discuss our proposed model for brain extraction in Sec. 2.4.

2.3.1. Low-Rank + Sparse (LRS)

An LRS decomposition aims at minimizing [38]

$$E(L,S) = \text{rank}(L) + \lambda\|S\|_0 \quad \text{s.t.} \quad D = L + S. \tag{1}$$

I.e., the goal is to find an additive decomposition of a data matrix $D = L + S$ such that L is low-rank and S is sparse. Here, $\|S\|_0$ denotes the number of non-zero elements in S and $\lambda > 0$ weighs the contribution of the sparse part, S, in relation to the low-rank part L. Neither rank nor sparsity are convex functions. Hence, to simplify the solution of this optimization problem it is relaxed: the rank is replaced by the nuclear norm and the sparsity term is replaced by the one-norm. As both of these norms are convex and $D = L + S$ is a linear constraint one obtains the convex approximation to LRS decomposition by minimizing the energy

$$E(L,S) = \|L\|_* + \lambda\|S\|_1, \quad \text{s.t.} \quad D = L + S \ , \tag{2}$$

where $\|\cdot\|_*$ is the nuclear norm (i.e., a convex approximation for the matrix rank). In imaging applications, D contains all the (vectorized) images: each image is represented as a column of D. The low-rank term captures common information across columns. The sparse term, on the other hand, captures uncommon/unusual information. As Eq. (2) is convex, minimization results in a global minimum.

In practice, applying the LRS model requires forming the matrix D from all the images. D is of size $m \times n$, where m is the number of voxels, and n is the number of images. For 3D images, $m \gg n$ (typically). Assuming all images are spatially well-aligned, L captures the quasi-normal appearance of the images whereas S contains pathologies which are not shared across the images. Of course, in practice, the objective is image alignment and hence the images in D cannot be assumed to be aligned a-priori. Hence, Liu et al. [26] alternate LRS decomposition steps with image registration steps. Here the registrations are between all the low-rank images (which are assumed to be approximately pathology-free) and an atlas image. This approach is effective in practice, but can be computationally costly, may require large amounts of memory, and has the tendency to lose fine image detail in the quasi-normal image reconstructions, L. In detail, the matrix D has a large number of rows for typical 3D images, hence it can be costly to store. Furthermore, optimizing the LRS decomposition involves a singular value decomposition (SVD) at each iteration with a complexity of $\mathcal{O}(min\{mn^2, m^2n\})$ [39] for an $m \times n$ matrix. While large datasets are beneficial to capturing data variation, the quadratic complexity renders LRS computationally challenging in these situations.

However, it is possible to overcome many of these shortcomings while staying close to the initial motivation of the original LRS approach. The following Section 2.3.2 discusses how this can be accomplished.

2.3.2. Joint PCA-TV model

To avoid the memory and computational issues of the low-rank + sparse decomposition discussed above, we previously proposed a joint PCA/Image-Reconstruction model [27] for improved and more efficient registration of images with pathologies. In this model, we have a collection of normal images and register all the normal images to the atlas *once*, using a standard image similarity measure. These normal images do not need to be re-registered during the iterative approach. We mimic the low-rank part of the LRS by a PCA decomposition of the atlas-aligned normal images from which we obtain the PCA basis and the mean image. Let us consider the case when we are now given a *single* pathological image I. Let $\hat{I}$ denote the pathological image after subtracting the mean image M and B the PCA basis matrix. $\hat{L}$ and T are images of the same size as I[4]. Specifically, we minimize

$$E(T,\hat{L},\boldsymbol{\alpha}) = \frac{1}{2}\|\hat{L} - B\boldsymbol{\alpha}\|_2^2 + \gamma\|\nabla T\|_{2,1}, \quad \text{s.t. } \hat{I} = \hat{L} + T \tag{3}$$

where $\|\nabla T\|_{2,1} = \sum_i \|\nabla T_i\|_2$ and i denotes spatial location. This model is similar to the Rudin-Osher-Fatemi

[4]Images are vectorized for computational purposes, but the spatial gradient ∇ denotes the gradient in the spatial domain.

(ROF) image denoising model [40]. It results in a total variation (TV) term, T, which captures the parts of $\hat{I}$ that are (i) relatively large, (ii) spatially contiguous, and (iii) cannot be explained by the PCA basis, e.g., pathological regions. The quasi-low-rank part $\hat{L}$ remains close to the PCA space but retains fine image detail. The quasi-normal image L can then be reconstructed as $L = M + \hat{L}$. We refer to this model as our joint PCA-TV model.

As in the LRS approach, we can register the quasi-normal image L to atlas space and alternate decomposition and registration steps. However, in contrast to the LRS model, the PCA-TV model registers only *one* image (L) in each registration step and consequently requires less time and memory to compute. Furthermore, the reconstructed quasi-normal image, L, retains fine image detail as pathologies are captured via the total variation term in the PCA-TV model.

2.4. Proposed brain extraction approach

The following sections describe how our proposed brain extraction approach builds upon the principles of the PCA-TV model (Section 2.4.1), and discusses image preprocessing (Section 2.4.2), the overall registration framework (Section 2.4.3), and post-processing steps (Section 2.4.4).

2.4.1. Joint PCA-Sparse-TV model

The PCA-TV model captures the pathological information well, but it does not model non-brain regions (such as the skull) appropriately. The skull is, for example, usually a thin, shell-shape structure and other non-brain tissue may be irregularly shaped with various intensities. The only commonality is that all these structures surround the brain. Specifically, if a test image is aligned to the atlas well, these non-brain tissues should *all* be located outside the atlas' brain mask. Hence, we reject these non-brain regions via a spatially distributed sparse term. We penalize sparsity heavily inside the brain and relatively little on the outside of the brain. This has the effect that it is very cheap to assign voxels outside the brain to the sparse term; hence, these are implicitly declared as brain outliers. Of course, if we would already have a reliable brain mask we would not need to go through any modeling. Instead, we assume that our initial affine registration provides a good *initial alignment* of the image, but that it will be inaccurate at the boundaries. We therefore add a constant penalty close to the boundary of the atlas brain mask. Specifically, we create two masks: a two-voxel-eroded brain mask, which we are confident is within the brain and a one-voxel-dilated brain mask, which we are confident includes the entire brain. We then obtain the following model:

$$\begin{aligned} E(S,T,\hat{L},\boldsymbol{\alpha}) = \frac{1}{2}\|\hat{L} - B\boldsymbol{\alpha}\|_2^2 + \gamma\|\nabla T\|_{2,1} \\ + \|\boldsymbol{\Lambda} \odot S\|_1, \\ \text{s.t. } \hat{I} = \hat{L} + S + T \end{aligned} \tag{4}$$

where $\boldsymbol{\Lambda} = \Lambda(\boldsymbol{x}) \geq 0$ is a spatially varying weight

$$\Lambda(\boldsymbol{x}) = \begin{cases} \infty, & \boldsymbol{x} \in \text{Eroded Mask (inside)} \\ \lambda, & \boldsymbol{x} \in \text{Dilated Mask and} \\ & \boldsymbol{x} \notin \text{Eroded Mask (at boundary)} \\ 0, & \boldsymbol{x} \notin \text{Dilated Mask (outside)} \end{cases} \tag{5}$$

with $\boldsymbol{x}$ denoting the spatial location. Further, in Eq. (4), $\odot$ indicates an element-wise product and $\gamma \geq 0$ weighs the total variation term.

We refer to this model as our joint PCA-Sparse-TV model. It decomposes the image into three parts. Similar to the PCA-TV model, the quasi-low-rank part $\hat{L}$ remains close to the PCA space and the TV term, T, captures pathological regions. Here, the PCA basis is generated from normal images that have been already brain-extracted. Therefore $\hat{L}$ only contains the brain tissue. Different from the previous model, we add a spatially distributed sparse term, S, which captures tissue outside the brain, e.g., the skull. In effect, since Λ is very large inside the eroded mask, none of the image inside the eroded mask will be assigned to the sparse part. Conversely, all of the image outside the dilated mask will be assigned to the sparse part. We then integrate this PCA-Sparse-TV model into the low-rank registration framework. This includes three parts: pre-processing, iterative registration and decomposition, and post-processing as we will discuss in the following.

2.4.2. Pre-processing

Fig. 3 shows a flowchart of our pre-processing approach as discussed in the following paragraphs.

Intensity normalization. Given a test image from which we want to extract the brain, we first affinely transform the image intensities to standardize the intensity range to $[0, 1000]$. Note that our PCA model of section 2.2.1 is build based on images with intensities standardized to $[0, 1]$. The different standardization is necessary here as the bias field correction algorithm removes negative and small intensity values (< 1) followed by a log transform of the intensities. Specifically, we first compute the 1st and the 99th percentile of the voxel intensities. We then affinely transform the image intensities of the entire image such that the intensity of the 1st percentile is mapped to 100 and of the 99th percentile to 900. As this may result in intensities smaller than zero or larger than 1000 for the extreme ends of the intensity distribution, we clamp the intensities to be within $[0, 1000]$.

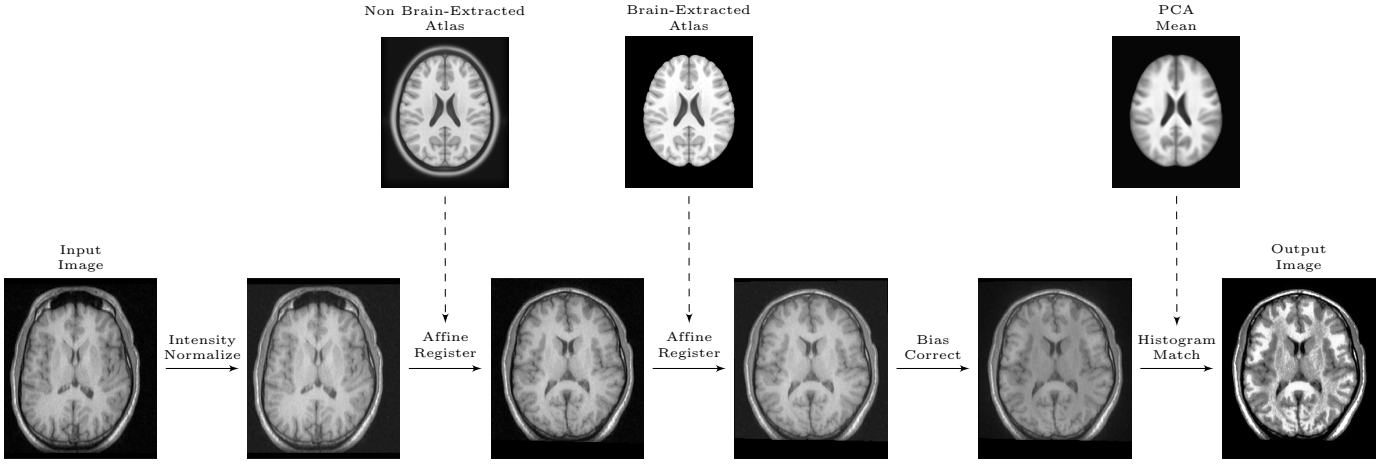


Figure 3: Preprocessing flow chart: Input image is the original image. Eventually, the output image will be fed into the registration/decomposition framework.

Atlas registration. Next, we first align the intensity-normalized input image to the non brain-extracted atlas. Then, we affinely register the result from the first step to the brain-extracted atlas, but this time using a one-voxel-dilated brain mask in atlas space; this step has the effect of ignoring parts of the image which are not close to the brain in the registration and it gives us a better alignment in the brain region. For both steps we use `reg_aladin` of `NiftyReg` [41] disabling symmetric registration (`-noSym`). The first registration initializes the transformation using the center of gravity (CoG) of the image. Note that the differing intensity range of the atlas and the image is immaterial in this step as the registration uses local normalized cross-correlation as the similarity measure.

Bias field correction. Next, we use N4ITK [42], a variant of the popular non-parametric non-uniform intensity normalization (N3) algorithm [43], to perform bias field correction. As the image has been affinely aligned to the atlas in the previous step, we use our two-voxel-eroded brain mask as the region for bias field estimation. Specifically, we use the `N4BiasFieldCorrection` function in `SimpleITK` [44], with its default settings.

Histogram matching: The final step of the preprocessing is histogram matching. We match the histograms of the bias corrected image with the histogram of the mean image of the population data only within the two-voxel-eroded brain mask. This histogram matched image is then the starting point for our brain extraction algorithm and it is now in an intensity range comparable to the PCA model.

2.4.3. Registration framework

Similar to the PCA-TV model, we alternate between *image decomposition* steps using the PCA-Sparse-TV model and *registration to the brain-extracted atlas*. We use a total of six iterations in our framework. In the first iteration ($k = 1$), the images are in the original space. We decompose the input image $I_1 = I$, into the quasi-normal ($L_1 = \hat{L}_1 + M$), sparse (S_1), and total variation (T_1) images by minimizing the energy from Eq. (4). We then obtain a pathology-free or pathology-reduced image, R_1, by adding the sparse and the quasi-normal images of the decomposition: $R_1 = L_1 + S_1$.

For the next two iterations ($k = \{2, 3\}$), we first find the affine transform Φ_k^{-1} by affinely registering the pathology-reduced images from the previous iteration, R_{k-1} (i.e., $R_{k-1} = L_{k-1} + S_{k-1}$), to the brain-extracted atlas[5]. We use the one-voxel-dilated brain mask for cost-function masking which allows the registration to focus only on the brain tissue. This is important as the first few registrations will not be very precise as they are only based on an affine deformation model. The main objective is to reduce the pathology within the brain. Only after these initial steps, when a good initial alignment has already been obtained, we use the quasi-normal image (excluding the non-brain regions) to perform the registration. We then apply the transform Φ_k^{-1} to transform the previous input images to atlas space and obtain new input images, I_k, (i.e., $I_k = I_{k-1} \circ \Phi_k^{-1}$). We minimize Eq. (4) again to obtain new decomposition results (L_k, S_k, T_k). These decomposition/affine-registration steps are repeated two times, which is empirically determined to be sufficient for convergence. These affine registration steps result in a substantially improved alignment in comparison to the initial affine registration by itself.

The last three iterations ($k = \{4, 5, 6\}$) repeat the same process, but are different in the following aspects: (i) we now use a B-spline registration instead of the affine registration; (ii) we use the pathology-reduced image and cost

[5]We follow standard image-registration notation. I.e., a map Φ^{-1} is defined in the space an image is deformed to. For us this is the space of the atlas image. Conversely, Φ maps an image from the atlas space back into the original image space and hence is defined in the original image coordinate space.

function masking only for the first B-spline registration step, as we did in the previous affine steps. For the remaining two steps, we use the quasi-normal images $L_{k:k=\{5,6\}}$ as the moving images and we do not use the mask during the registrations. The use of the mask is no longer necessary as registrations are now performed using the quasi-normal image; (iii) we use the non-greedy registration strategy of the original low-rank + sparse framework [45], in which we deform the quasi-normal image back to the image space of the third iteration (after the affine steps) in order to avoid accumulating deformation errors.

These steps further refine the alignment, in particular, close to the boundary of the brain mask. After the last iteration, the image is well-aligned to the atlas and we have all the transforms from the original image space to atlas space. As a side effect, the algorithm also results in a quasi-normal reconstruction of the image, L_6, an estimate of the pathology, T_6, and an image of the non-brain tissue S_6, all in atlas space.

2.4.4. Post-processing

Post-processing consists of applying to the atlas mask the inverse transforms of the affine registrations in the preprocessing step and the inverse transforms of the registrations generated in the framework described in section 2.4.3. The warped-back atlas mask is the brain mask for the original image. To extract the brain in the original image space, we simply apply the brain mask on the original input image. All subsequent validations are performed in the original image space.

Algorithm 1 summarizes these steps as pseudo-code.

3. Experimental results

The following experiments are for brain-extraction from T1-weighted MR images. However, our method can be easily adapted to images from other modalities, as long as the atlas image and the images from which the PCA basis is computed are from the same modality.

3.1. Experimental setup

We evaluate our method on all four evaluation datasets. For comparison, we also assess the performance of BET, BSE, ROBEX, BEaST, MASS and CNN on these datasets. We use BET v2.1 as part of FSL 5.0, BSE v.17a from BrainSuite, ROBEX v1.2, BEaST (`mincbeast`) v1.90.00, and MASS v1.1.0. We solve our PCA model via a primal-dual hybrid gradient method [46]. In addition, we implement the decomposition on the GPU and run it on an NVIDIA Titan X GPU [47] [48].

3.2. Evaluation Measures

We evaluate the brain extraction approaches using the measures listed below.

Dice coefficient. Given two sets X and Y (containing the spatial voxel positions of a segmentation), the Dice coefficient $D(X, Y)$ is defined as

$$D(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{6}$$

where $X \cap Y$ denotes set intersection between X and Y and $|X|$ denotes the cardinality of set X.

Average, maximum and 95% surface distance. We also measure the symmetric surface distances between the automatic brain segmentation and the gold-standard brain segmentation. This is defined as follows: the distance of a point x to a set of points (or set of points of a triangulated surface S_A) is defined as

$$d(x, S_A) = \min_{y \in S_A} d(x, y), \tag{7}$$

where $d(x, y)$ is the Euclidean distance between the point x and y. The average symmetric surface distances between two surfaces S_A and S_B is then defined as

$$ASD(S_A, S_B) = \frac{1}{|S_A| + |S_B|} \times \Big(\sum_{x \in S_A} d(x, S_B) + \sum_{y \in S_B} d(y, S_A) \Big), \tag{8}$$

where $|S_A|$ denotes the cardinality of S_A [49] (i.e., number of elements if represented as a set or surface area if represented in the continuum). To assess behavior at the extremes, we also report the maximum symmetric surface distance as well as the 95th percentile symmetric surface distance, which is less prone to outliers. These are defined in analogy, i.e., by computing all distances from surface S_A to S_B and vice versa followed by the computation of the maximum and the 95th percentile of these distances.

Sensitivity and specificity. We also measure sensitivity, i.e., true positive (TP) rate and specificity, i.e., true negative (TN) rate. Here TP denotes the brain voxels which are correctly labeled as brain; TN denotes the non-brain voxels correctly labeled as such. Furthermore, the false negatives (FN) are the brain voxels incorrectly labeled as non-brain and the false positives (FP) are the non-brain voxels which are incorrectly labeled as brain. Let V be the set of all voxels of an image, and X and Y the automatic brain segmentation and gold-standard brain segmentation, respectively. The sensitivity and specificity are then defined as follows [50] :

$$sensitivity = \frac{TP}{TP + FN} = \frac{|X \cap Y|}{|Y|} \tag{9}$$

$$specificity = \frac{TN}{TN + FP} = \frac{|V| - |X \cup Y|}{|V| - |Y|} \tag{10}$$

3.3. Datasets of normal images: IBSR/LPBA40

IBSR results: Fig. 4 shows the box-plots summarizing the results for the IBSR dataset. Overall, ROBEX,

Algorithm 1: Algorithm for Brain Extraction

Input: Image I, Brain-Extracted Atlas A, Atlas Mask A_M
Output: Brain-Extracted Image I_B and mask I_M

I_1, Φ_1^{-1} = pre-processing(I);
for $k \leftarrow 1$ **to** 6 **do**
  if $k \geq 2$ **then**
    if $k \leq 3$ **then**
      find Φ_k^{-1}, s.t., $R_{k-1} \circ \Phi_k^{-1} = A$ and Φ_k^{-1} is affine;
    else if $k == 4$ **then**
      find Φ_k^{-1}, s.t., $R_{k-1} \circ \Phi_k^{-1} = A$ and Φ_k^{-1} is B-spline;
    else
      find Φ_k^{-1}, s.t., $(L_{k-1} \circ \Phi_{k-1}) \circ \Phi_k^{-1} = A$ and Φ_k^{-1} is B-spline;
    end
    $I_k = I_{k-1} \circ \Phi_k^{-1}$;
  Decompose I_k, s.t., $I_k = L_k + S_k + T_k$;
  if $k \leq 3$ **then**
    $R_k = L_k + S_k$;
  end
end
I_B, I_M = post-processing(A_M, $\{\Phi_k^{-1}\}$).

BEaST*, BSE, BET and our model perform well on this dataset, with a median Dice coefficient above 0.95. BEaST does not work well when applied directly on the IBSR images. This is due to failures with the initial spatial normalization (in 5 cases the computations themselves fail and in 10 cases the results are poor). Therefore, in our experiment, we first applied the same affine registration to atlas space as in the pre-processing step for our PCA model for all images. This affine transformation corresponds to a composition of the two affine transformations in Fig. 3. BEaST is then applied to the affinely aligned images. We use the same strategy for BRATS. We refer to the resulting approach as BEaST*. BEaST* performs well on most cases with high Dice scores and low surface distances. MASS works well on some cases, but performs poorly on many cases. CNN does not perform satisfactorily, with low Dice scores, low sensitivity, large distance errors, and overall high variance. Our PCA model has similar performance to BEaST*, but does not result in extreme outliers and hence results in higher mean Dice scores than BEaST*. Both methods outperform all others with respect to Dice scores (median close to 0.97) and distance measures in most cases. BSE also works well on most cases, but it shows larger variability and exhibits two outliers which represent failure cases. ROBEX and BET show the highest sensitivity, but reduced specificity. Conversely, our PCA model, BEaST*, BSE, and CNN have high specificity but reduced sensitivity (the CNN model dramatically so).

Table 2 (top) shows medians, means and standard deviations for the test results on this dataset. Our PCA model achieves the highest median and mean Dice overlap scores (both at 0.97) with the smallest standard deviation. BEaST* also shows high median Dice scores, but results in reduces mean scores due to the presence of outliers. ROBEX and BET show slightly reduced Dice overlap measures (mean and median around 0.95). BSE also shows slightly reduced median Dice scores, but greatly reduced mean scores. MASS show reduced median Dice scores. CNN shows the lowest performance. Our PCA model also performs best for the surface distance measures; it has the lowest mean and median surfaces distances. Overall our PCA model performs best.

In addition, we perform a one-tailed paired Wilcoxon signed-rank test (to safeguard against deviations from normality) to compare results between methods. We test the null hypothesis that the paired differences for the results of our PCA model and of the compared method come from a distribution with zero median, against the alternative that the median of the paired differences is non-zero.[6]. Table 1 (top) shows the corresponding results. We apply the Benjamini-Hochberg procedure [51] for all the tests, in order to reduce the false discovery rate for multiple comparisons. We select an overall false discovery rate of 0.05 which results in an effective significance level of $\alpha \approx 0.0351$. Our model outperforms all other methods on Dice and surface distances except for BEaST* which is significant only in Dice and average surface distance. In addition, our approach performs better than MASS, BSE and CNN on sensitivity and better than ROBEX, BEaST*, MASS, and BET on specificity.

LPBA40 results: Fig. 5 shows the box-plots summarizing the validation results for the LPBA40 dataset. All

[6] We perform a one-tailed test, thus we test for greater than zero for the Dice overlap scores, sensitivity and specificity, and less than zero for the surface distances.

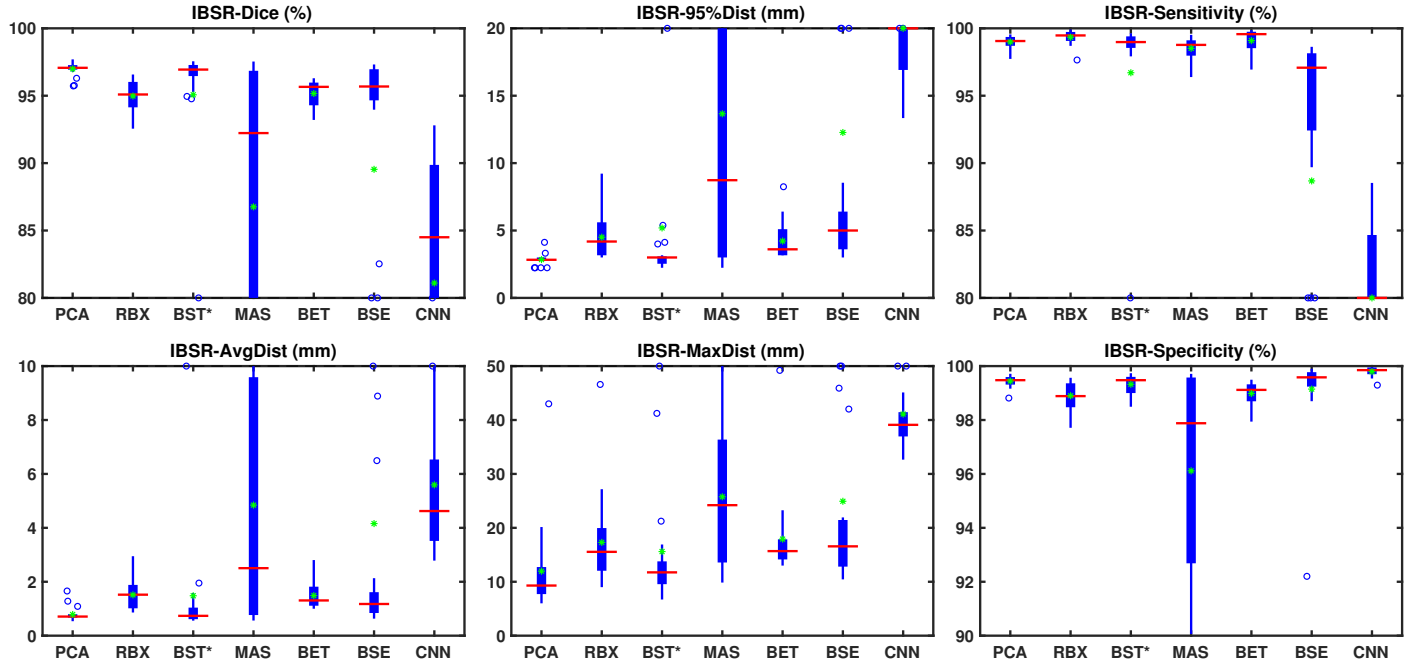


Figure 4: Box plot results for the IBSR normal dataset. We show the results from seven methods: PCA, RBX (ROBEX), BST* (BEaST*), MAS (MASS), BET, BSE and CNN. Due to the poor results of MASS and CNN, and the outliers of BSE on this dataset, we limit the range of the plots for better visibility. On each box, the center line denotes the median, and the top and the bottom edge denote the 75th and 25th percentile, respectively. The whiskers extend to the most extreme points that are not considered outliers. The outliers are marked with '+' signs. In addition, we mark the mean with green '*' signs. ROBEX, BET, and BSE show similar performance, but BSE exhibits two outliers. MASS works well on most images, but fails on many cases. BEaST fails on the original images. We therefore show the BEaST* results using the initial affine registration of our PCA model. BEaST* performs well with high Dice scores and low surface distances, but with low mean values. CNN performs poorly on this dataset. Our PCA model has similar performance to BEaST* but with higher mean values. Both methods perform better than other methods on the Dice scores and surface distances.

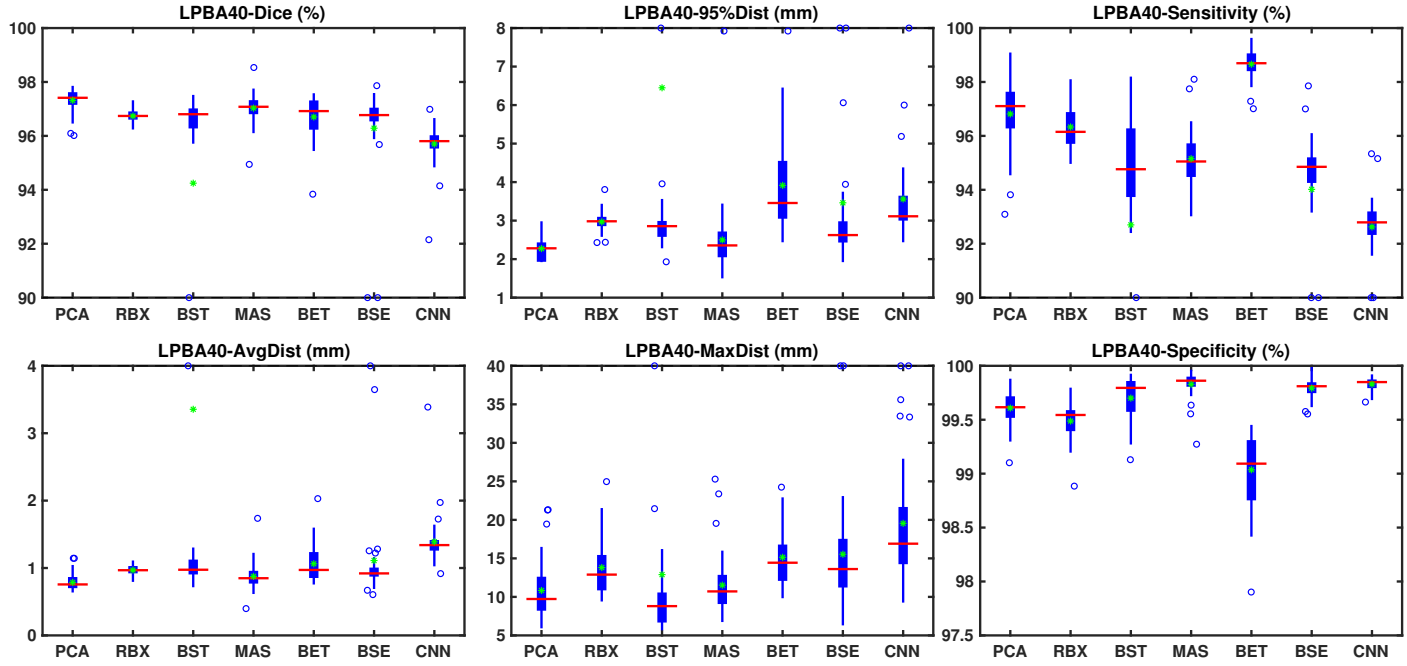


Figure 5: Box plot results for the LPBA40 normal dataset. All seven methods work well on this dataset. Our PCA model has the best Dice and surface distances. ROBEX, BEaST, MASS, BET and BSE show similar performance, but BET exhibits larger variance and BSE exhibits two outliers indicating failure. The CNN model shows overall slightly worse performance than the other methods.

seven methods perform well. ROBEX, BEaST, BET and BSE all have a median Dice score between 0.96 and 0.97. MASS has a median Dice score slightly above 0.97. Our PCA model obtains the highest median Dice score (0.974). All methods except for the CNN approach have a median average surface distance smaller than 1 mm. Table 2 (second top) shows the medians, means and standard deviations for all validation measures for this dataset. Again, all methods have satisfactory median, mean Dice scores and surface distances with low variances. Compared with other methods, the PCA model achieves the best results.

Table 1 (second top) shows the one-sided paired Wilcoxon signed-rank test results. Again we use the Benjamini-Hochberg procedure, resulting in a significance level $\alpha \approx 0.0351$. All methods perform well on this dataset, but our PCA approach still shows statistically significant improvement. We outperform other methods on Dice and all surface distances with statistical significance except for BEaST on maximum surface distance and for MASS on 95% surface distance. We perform better than all other methods except BET on sensitivity and better than BET and ROBEX on specificity.

Fig. 9 (left) visualizes the average brain mask errors for IBSR and LPBA40. All images are first affinely registered to the atlas. Then we transform the gold-standard expert segmentations as well as the automatically obtained brain masks of the different methods to atlas space. We compare the segmentations by counting the average over- and under-segmentation errors over all cases at each voxel. This results in a visualization for areas of likely mis-segmentation. Our PCA model, ROBEX, BEaST (BEaST*) and BET perform well on these two datasets. Compareed to our model, ROBEX, BEaST (BEaST*) and BET show larger localized errors, e.g., at the boundary of the parietal lobe, the occipital lobe and the cerebellum. While MASS, BSE and CNN perform well on the LPBA40 dataset, they perform poorly on the IBSR dataset. This is in particular the case for the CNN approach.

3.4. Datasets with strong pathologies: BRATS/TBI

BRATS results: Fig. 6 shows the box-plots for the validation measures for the BRATS dataset. BSE and CNN, using their default settings, do not work well on the BRATS dataset. This may be because of the data quality of the BRATS data. Many of the BRATS images have relatively low out-of-plane resolutions. BSE results may be improved by a better parameter setting. However, as our goal is to evaluate all methods with the same parameter setting across all datasets, we do not explore dataset specific parameter tuning. BEaST also fails on the original BRATS images due to the spatial normalization. As for the IBSR dataset, we therefore use BEaST*, our adaptation of BEaST using the affine transformation of our PCA model. BET shows good performance, but suffers from a few outliers. ROBEX and BEaST* work generally well, with a median Dice score around 0.95 and an average distance error of 1.3 mm. MASS also works well on most cases. However, as for IBSR and LPBA40, our PCA model performs generally the best with a median Dice score 0.96 and a 1 mm average distance error. The PCA model results also show lower variance, as shown in table 2 (second bottom), underlining the very consistent behavior of our approach.

Table 1 shows (via a one-sided paired Wilcoxon signed-rank test with a correction for multiple comparisons using a false discovery rate of 0.05) that our model has statistically significantly better performance than ROBEX, BEaST*, BET, BSE, CNN on most measures. The improvement over MASS, however, is not statistically significant.

TBI results: Fig. 7 shows the box-plots for the results on our TBI dataset. Our PCA model still outperforms all other methods. Our method achieves the largest Dice scores, and the lowest surface distances among all methods with best mean and lowest variance as shown in table 2 (bottom). Table 1 shows the one-sided paired Wilcoxon signed-rank test results with multiple comparisons correction with a false discovery rate of 0.05. Our model performs significantly better than ROBEX, BEaST, BET, BSE and CNN on most measures. The improvement over MASS is only statistically significant on Dice and 95% surface distance.

Finally, Fig. 9 (right) shows the average segmentation errors on the BRATS and TBI datasets: our PCA method shows fewer errors than most other methods in these two abnormal datasets. MASS also shows few errors, while ROBEX, BEaST (BEaST*) and BET exhibit slightly larger errors at the boundary of the brain. CNN and BSE particularly show large errors for the BRATS dataset presumably again due to the coarse resolution of the BRATS data.

In addition to extracting the brain from pathological datasets, our method also allows for the estimation of a corresponding quasi-normal image in atlas space, although this is not the main goal of this paper. Fig.8 shows an example of the reconstructed quasi-normal image (L) for an image of the BRATS dataset, as well as an estimation of the pathology (pathology image T and non-brain image S). Compared to the original image, the pathology shown in the quasi-normal image has been greatly reduced. Hence this image can be used for the registration with a normal image or a normal atlas. This has been shown to improve registration accuracy for the registration of pathological images [27]. Furthermore, an estimate of the pathology (here a tumor) is also obtained which may be useful for further analysis. Note that in this example image the total variation term captures more than just the tumor. This may be due to inconsistencies in the image appearance between the normal images (obtained from OASIS data) and the test dataset. As our goal is atlas alignment rather than quasi-normal image reconstruction or pathology segmentation, such a decomposition is acceptable, although we

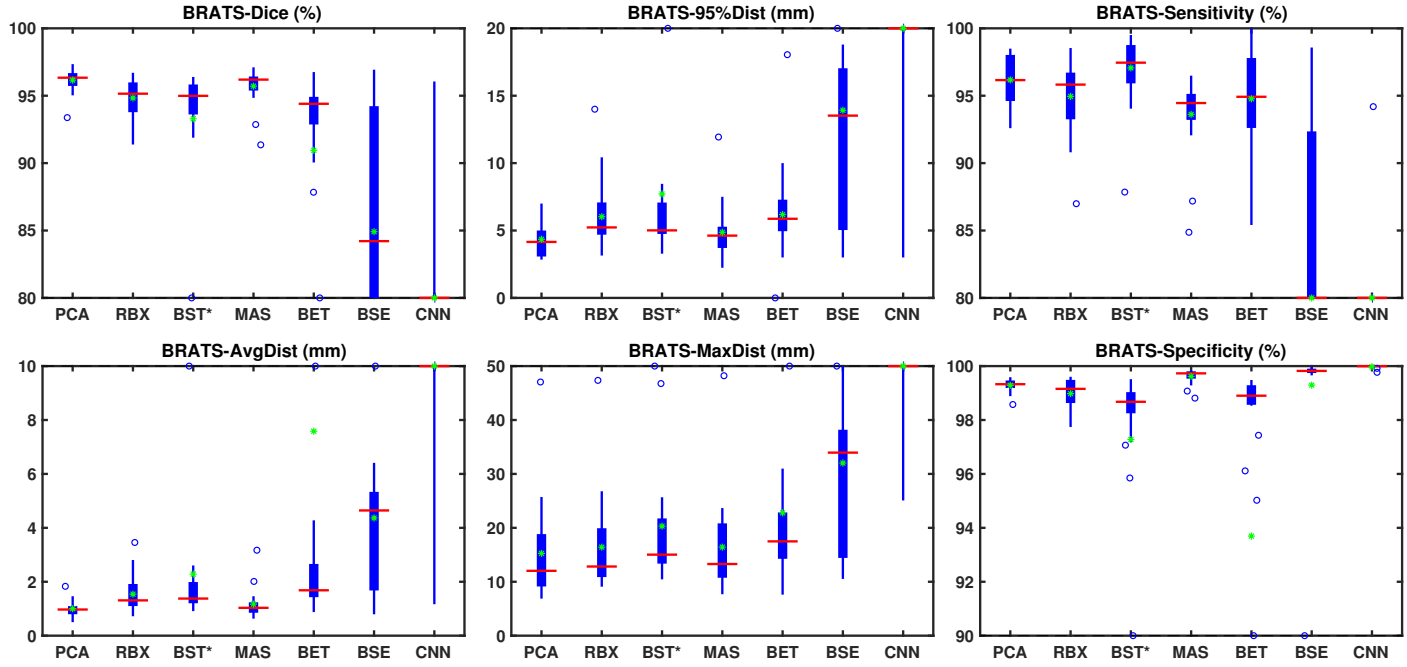


Figure 6: Box plot results for the BRATS tumor dataset. BSE and CNN fail on this dataset. BEaST also fails when applied directly to the BRATS dataset due to spatial normalization failures. We therefore show results for BEaST* here, our modification which uses the affine registration of the PCA model first. BET shows better performance, but also exhibits outliers. ROBEX, BEaST*, MASS, and our PCA model work well on this dataset. Overall our model exhibits the best performance scores.

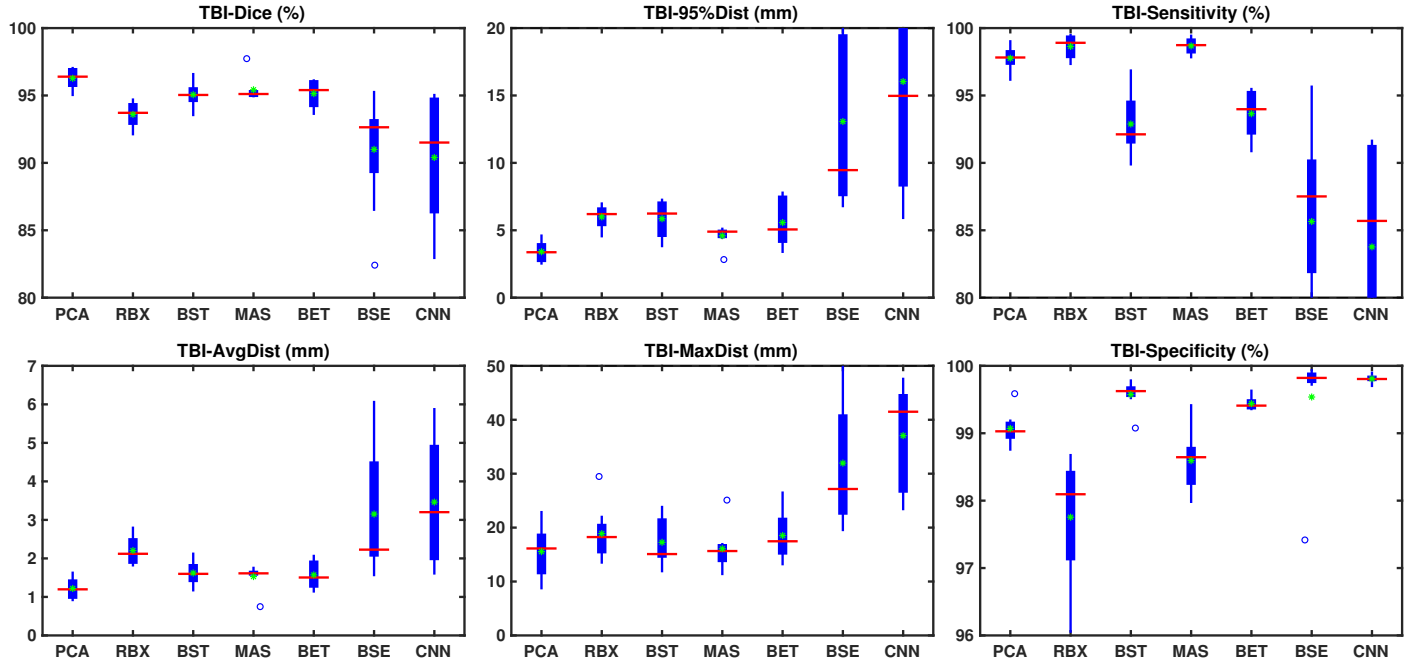
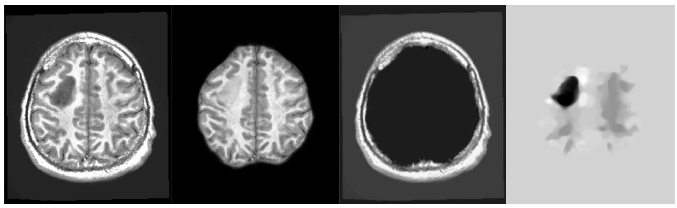


Figure 7: Box plot results for the TBI dataset. Our PCA model shows the best evaluation scores. BET, BEaST, MASS and ROBEX also perform reasonably well. BSE and CNN exhibit inferior performance on this dataset.

could improve this by tuning the parameters or applying regularization steps as in [27].

3.5. Runtime and memory consumption

Decomposition is implemented on the GPU. Each decomposition takes between 3 to 5 minutes. Currently, the registration steps are the most time-consuming parts of the overall algorithm. We use `NiftyReg` on the CPU for registrations. Each affine registration step takes less than 3 minutes and the B-spline step takes 5 minutes. However, in the current version of `NiftyReg` a B-spline registration can take up to 15 minutes when cost function



(a) (b) (c) (d)

Figure 8: Example BRATS image with its decomposition result in atlas space. (a) Input image after pre-processing; (b) quasi-normal image $L + M$; (c) non-brain image S; (d) pathology image T.

masking is used. Overall our brain extraction approach takes around 1 hour to 1.5 hours for each case, including the pre-processing step.

Storing the PCA basis requires the most memory. Each 197×232×189 3D image (stored as double) consumes about 66MB of memory. Hence it requires less than 7 GB to store the 100 PCA basis images, in addition to the atlases and masks. As our model only uses 50 PCA bases, stored in B, and requires two variable copies during runtime, our overall algorithm requires less than 7 GB of memory and hence can easily be run on modern GPUs.

4. Discussion

We presented a PCA-based model specifically designed for brain extraction from pathological images. The model decomposes an image into three parts. Non-brain tissue outside of the brain is captured by a sparse term, normal brain tissue is reconstructed as a quasi-normal image close to a normal PCA space, and brain pathologies are captured by a total-variation term. The quasi-normal image allows for registration to an atlas space, which in turn allows registering the original image to atlas space and hence to perform brain extraction. Although our approach is designed for reliable brain extraction from strongly pathological images, it also performs well for brain extraction from normal images, or from images with subtle pathologies.

This is in contrast to most of the existing methods, which assume normal images or only slight pathologies. These algorithms are either not designed for pathological data (BET, BSE, BEaST) or use normal data for training (e.g., ROBEX and CNN). Consequently, as we have demonstrated, these methods may work suboptimally or occasionally fail when presented with pathological data. While our PCA model is built on OASIS data which contains abnormal images (from patients with Alzheimer's disease), OASIS data does not exhibit strong pathologies as, for example, seen in the BRATS and the TBI datasets. However, as our algorithm is specifically modeling pathologies on top of a statistical model of *normal* tissue appearance, it can tolerate pathological data better and, in particular, does not require pathology-specific training.

In fact, one of the main advantages of our method is that we can use a *fixed* set of parameters (without additional tuning or dataset-specific brain templates) across a wide variety of datasets. This can, for example, be beneficial for small-scale studies, where obtaining dataset-specific templates may not be warranted, or for more clinically oriented studies, where image appearance may be less controlled. We validated our brain extraction method using four different datasets (two of them with strong pathologies: brain tumors and traumatic brain injuries). On all four datasets our approach either performs best or is among the best methods. Hence, our approach can achieve good brain extraction results on a variety of different datasets.

There are a number of ways in which our method could be improved. For example, our decomposition approach is a compromise between model realism and model simplicity to allow for efficient computational solutions. However, it may be interesting to explore more realistic modeling assumptions to improve its quality. While the total variation term succeeds at capturing the vast majority of large tumor masses and would likely work well for capturing volumes of resected tissue, the texture of pathological regions will not be appropriately captured and will remain in the quasi-normal image. To obtain a more faithful quasi-normal image reconstruction would require more sophisticated modeling of the pathology. A possible option could be to train a form of auto-encoder (i.e., a non-linear generalization of PCA) to remove the pathology as in our prior work [37]. A natural approach could also be to perform this in the setting of a general adversarial network [52] (GAN) to truly produce normal-looking quasi-normal images. As tumor images, for example, frequently exhibit mass effects, training and formulating such a model could be highly interesting as one could attempt to model the expected mass effect as part of the GAN architecture.

The way we integrate our PCA model into the decomposition could also be improved. Specifically, for computational simplicity we only use the eigenspace created by a chosen number of PCA modes, but we do not use the strength of these eigenmodes. This is a simple, yet reasonable strategy, to form a low-dimensional subspace capturing normal tissue appearance as long as a pathology remains reasonably orthogonal to this subspace and hence would get assigned to the total variation part of the decomposition.

We effectively constructed a form of robust PCA decomposition, which prefers outliers that jointly form regions of low total variation. Instead of modeling the decomposition in this way, it could be interesting to explore an LRS model which uses a partially-precomputed L matrix and gets adapted for a given single image. Such a strategy may allow more efficient computations of the LRS decomposition, but would require keeping the entire training dataset in memory (instead of only a basis of reduced dimension). Such an approach could likely also be extended to a form of low-rank-total variation decomposition if desired.

Regarding our PCA decomposition, it would be natural to use a reconstruction that makes use of a form of Mahalanobis distance [53]. This would then emphasize

the eigendirections that explain most of the variance in the training data. Note, however, that our model is relatively insensitive to the number of chosen PCA modes. In fact, while different numbers of chosen PCA modes may affect how well the quasi-normal image is reconstructed, the number of PCA modes has only slight effects on the brain extraction results.

Tumors or general pathologies may also affect some of the pre-processing steps. For example, we perform histogram matching over the entire initial brain mask which includes the pathology. In practice, we visually assessed that such a histogram matching strategy produced reasonable intensity normalizations. However, this step could be improved, for example, by coupling it or alternating it with the decomposition in such a way that regions that likely correspond to pathologies are excluded from the histogram computations for histogram matching.

While our model's simplicity allowed it to work well across a wide variety of datasets, this generality likely implies suboptimality. For example, a likely reason why the CNN approach performs poorly on some of the datasets is because these datasets do not correspond well to the data the CNN was trained on. Dataset-specific fine-tuning of the model would likely help improve the CNN performance. Similarly, approaches, including our own, relying on some form of registration and a model of what a well-extracted brain looks like would likely also benefit from a dataset-specific atlas (including a dataset-specific PCA basis in our case) or dataset-specific registration templates. Such dataset-specific templates can, for example, easily be used within MASS and improve performance slightly. Similarly, we observed that the performance for BEaST can be improved if we use dataset-specific libraries. In practice, large-scale studies may warrant the additional effort of obtaining dataset-specific manually segmented brain masks for training. However, in many cases such manual segmentation may be too labor-intensive. In this latter case our proposed approach is particularly attractive as it is only moderately affected by differing image appearances and works well with a generic model for brain extraction.

Runtime of the algorithm is currently still in the order of an hour. It could be substantially reduced by using a faster registration method. For example, it may be possible to use one of the recently proposed deep learning approaches for fast registration [54, 55]. Furthermore, to speed-up the decompositions one could explore numerical algorithms with faster convergence or reformulations of the decomposition itself, as discussed above.

Exploring formulations for different image sequences or modalities (or combination of modalities) would be interesting future work as well. It would also be interesting to explore if the generated quasi-normal image and the identified pathology could be used to help assess longitudinal image changes, for example for comparing the chronic and the acute phases of TBI.

Our software is freely available as open source code at `https://github.com/uncbiag/pstrip`.

Acknowledgements

Research reported in this publication was supported by the National Institutes of Health (NIH) and the National Science Foundation (NSF) under award numbers NIH R41 NS081792, NSF ECCS-1148870, and EECS-1711776. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the NSF.

Appendix A. `NiftyReg` settings

This section introduces the settings for `NiftyReg` used in this paper. We mainly use the affine registration `reg_aladin` and the B-spline registration `reg_f3d`.

Affine Registration:. For affine registration, we use `reg_aladin` in `NiftyReg`. The options for affine registration are `-ref, -flo, -aff, -res`, which stand for reference image, floating image, affine transform output, warped result image, respectively. If the symmetric version is disabled, we add "`-noSym`". If center of gravity is used for the initial transformation, we add "`-cog`".

B-spline Registration:. For B-spline registration, we use `reg_f3d` in `NiftyReg`. In addition to the options as shown in affine (except for `reg_f3d` we use `-cpp` for output transform), we also use options `-sx 10, --lncc 40, -pad 0`, which include local normalized cross-correlation with standard deviation of the Gaussian kernel of 40, grid spacing of 10 mm along all axes, and padding 0.

Appendix B. Methods settings

This section introduces the settings that are used for all methods.

PCA. We use $\lambda = 0.1$ for the sparse penalty and $\gamma = 0.5$ for the total variation penalty.

ROBEX/CNN. ROBEX and CNN do not require parameter tuning. Therefore, we use the default settings, and for ROBEX we add a seed value of `1` for all datasets.

BET. We use the parameter settings suggested in the literature [8][11] for the IBSR and LPBA40 datasets. For the BRATS and TBI datasets, we choose the option `"-B"` for BET, which corrects the bias field and "cleans-up" the neck.

BSE. We use the parameter settings suggested in the literature [8][11] for the IBSR and LPBA40 datasets. For the BRATS and TBI datasets, we use the default settings.

BEaST. We use the ICBM and ADNI BEaST libraries to run all our experiments. We first normalize the images to the `icbm152_model_09c` template in the BEaST folder. Then, we run BEaST with options `"-fill"`, `"-median"` and with configuration file `"default.1mm.conf"`. The spatial normalization step does not work reliably on the original IBSR and BRATS data. Thus, for these two datasets, we first apply the same affine transform as in our PCA pre-processing and then perform BEaST on the affine aligned images.

MASS. We use the default parameters for MASS and use the 15 anonymized templates, provided with the MASS software package.

References

[1] J. Muschelli, N. L. Ullman, W. A. Mould, P. Vespa, D. F. Hanley, C. M. Crainiceanu, Validated automatic brain extraction of head CT images, NeuroImage 114 (2015) 379–385.

[2] S. M. Smith, Fast robust automated brain extraction, Human Brain Mapping 17 (3) (2002) 143–155.

[3] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, S. M. Smith, FSL, Neuroimage 62 (2) (2012) 782–790.

[4] FMRIB Software Library (FSL), `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki`, accessed: 2018-01-30.

[5] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, R. M. Leahy, Magnetic resonance image tissue classification using a partial volume model, NeuroImage 13 (5) (2001) 856–876.

[6] D. W. Shattuck, R. M. Leahy, Brainsuite: an automated cortical surface identification tool, Medical image analysis 6 (2) (2002) 129–142.

[7] `http://brainsuite.org`, accessed: 2018-01-30.

[8] J. E. Iglesias, C.-Y. Liu, P. M. Thompson, Z. Tu, Robust brain extraction across datasets and comparison with publicly available methods, IEEE Transactions on Medical Imaging 30 (9) (2011) 1617–1634.

[9] ROBEX, `https://www.nitrc.org/projects/robex`, accessed: 2018-01-30.

[10] W. Speier, J. E. Iglesias, L. El-Kara, Z. Tu, C. Arnold, Robust skull stripping of clinical glioblastoma multiforme data, in: MICCAI, Springer, 2011, pp. 659–666.

[11] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, A. Biller, Deep MRI brain extraction: a 3D convolutional neural network for skull stripping, NeuroImage 129 (2016) 460–469.

[12] Deep MRI Brain Extraction, `https://github.com/GUR9000/Deep_MRI_brain_extraction`, accessed: 2018-01-30.

[13] The Internet Brain Segmentation Repository (IBSR) v2.0, `https://www.nitrc.org/projects/ibsr`, accessed: 2018-01-30.

[14] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, A. W. Toga, Construction of a 3D probabilistic atlas of human cortical structures, Neuroimage 39 (3) (2008) 1064–1080.

[15] LPBA40, `http://www.loni.usc.edu/atlases/Atlas_Detail.php?atlas_id=12`, accessed: 2018-01-30.

[16] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, R. L. Buckner, Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults, Journal of Cognitive Neuroscience 19 (9) (2007) 1498–1507.

[17] OASIS, `http://www.oasis-brains.org`, accessed: 2018-01-30.

[18] S. F. Eskildsen, P. Coupé, V. Fonov, J. V. Manjón, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Østergaard, D. L. Collins, A. D. N. Initiative, et al., Beast: brain extraction based on nonlocal segmentation technique, NeuroImage 59 (3) (2012) 2362–2373.

[19] BEAST, `https://github.com/BIC-MNI/BEaST`, accessed: 2018-01-30.

[20] J. Doshi, G. Erus, Y. Ou, B. Gaonkar, C. Davatzikos, Multi-atlas skull-stripping, Academic Radiology 20 (12) (2013) 1566–1576.

[21] MASS, `https://www.med.upenn.edu/sbia/mass.html`, accessed: 2018-01-30.

[22] Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Image processing portal - `https://ipp.cbica.upenn.edu/`, accessed: 2018-01-30.

[23] F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, B. Fischl, A hybrid approach to the skull stripping problem in MRI, NeuroImage 22 (3) (2004) 1060–1075.

[24] R. W. Cox, Afni: software for analysis and visualization of functional magnetic resonance neuroimages, Computers and Biomedical research 29 (3) (1996) 162–173.

[25] Analysis of Functional Neuro Images (AFNI), `https://afni.nimh.nih.gov`, accessed: 2018-01-30.

[26] X. Liu, M. Niethammer, R. Kwitt, M. McCormick, S. Aylward, Low-rank to the rescue–atlas-based analyses in the presence of pathologies, in: MICCAI, 2014, pp. 97–104.

[27] X. Han, X. Yang, S. Aylward, R. Kwitt, M. Niethammer, Efficient registration of pathological images: A joint pca/image-reconstruction approach, in: ISBI, 2017, pp. 10–14.

[28] V. S. Fonov, A. C. Evans, R. C. McKinstry, C. Almli, D. Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, NeuroImage 47 (2009) 39–41.

[29] The Internet Brain Segmentation Repository (IBSR), `https://www.nitrc.org/projects/ibsr`, accessed: 2018-01-30.

[30] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Transactions on Medical Imaging 34 (10) (2015) 1993–2024.

[31] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, C. Davatzikos, Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features, Scientific data 4 (2017) 170117.

[32] M. Brant-Zawadzki, G. D. Gillan, W. R. Nitz, MP RAGE: a three-dimensional, T1-weighted, gradient-echo sequence–initial experience in the brain., Radiology 182 (3) (1992) 769–775.

[33] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, S. Ourselin, Fast free-form deformation using graphics processing units, Computer Methods and Programs in Biomedicine 98 (3) (2010) 278–284.

[34] T. F. Cootes, G. J. Edwards, C. J. Taylor, Active appearance models, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 681–685.

[35] M. Brett, A. P. Leff, C. Rorden, J. Ashburner, Spatial normalization of brain images with focal lesions using cost function masking, NeuroImage 14 (2) (2001) 486–500.

[36] M. Niethammer, G. L. Hart, D. F. Pace, P. M. Vespa, A. Irimia, J. D. Van Horn, S. R. Aylward, Geometric metamorphosis, in: MICCAI, 2011, pp. 639–646.

[37] X. Yang, X. Han, E. Park, S. Aylward, R. Kwitt, M. Niethammer, Registration of pathological images, in: MICCAI SASHIMI workshop, 2016, pp. 97–107.

[38] J. Wright, A. Ganesh, S. Rao, Y. Peng, Y. Ma, Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization, in: NIPS, 2009, pp. 2080–2088.

[39] M. Holmes, A. Gray, C. Isbell, Fast SVD for large-scale matrices, in: Workshop on Efficient Machine Learning at NIPS, 2007, pp. 249–252.

[40] L. I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D: Nonlinear Phenomena 60 (1-4) (1992) 259–268.

[41] M. Modat, D. M. Cash, P. Daga, G. P. Winston, J. S. Duncan, S. Ourselin, Global image registration using a symmetric block-matching approach, Journal of Medical Imaging 1 (2) (2014) 024003–1 – 024003–6.

[42] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, J. C. Gee, N4ITK: improved N3 bias correction, IEEE Transactions on Medical Imaging 29 (6) (2010) 1310–1320.
[43] J. G. Sled, A. P. Zijdenbos, A. C. Evans, A nonparametric method for automatic correction of intensity nonuniformity in MRI data, IEEE Transactions on Medical Imaging 17 (1) (1998) 87–97.
[44] B. C. Lowekamp, D. T. Chen, L. Ibáñez, D. Blezek, The design of simpleitk, Frontiers in Neuroinformatics 7.
[45] X. Liu, M. Niethammer, R. Kwitt, N. Singh, M. McCormick, S. Aylward, Low-rank atlas image analyses in the presence of pathologies, IEEE Transactions on Medical Imaging 34 (12) (2015) 2583–2591.
[46] T. Goldstein, M. Li, X. Yuan, E. Esser, R. Baraniuk, Adaptive primal-dual hybrid gradient methods for saddle-point problems, `arXiv:1305.0546`.
[47] J. Nickolls, I. Buck, M. Garland, K. Skadron, Scalable parallel programming with CUDA, Queue 6 (2) (2008) 40–53.
[48] L. E. Givon, T. Unterthiner, N. B. Erichson, D. W. Chiang, E. Larson, L. Pfister, S. Dieleman, G. R. Lee, S. van der Walt, B. Menn, T. M. Moldovan, F. Bastien, X. Shi, J. Schlüter, B. Thomas, C. Capdevila, A. Rubinsteyn, M. M. Forbes, J. Frelinger, T. Klein, B. Merry, L. Pastewka, S. Taylor, A. Bergeron, N. H. Ukani, F. Wang, Y. Zhou, scikit-cuda 0.5.1: a Python interface to GPU-powered libraries, `http://dx.doi.org/10.5281/zenodo.40565`, accessed: 2018-01-30.
[49] V. Yeghiazaryan, I. Voiculescu, An overview of current evaluation methods used in medical image segmentation, Tech. Rep. CS-RR-15-08, Department of Computer Science, University of Oxford, Oxford, UK (2015).
[50] M. Sonka, J. M. Fitzpatrick, Handbook of medical imaging, Vol. 2, SPIE Publications, 2000.
[51] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society. Series B (Methodological) (1995) 289–300.
[52] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: NIPS, 2014, pp. 2672–2680.
[53] P. C. Mahalanobis, On the generalised distance in statistics, in: Proceedings of the national Institute of Science of India, Vol. 12, 1936, pp. 49–55.
[54] X. Yang, R. Kwitt, M. Styner, M. Niethammer, Quicksilver: Fast predictive image registration-a deep learning approach, NeuroImage 158 (2017) 378–396.
[55] B. Gutierrez-Becker, D. Mateus, L. Peter, N. Navab, Guiding multimodal registration with learned optimization updates, Medical Image Analysis 41 (2017) 2–17.

Dataset: IBSR						
	ROBEX	BEaST*	MASS	BET	BSE	CNN
Dice	4.78e-5	1.20e-2	2.77e-4	4.78e-5	7.55e-5	4.78e-5
Avg Dist	4.78e-5	2.73e-2	1.82e-4	4.78e-5	4.78e-5	4.78e-5
95% Dist	4.74e-5	5.91e-2	1.05e-4	4.71e-5	4.74e-5	4.78e-5
Max Dist	4.78e-5	5.36e-2	4.78e-5	4.78e-5	1.58e-4	5.58e-5
Sensitivity	0.994	0.448	3.40e-3	0.829	4.78e-5	4.78e-5
Specificity	5.58e-5	2.97e-2	2.41e-3	4.78e-5	0.894	1.000

Dataset: LPBA40						
	ROBEX	BEaST	MASS	BET	BSE	CNN
Dice	1.47e-7	2.51e-8	1.89e-3	9.58e-5	2.24e-7	1.85e-8
Avg Dist	1.36e-7	2.51e-8	2.75e-3	1.60e-6	6.31e-7	1.85e-8
95% Dist	2.90e-8	3.29e-7	5.69e-2	2.71e-8	1.02e-5	1.85e-8
Max Dist	2.16e-8	1.000	2.58e-2	2.92e-8	3.01e-5	2.51e-8
Sensitivity	4.13e-3	1.27e-7	1.60e-6	1.000	6.14e-8	1.85e-8
Specificity	5.70e-6	0.998	1.000	2.00e-8	1.000	1.000

Dataset: BRATS						
	ROBEX	BEaST*	MASS	BET	BSE	CNN
Dice	1.58e-4	3.18e-4	7.02e-2	4.78e-5	4.78e-5	4.78e-5
Avg Dist	1.36e-4	2.77e-4	9.89e-2	4.78e-5	4.78e-5	4.78e-5
95% Dist	8.41e-5	4.17e-4	0.266	1.53e-3	4.78e-5	7.15e-5
Max Dist	1.91e-2	7.38e-4	0.222	2.41e-4	1.18e-4	4.78e-5
Sensitivity	3.51e-2	0.981	2.09e-4	8.08e-2	5.58e-5	4.78e-5
Specificity	6.53e-2	1.82e-4	0.999	4.73e-3	0.999	1.000

Dataset: TBI						
	ROBEX	BEaST	MASS	BET	BSE	CNN
Dice	3.91e-3	1.95e-2	2.73e-2	7.81e-3	3.91e-3	3.91e-3
Avg Dist	3.91e-3	1.95e-2	3.91e-2	7.81e-3	3.91e-3	3.91e-3
95% Dist	3.91e-3	7.81e-3	7.81e-3	3.91e-3	3.91e-3	3.91e-3
Max Dist	1.17e-2	9.77e-2	0.344	5.47e-2	3.91e-3	3.91e-3
Sensitivity	0.980	3.91e-3	0.961	3.91e-3	3.91e-3	3.91e-3
Specificity	3.91e-3	1.000	2.73e-2	1.000	0.926	1.000

Table 1: p-values for all datasets, computed by the signed-rank test. We perform a one-tailed paired Wilcoxon signed-rank test, where the null-hypothesis ($\mathcal{H}_0$) is that the paired differences for the results of our PCA model and of the compared method come from a distribution with zero median, against the alternative ($\mathcal{H}_1$) that the paired differences have a non-zero median (greater than zero for Dice, sensitivity and specificity, and less than zero for surface distances). In addition, we use the Benjamini-Hochberg procedure to reduce the false discovery rate (FDR). We highlight, in green, the results where our PCA model performs statistically significantly better. The results show that our PCA model outperforms other methods on most of the measures.

Dataset: IBSR							
	PCA	ROBEX	BEaST*	MASS	BET	BSE	CNN
Dice(%)	97.07 96.99±0.53	95.09 94.98±1.17	96.94 95.07±7.50	92.23 86.76±11.06	95.66 95.16±0.96	95.68 89.54±21.76	84.50 81.10±12.07
Avg Dist(mm)	0.71 0.79±0.27	1.52 1.51±0.56	0.74 1.48±2.76	2.50 4.84±4.54	1.31 1.49±0.47	1.18 4.16±10.53	4.62 5.59±3.10
95% Dist(mm)	2.83 2.84±0.43	4.18 4.50±1.58	3.00 5.19±9.79	8.73 13.66±11.81	3.61 4.22±1.39	5.00 12.27±20.83	20.05 22.25±9.41
Max Dist(mm)	9.30 11.97±8.14	15.55 17.30±8.40	11.74 15.60±13.41	24.20 25.76±12.94	15.68 17.91±7.85	16.57 24.93±23.32	39.10 41.10±8.14
Sensitivity(%)	99.06 98.99±0.46	99.47 99.33±0.54	98.98 96.70±10.12	98.77 98.52±0.77	99.57 99.09±0.93	97.08 88.68±22.86	74.87 70.96±15.89
Specificity(%)	99.48 99.44±0.21	98.89 98.90±0.51	99.48 99.32±0.37	97.88 96.11±3.81	99.12 98.98±0.46	99.58 99.15±1.67	99.85 99.80±0.19
Dataset: LPBA40							
	PCA	ROBEX	BEaST	MASS	BET	BSE	CNN
Dice(%)	97.41 97.32±0.42	96.74 96.74±0.24	96.80 94.25±15.29	97.08 97.03±0.57	96.92 96.70±0.78	96.77 96.29±2.26	95.80 95.70±0.74
Avg Dist(mm)	0.76 0.79±0.12	0.97 0.97±0.07	0.97 3.36±14.83	0.85 0.88±0.21	0.97 1.06±0.27	0.92 1.11±0.81	1.34 1.39±0.37
95% Dist(mm)	2.28 2.27±0.32	2.98 2.97±0.26	2.86 6.45±23.19	2.36 2.50±0.97	3.46 3.92±1.24	2.62 3.46±3.38	3.11 3.56±1.56
Max Dist(mm)	9.73 10.83±3.76	12.89 13.81±3.47	8.80 12.89±24.44	10.71 11.53±4.04	14.44 15.14±3.75	13.61 15.54±7.74	16.91 19.55±8.17
Sensitivity(%)	97.10 96.81±1.23	96.15 96.33±0.85	94.76 92.70±15.12	95.05 95.15±1.08	98.70 98.66±0.54	94.85 94.02±4.10	92.79 92.62±1.46
Specificity(%)	99.62 99.61±0.16	99.54 99.49±0.16	99.80 99.70±0.21	99.86 99.83±0.12	99.09 99.04±0.34	99.81 99.79±0.09	99.85 99.83±0.07
Dataset: BRATS							
	PCA	ROBEX	BEaST*	MASS	BET	BSE	CNN
Dice(%)	96.34 96.16±0.92	95.15 94.83±1.49	94.99 93.29±7.00	96.20 95.71±1.39	94.40 90.95±13.41	84.21 84.91±8.89	1.75 21.89±29.54
Avg Dist(mm)	0.97 1.00±0.31	1.31 1.54±0.70	1.38 2.28±3.34	1.03 1.17±0.56	1.68 7.58±25.30	4.65 4.37±3.61	55.88 44.87±29.05
95% Dist(mm)	4.15 4.35±1.27	5.23 6.03±2.50	5.01 7.71±9.29	4.62 4.87±2.04	5.87 6.18±3.53	13.52 13.92±13.00	78.45 73.85±38.77
Max Dist(mm)	12.03 15.26±9.32	12.83 16.42±8.80	15.03 20.32±14.57	13.29 16.43±8.97	17.49 22.78±22.61	33.95 32.02±22.38	87.73 86.60±36.92
Sensitivity(%)	96.16 96.17±1.84	95.82 94.95±2.88	97.45 97.06±2.66	94.46 93.62±2.87	94.92 94.77±3.82	74.28 77.80±13.43	0.89 16.17±24.73
Specificity(%)	99.33 99.29±0.25	99.16 98.98±0.65	98.68 97.28±5.46	99.73 99.62±0.28	98.90 93.69±22.08	99.82 99.29±2.38	100.00 99.97±0.05
Dataset: TBI							
	PCA	ROBEX	BEaST	MASS	BET	BSE	CNN
Dice(%)	96.40 96.28±0.85	93.71 93.60±1.00	95.04 95.06±0.96	95.11 95.42±0.96	95.40 95.14±1.12	92.64 91.00±4.31	91.51 90.40±5.07
Avg Dist(mm)	1.20 1.22±0.30	2.12 2.20±0.40	1.60 1.62±0.33	1.61 1.53±0.33	1.50 1.57±0.40	2.23 3.15±1.66	3.20 3.46±1.75
95% Dist(mm)	3.37 3.41±0.85	6.20 5.99±0.97	6.24 5.86±1.44	4.90 4.59±0.77	5.06 5.57±1.91	9.46 13.07±7.11	14.97 16.04±8.72
Max Dist(mm)	16.13 15.54±5.03	18.25 18.89±5.12	15.09 17.27±4.60	15.65 16.08±4.16	17.46 18.53±4.59	27.16 31.96±12.71	41.49 37.06±10.09
Sensitivity(%)	97.82 97.76±0.92	98.91 98.64±0.93	92.12 92.89±2.44	98.74 98.68±0.66	93.98 93.65±1.87	87.51 85.65±8.17	85.69 83.77±8.58
Specificity(%)	99.03 99.07±0.26	98.09 97.75±0.93	99.63 99.57±0.22	98.64 98.59±0.47	99.41 99.44±0.11	99.82 99.54±0.86	99.80 99.81±0.07

Table 2: Medians (top), and means with standard deviations (bottom) for validation measures for all the methods and all the datasets. We highlight the best results in green based on the *median* values. Among all datasets, our PCA model has the best median on Dice overlap scores and generally on surface distances. Exception is BEaST which achieves a lower maximum surface distances on the LPBA40 and the TBI datasets. In addition, our model also has the best mean and variance for the Dice overlap scores and the surface distances on most of these datasets.

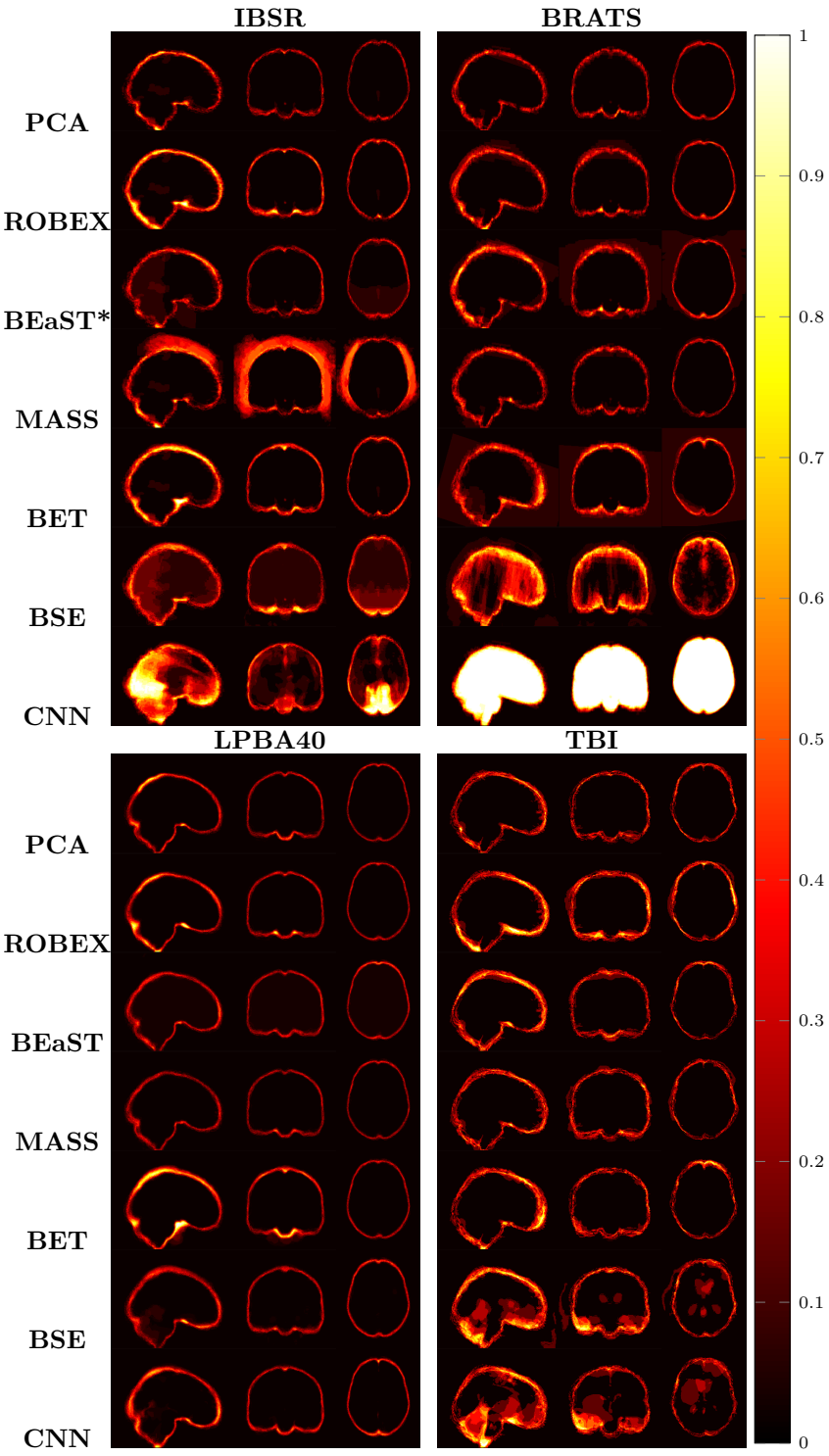


Figure 9: Examples of 3D volumes of average errors for the normal IBSR and LPBA40 datasets, as well as for the pathological BRATS and TBI datasets. For IBSR/BRATS, we show results for BEaST*. Images and their brain masks are first affinely aligned to the atlas. At each location we then calculate the proportion of segmentation errors among all the segmented cases of a dataset (both over- and under-segmentation errors). Lower values are better (a value of 0 indicates perfect results over all images) and higher values indicate poorer performance (a value of 1 indicates failure on all cases). Clearly, BSE and CNN struggle with the BRATS dataset whereas our PCA method shows good performance across all datasets.