Landmark Localization, Feature Matching and Biomarker Discovery from Magnetic Resonance Images

Ricardo Enrique Guerrero Moreno

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
of
Imperial College London

March 2013
Department of Computing
Imperial College London

To my family

Declaration of originality

I hereby declare that the work described in this thesis is my own, except where specifically acknowledged.

Ricardo Enrique Guerrero Moreno

Copyright Declaration

Abstract

The work presented in this thesis proposes several methods that can be roughly divided into three different categories: I) landmark localization in medical images, II) feature matching for image registration, and III) biomarker discovery in neuroimaging.

The first part deals with the identification of anatomical landmarks. The motivation stems from the fact that the manual identification and labeling of these landmarks is very time consuming and prone to observer errors, especially when large datasets must be analyzed. In this thesis we present three methods to tackle this challenge: A landmark descriptor based on local self-similarities (SS), a subspace building framework based on manifold learning and a sparse coding landmark descriptor based on data-specific learned dictionary basis.

The second part of this thesis deals with finding matching features between a pair of images. These matches can be used to perform a registration between them. Registration is a powerful tool that allows mapping images in a common space in order to aid in their analysis. Accurate registration can be challenging to achieve using intensity based registration algorithms. Here, a framework is proposed for learning correspondences in pairs of images by matching SS features and random sample and consensus (RANSAC) is employed as a robust model estimator to learn a deformation model based on feature matches.

Finally, the third part of the thesis deals with biomarker discovery using machine learning. In this section a framework for feature extraction from learned low-dimensional subspaces that represent inter-subject variability is proposed. The manifold subspace is built using data-driven regions of interest (ROI). These regions are learned via sparse regression, with stability selection. Also, probabilistic distribution models for different stages in the disease trajectory are estimated for different class populations in the low-dimensional manifold and used to construct a probabilistic scoring function.

Acknowledgments

I would first of all like to thank my supervisor Daniel Rueckert for his help and support throughout this thesis. His inspiration and constant motivation made sure that working on this project always was both a scientific challenge and a pleasure. I am very grateful to Robin Wolz for many fruitful discussions and his patience with all my questions.

I also want to thank CONACyT, SEP-DGRI and the Rabin Enzra Trust for the funding they have provided, the whole BioMedIA group for a great atmosphere, and finally, my friends and family for constant motivation.

Contents

List of Tables

List of Figures

List of acronyms

Notation

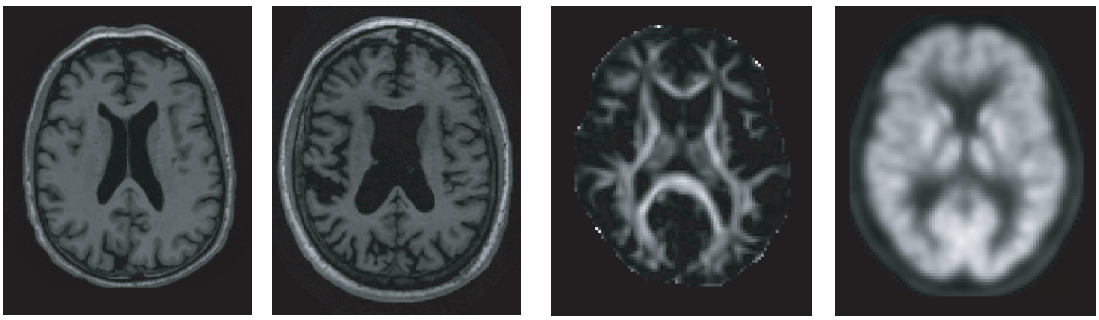
$\mathbf{X}$	Matrix of set of feature vectors $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \in \mathbb{R}^D$
$\mathbf{x}$	Feature vector $\mathbf{x} = \{x_1, ..., x_D\}$, a row in $\mathbf{X}$
x	Single feature, an element of $\mathbf{x}$
D	Feature domain dimensionality
$\mathbf{Y}$	Matrix of low dimensional feature vectors $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\} \in \mathbb{R}^d$ where $d << D$
$\mathbf{y}$	Low dimensional feature vector $\mathbf{y} = \{y_1, ..., y_d\}$, a row in $\mathbf{Y}$
y	Single feature, an element of $\mathbf{y}$
d	Manifold domain dimensionality $d \ll D$
I_i	Image i as matrix
$\mathbf{p}.\mathbf{q}$	Vector of voxel locations $\mathbf{p} = \{p_1, ..., p_D\}$
p_j, q_j	Euclidean space location in image $p_j = \{x, y, z\}$
$I_i(p_j)$	Intensity of image I_i at location p_j
T	Transformation
T^{-1}	Inverse transformation
$T(p)$	Transformed position of location p
$I(T(p))$	Intensity of transformed location p
$\mathbf{l}$	Set of labels $\mathbf{l} = \{l_1, ..., l_n\}$
$\mathbf{K}$	Kernel matrix
P	Image patch

Chapter 1

Introduction

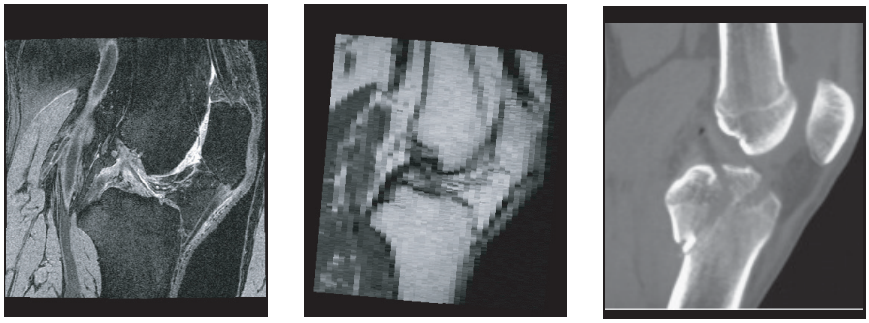
As new imaging techniques are developed and improvements of existing imaging modalities offer lower signal-to-noise ratio, higher resolution and better contrast, three-dimensional (3D) medical imaging offers a huge potential for advances in science and medicine [58]. Different imaging systems for biomedical applications produce mappings of several physical attributes in various ways [208], by measuring either directly or indirectly certain anatomical or physiological properties of tissues. Medical imaging modalities can be broadly divided into ionizing and nonionizing, according to the radiation technique used. Amongst the several medical imaging modalities that exist, some commonly used techniques include:

- Ionizing:
  - (a) X-rays
  - (b) Computed tomography (CT)
  - (c) Positron emission tomography (PET)
  - (d) Single-photon emission computed tomography (SPECT)
- Nonionizing:
  - (e) Magnetic resonance (MR)
  - (f) Optical tomography
  - (g) Ultrasonography (US)



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

Figure 1.1: (a) MR image of a healthy volunteer, (b) MR image of AD patient, (c) fractional anisotropy MR image and (d) PET image of a healthy volunteer, of comparable regions of the brain in different subjects.



(a) (b) (c)

Figure 1.2: (a) MR double echo steady state image, (b) MR turbo spin echo image and (c) CT image, of comparable regions of the knee in different subjects.

3D biomedical imaging systems allow the visualization of specimens in all three dimensions. This not only helps in providing a more complete understanding of the issue, but can also be essential in diagnosing conditions that may not be clearly visible using two-dimensional (2D) imaging systems. Despite this, 2D imaging techniques are still widely used due to their low cost, high resolution, and lower radiation dosages for ionizing modalities.

1.1 Challenges in medical image analysis

Images of the same structure captured with different imaging modalities reveal different aspects of the imaged tissue, and hence produce largely different images. Additionally, anatomical variations between subjects further contributes to image variability. Figures 1.1 and 1.2 show some examples of brain and knee images captured with various imaging

modalities and from different subjects. As can be seen, different imaging techniques reveal substantially different properties of the structures, e.g. MR images give a more detailed contrast in soft tissues, while CT images offer a higher contrast of bony structures. This, coupled with higher levels of noise, low contrast and complex 3D information, among other issues, make the analysis of medical images a challenging task. In many medical imaging applications the assumption is made that image features, especially step changes or edges in intensity, are interesting features [46]. Consider a pair of images $\{I_A, I_B\} \in \mathbb{R}^D$, where all of their voxels, $\mathbf{p}$ and $\mathbf{q}$, have unique labels, $\mathbf{l}_{\mathbf{p}_A}$ and $\mathbf{l}_{\mathbf{p}_B}$, associated to them. Since both images lie in $\mathbb{R}^D$, then, in principle, correspondence between $\mathbf{p}$ and $\mathbf{q}$ can be fully established. However, in practice the full correspondence assumption does not hold. This is in part because voxels at different locations often have very similar intensities, signal degradation due to partial volume effects, and corruption of the signal by imaging noise. Moreover, the imaged features may not present in both images (e.g. due to anatomical variability or pathology). If a full correspondence cannot be established, then a unique spatial correspondence between the images cannot be established. Generally, unique voxel-level mappings cannot be derived based purely on intensities alone [46]. The absence of one-to-one correspondences underlines some of the challenges in medical image analysis.

The analysis of medical images can be broadly subdivided in to three categories: Registration, segmentation and extraction of clinically relevant information. In the following sections we will describe each of these categories.

1.1.1 Registration

Image registration is the process of finding a spatial transformation that maps points from one image to the corresponding points in another image (see Figure 1.3). Chapter 2 gives a more detailed description of registration and its components. Medical image registration has many clinical and research applications [13, 82, 52, 74, 182]. For example, images of the same subject acquired at different times or at the same time but from different views, will have inherently different coordinate changes at different locations. Image registration enables the detection of subtle differences by eliminating the effect of the subject's position

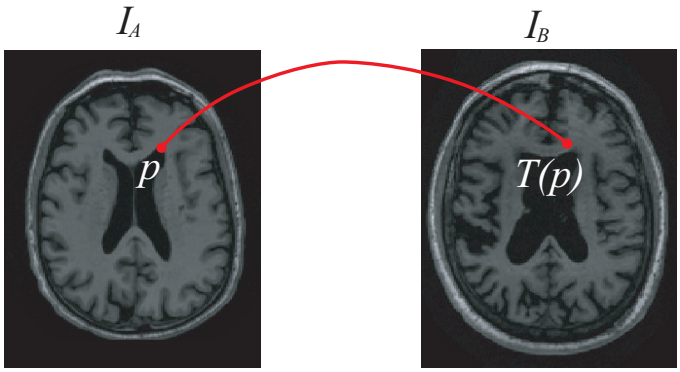


Figure 1.3: Basic registration example. Transformation T maps every point in I_A to I_B.

and orientation. Once the images have been registered they can be subtracted or compared in some other way to visualize and quantify the changes that have occurred [182, 52, 74].

The vastness of applications implies that it is unfeasible to have a single registration method that is optimal for every use. Nonrigid registration is particularly useful when working with medical images, as it is desirable to have images in the same space for analysis and comparison. This usually requires establishing correspondence between different images, due to the fact that tissue may have deformed between taking one image and another, of the same subject (intrasubject registration) or when establishing correspondence between an individual and an atlas, computer model, or another individual (intersubject registration).

The motivation to have more accurate registration algorithms stems from applications in medical imaging, where high accuracy and low uncertainty are very desirable traits. This is because a number of diseases cause subtle changes in the anatomy over time, e.g. dementia patients show changes in different brain structures (the ventricles expand, the hippocampus and the cerebral cortex shrink) while osteoarthritis patients display a gradual degradation of the joints (like the articular cartilage and subchondral bone). Assessing this changes accurately may allow the monitoring of the treatment of patients and the early detection of abnormalities in normal subjects.

1.1.2 Segmentation

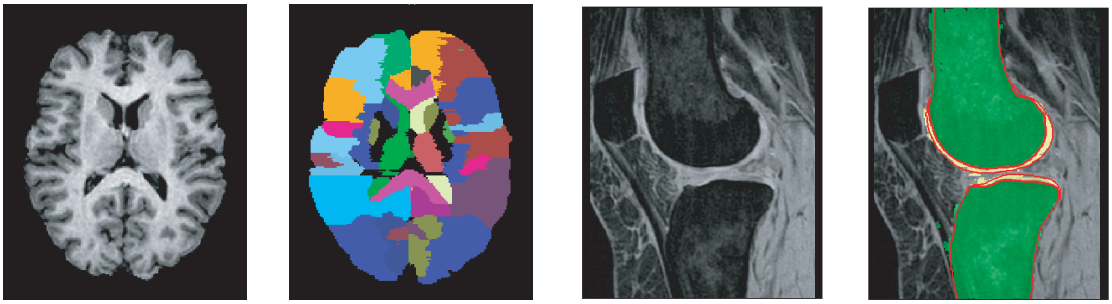
Segmentation refers to the partitioning process of the image domain into regions that correspond to specific anatomical or functional structures and background. Although segmenta-

tion does not form an integral part of the work presented in this thesis, a brief description is included for completeness. The segmentation of medical images is a complicated task and several algorithms have been developed in the field of image processing to tackle it [13, 224, 100, 78, 60]. Figure 1.4 shows an example of a brain MR image segmented into several structural and functional regions, and a knee MR image segmented into anatomical regions. Medical image segmentation is also used to analyze anatomical structures and tissue types, functional regions and regions associated with pathology. Additionally, segmentation can be a useful tool to aid the visualization of structures, e.g. via volume or surface rendering. The segmentation of structures or regions can be achieved by grouping together all voxels that belong to the structure or region [236, 224, 100]. Grouping can be carried out by analyzing intensity, texture or shape features, although other types of attributes might be used. Another way to achieve segmentation is to locate only those voxels that lie on or near the boundary of a structure or region. Typically this is done via edge detection techniques [1, 201].

Since segmentation requires classification of voxels into regions, it is often regarded as a machine learning problem and tackled with learning-based methods [229, 174]. Medical images tend to be highly variable and are often of poor quality, making their structural segmentation from the background very difficult. Moreover, the boundaries between structures or regions may be diffuse, and segmentation techniques often have to rely on prior information in those regions. Machine learning techniques provide an interesting approach to learn the necessary prior knowledge directly from the data.

1.1.3 Extraction of clinically relevant information

In the context of medical image analysis, the definition of what is considered clinically relevant information, can be ambiguous and subjective. In general, it refers to information that aids clinicians to objectively reach a certain diagnosis or conclusion. Types of clinically relevant information can include anatomical landmark locations, volume measurements, or shape analysis, to name a few. Figure 1.5 shows some examples of clinically relevant information extracted from medical images.



(a) (b)

Figure 1.4: Examples of image segmentation: An MR image of a brain segmented into several structural and functional regions (a), and an MR image of a knee segmented into anatomical regions (b).

An intuitive definition of feature points in medical images is that of salient anatomical points. Such points can be defined in the image space by an expert via visual inspection. In this thesis we refer to landmarks (or anatomical landmarks) as such identifiable anatomical points, while reserving the term features for a more generic definition of saliency. Several algorithms have been proposed for the location of anatomical landmarks [217, 108, 136, 134]. Volume measurements generally refer to that of a specific anatomical structure, e.g. the hippocampus or amygdala in the brain [111, 223, 231], or the articular cartilage, femur or tibia in the knee [60, 228]. Volume measurements are often clinically relevant measurements, as they offer an intuitive insight into anatomy. In many cases landmark or volume measurements do not suffice to properly quantify structure variability or pathology. Shape and morphological pattern analysis [71, 200, 235, 66] can model structures or pathologies to track changes across time or within a population. These pattern recognition techniques can be applied to different populations to gain insight into differences between populations as well as diseases.

1.2 Machine learning in medical imaging

Machine learning is a branch of artificial intelligence involving the design of computer systems whose performance can automatically adapt and learn through experience. Such systems can learn to make intelligent decisions based on their recognition and interpretation of complex patterns, with applications including but not limited to stock market analysis, email

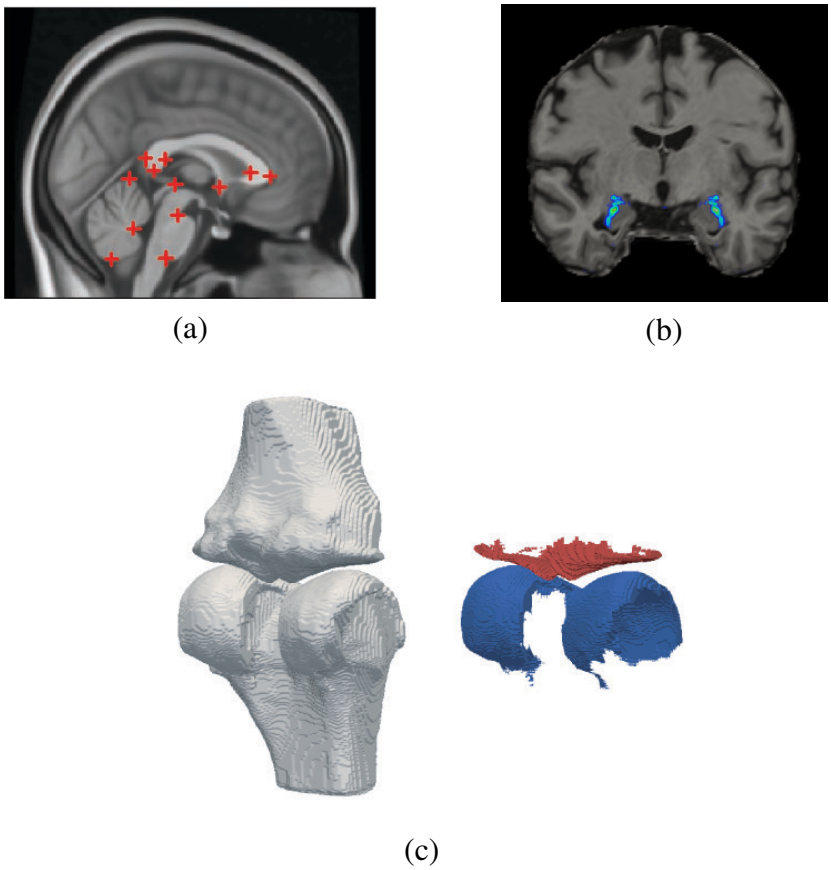


Figure 1.5: (a) Brain anatomical landmarks, (b) regions associated with AD and (c) knee joint/cartilage volume renderings. Best seen in color.

filtering, security, and medical image analysis. These applications involve the use of general models for recognition, diagnosis, planning, prediction, etc. Given a set of empirical pairs of inputs and outputs, machine learning methods have the ability to generate general models from complex patterns that might lay hidden among large amounts of data. This makes machine learning very suitable to solve medical image analysis problems. The main purpose of this thesis is to explore and develop methods based on machine learning that aim at tackling some of the previously mentioned challenges presented by medical image analysis.

1.3 Contributions

The main contributions presented in this thesis can be found in Chapters 4 to 8, and can be divided into three main categories: Landmark localization, registration via feature matching and Alzheimer's disease (AD) biomarker discovery. The contributions are as follows:

- A method is developed for the characterization of landmarks in brain MR images using local structure information in the form of 3D self-similarity (SS) feature descriptors. To locate the position of a landmark in an unseen image, several template images in which the landmark position is known are used for training. Matches between landmarks positions in the template images and the unseen image are found. The final landmark location in the unseen image is estimated using a voting scheme using all templates matches.

- An affine registration framework that builds on the 3D SS feature descriptor developed is presented. Feature correspondences are used to register knee MR images from patients with osteoarthritis, in a more robust and accurate way. Significant improvements in registration accuracy are achieved compared to existing registration approaches.

- A manifold learning method that uses Laplacian eigenmaps to learn a low-dimensional subspace representation of the local anatomy around a specific landmark is presented. The method is applied to brain MR images. The landmark-specific, low-dimensional manifolds are learned using image patches, around the vicinity of the landmark, using brain MR images from the Alzheimers Disease Neuroimaging Initiative (ADNI)[1] dataset. To demonstrate the method's versatility the approach has also been applied to images of the face. Prior knowledge of the spatial distribution of the landmarks is additionally used to reduce the search space, and hence, the size of the manifolds.

- A framework that combines dictionary learning and sparse coding in order to create data-specific feature descriptors is developed. Using a learned dictionary basis to reconstruct image patches using sparse coding, feature descriptors can be specially tailored to represent the image dataset (in this case brain MR images). The method has been used in conjunction with an online learned graphical model to regularize the landmark's location. The combined spatial information from several feature points permits a robust localization. The results demonstrate that landmark localization accuracy is improved when the approach is used in conjunction with a graphical model.

[1] http://adni.loni.ucla.edu

- A framework is developed for feature extraction from learned low-dimensional subspaces that represent inter-subject variability for a group of subjects with AD and a group of matched controls. The manifold subspace is constructed using a data driven region of interest (ROI), defined using an elastic net sparse regression technique. The learned manifold is used to perform classification of the subjects. Classification results improve significantly when using the learned regions compared to anatomical ROI, e.g. the hippocampus. Also, a new metric, the MR imaging disease-state-score (MRI-DSS), is proposed. Results of the proposed approach are shown using the ADNI and ADNI Grand Opportunity (ADNIGO) datasets.

1.4 Outline of the thesis

The thesis is organized as follows: Chapters 2 and 3 introduce the background of the most important techniques used in registration and machine learning, respectively, as they will form an integral part of the methods proposed in this thesis. Chapter 4 proposes a feature descriptor that is based on SS. This feature descriptor is used to find matching landmark locations in unseen brain MR images. In Chapter 5 the SS features introduced in Chapter 4, in combination with random sample and consensus (RANSAC), are used to automatically find matching features in knee MR images using a forward-backward matching algorithm, while at the same time estimating the parameters of an affine transformation model between images. In Chapter 6 manifold learning is used on patches from brain MR images to learn a subspace representation, where a regression function is used to estimate landmarks locations of unseen images. Chapter 7 presents a data specific feature descriptor that is learned specifically to represent the data at hand using dictionary learning. This type of descriptor is used in combination with a graphical model to find matching landmarks between a pair of brain MR images. In Chapter 8 a combination of machine learning techniques are used to derive a continuous metric that aims to act as an imaging biomarker for AD. Using sparse regression with stability selection, relevant features are identified in brain MR images, thus defining a ROI. Manifold learning (on the defined ROI) and nonparametric density estima-

tion are used to model different populations into the mentioned metric. Finally, Chapter 9 contains a discussion and concluding remarks of the work presented in this thesis, as well as potential future work research directions.

Chapter 2

Background: Registration

As mentioned before, image registration maps points in one image (target) to their corresponding points in another image (source). Let us consider each image that is involved in a registration as a coordinate system, which defines a space for that image. Image registration is defined as the estimation of the geometrical transformations [69] which map points from the space of an image I_A to the space of a second image I_B. The transformation T applied to a point $p \in I_A$, produces a transformed point p', such that $p' = T(p)$. If a point $q \in I_B$ corresponds to a point $p \in I_A$, then a successful registration will find the geometric transformation, such that $p' = q$ (Figure 1.3 shows a basic example). In this context correspondence can refer to anatomical or functional correspondence. In Figure 2.1 we can see that image registration consists of different components: Transformation, optimization, interpolation and a similarity (or dissimilarity) metric. A good overview of registration can be found in Sotiras [195] and Hajnal et al. [95]. For an additional overview of available registrations techniques see [238].

2.1 Transformation models

A transformation (or deformation, or spatial mapping) model T describes the relationship between corresponding locations in a pair of images, the target and source. The choice of transformation model is of great importance for the registration process as it entails an important compromise between computational efficiency and flexibility of the transformation

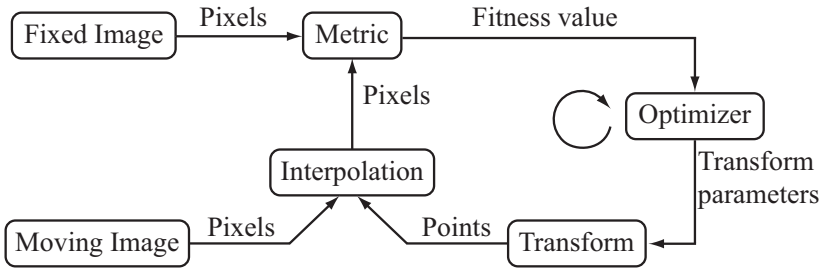


Figure 2.1: The basic components of a registration procedure (fixed and moving image, a transform, a metric an interpolation and an optimizer).

model, e.g. rigid, affine and nonrigid. The most common applications of medical image registration involve aligning pairs of 3D images. The number of parameters that are needed to describe a transformation model and hence need to be estimated through an optimization strategy (discussed in Section 2.3) defines the associated degrees of freedom (DOF) of the registration. The number of parameters hugely varies between transformation models, with the simplest ones requiring 6 (rigid) or 12 (affine) to parametrize global transformations, up to the number of voxels in the image in the case of non-parametric local deformations. The higher the number of parameters, the more descriptive and flexible the model will be, but also the higher the computational cost required to estimate the parameters. The choice of transformation model often implies prior knowledge about the nature of the objects being registered.

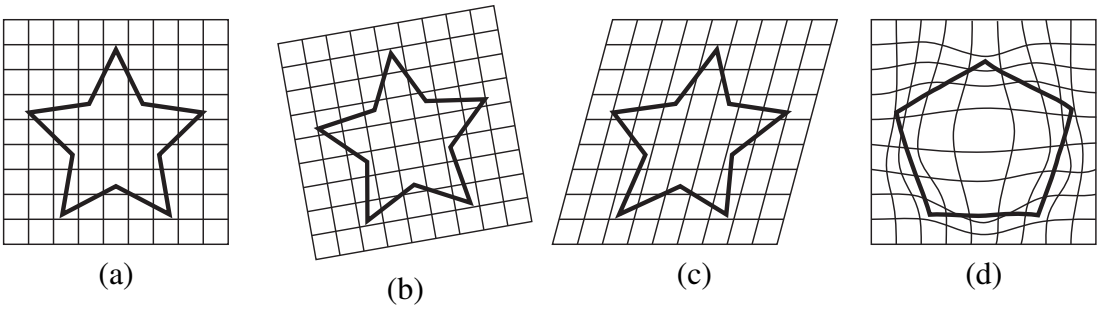


Figure 2.2: Illustration of transformation types: (a) identity, (b) rigid, (c) affine and (d) nonrigid.

Transformation models can be classified as global or local: Global transformation (rigid and affine) preserve the straightness of lines while local deformations (nonrigid) do not. Figure 2.2 depicts a basic example of rigid, affine and nonrigid transformations. Nonrigid deformation models capture more detailed and local variability, but as stated before they

have a higher computational cost associated. Global transformations are of particular use to provide the initial estimates for a nonrigid registration or in applications where intrasubject registrations are performed. In applications where intrasubject registrations are performed, local transformations are generally favored due to their flexibility.

2.1.1 Rigid and affine transformations

Rigid transformations preserve distances, straightness of lines and angles in the space to which they are applied. They therefore only allow for rotations and translations. A rigid transformation T in 3D has six DOF: Three parameters describe the rotation along the three axes and three parameters describe translation along the x, y and z axes. Using orthogonal matrices, rigid transformations can be represented by a 3x3 rotation matrix $\mathbf{R}$ and a 3x1 translation vector $\mathbf{t}$ as $T(p) = \mathbf{R}p + \mathbf{t}$ where $p = (x, y, z)$ and $T(p) = p' = (x', y', z')$. Alternatively, the rotational component of a rigid transformation can be parametrized using the axis-and-angle or quaternion parametrization.

Rigid transformations cannot accommodate shear (skew) or scale deformations. For this, a more general representation, namely affine transformations, is needed. Affine transformations preserve parallel lines as well as the straightness of the lines. An affine transformation in 3D has 12 DOF that describe rotation, scaling, shearing and translation. Figure 2.3 shows an illustration of the type of transformations that can be achieved with both rigid and affine transformation models. As can be seen such models are able to capture only global variation between images and as such they do not fully account the variability present in complex scenes. This shortcoming can be addressed using more flexible, nonlinear transformation models. Some of the most common nonlinear transformation models are described in the following sections.

2.1.2 Polynomial-based deformations

By adding even more DOF to linear models the transformation model can be extended to include nonlinear deformations. An example is the quadratic transformation model, which is defined by second order polynomials and can be expressed as:

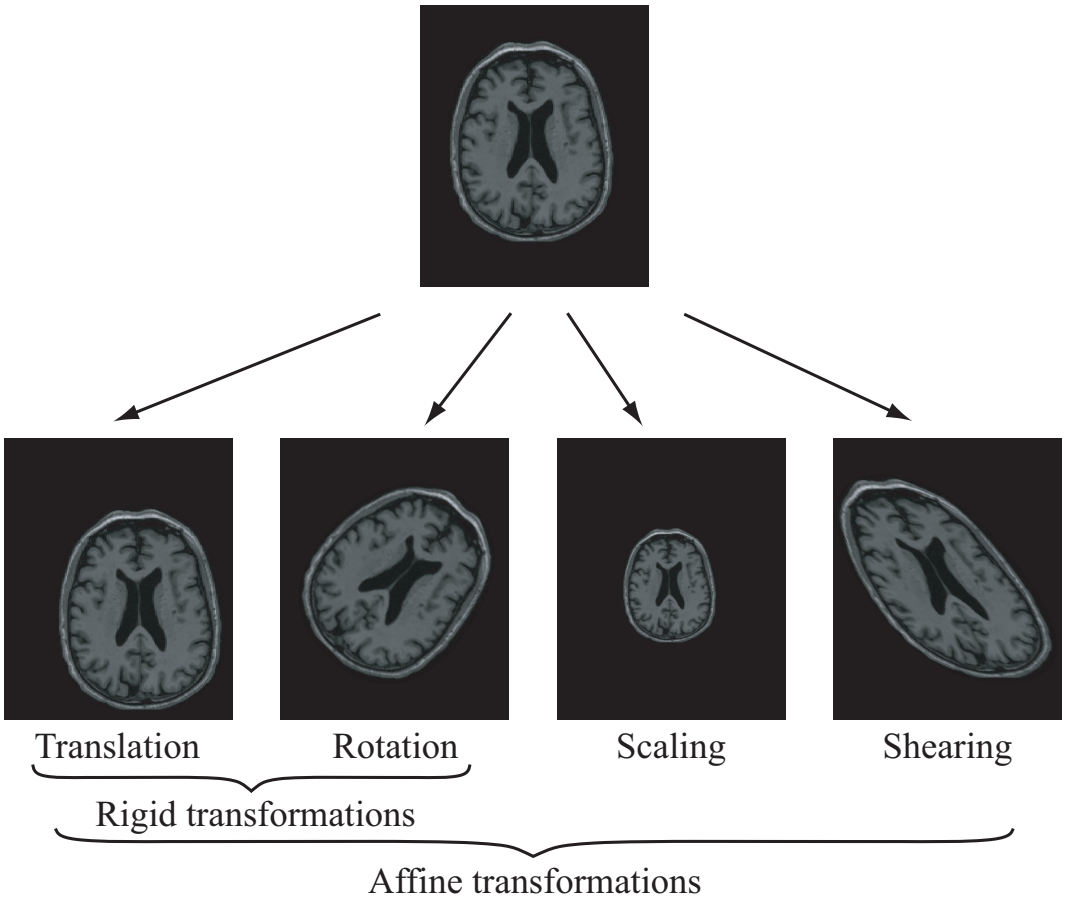


Figure 2.3: Illustration of possible rigid and affine transformations.

$$T(x,y,z) = \begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{19} & a_{1n} \\ a_{21} & \dots & a_{29} & a_{2n} \\ a_{31} & \dots & a_{39} & a_{3n} \\ 0 & \dots & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x^2 \\ y^2 \\ \vdots \\ 1 \end{bmatrix} \tag{2.1}$$

Here n is the order of the polynomial and determines the number of DOF of the transformation, e.g. if $n = 10$ then a 30 DOF quadratic transformation is defined. In a similar way the transformation model can be extended to higher order polynomials. However this is rarely done since higher order polynomials tend to introduce unrealistic oscillations in the transformation model.

2.1.3 Spline-based deformations

The oscillations created by higher order polynomials can be avoided by using piecewise low-order polynomials. Spline-based deformations address this issue and typically require the localization of control points or ”knots“ in both images. The selected points can be corresponding anatomical landmarks [19], or image features [178, 177, 179], or external markers. Another way of defining these points is to simply place evenly spaced points (pseudo-landmarks) throughout the image [188, 53, 182], regardless of whether they are located at anatomically relevant or geometrically salient regions. Most registration techniques using splines assume that this set of control points can be found in both the target and source images. The transformation aims at matching these control points while splines are used to interpolate between them while creating a smoothly-varying displacement field. The interpolating condition can be written as:

$$T(\phi_i) = \phi_i^{'} \quad i = 1, ..., n \, . \tag{2.2}$$

Here n is the number of control points, ϕ_i denotes the location of i-th control point in the target image and $\phi_i^{'}$ represents its corresponding location the source image.

Thin-plate splines

Thin-plate splines are based on radial basis functions. Originally formulated for the surface interpolation of scattered data by Duchon [59] and Meinguet [144], they can be defined as a linear combination of n radial basis functions $\theta(s)$,

$$t(p) = a_1 + a_2 x + a_3 x + a_4 x + \sum_{i=1}^{n} b_j \theta(|\phi_j - p|) \, . \tag{2.3}$$

The transformation model $T(p)$ can be expressed as three separate thin-plate spline functions, one along each dimension, as $T(p) = (t_1, t_2, t_3)^T$, where the coefficients a define an affine component and b the nonrigid component of the deformation model. Thin-plate splines offer the freedom of placing the control points anywhere in the image. However this also implies that the corresponding control points must be first identified. Also, radial basis

functions have an infinite support. This means that each control point has a global effect on the transformation.

B-spline deformations

As mentioned above, a drawback of using radial basis functions is their global support. This means that each control point influences the transformation everywhere in the image. This complicates the modeling of local deformations while at the same increasing the computational complexity required to solve such model. This makes it unfeasible to use the thin-plate spline model with large numbers of control points.

An alternative is to use free-form deformations (FFD), originally proposed by Sederberg and Parry [188] in the computer graphics community. FFD deforms a mesh of regularly arranged control points using local blending functions to produce a smooth transformation. B-spline basis functions have a limited support range as opposed to thin-plate splines. However, control points must be arranged in a regular grid. The limited support of B-spline basis functions means that transformations can be computed efficiently, even for large numbers of control points. A B-spline FFD is defined on the image domain $\Omega = \{(x,y,z)|0 \leq x < X,\ 0 \leq y < Y,\ 0 \leq z < Z\}$. Let Φ denote a n_x x n_y x n_z denote a mesh of control points $\phi_{i,j,k}$ with uniform spacing δ. The FFD can be written as a 3D tensor product of 1D cubic B-splines:

$$T(p) = \sum_{l=0}^{3}\sum_{m=0}^{3}\sum_{n=0}^{3} B_l(u)B_m(v)B_n(w)\phi_{i+l}\phi_{j+m}\phi_{k+n} \tag{2.4}$$

Here $i = \lfloor x/n_x \rfloor - 1,\ j = \lfloor y/n_y \rfloor - 1,\ k = \lfloor z/n_z \rfloor - 1,\ u = x/n_x - \lfloor x/n_x \rfloor,\ v = y/n_y - \lfloor y/n_y \rfloor,\ w = z/n_z - \lfloor z/n_z \rfloor$ and B_l represents the l-th basis function of cubic B-splines [182]:

$$\begin{aligned} B_0(u) &= (1-u)^3/6 \\ B_1(u) &= (3u^3 - 6u^2 + 4)/6 \\ B_2(u) &= (-3u^3 + 3u^2 + 3u + 1)/6 \\ B_3(u) &= u^3/6\,. \end{aligned} \tag{2.5}$$

2.1.4 Physical models of deformation

In this section we will describe physical deformation models, which constrain a deformation field using elastic, fluid or diffusion models. These models provide quantitative, and physically interpretable estimates of 3D deformation fields.

Elastic deformations

Elastic deformations were originally proposed by Bajcsy et al. [12] as a model for matching a brain atlas to CT images of a new subject. Here, the deformation of the source image to the target image is modelled as the physical process of stretching an elastic membrane. As with any elastic material, the process is governed by two forces: an external force that stretches the membrane and an internal force that counteracts any change from its equilibrium state. In this case, the image undergoing deformation is modeled using the Navier-Cauchy equation:

$$\mu\nabla^2\mathbf{u} + (\lambda+\mu))\nabla(\nabla\cdot\mathbf{u}) + \mathbf{b}) = 0 \tag{2.6}$$

Here ∇^2 is the Laplacian operator, μ and λ are Lamé's elasticity constants, $\mathbf{b}$ is the external force applied to the elastic body that drives the registration and $\mathbf{u}$ is the displacement field of a point p. Lamé's elasticity constants can be combined to give Young's modulus and Poisson's ratio of the modeled material.

A popular choice for the external force is the gradient of a similarity measure between the two images. Several similarity measures have been proposed for this purpose, e.g. local intensity correlation [12], intensity differences [35], or intensity features (such as edges and curvature) [83]. Alternatively, similarity measures may be based on distances between corresponding features based on anatomical structures, such as curves [51] and surfaces [205]. Davatzikos [49] proposed an extension to the original elastic registration method that adds spatially-varying elasticity parameters that allow organ-specific elasticity modelling.

Fluid deformations

Elastic deformations are limited by the fact that the deformation energy is increasing proportionally with respect to the applied force $\mathbf{f}$ and hence very localized deformations cannot be properly modeled. Fluid registration techniques allow this constraint to be relaxed by using time as an additional parameter for the deformation. This permits modeling large and localized deformations. This added flexibility increases the risk of misregistration as fluid registration models have a large amount of DOF. Fluid deformations are commonly formulated in a Eulerian reference framework and are modelled by the Navier-Stokes partial differential equation [36] as:

$$\mu\nabla^2\mathbf{v} + (\lambda+\mu))\nabla(\nabla\cdot\mathbf{v}) + \mathbf{f} = 0 \tag{2.7}$$

Here $\mu\nabla^2$ is the Laplacian operator, μ and λ are viscosity constants, $\mathbf{f}$ is the force that drives the registration and $\mathbf{v}$ is the velocity field of a point passing through $\mathbf{x}$. The relation between the displacement field $\mathbf{u}$ and velocity field $\mathbf{v}$ is given by:

$$\mathbf{v} = \frac{\partial\mathbf{u}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{u}\,. \tag{2.8}$$

Christensen et al. [36] estimated the velocity fields iteratively by solving a system of nonlinear partial differential equations through successive over relaxation (SOR). The implementation of this approach is not very computationally efficient and requires a significant amount of time. Bro-Nielsen and Gramkow [27] proposed a faster implementation technique based on a convolution filter in scale-space. However this requires the assumption that the viscosity is constant which is not necessarily the case. Spatially-varying viscosity models have been proposed [131] but their solution require numerical schemes like SOR.

Diffeomorphic flow deformations

Diffeomorphic flows are smooth and invertible transformations that allow connected sets to remain connected, disjoint sets to remain disjoint and preserve the smoothness of features such as curves and surfaces. Large deformation diffeomorphic metric mapping (LDDMM)

[16] is a technique that defines a distance between images or sets of points [117, 141] as a geodesic flow. In a variational framework, LDDMM can be estimated as:

$$\hat{\mathbf{v}} = \underset{\mathbf{v}}{\operatorname{argmin}} \left(\int_0^1 \|L\mathbf{v}_t\|_v^2 \, dt + \frac{1}{\sigma^2} \left\| I_1(T^{-1}) - I_2 \right\|_{L_2}^2 \right) \tag{2.9}$$

Here $\|\cdot\|_v$ is a norm on the space v that constrains the velocity field $\mathbf{v}$ to be smooth (regularization term), L is a differential operator and $\|\cdot\|_{L_2}$ is the L_2 norm of square integrable functions. Choosing an appropriate kernel associated with v allows for modeling of different levels of spatial regularization. The fact that the velocity varies over time allows for the estimation of large deformations. On the other hand, integrating the velocity field over time leads to significant computational and memory costs.

Optical flow and Demons algorithm

Optical flow techniques [104, 15] have been originally developed in the computer vision community as a tool to recover the relative motion of an object and the viewer in between frames in image sequences. The fundamental assumption behind optical flow is that the image brightness is constant over time and hence optical flow represents the distribution of velocities of movement of brightness patterns in an image. For a volumetric image sequence this assumption can be expressed as:

$$I(p,t) = I(x+\delta x, y+\delta y, z+\delta z, t+\delta t) \; . \tag{2.10}$$

Using a Taylor expansion (and ignoring high order terms), Equation (2.10) can be rewritten as

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial z}\frac{dz}{dt} + \frac{\partial I}{\partial t} = 0 \; . \tag{2.11}$$

A more compact representation of Equation (2.11) is

$$\Delta I + \nabla I \cdot \mathbf{u} = 0 \; , \tag{2.12}$$

where ΔI denotes the temporal difference between frames, ∇I is the spatial gradient of the image and $\mathbf{u}$ describes movement between images. Equation (2.12) is ill-posed. Therefore, additional smoothness constraints are added to the motion field $\mathbf{u}$.

A closely related technique to optical flow is the Demons algorithm, originally formulated by Thirion [204] (inspired by Maxwell's Demons from thermodynamics). In his work, Thiron proposes to register two images by first modelling objects boundaries in one image as semi-permeable membranes and then allowing the other image (considered as a deformable grid model) to diffuse through the membranes by the action of Demons placed inside the membranes. The optical flow constraint is used to calculate a Demon force and Gaussian filtering is used as regularization. This algorithm is an effective way to establish dense image correspondences. However, it lacks strong theoretical foundations. Several attempts have been made to give theoretical insight to the Demons algorithm [156, 65].

2.2 Similarity measures

The purpose of the similarity measure is to quantify the degree of alignment between images. The similarity measures can be subdivided into three categories: point-based (features, surfaces or curves), voxel-based (raw intensity) and entropy-based (information theory). There is a wide range of similarity measures to choose from and some of the most popular choices are described in the following sections.

2.2.1 Point-based methods

Assuming that a set of corresponding points can be found in a pair of images, a similarity metric can be defined to act on the alignment of these points. In order for points to be reliable enough for registration, they must be clearly identifiable image features or landmarks. The subject of identifying these points can be interactive (where a user annotates matching landmarks) or automatic (where an algorithm locates anatomical landmarks in a pair of images or extracts feature point descriptors from both images and then matches them). Alternatively, external markers (specifically designed to be easily localizable) can be attached

to the subject before being imaged. For example, markers can be placed directly on bone structures or on the skin. In either case the quality of the registration, will be strongly correlated to the reliability of the matched points (features, landmarks or markers). A more detailed analysis of point-based methods can be found in [115]. Consider a set of n matching points $\{p_i | i \in 1,...,n\}$ from image I_A and $\{q_i | i \in 1,...,n\}$ from image I_B: Any non-zero displacement $T(p_i) - q_i$ between a transformed point $T(p_i)$ and its corresponding point q_i can be viewed as a registration error. A common approach of measuring point misalignment is the root mean squared (RMS) error. Here the aim is to minimize the distance between matching points:

$$S = \frac{1}{n} \sum_{i=1}^{n} w_i^2 \, \|p_i - T(q_i)\|^2 \ , \tag{2.13}$$

where w_i^2 is a weighting factor that relates to point's p_i localization confidence.

Point correspondences can also be used in nonrigid registration. In this case it might be possible to perfectly align all points depending on their distribution and transformation model chosen. For example, if a thin-plate spline model is used then all points can be matched exactly while this is generally not the case if a B-spline model is used. The resulting transformation in other regions of the images (e.g. away from features, landmarks or markers) will depend very strongly on the transformation model used. The exact alignment of points through a nonrigid registration is generally not the most useful approach as the point localizations usually contain a certain amount of error. Thus, the exact alignment of inaccurately located points will result in an erroneous alignment of the images.

2.2.2 Voxel-based similarity metrics

Voxel-based similarity metrics measure differences between intensities or their distribution. If both images have been acquired using the same modality and only differ by image noise the sum of squared of differences (SSD) measure is a good choice [220]. The SSD between images I_A and I_B can be expressed as

$$SSD = \frac{1}{n}\sum_{p}^{n} |I_A(p) - I_B(p))|^2 \,, \tag{2.14}$$

where n is the total number of voxels and p are the voxels in I_A and I_B.

The SSD metric can be very sensitive to outlier values that might arise from subtracting I_A and I_B. To reduce the impact of outliers the sum of absolute differences (SAD) can be used as a similarity measure:

$$SAD = \frac{1}{n}\sum_{p}^{n} |I_A(p) - I_B(p)| \,. \tag{2.15}$$

SSD and SAD make the assumption that images I_A and I_B vary only by Gaussian noise. The correlation coefficient (CC) can be used as a similarity metric under the assumption that the intensities in images I_A and I_B are linearly related. CC can be defined as:

$$CC = \frac{\sum_p (I_A(p) - \bar{I_A})\,(I_B(p) - \bar{I_B})}{\left\{\sum_p (I_A(p) - \bar{I_A})^2 \sum_p (I_B(p) - \bar{I_B})^2\right\}^{1/2}} \tag{2.16}$$

Here $\bar{I_A}$ and $\bar{I_B}$ are the mean voxel values of images I_A and I_B respectively.

2.2.3 Entropy-based metrics

As stated before voxel-based similarity measures like SSD, SAD or CC operate on voxel intensity values and as consequence they are only suitable for mono-modal image registration. Entropy-based metrics measure the amount of shared information between images rather than directly comparing intensity values. This permits images to be registered even when they originate from different imaging modalities (multi-modal registration). A survey on entropy-based medical image registration techniques can be found in Pluim et al. [164].

Joint entropy measures the amount of information in the combined images [189], which can be seen as a measurement of image alignment by constructing the joint histogram of two images. The concept of joint entropy can be visualized using a joint histogram. If the joint histogram is normalized, then an estimate of the joint probability distribution function (PDF) of the intensities in the images can be obtained. The joint entropy of two images is defined

as:

$$H(I_A, I_B) = -\sum_a \sum_b \mathrm{PDF}(a,b) \log \mathrm{PDF}(a,b) \ . \tag{2.17}$$

Mutual information (MI) [39, 219] normalizes the joint entropy with respect to the marginal entropies of the contributing signals. In terms of image registration, this takes into account the change of the marginal entropies of both images as a result of the transformation. Mutual information is defined as:

$$\begin{aligned} MI(I_A, I_B) &= H(I_A) + H(I_B) - H(I_A, I_B) \\ &= \sum_a \sum_b \mathrm{PDF}(a,b) \log \frac{\mathrm{PDF}(a,b)}{\mathrm{PDF}(a)\mathrm{PDF}(b)} \ . \end{aligned} \tag{2.18}$$

Normalized mutual information (NMI) is defined by the ratio between the sum of images I_A and I_B marginal entropies, and their joint entropy. Originally proposed by Studholme et al. [198] it has been shown to be more robust to variations in image overlap:

$$NMI(I_A, I_B) = \frac{H(I_A) + H(I_B)}{H(I_A, I_B)} \ . \tag{2.19}$$

2.3 Optimization

As mentioned previously, the deformation model used in a registration application can vary significantly in its complexity, mainly due to the number of parameters or DOF of the transformation model. The aim of the optimization is to find the transformation parameters that maximize the similarity (or minimize the distance) of the two images. In general, the more complex the transform, the harder it is for the optimization to find the ideal set of parameters. The most general form of the objective function that is optimized in image registration is

$$C = C_{\text{similarity}} - C_{\text{deformation}} \ . \tag{2.20}$$

Here the first term (also known as data term) characterizes the similarity between the source and target image and the second term (regularization, penalty or smoothness term) associates a cost to a particular deformation. When simple transformations are considered, e.g. rigid or affine transformation, the regularization term is often omitted.

The optimization procedure in image registration generally requires iterative methods that gradually minimize a cost function. In the case of point-based rigid registration, the process of finding the optimal solution has a closed formed solution and it is generally known as the ”Orthogonal Procustes“ problem [106]. Another way to estimate the parameters of a rigid or affine transformation model in a point-based registration procedure is using RANSAC [67]. This permits an estimation that is robust against outliers (point localization errors). In broad terms, optimization techniques can be subdivided in to two categories: Continuous and discrete. Figure 2.4 shows a summary of some of the available optimization methods.

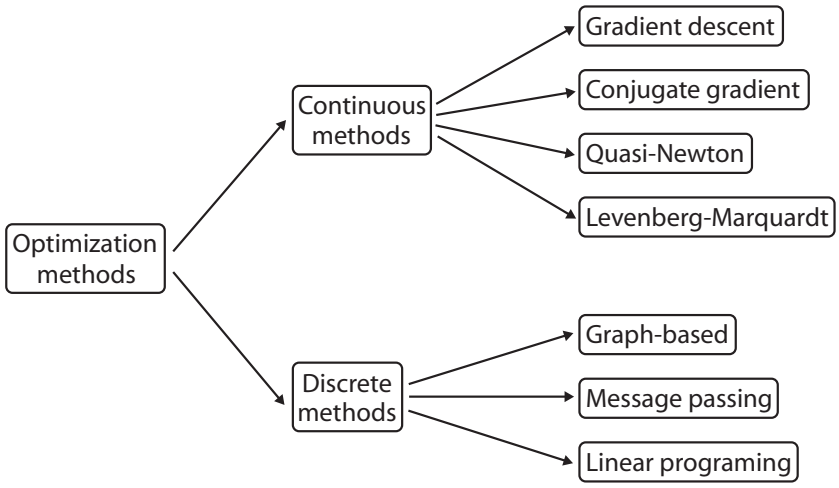


Figure 2.4: Diagram of optimization methods.

2.3.1 Continuous optimization methods

Continuous optimization methods are generally used when the registration/transformation parameters are assumed to be continuous and the associated cost function is differentiable. Assuming that θ is the vector of the transformation parameters, t denotes an iteration index, α_t is a step size and g_t defines a search direction, then continuous optimization iteratively

looks for the ”best“ solution using an update rule of the form: $\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t)$. The α_t and g_t parameters can be defined in several ways. The choice in setting these parameters differentiates between continuous optimization methods. For example, the search direction g can be specified using first-order information. The step size α_t can be constant, change with each iteration or be such that minimizes the objective function. Variations of the behavior of these parameters is what distinguishes between optimization methods. Some continuous optimization approaches commonly used in medical image registration are: Gradient descent, conjugate gradient descent, quasi-Newton and Levenberg-Marquardt methods, to name a few. For a comprehensive overview we refer the reader to [167].

Gradient descent methods traverse the search space in the direction of the negative gradient of the function $g = -\nabla_\theta(\theta)$. Gradient descent optimization is often used in registration algorithms like LDDMM [16], FFD [182] or [30, 137], to name a few. A drawback is that the gradient direction is local, and while it locally decreases the value of the function, it may not be moving in the best direction. For parameter spaces with two or more dimensions, this can lead to a slow convergence to the optimum.

Conjugate gradient descent algorithms guarantee convergence in a finite number of steps for quadratic functions. Instead of moving in the direction of the gradient, it does in the direction of the conjugate gradient. This linearly combines directions of previous steps with the current as $g = f(\nabla_\theta(\theta_t), g_{t-1})$. Some examples of registration methods that use conjugate gradient optimization are [145, 86, 166]. In general conjugate gradient methods converge faster, although computing the conjugate direction is slightly more complicated.

Similar to conjugate gradient descent, quasi-Newton methods accumulate information from the previous iterations in order to achieve better convergence. Their goal is to estimate the inverse Hessian matrix $\mathbf{H}^{-1}(\theta)$ to define the search direction. Thus, the search direction is defined as $g = -\hat{\mathbf{H}}^{-1}(\theta)\nabla_\theta(\theta)$ where the $\hat{}$ denotes the type of approximation that is used. Quasi-Newton optimization methods have been tested in several registration applications [45, 135, 214].

Another optimization method is the Levenberg-Marquardt algorithm. In this method the search direction is given by $g = -\left(\hat{\mathbf{H}}^{-1}(\theta) + \zeta\mathbf{I}\right)\nabla_\theta(\theta)$ where $\mathbf{I}$ is the identity matrix

and ζ is a computational cost vs. stability weight factor. Smaller ζ values achieve greater speeds while higher values increase stability. If ζ equals to zero, then g is the same as in the quasi-Newton algorithm. Some applications of the Levenberg-Marquardt optimization can be found in [234, 128, 11].

2.3.2 Discrete optimization methods

One of the shortcoming of continuous optimization methods is that they usually perform a local search in the parameter space for the optimal solution. This makes such methods sensitive to the initial estimate of the transformation parameters and they can easily get trapped in local minima [195]. The fact that the methods require the computation of the gradient of the cost functions limit their use to cost functions which are differentiable. Furthermore, the assumption that the design parameters are continuous does not always hold. Discrete methods are less sensitive to the initial conditions and often converge faster compared with continuous methods [88]. In many practical problems in engineering the design parameters can be modelled as discrete variables [170], and image registration is no exception. The main limitation of discrete methods is their lack of precision due to the fact that they quantize the search space. Therefore, a trade-off between computational speed and precision exists in discrete methods. If precision is desired a denser sampling of the parameter space is required, however, higher computational costs will be incurred. An additional trait of discrete methods is the possibility to introduce knowledge about the expected location of the solution through the quantification of the parameter space [195]. Discrete optimization techniques can be separated into three classes: Graph-based, message passing and linear programing methods.

Graph-based methods are based on the max-flow min-cut principle [73] that states that the maximum amount of flow that can pass from the source to the sink is equal to the minimum cut that separates the two terminal nodes. The α-expansion optimization technique [22] is a multi-label extension of the maximum a posteriori estimation [90] algorithm. Some applications of medical image registration based α-expansion optimization include [202, 194].

Message passing methods are based on belief propagation [155] where messages are locally exchanged between the nodes of a graph and then backtracking is used to recover the best solution to the problem. Each message conveys the belief of a node to its neighboring node regarding each solution. Belief propagation methods can provide an exact inference for graphs with a chain or tree-shaped topology, however this not the case for graphs that contain loops. In this case loopy belief propagation [77, 151] must be used. Image registration techniques that use message passing optimization can be found in [191, 133].

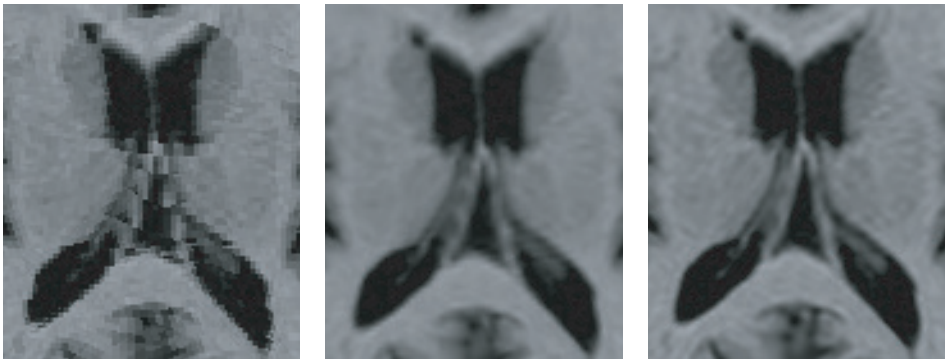
Another class of discrete optimization methods are based on linear programing, which provides better theoretical properties. Linear programing methods avoid solving the original, generally N-P hard problem, in favor of an LP relaxation solution of the problem. Some examples of linear programing based optimization methods in image registration can be found in [125, 126, 87].

2.4 Interpolation

The intensity correspondences between target voxel locations $\{p_1, ..., p_n\}$ and source locations $\{T(p_1), ..., T(p_n)\}$, that were obtained through the estimated transformation $T(p)$, are unlikely to coincide with voxel centers. Consequently, the transformed source image intensities need to be interpolated from the sampled source image values prior to evaluation of the similarity metric.

An ideal interpolation method involves multiplication with a rectangular function in the Fourier domain. This can be realized in the spatial domain by a convolution with the sinc function [129]. From sampling theory we know that sinc interpolation allows loss-less reconstruction. However, the sinc function cannot be applied to real images as it has an infinite support range. Generally, a kernel function that limits the support of the sinc function is used. Different interpolation methods make use of different kernel functions: One of the main trade-offs of different methods is between computational cost and accuracy. One of the simplest methods is based on nearest neighbor interpolation, in which the intensity value that is closest to the transformed location is assigned. Albeit this is a very computationally

efficient procedure, it can lead to data loss or block artifacts. Tri-linear interpolation involves performing a linear interpolation along each dimension. Although this approach can produce more accurate results than nearest-neighbor interpolation, the image is blurred and it is generally slower. Higher-order interpolation methods are used to improve re-sampling quality. A popular type of higher order interpolation is B-splines, which are derived by several self-convolutions of a basis function [211]. Another popular approach is cubic interpolation [120]. This method uses cubic polynomials to construct an interpolation kernel. Figure 2.5 shows some examples of the results produced by some interpolation methods. See Lehmann et al. [129] for a comprehensive overview on interpolation methods for medical image registration.



(a) (b) (c)

Figure 2.5: Results of using different interpolation methods on a ROI of a brain MR image: (a) nearest neighbor, (b) linear and (c) B-spline interpolation.

2.5 Applications of image registration

Applications of image registration in medical imaging can be broadly categorized as intra-subject (registering images of the same subject) or inter-subject (registering images from different subjects) registration.

2.5.1 Intra-subject registration

Longitudinal studies

Subjects can be imaged several times over a time in order to assess disease progression, response to therapy or follow-up. Registering images from the same subject at several different time points allows a quantitative comparison, e.g of tumor growth or shrinkage. For any of these examples, identifying patterns of longitudinal change can provide clinically useful information. Another example is to decide if a particular clinical intervention is appropriate, or to assess whether a patient with a neurodegenerative condition is responding to a particular drug or treatment [5].

Multi-modal image fusion

Multi-modal image fusion can be defined as the process of combining (via registration) information from multiple modalities of the same subject into a single fused image. Image fusion plays an important role in many clinical applications, such as the combination of MR and CT images, to give clear visualization of the relative position of bone and soft tissue for use in surgical planning of the skull base [82]. Another example is the combination of structural information provided by CT or MR images with the functional information provided by PET images [199]. Radiotherapy planning [119] is yet another example, where radiation doses need to be calculated to maximize tumor exposure, while minimizing over all patient radiation exposure [168].

2.5.2 Inter-subject registration

Cohort studies

The great variability of anatomical structures across subjects makes comparisons across populations a very difficult task. Registration-based comparison methods are based on spatial normalization to a common atlas. The development of atlases representing average models of the anatomy are therefore a critical part of cohort studies. Subjects from a population can be registered to as reference subject to create an atlas that captures the particular structural

characteristics of the population. It is often the case that in the atlas' space a group of test subjects is compared to a control group of healthy subjects in order to gain insight about the test group. Furthermore, the creation of atlases of different populations of subjects allows the comparison of typical anatomies for each group. For example, in [229] hippocampal volumes are extracted from healthy subjects and AD patients in order to differentiate them. Davatzikos et al. [52] performed morphological analysis of the corpus callosum to determine structural differences between males and females in an elderly population. Nicolson et al. [152] and Csernansky et al. [47] performed a hippocampal morphometry study in autism and schizophrenia patients, respectively.

Segmentation

Another use of inter-subject registration is image segmentation which refers to the identification of anatomical or functional structures in images. Some of the most common approaches for image segmentation involve the use of either the expectation maximization (EM) algorithm [236, 224, 174] or atlas propagation [13, 100, 229]. In EM segmentation, an image can be registered to an atlas containing prior information about the segmentation. Tissue classes are modeled based on their intensities or other properties and the EM algorithm is used to search for the best possible set of parameters of such model. On the other hand, label propagation segmentation uses one or several presegmented image atlases, each of which is nonrigidly registered to the unseen image. Labels are then inferred from the atlases using a label fusion strategy (e.g. majority voting).

2.6 Evaluation of image registration

As stated before, image registration is based on the identification of corresponding point landmarks or fiducial markers in the two images. Consequently, corresponding landmarks can be used to evaluate an image registration. The fiducial registration error (FRE), can be a useful metric to evaluate registration errors in landmark correspondence. Given a set of N anatomical landmarks $p_1, ..., p_N$ in the target image I_A and their corresponding locations

$q_1, \ldots, q_N$ in the source image I_B, the FRE is calculated as:

$$\mathrm{FRE}^2 \equiv \frac{1}{N} \sum_{i=1}^{N} |T(p_i) - q_i|^2 \ . \tag{2.21}$$

However, the FRE does not directly measure registration accuracy, as changing the positions of the registration landmarks in order to reduce FRE can increase the error in correspondence between other points or structures in the images that did not contribute to the registration. A better measure of registration error is the accuracy with which points that did not contributed to the registration in the two images can be aligned. This error is normally position-dependent in the image and in rigid registration is called the target registration error (TRE). In practical terms, TRE, and how it varies over the field of view, is the most important parameter determining image registration quality. Fitzpatrick [68] describes TRE prediction based on a distribution of identified corresponding points and the estimate of error in identifying correspondence at each point, the fiducial localization error (FLE). The squared expectation value of TRE at position p is then expressed as:

$$\left\langle \mathbf{TRE}(\mathbf{p})^2 \right\rangle \cong \left\langle \mathbf{FLE}^2 \right\rangle \left(\frac{1}{N} + \frac{1}{D} \sum_{i}^{D} \sum_{j \neq i}^{D} \frac{p(i)^2}{\Lambda_i^2 + \Lambda_j^2} \right) \ . \tag{2.22}$$

Here the number of dimensions $D = 3$, Λ are the singular values of the landmark locations, and are related to the distribution of landmarks with respect to the principal axes of the point distribution. Assuming all markers are identified with the same accuracy, the registration error as measured by TRE can be reduced by increasing the number of fiducial markers. If the error in landmark identification or FLE is randomly distributed about the true landmark position, the TRE reduces as the square root of the number of points identified, for a given spatial distribution of points [95].

2.7 Summary

This chapter has presented a review of the methods used as different components of image registration: Transformation, similarity metric, optimization and interpolation. In addition,

applications of medical image registration in longitudinal studies, multi-modal image fusion, cohort studies and segmentation have been discussed. In the following chapter we will discuss the state-of-the-art machine learning techniques for dimensionality reduction, classification and regression.

Chapter 3

Background: Machine learning

Machine learning, techniques and methods are vast and span several fields such as stock market analysis, email filtering, security, and medical image analysis. An in-depth presentation of all methods in this field is neither possible nor of interest here. In this chapter an overview of the main methods in machine learning along with their basic concepts will be presented. We will focus on dimensionality reduction, classification and regression methods as these methods are used throughout this thesis.

3.1 Dimensionality reduction

In this section we briefly describe widely used techniques for dimensionality reduction. We follow the description of manifold learning techniques given by Aljabar et al. [8] and van der Maaten et al. [213].

Manifold learning in general refers to a set of machine learning techniques that aim at finding a low-dimensional representation of high dimensional data while trying to faithfully describe the intrinsic geometry of the data. A simplified schematic overview of manifold learning techniques is given in Figure 3.1. For example, an image can be considered a single point in a very high dimensional space. However, if we consider all images of an anatomical structure like the brain, these images only occupy a small part of this high-dimensional space. Recently, several new manifold learning algorithms have been proposed and applied to solve different problems in the field of medical image analysis such as morphological

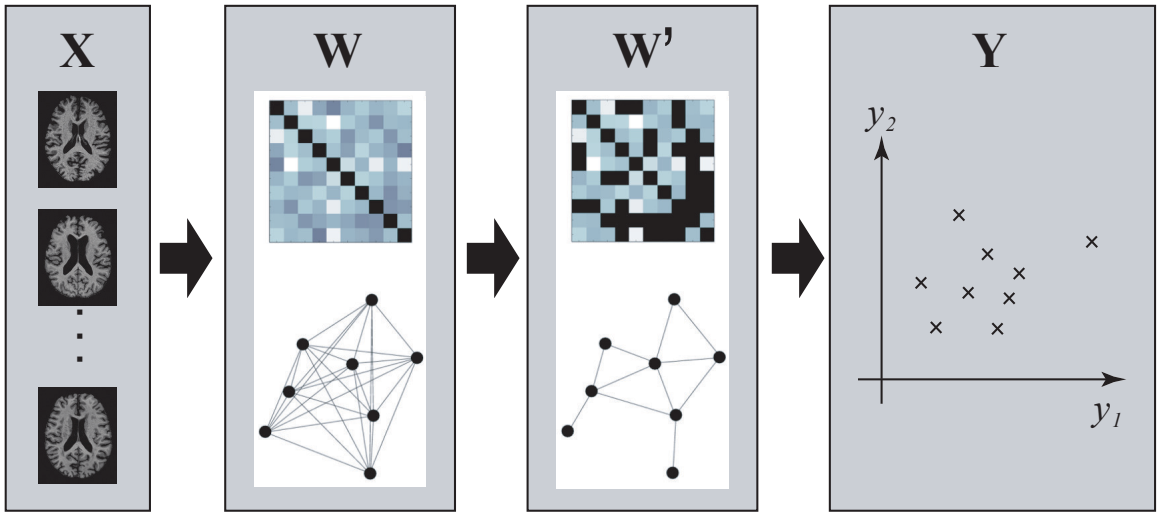


Figure 3.1: An example of manifold learning with brain MR images: The images $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ are compared in pairs and measures of similarity between them are obtained. The measures define a nxn similarity matrix that encodes the edge weights in the graph model representation of the data. The graph/matrix representation may be either full (dense, $\mathbf{W}$) or sparse ($\mathbf{W}'$). Typically, the eigenvalue-eigenvector structure of the matrix $\mathbf{W}$ (or $\mathbf{W}'$) is used to derive a coordinate representation for an embedded manifold representation $\mathbf{y}_i$ of the original data. Only two dimensions of $\mathbf{y}_i$ are shown above.

analysis [7], segmentation [229], landmark localization [92] and classification [230].

In the following we give a more formal description of manifold learning: Consider a set of N images $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \in \mathbb{R}^D$ with $i \in \{1, 2, ..., N\}$, where each image $\mathbf{x}_i$ is arranged as a vector of its voxel intensities and D is the number of voxels per image. Assuming that images $\mathbf{x}_1, ..., \mathbf{x}_N$ lie on or near an d-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$, it is possible to learn a low-dimensional representation of the input images in $\mathcal{M}$, such that $f : \mathbf{X} \rightarrow \mathbf{Y}$, $\mathbf{y}_i = f(\mathbf{x}_i)$ with $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\} \in \mathbb{R}^d$, where $d << D$.

Several manifold learning techniques use a graph to represent the relation between pairs of data points. In the following we assume that the data points are images. The graph in turn, may be viewed as a representation where each node is an image and the weight of each edge denotes the similarity or distance between the image pair it joins. Manifold learning techniques can be separated into two broad categories: Methods that use a fully connected graph to model the relations among data points and methods that use a sparse representation of the graph with a smaller number of edges, around local neighborhoods. In general, different manifold learning techniques seek to optimize different criteria as functions of the matrix

representation (see Figure 3.1). Many of the manifold learning techniques can be described as spectral since the optimization is often carried out using the eigenvalue-eigenvector structure of the associated matrix.

3.1.1 Dense spectral techniques

This section describes dense (or full spectral) techniques for manifold learning. These methods use a full matrix that measures all pairwise relations between data points to learn their low-dimensional representation.

Principal component analysis (PCA)

PCA [116] aims to build a low-dimensional representation of the data to describe as much of the variance in the data as possible using only a few principal components. This is done by finding a linear basis of reduced dimensionality for the data, in which the amount of variance in the data is maximal. The problem is described as finding the linear mapping function $\mathbf{M}$ that optimizes the objective function

$$\max_{\mathbf{M}} \operatorname{trace}\left(\mathbf{M}^T \operatorname{cov}(\mathbf{X})\,\mathbf{M}\right) \tag{3.1}$$

where $\operatorname{cov}(\mathbf{X})$ is the sample covariance matrix of $\mathbf{X}$. The linear mapping is defined by the d principal eigenvalues λ of the eigenproblem

$$\operatorname{cov}(\mathbf{X})\,\mathbf{M} = \lambda\mathbf{M}\,. \tag{3.2}$$

Using the mapping function $\mathbf{M}$, the low-dimensional space is defined as $\mathbf{Y} = \mathbf{XM}$.

Kernel PCA

Kernel PCA [185] is a dimensionality reduction technique that uses the “kernel trick“ [4] to formulate a nonlinear extension of classic PCA. In Kernel PCA the principal eigenvectors are computed using the kernel matrix, rather than the covariance matrix as in PCA. The kernel matrix $\mathbf{K}$ is defined from the data points in D-dimensional space with

$$\mathbf{k}_{ij} = \kappa\left(\mathbf{x}_i, \mathbf{x}_j\right), \tag{3.3}$$

where κ is a kernel that can be any function that results in a positive-semidefinite kernel matrix $\mathbf{K}$. A centering operation is performed on the new features so that they have zero-mean. The d principal eigenvectors $\mathbf{v}_i$ and eigenvalues λ_i of $\mathbf{K}$, can be computed using the relation between the eigenvectors $\mathbf{a}_i$ of the associated covariance matrix and the eigenvectors $\mathbf{v}_i$ of the kernel matrix through

$$\mathbf{a}_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{v}_i. \tag{3.4}$$

The low-dimensional embedding of image $\mathbf{x}_i$ is then defined through a projection onto the eigenvectors $\mathbf{a}_i$ of the covariance matrix as

$$\mathbf{y}_i = \left\{ \sum_{j=1}^{N} \mathbf{a}_1^{(j)} \kappa\left(\mathbf{x}_j, \mathbf{x}_i\right), \ldots, \sum_{j=1}^{N} \mathbf{a}_d^{(j)} \kappa\left(\mathbf{x}_j, \mathbf{x}_i\right) \right\}, \tag{3.5}$$

where $\mathbf{a}_i^{(j)}$ is the j-th entry of vector $\mathbf{a}_i$.

Multidimensional scaling (MDS)

MDS [44] is a linear technique closely related to PCA. It is based on a distance matrix $\mathbf{W}$ with w_{ij} representing the distance between two data points $\mathbf{x}_i$ and $\mathbf{x}_j$. MDS seeks to find the low-dimensional representation that best preserves the pairwise distances in the high-dimensional space. This is carried out by minimizing the objective function

$$\phi(\mathbf{Y}) = \sum_{ij} \left(w_{ij}^2 - \left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2 \right), \tag{3.6}$$

where $\left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2$ is the Euclidean distance between two data points in d-dimensional space, $d \ll D$. The optimal embedding for Equation (3.6) can be obtained through the eigendecomposition of the Gram matrix $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ of the data in the D-dimensional space. There is a duality between PCA and MDS when similarities w_{ij} are measured by the Euclidean distance [44].

Isomap

Isomap [203] is a nonlinear embedding technique that builds upon the MDS approach. Note that MDS does not take into account the distribution of the neighboring data points. For instance, if the high-dimensional data lies on or near a curved manifold, MDS might treat data points as being close, even if their distance in the manifold is large. In Isomap pairwise distances (or weights) w_{ij} are not calculated directly between data points $\mathbf{x}_i$ and $\mathbf{x}_j$ but using a neighborhood graph G that connects all N data points. This graph is build by either connecting all data points $\mathbf{x}_i$ to its k closest neighbors or to all subjects within some radius ε. After constructing G, the weights w_{ij} are estimated as the shortest path distances w_{ij}^G within the graph. The final embedding coordinates $\mathbf{y}_i$ are obtained by applying classical MDS to the distance matrix $\mathbf{W}^G = \left\{ w_{ij}^G \right\}$.

3.1.2 Sparse spectral techniques

In this section, some of the available sparse techniques for manifold learning are described. These techniques focus on retaining the local similarities measured in the input space via the solution of a sparse (generalized) eigenproblem.

Locally linear embedding (LLE)

A low-dimensional manifold constructed with LLE [180] aims to preserve the local neighborhoods of the high-dimensional data in the low-dimensional learned space. LLE is similar to Isomap considering that both approaches construct a graph representation of the data points. However, Isomap solely attempts to preserve the local properties of the data, assuming locally linear relationship between neighboring data points. It represents every data point $\mathbf{x}_i$ as a weighted combination of its k nearest neighbors in the high-dimensional space. This defines a set of weights w_{ij} for the k neighbors of $\mathbf{x}_i$ and the aim is to find a low-dimensional representation $\mathbf{y}_i$ that respects this weighting. The LLE objective function is defined as:

$$\phi(\mathbf{Y}) = \sum_i \left\| \mathbf{y}_i - \sum_{j=1}^{k} w_{ij} y_{ij} \right\|^2 \text{ subject to } \left\| \mathbf{y}^{(k)} \right\|^2 = 1 \text{ for } \forall k \,, \tag{3.7}$$

where $\mathbf{y}^{(k)}$ represents the kth column of the solution matrix $\mathbf{Y}$. The constraint on the covariance of the columns of $\mathbf{Y}$ is required to exclude the trivial solution $\mathbf{Y} = 0$. Using a sparse weight matrix $\mathbf{W}$, it can be shown that the embedding can be obtained from the d eigenvectors corresponding to the smallest nonzero eigenvalues of $(\mathbf{I} - \mathbf{W})^T (\mathbf{I} - \mathbf{W})$ [180].

Hessian LLE

Using the same concept of local linearity, Hessian LLE [57] minimizes the curvature of the high-dimensional manifold when learning the low-dimensional representation. The method constrains the distances in both spaces to be locally isometric. Applying PCA to every data point $\mathbf{x}_i$ and its k nearest neighbors gives an approximation of the local tangent space at every data point. The mapping function $\mathbf{M}$ obtained from the d principal components at every point $\mathbf{x}_i$ is then used to obtain an estimator for the Hessian $\mathbf{H}_i$ of the manifold at that data point [57]. From the Hessian estimators in tangent space, a matrix $\mathcal{H}$ is constructed with entries:

$$\mathcal{H}_{lm} = \sum_i \sum_j \left((\mathbf{H}_i)_{jl} \times (\mathbf{H}_i)_{jm} \right) \,. \tag{3.8}$$

The eigenvectors that correspond to the d smallest eigenvectors of $\mathcal{H}$ are used to define the low-dimensional embedding $\mathbf{Y}$ that minimizes the curvature of the manifold.

Laplacian eigenmaps

Laplacian eigenmaps can be used to find a low-dimensional representation of the data while preserving the local geometric properties of the manifold [17]. Laplacian Eigenmaps uses a local neighborhood graph to approximate geodesic distances between data points. This graph is defined by either connecting every data item $\mathbf{x}_i$ to its k closest neighbors or to all subjects within some fixed radius ε. From these distances a sparse neighborhood graph G is constructed. Furthermore, a weight matrix $\mathbf{W}$ that assigns a value to each edge conect-

ing points $\mathbf{x}_i$ and $\mathbf{x}_j$ in G (zero elsewhere) according to the distance between the points is computed using a Gaussian heat kernel:

$$w_{i,j} = \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2}\right) . \tag{3.9}$$

Here σ is the standard deviation of the Gaussian kernel. Laplacian eigenmaps aims to place points $\mathbf{x}_i$ and $\mathbf{x}_j$ close together in the low-dimensional space if their weight $w_{i,j}$ is high, e.g. if they are close in the original, high-dimensional space. This is done by means of minimizing the cost function given by

$$\phi(\mathbf{Y}) = \operatorname{argmin} \sum_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 w_{i,j} , \tag{3.10}$$

under the constraint that $\mathbf{y}^T\mathbf{D}\mathbf{y} = 1$ which removes an arbitrary scaling factor in the embedding and prevents the trivial solution where all $\mathbf{y}_i$ are zero. The minimization of Equation (3.10) can be formulated as an eigenproblem [10] through the computation of the degree matrix $\mathbf{M}$ and the Laplacian $\mathbf{L}$. The degree matrix $\mathbf{M}$ is a diagonal matrix that contains information about the degree of each vertex of $\mathbf{W}$, where $m_{i,i} = \sum_j w_{i,j}$ and the Laplacian $\mathbf{L} = \mathbf{M} - \mathbf{W}$. Hence the low-dimensional manifold $\mathbf{Y}$ that represents all the data points can be obtained via solving a generalized eigenproblem

$$\mathbf{L}\nu = \lambda\mathbf{M}\nu , \tag{3.11}$$

where ν and λ are the eigenvectors and eigenvalues, and in turn the d eigenvectors ν corresponding to the smallest (non-zero) eigenvalues λ represent the new coordinate system.

3.1.3 Summary

Some of the advantages of the dense spectral techniques is that they can achieve a more faithful representation of the data's global structure and that their metric-preserving properties are better understood theoretically [54]. Sparse spectral techniques have two main advantages: Computational efficiency as they involve only sparse matrix computations which may

yield a polynomial speedup and representational capacity as they may give useful results on a broader range of manifolds, whose local geometry is close to Euclidean, but whose global geometry may not be [54]. An additional consideration that must be taken into consideration when choosing a dimensionality reduction technique is that some application require the mapping of new points into the learned manifold: Linear dimensionality reduction, such as PCA, provides a projection matrix for exact transformation between the original and the embedded space. This is not the case for most non-linear methods and approximation techniques must be used. Bengio et al. [18] proposed an out of sample embedding technique, that employs the Nyström approximation [163], for dimensionality reduction techniques that rely on an eigendecomposition.

3.2 Classifiers

Classifiers are a group of machine learning techniques that aim to predict group membership of data instances. Constructing a general model based on training data, for which the group membership is known, to infer membership (class) of unseen data instances (or samples) is known as supervised learning. Alternatively, if no training samples are available clustering techniques (unsupervised learning) can be used to determine sample class membership. Here we will focus on supervised learning techniques as the bulk of work presented in this thesis relies on them.

Consider a set of feature vectors for training, $\mathbf{X} = \{\{\mathbf{x}_1, l_1\}, ..., \{\mathbf{x}_N, l_N\}\}$ where each sample $\mathbf{x}_i$ has a known class l_i. The classification problem is then to find a good prediction function, given a set of observed features $\mathbf{X}$ for the class label l_i of any sample belonging to same distribution as $\mathbf{X}$, e.g. our training dataset is of the form:

$$\{\mathbf{x}_i, l_i\} \quad \text{where} \quad i = 1, ..., N, \;\; l_i \in \{k_0, ..., k_m\}, \;\; \mathbf{x} \in \mathbb{R}^D . \tag{3.12}$$

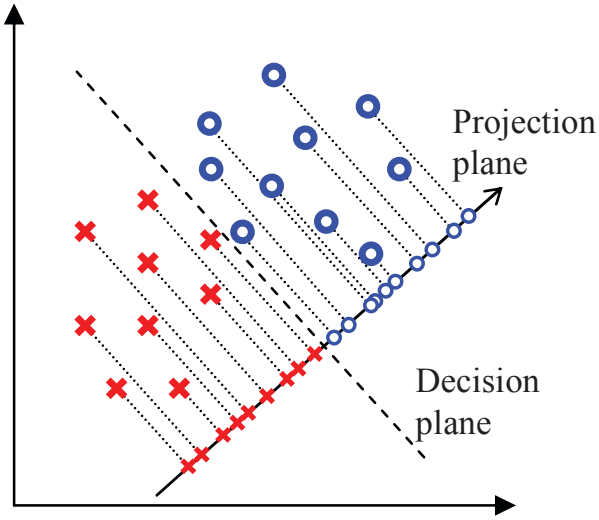


Figure 3.2: Decision and projection planes from linear discriminant analysis. Feature vectors belonging to class k_0 are shown in red, and those belonging to class k_1 are shown in blue. The decision plane is defined by its orthogonality to the projection plane.

3.2.1 Linear and quadratic discriminant analysis

In discriminant analysis the aim is to find an optimal low-dimensional space such that when data points are projected, data from different classes are well-separated (see Figure 3.2). This method maximizes the ratio of between-class variance to within-class variance for any particular data set thereby guaranteeing maximal separability.

Let us assume a normal distribution of the class-likelihood density functions $f_k(\mathbf{x})$, with mean and covariance parameters $(\boldsymbol{\mu}_k, \Sigma_k)$ for class k. Under this assumption, the Bayes optimal solution is to predict the class-likelihood, which can be done by maximizing the posterior probability:

$$\begin{aligned} \hat{l}(\mathbf{x}) &= \underset{k}{\operatorname{argmax}} \; p(k|\mathbf{x}) \\ &= \underset{k}{\operatorname{argmax}} \; f_k(\mathbf{x})\pi_k \, . \end{aligned} \tag{3.13}$$

Here π_k is the class prior probability and is defined as the ratio between the number of samples in class k and the total number of samples. Defining $f_k(\mathbf{x})\pi_k$ as the linear discriminant function $\delta_k(\mathbf{x})$, then the two class decision boundary between classes k_0 and k_1 is

defined as

$$\{\mathbf{x} : \delta_{k_0}(\mathbf{x}) = \delta_{k_1}(\mathbf{x})\} \ . \tag{3.14}$$

If no assumption is made about the covariances Σ_{k_0} and Σ_{k_1}, of distributions $f_{k_0}(\mathbf{x})$ and $f_{k_1}(\mathbf{x})$, then the resulting classifier is known as quadratic discriminant analysis (QDA) [99]. Simplifying the relationship in Equation (3.14) by making the assumption that $\Sigma_{k_0} = \Sigma_{k_1} = \Sigma$ yields a linear discriminant analysis (LDA) classifier [99]. Then, without loss of generality, it can be shown that the relationship in Equation (3.14) can be expressed as

$$\log \frac{\pi_{k_0}}{\pi_{k_1}} - \frac{1}{2} \left(\boldsymbol{\mu}_{k_0} + \boldsymbol{\mu}_{k_1}\right)^T \left(\boldsymbol{\mu}_{k_0} - \boldsymbol{\mu}_{k_1}\right) + \mathbf{x}^T \Sigma^{-1} \left(\boldsymbol{\mu}_{k_0} - \boldsymbol{\mu}_{k_1}\right) = 0 \ . \tag{3.15}$$

3.2.2 Support vector machines

Support vector machines (SVM) where originally proposed by Vapnik and Lerner [215] as a two-class linear classifier. A SVM aims to construct a hyperplane that maximizes the margin between the hyperplane and the closest points (support vectors) on either side of the boundary. Reformulations of the original SVM to deal with data that are not linearly separable have also been proposed: Cortes et al. [42] give a soft-margin SVM formulation that allows for mislabeled data. Also, making use of the "kernel trick" [4], Boser et al. [21] developed nonlinear SVM classifiers. The following sections will describe in more detail these three formulations of SVM.

Linear SVM

Consider a set of N training samples $\mathbf{x}_i$ where each sample is of dimensionality D and has an associated binary label l_i. That is, the training samples can be expressed as:

$$\{\mathbf{x}_i, l_i\} \quad \text{where} \quad i = 1 \ldots N, \ l_i \in \{-1, 1\}, \ \mathbf{x} \in \mathbb{R}^D \ . \tag{3.16}$$

Let us assume that the data is linearly separable, that is, there is a separating surface of the form $y(\mathbf{x}) = \mathbf{x} \cdot \mathbf{w} - b = 0$ that perfectly separates both classes, where $\mathbf{w}$ is normal to the

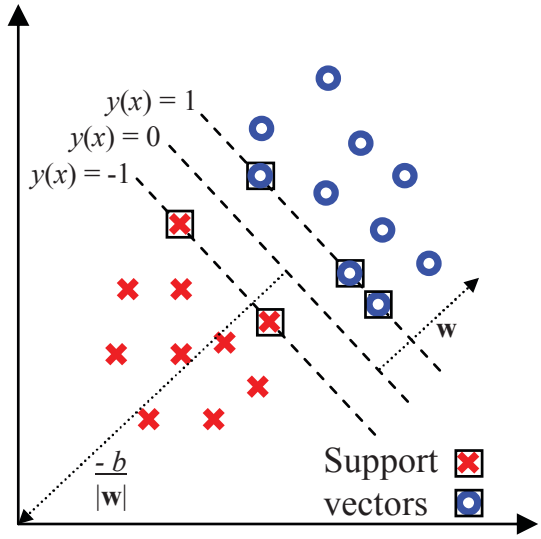


Figure 3.3: 2-D illustration of maximum-margin hyperplane and margins for a linear SVM.

separating surface and $\frac{b}{\|\mathbf{w}\|}$ is the perpendicular distance from the separating surface to the origin. The support vectors are the samples that are closest to this surface and, as can be seen in Figure 3.3, they lie on the planes $y(\mathbf{x}) = 1$ and $y(\mathbf{x}) = -1$. The values of $\mathbf{w}$ and the threshold b are then chosen as to maximize the distance between the support vectors and the separation surface, which can be expressed as

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \quad \text{for } y_i = +1 \\ \mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \quad \text{for } y_i = -1 \end{aligned} \tag{3.17}$$

or equivalently as:

$$y_i(\mathbf{x}_i \cdot w + b) - 1 \geq 0 \ \forall_i \ . \tag{3.18}$$

Therefore, the margin maximization can be expressed as a constrained optimization problem of the form:

$$\min_{\mathbf{w}} \|\mathbf{w}\| \quad \text{s.t.} \quad y_i(\mathbf{x}_i \cdot w + b) - 1 \geq 0 \ \forall_i \ . \tag{3.19}$$

Alternatively, one could minimize $\frac{1}{2}\|\mathbf{w}\|^2$ instead of $\mathbf{w}$ in order to allow the use of

quadratic programing optimization. Hence, Equation (3.19) can be rewritten in the form:

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{s.t.} \quad y_i(\mathbf{x}_i \cdot w + b) - 1 \geq 0 \ \forall_i \ . \tag{3.20}$$

Adding Lagrange multipliers α to the constraints to force that no feature vectors lie within the margin, Equation (3.20) may be rewritten as

$$\min_{\mathbf{w},b} \max_{\alpha} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i(\mathbf{x}_i \cdot \mathbf{w} - b) - 1 \right] \right\} \quad \text{s.t.} \quad \alpha_i \geq 0 \ \forall_i \ . \tag{3.21}$$

Differentiating Equation (3.21) with respect to $\mathbf{w}$ and setting it to zero, allows us to find the value of $\mathbf{w}$ that maximizes the equation:

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \ . \tag{3.22}$$

The solution for b can be found by averaging over the support vectors N_{sv},

$$b = \frac{1}{N_{sv}} \sum_{i=1}^{N_{sv}} (\mathbf{x}_i \cdot \mathbf{w} - y_i) \ . \tag{3.23}$$

Substituting equations (3.22) and (3.23) in Equation (3.21), the dual form of the Lagrangian can be expressed as

$$\begin{aligned} \max_{\alpha} \tilde{L}(\alpha) = \max_{\alpha} & \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right\} \\ & \text{s.t.} \quad \alpha_i \geq 0 \quad \forall_i \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \ , \end{aligned} \tag{3.24}$$

which expresses an optimization criterion based only on terms of inner products of the feature vectors.

Soft margin SVM

Soft margin SVM is a reformulation of linear SVM that allows the handling of non-linearly separable data. This is done by relaxing the constraints imposed on linear SVM by adding

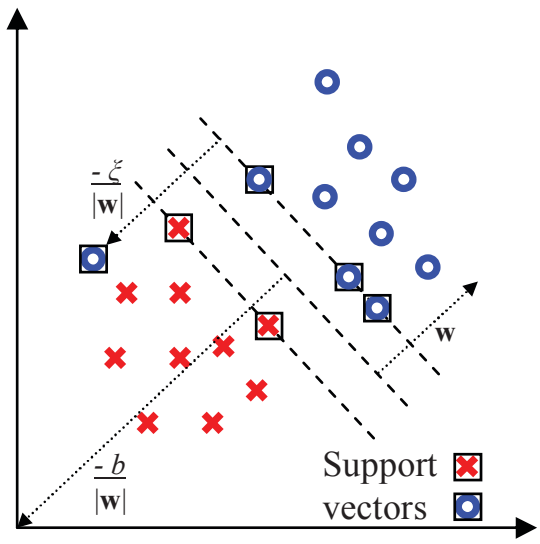


Figure 3.4: 2-D illustration of maximum-margin hyperplane and margins for a soft margin SVM. The slack parameter ξ measures the degree of misclassification.

slack variables ξ_i, $i = 1, \ldots, N$ that measures the misclassification of points. Figure 3.4 shows a visual example of this. Data points incorrectly classified will incur a penalty proportional to their distance to the decision surface. Hence, the optimization problem now becomes one where there is a trade-off between maximum margin and minimum misclassification. The trade-off parameter C acts in such way that the optimization equation takes the form:

$$\min_{\mathbf{w},\xi,b} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i \right\} \quad \text{s.t.} \quad y_i(\mathbf{x}_i \cdot \mathbf{w} - b) \geq 1 - \mathbf{x}_i \quad \text{and} \quad \xi_i \geq 0 \quad \forall_i \,. \tag{3.25}$$

In a similar way as in the linear SVM case, Lagrange multipliers can be used to rewrite Equation (3.25) as an unconstrained optimization problem in the form:

$$\min_{\mathbf{w},\xi,b} \max_{\alpha,\xi,\beta} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \left[y_i(\mathbf{x}_i \cdot \mathbf{w} - b) - 1 + \xi_i \right] - \sum_{i=1}^{N} \beta_i \xi_i \right\} \quad \text{s.t.} \quad \alpha_i \beta_i \geq 0 \,. \tag{3.26}$$

Differentiating Equation (3.26) with respect to $\mathbf{w}$, b and ξ, setting the derivatives to zero and then replacing in Equation (3.25), yields the dual form:

$$\max_{\alpha} \tilde{L}(\alpha) = \max_{\alpha} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right\}$$
$$\text{s.t.} \quad 0 \leq \alpha_i \leq C \quad \forall_i \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \,, \tag{3.27}$$

where the only difference between equations (3.27) and (3.24) is the upper limit C imposed on α.

Nonlinear SVM

In their initial development, SVMs were proposed as a linear classifier. In the previous section we saw how to extend SVM to the case where the data is not fully linearly separable using slack variables. Another approach to classify nonlinearly separable data using SVM is to apply the kernel trick. Using a nonlinear function $\phi(\mathbf{x})$ data points are mapped to a higher (and in most cases much higher) dimensional space were the data is linearly separable, as illustrated in Figure 3.5. In the same way as linear SVM, the nonlinear case can be solved using the kernel trick to transform the input feature vectors in the high dimensional space by optimizing the dual form Lagrangian:

$$\max_{\alpha} \tilde{L}(\alpha) = \max_{\alpha} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) \right\} \,. \tag{3.28}$$

Here the optimization criterion is expressed in terms of inner products of the transformed feature vector. If a nonlinear mapping function $\phi(\mathbf{x})$ that allows the inner products to be expressed in terms of a kernel function $\mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ is used, then it is not necessary to perform an explicit mapping of the feature vectors into the high dimensional space. Some popular choices of kernels include Gaussian, polynomial and hyperbolic tangent kernels.

3.2.3 Artificial neural networks

An efficient way of solving complex problems is by subdividing them into smaller, simpler and more manageable problems. Artificial neural networks can be seen as such a system.

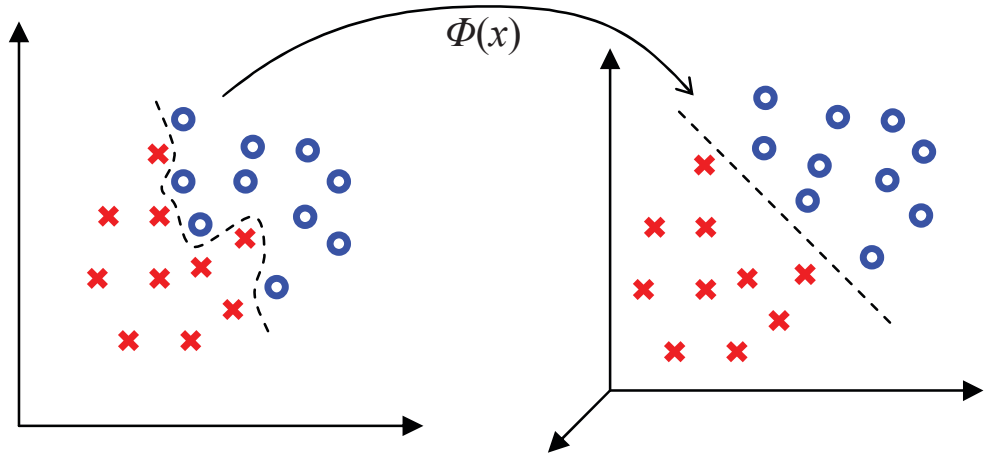


Figure 3.5: Illustration of how a nonlinear decision boundary can become linear in a higher dimensional space.

There are various types of networks with different attributes, but they all share the same basic components: a set of nodes and connections between them. In an artificial neural network nodes are seen as "artificial neurons". McCulloch and Pitts [143] were the first to propose a computational model of "nervous activity", where the neurons act as binary devices with a fixed threshold logic.

An efficient technique for evaluating the gradient of the error function of a layered feed-forward neural network (also known as error backpropagation [183], Figure 3.6), can be achieved using local message passing of information alternately forwards and backwards through the network. Using supervised learning, the error backpropagation algorithm calculates the network's error based on training input and output examples. The idea of the error backpropagation algorithm is that the artificial neural network learns the training data via a minimization of this error. The weights in the network are randomly initialized and the goal is to optimize their values in order to minimize the error. The error backpropagation algorithm can therefore be summarized in the following four steps:

- Apply an input vector $\mathbf{x}_i$ to the network and forward propagate through the network using $a_j = \sum_i w_{ji} z_i$ to find the activations of all the hidden and output units.

  Here z_i is the activation of a unit i, or input, that sends a connection to hidden unit j, and w_{ji} is the weight associated with that connection. This sum can be transformed by a nonlinear activation function $h(\cdot)$ to give the activation z_j of unit j in the form

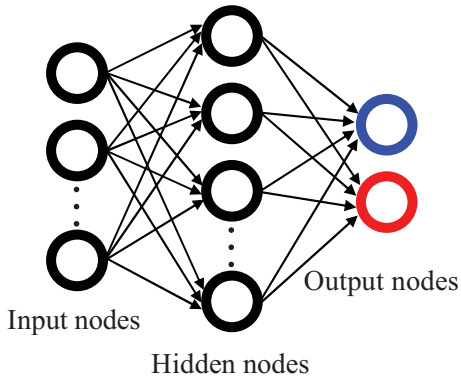


Figure 3.6: Three layer artificial neural network.

$z_j = h(a_j)$.

- Evaluate the $\delta_k = y_k - t_k$ for all the output units, where the outputs y_k are linear combinations of the input variables $\mathbf{x}_i$ so that $y_k = \sum_i w_{ki}\mathbf{x}_i$, t_k is the associated binary class and w_{ki} are the associated weights.

- Backpropagate the δ's using $\delta_j = h'(a_j)\sum_k w_{kj}\delta_k$ to obtain δ_j for each hidden unit in the network.

- Use $\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$ where $E_n = \frac{1}{2}\sum_k (y_k - t_k)^2$ to evaluate the required derivatives.

One of the main advantages of artificial neural networks is the ability to learn a mapping function based on a large training set comprised of input and output data. Artificial neural networks are also good when dealing with noisy or incomplete data. However, some key disadvantages are that the mapping function that emerges from artificial neural networks weights can be difficult to interpret and that their training can take significantly longer than other machine learning methods, e.g. LDA or SVM.

3.2.4 Boosting and AdaBoost

Ever since the mid 1990's boosting has received a significant amount of attention as an effective classification tool and more generally as a regressor. Boosting is a meta-algorithm that seeks to combine weak learners (classifiers) into one single strong classifier, which produces much more accurate results than any of the single weak ones. Assuming that we

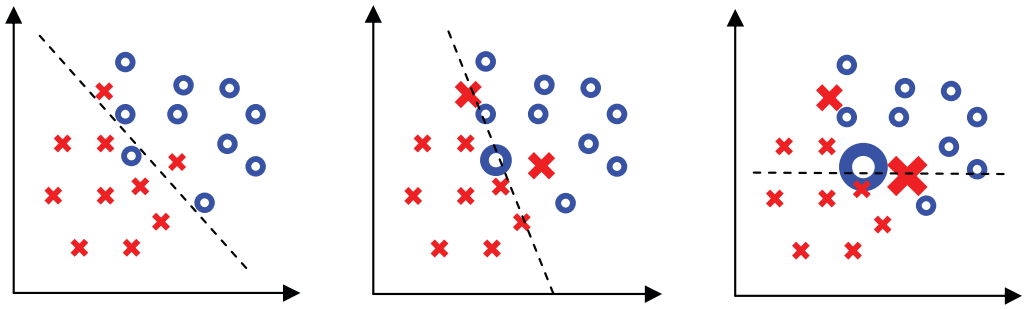
have a simple classifier learning algorithm that produces very modest results based on the observed training data, boosting works by employing this black box classifier learning algorithm several times with different subsets of the training data, or rather, differently weighted versions of the training data. At each round or iteration of boosting the aim is to find a weak classifier to separate the training data (based also on its weights). The only requirement of this weak classifier is that it has to predict the training data's labels slightly better than random. This is a very relaxed constraint since in case of a weak classifier performs worse than random the classification can be simply inverted. After a weak classifier is learned, a weight is assigned to this classifier according to its accuracy. The training data is re-weighted based on the performance of the classifier on each sample: a higher weight is given to samples that where misclassified, while a low weight is given to those correctly classified. Figure 3.7 illustrates an example (three iterations), where at each round a linear classifier is learned and the samples re-weighted according to the classifier's output. The weighting reflects the "focus" a particular sample should receive in the next round of boosting.

Freund and Schapire [76] describe a series of classifiers (for two or more classes) and regressors, as well as a mathematical proof of their guaranties and properties. They introduced the powerful and very popular algorithm called adaptive boosting (AdaBoost), which has been studied and tested in detail. Given a set of N training samples $\mathbf{x}$, where $\{\mathbf{x}_i, y_i\}$ and $i = 1,...,N\,,\;\; y_i \in \{-1,1\}$, AdaBoost first initializes a vector of weights for each sample, such that $D_1(i) = 1/N$. The objective at each iteration $t = 1,...,T$ is to choose the weak classifier $h_t \in \mathcal{H}$ that minimizes the classification error ε_t based on the sum of misclassified samples' weights D_t. The classifier is stored along with a weighting α_t that relates to its performance in the form:

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)\,. \tag{3.29}$$

Here ε_t is the sum of weights that where incorrectly classified:

$$\varepsilon_t = \sum_{i=1}^{N} D_t\left[h_t(\mathbf{x}_i) - y_i\right]\,. \tag{3.30}$$



(a) (b) (c)

Figure 3.7: Three iterations of a boosting procedure: The marker size represents the weight at each iteration. (a) All samples have the same weight and the best classifier is chosen. (b) Based on re-weighted samples, the best classifiers is chosen, (c) Samples that have been consistently been miss-classified have an even higher weight.

Using this new weak classifier h_t the distribution model D is then updated by

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(\mathbf{x}_i))}{Z_t} , \tag{3.31}$$

where Z_t is a normalization factor. The output (strong) classifier is then formed from the combination of the selected classifiers multiplied by their weight as

$$H(\mathbf{x}) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{x}_i)\right) . \tag{3.32}$$

3.2.5 Bagging and random forests

Bagging (bootstrap aggregating) [24] is a useful meta-algorithm designed to improve the stability and the predictive performance of tree models. Consider a base classification or regression algorithm that produces a mapping function $f(\cdot) : \mathbf{X} \rightarrow \mathbf{Y}$. Bagging generates B bootstrapped samples $(X_1^*, Y_1^*), \ldots, (X_B^*, Y_B^*)$ (randomly subsampled with replacement), uses the base algorithm to find a mapping function $f_b^*(\cdot)$, where $b \in B$, and finally aggregates the bootstrap estimates as

$$F(\cdot) = B^{-1} \sum_{b=1}^{B} f_b^*(\cdot) . \tag{3.33}$$

Bagging has been proven to improve unstable procedures, e.g. classification and regres-

sion trees (CART) [26] and artificial neural networks, while providing little or no performance improvement to more stable approaches [24], e.g. K-nearest neighbors.

Random forests are a powerful approach to data exploration, data analysis, and predictive modeling. Originally proposed by Breiman [25], random forests have their roots in bagging and random feature selection [9, 102]. They have been shown to have good performance in classification, regression, clustering and density estimation problems. High levels of predictive accuracy are achieved automatically, with only a few control parameters to tune, while remaining resistant to over-fitting (good generalization to new data).

Random forests generate a large number of different tree models that are grown using binary partitioning (see Figure 3.8). Randomness is introduced in the trees in two simultaneous ways. First, growing each tree on a different random subsample of size M from the training data of size N without replacement (bootstrap). Second, by selecting the best splitter at any node using only $d << D$ features chosen at random, where D is the total number of features and typically $d = \sqrt{D}$ or $d = \log_2 D$. The motivation for generating multiple tree models is that by combining different models the results will be better than if we relied on a single model. In classification problems the outputs generated by the multiple models are typically combined by majority voting (aggregation). Combining trees via voting will only be beneficial if the trees are different from each other. A reduction of the number of features available at each split, d, corresponds to a reduction in the correlation between trees, $\bar{\rho}$, and the strength of the trees, s. The error rate depends on both $\bar{\rho}$ and s, such that an upper bound for the generalization error is given by $\bar{\rho}(1-s^2)/s^2$. In random forests, bagging is improved by minimizing the model inter-dependence by forcing splits to be based on different predictors.

3.3 Regression overview

Regression is a technique that allows the modelling and analysis of several variables. A regression model estimates the relationship between one or more dependent variables $\mathbf{l}$ and the observed independent variables $\mathbf{X}$ through the unknown coefficients $\boldsymbol{\beta}$:

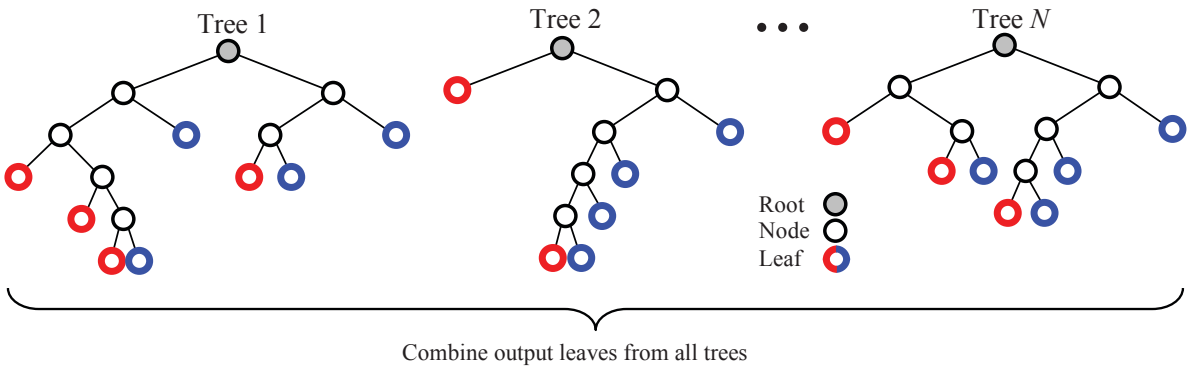


Figure 3.8: Illustration of a random forest.

$$\mathbf{l} = f(\mathbf{X}, \boldsymbol{\beta}) \ . \tag{3.34}$$

Contrary to classification where the outputs are categorical, a regression model estimates a continuous function.

3.3.1 Ordinary least squares regression

One of the most popular regression techniques is ordinary least squares regression (OLSR). Here the aim is to estimate the unknown parameters of a linear regression model via minimizing the axis-aligned squared error between the observed or measured data (predictors) and the predictions made by the linear model approximation. This simple estimator can be represented as follows:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|l - \mathbf{X}\boldsymbol{\beta}\|_2^2 \right\} \ . \tag{3.35}$$

Here l is a dependent variable, $\mathbf{X}$ are independent variables and $\boldsymbol{\beta}$ are the estimated model coefficients. The prediction power of a model found via ordinary least squares can be somewhat poor in the presence of outliers and noise in the independent variables, unequal training point variances, dependance among variables or too many variables. Most real life applications might contain some of these problems. Additionally, OLSR tends to produce rather complex and not very intuitive models since all predictor variables are used in the model regardless of their contribution.

3.3.2 Ridge regression

Ridge, or Tikhonov, regression [103] attempts to address some of the drawbacks of OLSR. This is done via adding an L_2-norm regularization term (that bounds the model's coefficients) to the sum of squared residuals. Hence it requires minimizing

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|l - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_R \|\boldsymbol{\beta}\|_2^2 \right\} , \tag{3.36}$$

where λ_R is the ridge regression penalty.

Ridge regression achieves better prediction performance than OLSR. However, the model complexity, and hence its intuitivity, is not addressed since it also keeps all available predictors in the model.

3.3.3 LASSO regression

Another approach that seeks to simplify the estimated model is the so called least absolute shrinkage and selection operator (LASSO) technique [206]. The LASSO is a least squares method that penalizes the L_1-norm of the regression coefficients. Formally we can write the LASSO model as:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|l - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_L \|\boldsymbol{\beta}\|_1 \right\} . \tag{3.37}$$

Due to the nature of the L_1-norm penalty sparse solutions to the problem are favored. Hence this approach acts as an automatic variable selector. Comparisons made with other techniques (ridge and bridge regression, [206] and [80]) found neither ridge, bridge or LASSO outperforms the other. Due to the exponential growth of data in many applications, variable selection is becoming ever more important for modern data analysis problems. Although the LASSO technique has been proven successful in many applications it is not without its drawbacks. For example, if the amount of variables is larger than the amount of samples, the LASSO will select only a number of variables that is equal to the number of samples. This limitation is a major constraint when dealing with very high dimensional data, e.g. medical images, where the number of samples can be in the thousands while the

dimensionality can easily be in the order of millions. Also, if a group of variables has high pairwise correlations, the LASSO selects only one of these variables ignoring the rest and hence potentially ignoring other important variables.

3.3.4 Elastic net regression

The elastic net regression technique [239] seeks to fix the drawbacks of the LASSO, while still maintaining its high performance. This is done by adding an additional L_2 penalty term on the model's coefficients. Similar to the LASSO, the elastic net performs automatic variable selection while encouraging the grouping of highly correlated variables. The elastic net is formulated as follows:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|l - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_R \|\boldsymbol{\beta}\|_2^2 + \lambda_L \|\boldsymbol{\beta}\|_1 \right\} . \tag{3.38}$$

Here $\mathbf{X}$ is a n by D matrix containing N vectorized images, $\boldsymbol{\beta}$ is a D long vector of the regression coefficients, l is the response variable, λ_R and λ_L are the ridge and the LASSO regression penalty weights, respectively. In Equation (3.38), the L_1 term encourages solutions that are sparse, while the L_2 term promotes the grouping of correlated variables.

3.4 Performance and fit measures

3.4.1 Classifier performance

A confusion matrix allows the performance of a binary classifier to be characterized. The columns of the matrix represent the class predictions, while rows represent the true classes. Table 3.1 shows an example of a confusion matrix. Correctly classified instances are located along the diagonal of the matrix. The true positives (TP) represent the correctly identified instances, while true negatives (TN) represent the correctly rejected instances. In a similar way, outside the diagonal line of the confusion matrix lie the false positives (FP) and the false negatives (FN), which represent incorrectly identified and the incorrectly rejected instances, respectively.

True \ Predicted	Class A	Class B
Class A	TP	FN
Class B	FP	TN

Table 3.1: Confusion matrix for a binary classifier.

A common metric to measure a classifier's performance is accuracy (ACC) which measures the rate of correctly classified examples as:

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \,. \tag{3.39}$$

ACC may not be a good metric of performance if the class distribution of the dataset is unbalanced. For example, if the dataset consists of a larger number of instances labeled as class A than B, a high accuracy can be achieved by a classifier that simply labels all instances as class A. The Sensitivity (SEN) and specificity (SPE) measures provide an overall assessment of the classifiers performance. SEN measures the ratio of correctly classified instances, while SPE measures the ratio of correctly rejected instances:

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad \text{and} \quad \mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \,. \tag{3.40}$$

Another way to measure ACC is through the balanced ACC, in which both classes have an equal weight on the output. It can be expressed as:

$$\text{balanced ACC} = \frac{\mathrm{SEN}+\mathrm{SPE}}{2} \,. \tag{3.41}$$

A receiver operating characteristic (ROC) is a graph that provides visualization of the performance of a binary classifier. ROC depicts the inherited trade-off that exists in binary classifiers between the true positive rate (SEN) and the false positive rate (1-SPE) as the discrimination threshold varies (see figure 3.9). The area under a ROC curve (AUC) may be interpreted as an aggregated measure of classifier performance [70].

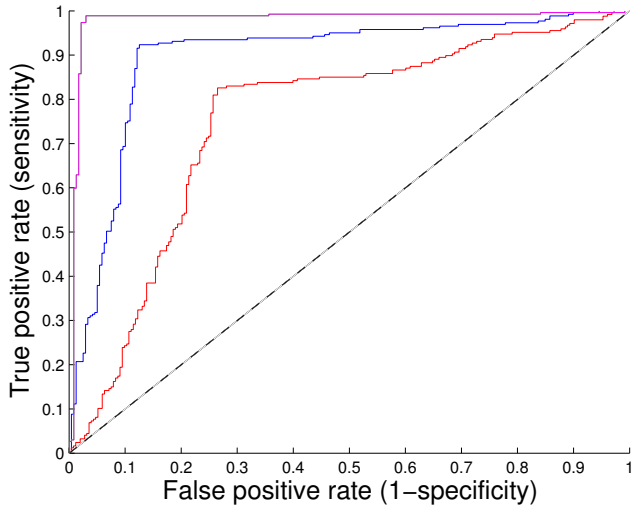


Figure 3.9: Illustration of the ROC curves of three binary classifiers. Each solid line shows the relationship between sensitivity and specificity of the classifier as the discrimination threshold is varied. The dashed line depicts a random classifier (best seen in color).

3.4.2 Cross-validation

The parameters of a model, e.g. a classifier, are usually optimized based on the available training data. An independent test set is therefore required for making a reliable assessment of the applicability of the model to unseen data. Cross-validation provides a statistical procedure to evaluate and compare models by partitioning data into two segments: one is used to learn or train the model and the other is used to validate the model. In a typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against. A commonly used method is k-fold cross-validation [99], in which the data is partitioned into k equally sized segments or folds. Subsequently, k iterations of training and validation are performed. Each iteration trains a model based on $k-1$ folds and then validates the trained model on the reminding fold. The performance of the model on each iteration can be tracked using performance metrics, e.g. ACC, which upon completion can be averaged. Another common cross-validation method is repeated random subsampling [161]. This method randomly splits the dataset into training and validation sets of fixed sizes. This process is repeated k times, and in the same way as in k-fold cross-validation the results are averaged over all folds. The advantage of repeated random subsampling is that the proportions of the training and validation sets are not de-

pendent on the number of iterations. However, due to its random nature some observations may never be selected in the validation subsample, whereas others may be selected more than once. In stratified cross-validation the training and validation sets for a classifier are selected such that they contain observations from each class in equal proportions to the full dataset. Stratified cross-validation has been shown to produce results with a lower variance than regular cross-validation [123]. Nested cross-validation or ”double-cross“ [196] is used to estimate the performance of a model for which the training includes selecting parameters or attributes. First, the attributes or parameters of a model are selected using an inner cross-validation loop. Then, an outer cross-validation loop is used to evaluate the performance of the selected model in the inner loop.

3.4.3 Coefficient of determination

The coefficient of determination, also known as R^2, is a measure of the goodness of fit of a regression model to the data. The higher the $R^2 \in [0,1]$, the better the variance of the dependent variable is explained by the independent variable. For the standard linear regression model the R^2 is a widely used goodness of fit measure and can be calculated as the relation between the total sum of squares and the residual sum of squares. If we denote y_i as the observed values of the dependent variable, $\bar{y}_i$ as its mean, and f_i as the fitted value, then R^2 is defined as:

$$R^2 = 1 - \frac{\sum_i (f_i - y_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \,. \tag{3.42}$$

Using this metric for nonlinear regression models can lead to $R^2 \ni [0,1]$. However, some applications of this measure to particular nonlinear models have been constructed using a variety of methods [138, 227, 38].

3.5 Summary

Dimensionality reduction, classification and regression are important tools in machine learning and they can form an important part of medical image analysis. This chapter has pro-

vided details on some of the most relevant machine learning algorithms, as well as some performance evaluation techniques, related or used as part of the work presented in later chapters of this thesis. Careful consideration, with the application in mind, must be given when choosing an algorithm for a specific task, i.e. linear or nonlinear data/features, known or unknown labels, discrete or continuous labels, feature space size, etc. The purpose of this chapter was to give the reader an insight into the state-of-the-art in dimensionality reduction, classification and regression techniques available, as well as the know how to select the most appropriate technique. The following five chapters will give a detailed description of the the main research contributions of the thesis.

Chapter 4

Landmark localization in brain MR images based on 3D local self-similarities

This Chapter is based on:

- R. Guerrero, L. Pizarro, R. Wolz, and D. Rueckert. Landmark localisation in brain MR images using feature point descriptors based on 3D local self-similarities. In IEEE International Symposium on Biomedical Imaging (ISBI), pages 1535-1538, 2012.

Abstract

The identification of anatomical landmarks in the brain is an important task in registration and morphometry. The manual identification and labeling of these landmarks is very time consuming and prone to observer errors, especially when large datasets must be analysed. In this chapter we present an approach that describes and locates landmarks based on their intrinsic geometry, rather than their intensity patterns. As the proposed approach moves away from intensity-based landmark description, we show that descriptors of intrinsic geometry are well suited for the landmark localization problem in MR brain images since the intensity information in these images is not quantitative (and intensity normalization is not straight forward). Our results show that for the task of localizing 20 anatomical landmarks in brain MR images, the proposed descriptor performs better in 75% of cases when compared with a sliding window with Haar features detector and in 100% of cases when compared to non-rigid registration.

4.1 Introduction

In recent years several algorithms for landmark localization have been proposed independently in the medical image analysis, computer vision and machine learning communities, each with specific advantages and disadvantages. The detection of landmarks is a crucial step in many medical imaging applications, including registration, shape modelling and morphometry.

In this chapter we propose the use of descriptors that define a landmark based on the structural pattern of its neighborhood. As descriptors we use a modified version of the local self-similarities described in [190]: First, descriptors are found on several training brain MR images. Then, when an unseen query image is presented, each training image's descriptor votes on the position of the landmark. Finally, a consensus for the landmark's location is estimated by fusing all the predictions available. In our evaluations, the presented approach has been trained on a large dataset of 100 brain MR images from cognitively normal (CN) subjects, patients with mild cognitive impairment (MCI) and AD from the

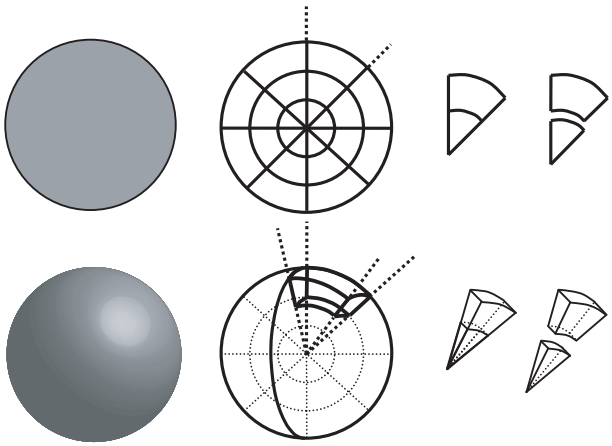


Figure 4.1: Radial bins used to construct the self-similarity (SS) descriptor, in 2D (top) and 3D (bottom). From left to right, complete area covered by the descriptor, partitioning of the area in radial bins and individual bins.

ADNI study database. A different set of 100 images from ADNI is used for testing the proposed approach.

4.2 Method

4.2.1 3D local self-similarity (SS) landmark descriptor

In order to characterize landmarks based on their surrounding (local) structures we propose an extended and enhanced SS descriptor. The local SS descriptor was recently described as an approach for measuring similarity between two visual entities in either images or video [190].

For every pixel in an image a local SS descriptor can be computed. In 2D this can be done by computing the similarity between a small square patch around the pixel and every other point (another small square patch) in a larger surrounding circular image region, which results in an internal similarity map. This similarity map is then binned into a log-polar representation (Figure 4.1, top row). Each bin is filled with the highest similarity that falls within its supported range. This representation yields three benefits: It compresses the descriptor's length for the pixel. It also accounts for radially increasing affine deformations allowing for invariance to small rotations, shears and scales. Finally, since only the largest similarity is used and bin sizes allow for some leeway, small local non-rigid deformations

can be tolerated.

In our work we extend this approach to 3D volumetric images and borrow ideas from the non-local means algorithm [29] to improve the descriptors' robustness. We define spherical regions around each voxel and find matches in a way similar to the approach previously described for 2D. We use SSD as a similarity measure between patches, and bin the results according to a log-spherical representation (Figure 4.1, bottom row). The small patches' descriptors, used to compute the SS, are defined as spherical regions surrounding the voxel (typically with radius of 3 or 5 voxels) and each voxel within the sphere is weighted individually using a radially decreasing Gaussian kernel **K**:

$$SSD_q(p) = \sum K_i \left(P_p - P_q\right)^2 . \tag{4.1}$$

Here P_p and P_q are the spherical patches at locations p and q, and the sum is over all voxels in the patches. This has been shown to improve patch similarity measurements [162].

The calculated SSD is then normalized to form a "correlation volume" S_q, that is associated with any voxel $q \in \mathbb{R}^3$, and can be written as

$$S_q(p) = \exp\left(-\frac{SSD_q(p)}{\max(var_{noise}, var_{auto}(q))}\right) \tag{4.2}$$

for all $p \in \mathbb{R}^3$ such that $\|p-q\|^2 < \rho^2$. This is computed for all points within a distance ρ from pivot point q at which the descriptor is being calculated. $var_{noise} = 2\rho^2 * var(I)$ is the estimated photometric image variance, and $var(I)$ is the variance in image I. var_{auto} is the maximal variance of the difference of all patches within a radius of one voxel relative to the patch centered at q. The correlation volume, which is defined in Cartesian space, is mapped to a spherical coordinate system $S_q(x,y,z) \rightarrow \tilde{S_q}(r',\theta',\phi')$. Thus, the SS descriptor SS_q is given by the maximum correlation value within each bin (r_i, θ_j, ϕ_i):

$$SS_q(r_i, \theta_j, \phi_i) = \max_{(r',\theta',\phi') \in \mathbb{R}^3} \tilde{S_q}(r',\theta',\phi') \tag{4.3}$$

where

$$
\begin{aligned}
r' &\in [r_i, r_{i+1}], \qquad r_i \in R = \{r_1, ..., r_L\} \\
\theta' &\in [\theta_j, \theta_{j+1}], \qquad \theta_j \in \Theta = \{\theta_1, ..., \theta_M\} \\
\phi' &\in [\phi_k, \phi_{k+1}], \qquad \phi_k \in \Phi = \{\phi_1, ..., \phi_N\}
\end{aligned} \tag{4.4}
$$

and R, Θ, Φ denote the sets of radii, elevation and azimuth angles, each discretized into L, M and N values, respectively.

This type of descriptors are specially suited to work with imaging modalities where there is no consistent intensity scaling between images (e.g. MR images) or different modalities, since they encode the intrinsic surrounding geometry of the point of interest, and not the intensity distribution. In this way we can characterize anatomical landmarks based on their surrounding structures.

4.2.2 Landmark localization

With a set of annotated images, in which landmark location and landmark descriptors have previously been determined, the landmark can be localized in unseen images: First, we assume that the brain is in some approximately known orientation and position. Thus, a landmark's spatial location is likely to fall inside a particular volume within the brain. We can reduce the search space to a limited ROI that is defined as a non-zero probability volume of where we expect to find the landmark. This could be done in two different ways: One can define a box in which one could expect to observe the landmark or one can learn the spatial prior probabilities from the training set (atlases). Here, the latter approach is used. The spatial prior probabilities are estimated using kernel (parzen window) density estimation. This can be formulated as

$$
PDF(p) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n^d} \mathbf{K}_p \left(\frac{p - \iota_i}{h_n} \right) \tag{4.5}
$$

where p denotes the 3D voxel coordinates (x, y, z), $\iota_i \in \{\iota_1, ..., \iota_n\}$ refers to the i-th landmark position, $\mathbf{K}_p(\cdot)$ is the window function or the kernel in a 3D space such that $\int_{\Re^d} \mathbf{K}_p(\mathrm{x}) d\mathrm{x} =$

1. Here n is the number of observations and $h_n > 0$ is the bandwidth parameter that corresponds to the width of the kernel. The kernel function $\mathbf{K}_p(\cdot)$ is modelled as a Gaussian function.

In the search space, defined by thresholding the prior probability $PDF(p)$, we calculate a descriptor for every voxel contained in this ROI, and then find the best possible match among the set of training descriptors. Each training descriptor provides an estimate (vote) for the spatial localization of the landmark. Each vote will be associated with a certain weight w. For an unseen image i, we estimate the localization $\mathfrak{l}_i^l$ of the landmark l by fusing all votes using the following equation:

$$\mathfrak{l}_i^l = \frac{\sum_j I_{i,j}^l \cdot w_{i,j}^l}{\sum_j w_{i,j}^l} \,. \tag{4.6}$$

Here $I_{i,j}^l$ represents the estimated position of the landmark l in the image i as voted by template j. In this work we explore several strategies to select the weights $w_{i,j}^l$:

- Simple average vote (SAV): All votes have the same weight.
- Sparsity (S): The weighting is calculated according to the sparsity of the likelihood output, see Equation (4.7). This favors descriptors with concentrated "energy" peaks, which are considered more informative that evenly distributed descriptors.
- Normalized sparsity (NS): Same as S, but with sparsity values normalized to the range [0,1].
- Descriptor similarity (DS): Similarity between template and test descriptors. This favors descriptors with a lower SSD.
- k most similar descriptors (kSD): Only the k most similar templates, according to their SSD, are considered.

As a metric for the sparsity we used a measure based on the relationship between the L_1 and L_2 norms (as described in [105]):

$$\sigma\left(SS_q\right) = \frac{\sqrt{n} - \left(\sum \left|SS_{q_i}\right|\right)\sqrt{\sum SS_{q_i}^2}}{\sqrt{n}-1} \tag{4.7}$$

where n is the number of bins of the descriptor SS_{q_i}, and $0 \leq \sigma \leq 1$.

4.3 Comparison to other landmark detection approaches

In this section we briefly describe two other different methods commonly used for landmark localization: (a) Sliding window detector with Haar feature (SW) [218] and (b) non-rigid image registration (REG). These two methods were used as a comparison to the method proposed.

4.3.1 Sliding window with Haar features detector (SW)

A detector for each landmark point was built using a variation of the Viola-Jones face detector [218]. In our implementation, Haar-like features are calculated within a cuboid region in 3D space for the MR brain images. Haar-like features offer the advantage of being very computationally inexpensive. Figure 4.2 illustrates some of the Haar-like features in 2D and 3D space. The computational efficiency stems from the usage of integral images, where each value in an integral image takes the value of the sum of pixels above and to the left of the pixel,

$$II(x,y) = \sum_{x'<x}\sum_{y'<y} I_i(x',y') \tag{4.8}$$

where II is the integral image of I_i, which is the original image. Extending this idea to 3D, the integral image at location $p = \{x,y,z\}$ takes the value

$$II(x,y,z) = \sum_{x'<x}\sum_{y'<y}\sum_{z'<z} I_i(x',y',z'). \tag{4.9}$$

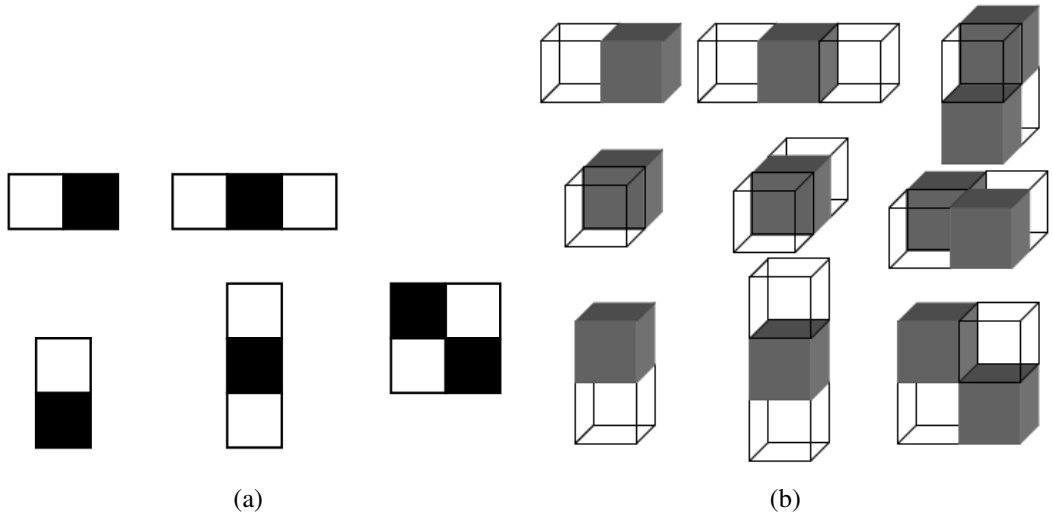


Figure 4.2: Haar-like features in (a) 2D space are rectangles and in (b) 3D are cuboids. In columns, from left to right: two, three and four cuboid features.

Using the following set of recurrences, the integral image can be calculated in one single pass over the original image

$$\begin{aligned} \mathbf{s}(x,y,z) &= \mathbf{s}(x,y-1,z) + p_i(x,y,z) \\ \mathbf{s}_2(x,y,z) &= \mathbf{s}_2(x-1,y,z) + \mathbf{s}(x,y,z), \\ II(x,y,z) &= II(x,y,z-1) + \mathbf{s}_2(x,y,z). \end{aligned} \tag{4.10}$$

where s and s_2 are sum accumulators.

Boosting (Section 3.2.4) is then used to perform classification based on all calculated Haar-like features. In each round of boosting, the algorithm picks a single feature that best classifies the data. For each feature, the algorithm determines the optimal threshold classification function using QDA, such that the minimum number of examples are misclassified. A weak classifier $h_j(\mathbf{x}_i)$ consists of a feature f_j, a threshold θ_j and a parity τ_j indicating the direction of the inequality sign (since we do not know if a high or low feature response value is desired):

$$h_j(\mathbf{x}_i) = \begin{cases} 1 & \text{if } \tau_j f_j(\mathbf{x}_i) < \tau_j \theta_j \\ 0 & \text{otherwise} \end{cases} \tag{4.11}$$

where $\mathbf{x}$ is a positive or negative training instance.

No single feature is capable of correctly classifying the whole dataset, e.g. patches

where the landmark is present and patches where it is not. However, the aim is to distinguish between only two classes (landmark present and not present) and even a naive random classification scenario should achieve an accuracy of 50%, which is unlikely to occur since the best feature was chosen. All that is needed is that the simple classifier based on a single feature, perform slightly better than random (which is guaranteed since we can always change the direction of the threshold by changing the sign of the parity function). At each round of boosting the algorithm picks the best weak classifier/feature and reweights the training examples according to the performance of the chosen weak classifier. This process is repeated for T rounds, or until an early termination criteria is met, such as the error on the training set being below a threshold. Instead of forming a final "strong" classifier using a weighted combination of all the selected features and thresholding the response, when evaluating the classifier response in the test images, we opted not to threshold the result. By not thresholding the output of the classifier we can retain information about the confidence of the classifier. To localize the landmark we can select the voxel which yields the highest confidence from the classifier. A drawback of a sliding window detector is that it does not allow for sub-voxel landmark localization accuracy.

4.3.2 Non-rigid image registration (REG)

Intensity-based image registration establishes dense point correspondences across images by computing a transformation that maps points from one image I_A to corresponding points in a second image I_B. The transformation T applied to a point p in I_A, represented by the column vector $p = \{x, y, z\}$, yields a transformed point $p' = \{x', y', z'\}$ in I_B, such that, $p' = T(p)$. Intensity-based image registration can be used to propagate the annotation (e.g. landmarks) from a reference image to a new image. In this case the annotated reference image acts as a template. Hence, landmarks that are annotated in the template can be propagated to new images. Figure 4.3 illustrates the landmark propagation procedure. In this work we used the non-rigid FFD registration algorithm proposed in [182]. Since correspondences between images are not constrained to voxel centroids, a REG approach to landmark localization allows for sub-voxel accuracy.

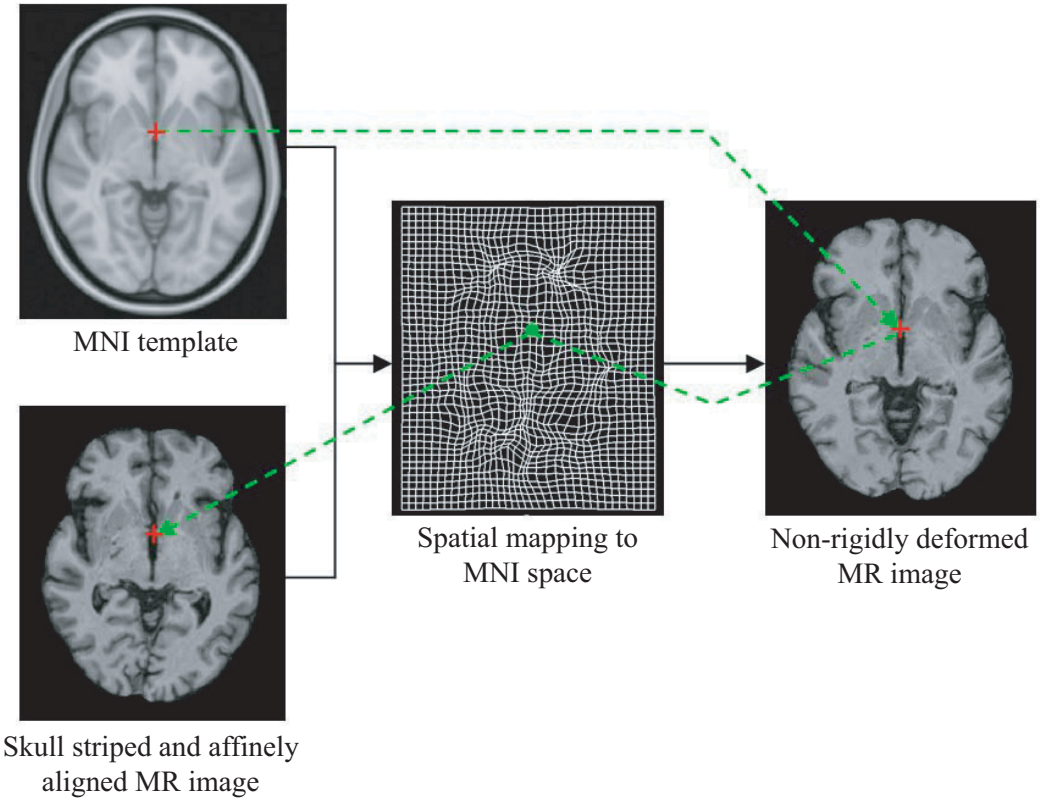


Figure 4.3: Diagram showing the landmark annotations propagated from the MNI template to the skull striped and affinely aligned images. First, the image is affinely and non-rigidly aligned to MNI space (see the solid black arrows). The landmarks are manually located in MNI space, then propagated to the affine images (see the dashed green arrows).

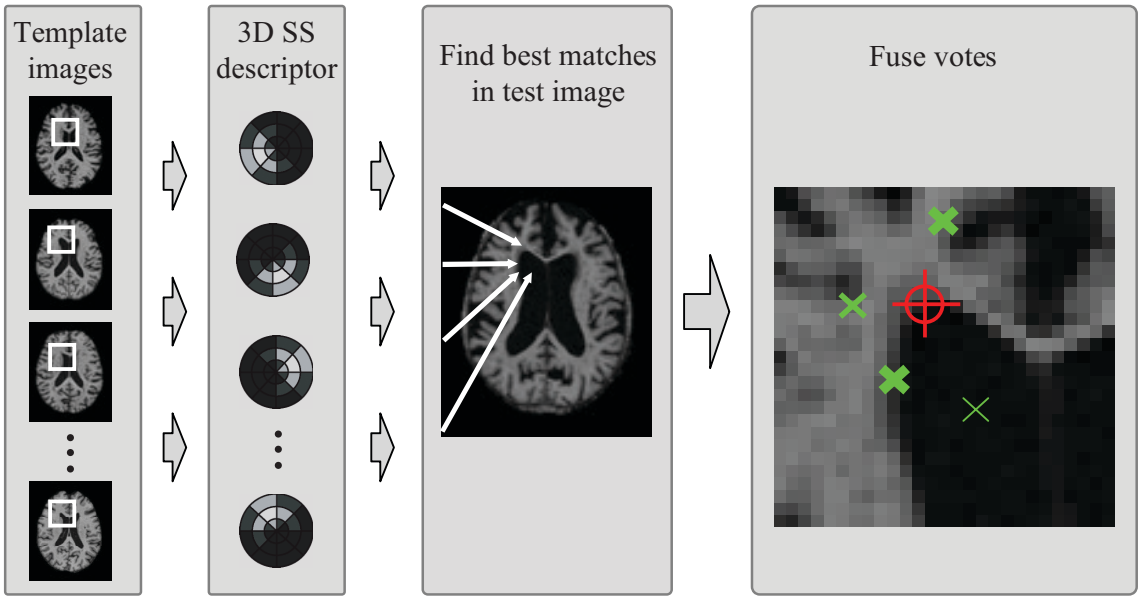


Figure 4.4: Diagram of proposed landmark localization method.

4.4 Experiments and results

The images that were used to evaluate the proposed method were obtained from the ADNI database [148]. In the ADNI study, brain MR images were acquired at regular intervals after an initial baseline scan from approximately 200 CN older subjects, 400 subjects with MCI, and 200 subjects with early AD. In this work, we used a subset of 1.5T T1-weighted baseline images of 100 randomly chosen subjects for training and another 100 randomly chosen subjects for testing. In both (training and testing) datasets there are 24 AD, 48 MCI and 28 healthy subjects, to faithfully represent the full ADNI dataset. All brain MR images were skull stripped and affinely aligned to the Montreal Neurological Institute (MNI) space.

Figure 4.4 shows the pipeline of the proposed landmark localization method. For both the training and testing datasets a total of 20 landmarks (Figure 4.5) were manually selected by an expert observer using three orthogonal views. See Appendix C.1 for the description of the landmarks. As mentioned before, two other commonly used methods for landmark localization were used as a comparison to the proposed one: (a) SW and (b) REG. In total 20 different landmark specific SW detectors where learned. 3D image cubic patches of 31^3 voxels centered at the landmark where used as positive training samples, while patches extracted from the vicinity of the landmark where used as negative training samples.

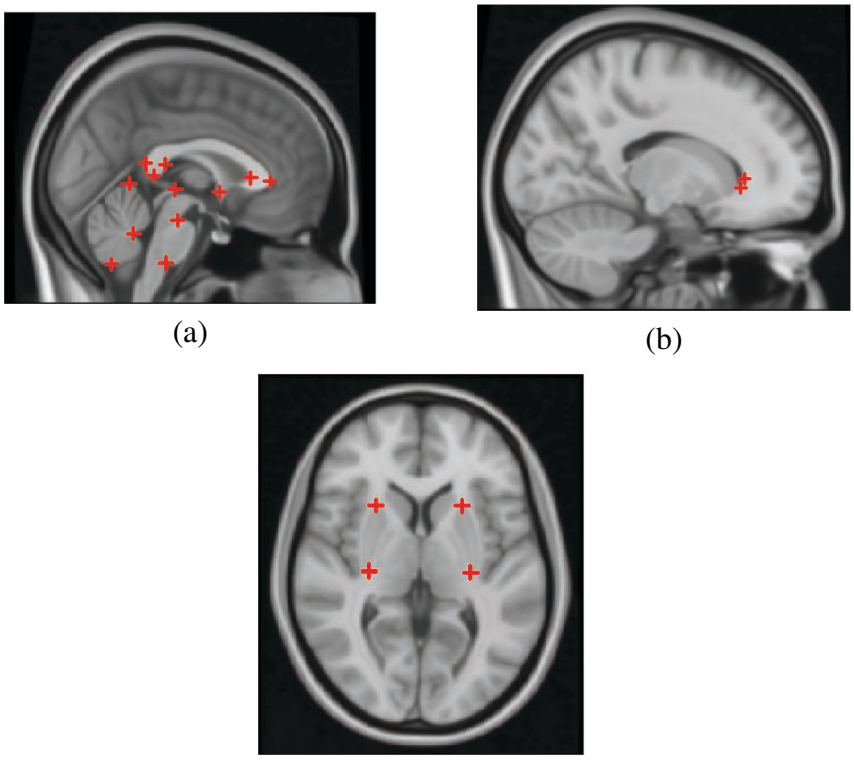
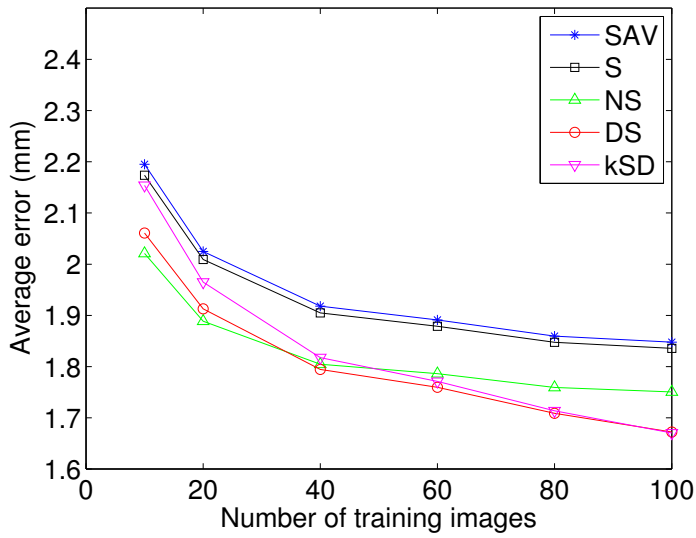


Figure 4.5: Anatomical landmarks in the MNI152 atlas. (a) Splenium and genu of corpus callosum, superior and inferior tip of the cerebellum, fourth ventricle, anterior and posterior commisure, and superior and inferior aspect of the pons. (b) Anterior and inferior tip lateral ventricle (only left side shown). (c) Superior and inferior tip of the putamen (left and right).

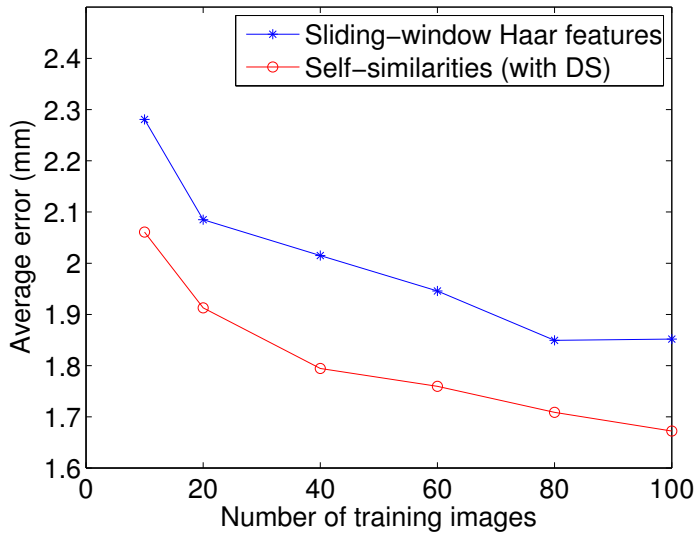
From the training set we calculate 3D SS descriptors at each landmark's position. To calculate the 3D SS descriptors we used small spherical kernels with a radius of three voxels, that were weighted using a second Gaussian kernel (same size) with $\sigma = 3$, and a larger correlation sphere of radius 10. The similarity results were binned in a partitioned sphere with four radial intervals, five elevation angles and 10 azimuth angles (see Figure 4.1, bottom row for visualization). We tested the performance of classifiers built from different numbers of training images (10, 20, 40, 60, 80 and 100). During testing, we look for the best matches of the training descriptors in the test image and then form a consensus on the final landmark location by fusing the results using several techniques. Figure 4.6 (a) shows the average (over the 20 landmarks shown in Table 4.4) landmark error. Using descriptor similarity weighting (most robust results, Figure 4.6 (a)), a comparison with the boosted Haar features classifier, was carried out (also using the 20 landmarks from Table 4.4). We show that using less training images we can obtain the same level of accuracy, which is a highly desired as annotated images are scarce and expensive in terms labor involved. Using the same amount of images for training the proposed method achieves better results. Figure 4.6 (b) illustrates this even better. Another desirable feature of the proposed method is that it does not require any intensity normalization step, as is required by intensity based approaches (such as SW). An unfeasible computational cost is associated to carrying out the number of registrations needed to do a several-to-one landmark localization, e.g. to use vote fusion. Hence, registration was used as a one-to-one landmark propagation tool and a direct comparison is not possible.

Table 4.4 shows the result of individual landmark localization error, using the proposed method and both comparison techniques. A five-fold cross validation of the method was carried out in order to asses the results and ensure reproducibility (the average of the five tests is shown, with a variability among tests of $\sim$0.15mm).

The proposed method was implemented in Matlab, and the implementation of SS is based on the code provided by Varun Gulshan [94]. Training time depends on the number of template images used: Typically each image requires $\sim$0.11 seconds on a 3.00GHz 2-core machine with 8Gb of RAM. Testing takes about 16 seconds per landmark per image, if



(a)



(b)

Figure 4.6: (a) Performance of the proposed classifier built with different amounts of training images and using different fusion techniques. (b) Comparison of the proposed method vs SW.

Landmark	SS	SW	REG
Splenium of corpus callosum (outer aspect)	**1.29(0.99)**+*	1.79(1.07)	3.95(1.43)
Splenium of corpus callosum (inferior tip)	**1.27(0.88)**+*	1.66(0.96)	2.10(0.89)
Splenium of corpus callosum (inner aspect)	2.28(1.54)+	**1.81(0.82)**	2.31(1.31)
Genu of corpus callosum (outer aspect)	**1.08(0.78)**+*	1.87(1.08)	1.73(1.01)
Genu of corpus callosum (inner aspect)	2.56(2.17)+*	**1.32**(0.73)	1.47**(0.64)**
Superior aspect of pons	**1.06(0.77)***	1.12**(0.77)**	2.79(1.26)
Inferior aspect of pons	**1.26**(2.29)*	1.45**(0.72)**	1.70(0.85)
Superior aspect cerebellum	3.92**(1.02)**+*	**2.83**(1.75)	2.99(1.64)
Fourth ventricle	**0.83(0.78)**+*	1.19(0.80)	5.57(2.70)
Putamen posterior (left)	**2.24(1.16)***	2.43(1.41)	4.36(1.81)
Putamen anterior (left)	**1.78(1.14)**+*	2.14(1.27)	2.48(1.29)
Putamen posterior (right)	2.28(1.33)*	**2.25(1.13)**	3.53(1.78)
Putamen anterior (right)	**1.90**(1.24)+*	2.28**(1.23)**	2.79(1.43)
Anterior commissure	**0.67(0.59)**+*	1.16**(0.59)**	1.05(1.42)
Posterior commissure	**0.64(0.31)***	0.69(0.59)	1.85(0.48)
Inferior aspect cerebellum	2.87(2.02)+*	**2.39**(1.89)	3.71**(1.68)**
Anterior tip of lateral ventricle (left)	**1.31(0.89)**+*	2.22(1.24)	3.67(1.72)
Anterior tip of lateral ventricle (right)	**1.14(0.71)**+*	1.78(0.98)	3.65(1.73)
Inferior tip of lateral ventricle (left)	**1.76(1.34)**+*	2.80(1.35)	4.44(2.07)
Inferior tip of lateral ventricle (right)	**1.27(0.78)**+*	2.20(1.11)	4.01(1.79)

Table 4.1: Accuracy of the proposed method (using 100 training images) on the ADNI database, for the 20 landmarks listed. Errors in mm with standard deviation in brackets. Best results shown in bold numbers. Statistical significance (to %5, results not corrected for multiple comparisons) is indicated by + and *, for comparisons between SS and SW or REG, respectively.

only one image is tested. Time can be reduced to about 11 seconds per landmark per image if several images are tested at the same time due to computational overhead.

4.5 Conclusions

We have proposed a method that localizes landmarks in brain MR images using a 3D local SS descriptor that is not intensity dependent and describes the self-similarity around the vicinity of the landmark. Using this landmark representation we search for and obtain votes on where the landmark is located. Several vote fusion strategies have been tested. Also, prior knowledge of the spatial distribution of the landmarks was used to reduce the search space. Our results show that the proposed method outperformed a sliding window with Haar features detector in 15 out of 20 cases and non-rigid image registration in every case in the landmark localization task. Preliminary comparison experiments with 3D Scale-invariant feature transform (SIFT) feature descriptors [187] where tried using an implementation provided by Scovanner, but the computational burden of the SIFT descriptor calculation was deemed to high. It should be noted that a disadvantage of both the 3D local SS descrip-

tors and SW approaches to landmark detection can only offer localization accuracies at the voxel level as they find the best matching location. This is in contrast to non-rigid image registration, which locates landmarks with a subvoxel accuracy. However, it may be possible to achieve subvoxel localization accuracies using both 3D local SS descriptors and SW approaches by modeling their output as a continuous function. Heinrich et al. [101] have successfully used SS maps as a similarity metric in multimodal lung image registration. However, due to the metric's associated high computational expense the registration run time was very high, even though the similarity maps where kept small. Which points out the clear computational expense disadvantage of the proposed descriptor, which limits its scalability.

This chapter introduced the use of 3D local SS features as a tool for landmark localization in brain MR images. In the following chapter we extend this idea to automatic feature matching and combine it with RANSAC for affine image registration of knee MR images.

Chapter 5

Learning correspondences in knee MR images using 3D local self-similarities

This Chapter is based on:

- Ricardo Guerrero, Claire Donoghou, Luis Pizarro, Daniel Rueckert. ”Learning correspondences in knee MR images from the Osteoarthritis Initiative”. Machine learning in medical imaging - Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 7588, pages 218-225, 2012.

Abstract

Registration is a powerful tool that allows mapping of images into a common space in order to aid in their analysis. Accurate registration of images that contain articulate structures such as the knee is challenging to achieve using intensity-based registration algorithms. Problems arise due to potentially very large inter-subject differences in both articulation and anatomy. This can cause intensity-based registration algorithms to fail to converge to an optimal solution. In this work we propose a method for learning correspondences in pairs of images in order to match the self-similarity features introduced in Chapter 4. These features where used in Chapter 4 to find anatomical landmarks in brain MR images. We use RANSAC, in combination with the automatically obtained feature matches, to robustly estimate the parameters of an affine transformation model. We show a substantial improvement in terms of mean error and standard deviation of 2.13mm and 2.47mm compared to intensity-based registration methods when comparing target registration error.

5.1 Introduction

In many medical image analysis applications it is important to estimate spatial transformations to a common space, e.g. in registration [238, 195], statistical shape modelling [40, 35, 41], atlas construction [93, 197, 31], segmentation [224, 174, 13, 100, 6] and computer aided diagnosis [55, 43, 192]. Many of these methods rely on global registration (e.g. rigid or affine) as an initialization for a more local alignment. In some cases this global registration provides a very good initialization, e.g. in brain imaging where the global variations in head shape, position and orientation are generally limited when compared to other anatomical structures. Unfortunately intensity based affine registrations does not suffice for datasets with large inter-subject anatomical variability or for articulated structures. In these cases the global registration can fail completely.

One such example are MR images of the knee: Obtaining good global registrations in knee MR images can be particularly challenging due to the aforementioned reasons. This means that intensity-based affine registration may converge to a poor local minimum and

hence fail to align the images correctly. Fine tunning registration methods often solves miss-registration problems, but migth also introduce an unwanted bias in the "fixed" registrations. Several methods have been shown to be successful at registering knee MR images, mainly in cartilage and bone segmentation [78, 113]. For very large population studies manually fine tuning algorithms to correct registration errors is a very time-consuming and tedious task. The Osteoarthritis Initiative (OAI) is a multi-center, longitudinal, prospective observational study of knee osteoarthritis, and provides public access to MR images and clinical data. The OAI is a large-scale study into this disease with a large number of participants (4796 men and women aged 45-79). A robust and automated registration method that yields accurate and consistent results is essential for such a large-scale study.

Several methods have been proposed to address the problem of registration in the presence of large intra-subject variability. Recently graph-based registration methods [96, 114] that aim to find geodesic path across a similarity graph between images have emerged, with Donoghue et al. [56] applying the concepts of graph-based registration to a large population knee MRI study. This allowed the registration of two images in the graph to be expressed as a composition of incremental transformations along the shortest path between images. The main advantage of these techniques is that the incremental transformations can avoid getting trapped in local minima, thus achieving a more accurate registration, but at a higher computational cost. Moreover, it is not straightforward how to deal with images that were not used in the initial construction of the graph. In general, feature-based registration methods [178, 157, 176] aim to find and match features in a pair of images, and use these obtained correspondences to define a transformation from one image to another at these landmarks while interpolating the transformation between the landmarks. This can lead to a faster registration, while at the same not requiring to learn a graph. However, as stated in [176], the success of point- or feature-based image registration highly depends on the representative power and accuracy of the feature matching.

In this work we propose an approach to the problem of image registration in the presence of large-scale variations that explores feature matching. We use the recently proposed 3D local SS features [91]. Through saliency measures we compute and match features in each

pair of images. These matched features are used in turn for a robust affine transformation parameter estimation that minimizes the feature alignment error. We show results of using the proposed method on a subset of knee MR images of the OAI cohort although the method is generalizable to other types of images.

5.2 Method

To register a pair of images we use the structure tensor [176] to filter out image regions that contain no structure and therefore are considered uninformative. Then we calculate 3D local SS features [91] for all remaining voxels. After this, the feature list is reduced by measuring the energy distribution and level of similarity of the descriptors. The remaining features are used in a forward-backward matching algorithm that further reduces the list of features by finding and matching stable points in both images. Finally, the parameters of an affine transformation are estimated using RANSAC [67], which again reduces the list of matching points by removing outliers. RANSAC is a popular model estimation algorithm that is robust against outliers that has been used before to estimate transformation model between medical images [175, 184, 140]. Figure 5.1 shows the pipeline describing the proposed feature analysis and matching methods.

5.2.1 Dense 3D local SS feature descriptors

As mention in Chapter 4, for every voxel in an image I a local SS descriptor can be computed [91]. This can be done by calculating the similarity (using SSD) between a small spherical patch around the voxel and every other point (another small spherical patch) in a larger surrounding spherical image region. This results in an similarity map between a voxel and its neighborhood that is then binned into a log-spherical representation. Each bin is populated with the highest similarity value that falls within its supported range. This representation has three benefits: It leads to very compact descriptors for each voxel, it accounts for radially increasing affine deformations and it can handle small amounts of local non-rigid deformations. Following the same description as in Chapter 4, the calculated similarities are then

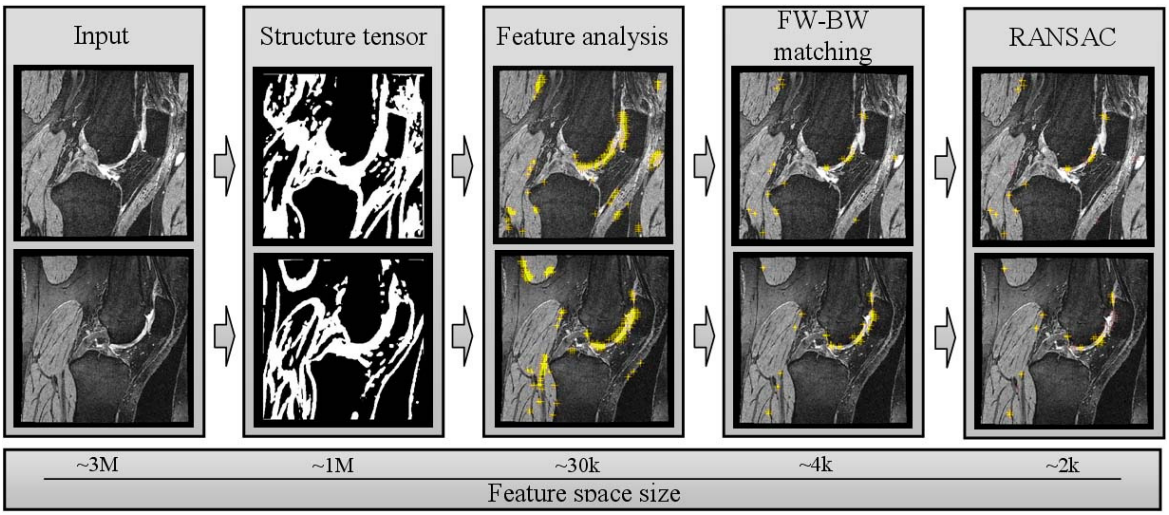


Figure 5.1: Proposed feature analysis and matching methods to identify potentially stable feature points: First, the structure tensor identifies regions that contain no structure and therefore are considered uninformative (second column). Then, 3D local SS features are computed (third column). After forward-backward matching the number of matches is further reduced (fourth column). In the final stage an affine transformation model is estimated using RANSAC (last column).

normalized to form a *correlation volume* S_q, that is associated with any voxel $q \in \mathbb{R}^3$, and can be written as

$$S_q(p) = \exp\left(-\frac{SSD_q(p)}{\max(var_{noise}, var_{auto}(q))}\right) \tag{5.1}$$

for all $p \in \mathbb{R}^3$ such that $\|p-q\|^2 < \rho^2$. That is, for all points within a distance ρ from pivot point q at which the descriptor is being calculated. $var_{noise} = 2\rho^2 * var(I)$ is the estimated photometric image variance, where $var(I)$ is the variance in image I. var_{auto} is the maximal variance of the difference of all patches within a radius of one voxel relative to the patch centered at q. The correlation volume, which is defined in Cartesian space, is mapped to a spherical coordinate system $S_q(x,y,z) \rightarrow \tilde{S}_q(r',\theta',\phi')$. Thus, the SS descriptor SS_q is defined by the maximum correlation value within each bin (r_i, θ_j, ϕ_k):

$$SS_q(r_i,\theta_j,\phi_i) = \max_{(r',\theta',\phi')\in\mathbb{R}^3} \tilde{S}_q(r',\theta',\phi') \ , \tag{5.2}$$

where

$$
\begin{aligned}
r' &\in [r_i, r_{i+1}], \qquad r_i \in R = \{r_1, ..., r_L\} \\
\theta' &\in \left[\theta_j, \theta_{j+1}\right], \qquad \theta_j \in \Theta = \{\theta_1, ..., \theta_M\} \\
\phi' &\in [\phi_k, \phi_{k+1}], \qquad \phi_k \in \Phi = \{\phi_1, ..., \phi_N\}
\end{aligned}
\tag{5.3}
$$

and R, Θ, Φ denote the sets of discretized radii, elevation and azimuth angles, each with L, M and N values, respectively. This type of descriptor is especially well suited for image modalities where there is no intensity scale consistency across images (e.g. MR images), since they encode the intrinsic surrounding geometry of a point, rather than their absolute intensity values. In this way we move away from an intensity-based characterization towards a geometric-based characterization of feature points.

5.2.2 Feature analysis

Since not all descriptors are informative, we have to remove non-informative ones. Therefore, we aim to identify which parts of the image do not contain any informative features at all (e.g. structureless regions). We employ two different techniques to reduce the feature space: (1) using the 3D structure tensor of the image we define regions that contain potentially relevant and stable features, hence reducing the feature calculation burden and (2) we measure the energy distribution and level of SS of the calculated features and ignore features for which the energy distribution or level of SS falls below a certain threshold in order to further reduce the number of potential stable feature points. Using a subset of salient features determined by the two mentioned tests, we employ a forward-backward feature matching algorithm [81] to determine feature correspondences of stable points.

Image structure tensor

The image structure tensor or matrix of second-order moments defines the predominant directions of the image gradient around a particular point q [176]. The discrete version of the 3D structure tensor Γ can be written as $\Gamma_w[q] = \sum_r w[r]\Gamma_0[q-r]$ where r defines a set of

indices centered around q, $w[r]$ is a weight within the window such that $\sum w[r] = 1$ and $\Gamma_0[q]$ is the matrix given by:

$$\Gamma_0[q] = \begin{bmatrix} (I_x[q])^2 & I_x[q]I_y[q] & I_x[q]I_z[q] \\ I_x[q]I_y[q] & (I_y[q])^2 & I_y[q]I_z[q] \\ I_x[q]I_z[q] & I_y[q]I_z[q] & (I_z[q])^2 \end{bmatrix} \tag{5.4}$$

Here I_x, I_y and I_z are the partial derivatives of image I. The eigenvalues λ_1, λ_2 and λ_3 of $\Gamma_w[q]$ and their corresponding eigenvectors e_1, e_2 and e_3 describe the distribution of gradients of the image within a small, pre-specified region around q. The values obtained from the structure tensor can be then used to define regions that contain structure in a robust way, and hence might contain stable features. Potentially stable features will lie in regions where the structure tensor's eigenvalues are not zero, i.e. in regions where there is high contrast.

Measures on the descriptors

Once the 3D self similarity features are calculated, there are two tests that are applied directly to the descriptor vectors in order to assess if they are considered informative. First, we only consider feature vectors that contain certain level of SS, that is, the similarity between the patch around the voxel for which the descriptor is calculated and the patches in the lager volume being considered, should be above a certain threshold $0 \leq C \leq 1$. Secondly, we also evaluate the energy distribution of the feature vectors. Specifically, we use a sparsity measure to check whether the energy distribution of the feature descriptor contains peaks or is homogeneous. The sparsity metric [105] used is defined as:

$$\sigma\left(SS_q\right) = \frac{\sqrt{n} - \left(\sum \left|SS_{q_i}\right|\right) \sqrt{\sum SS_{q_i}^2}}{\sqrt{n} - 1}, \tag{5.5}$$

where n is the number of bins of the descriptor SS_{q_i}, and $0 \leq \sigma \leq 1$. If the descriptor does not meet both criteria the feature vector is considered non-informative and is ignored.

Finding feature matches

The forward-backward matching algorithm (see Algorithm 1) was originally designed to find distinctive and stable point between two stereoscopic images [81]. Given a set of points $\mathbf{p} = \{p_1, p_2, ..., p_n\}$ belonging to image I_A, point p_i is considered stable if its best match in a set of points $\mathbf{q} = \{q_1, q_2, ..., q_n\}$ belonging to image I_B, say q_j (forward), also has as best match the original point p_i in image I_A (backward). If the previous condition is met, then both points considered stable, if not, the points are discarded. The Euclidean distance is used as a measure of similarity between feature point descriptors.

Algorithm 1 Forward-backward matching.

Input: $\mathbf{p}, \mathbf{q}$
Output: $StablePoints\_\mathbf{p}, StablePoints\_\mathbf{q}$
  $StablePoints\_\mathbf{p} \leftarrow \emptyset$
  $StablePoints\_\mathbf{q} \leftarrow \emptyset$
  for all p in $\mathbf{p}$ **do**
    $\mathbf{q} \leftarrow$ findBestFeatureMatch of p in $\mathbf{q}$
    $\mathbf{p}' \leftarrow$ findBestFeatureMatch of q in $\mathbf{p}$
    if $\mathbf{p} = \mathbf{p}'$ **then**
      $StablePoints\_\mathbf{p} \leftarrow StablePoints\_\mathbf{p} \cup p$
      $StablePoints\_\mathbf{q} \leftarrow StablePoints\_\mathbf{q} \cup q$
    end if
  end for

5.2.3 Point-based affine registration

Feature correspondences that have been established in the previous stages are used as input for a point-based image registration algorithm that fits an affine transformation model to the set of features by minimizing the RMS error between feature correspondences. The image is then transformed and interpolated according to this affine transformation model. It is worth noting that only the RMS error between the feature correspondences drives the registration procedure and not the image intensities.

A reasonable assumption is that feature correspondences are noisy. That means that they are likely to be contaminated by outliers, e.g. matching features do not correspond to matching anatomical structures. Using RANSAC we can learn the parameters of an affine transformation model that is robust against outliers. Initially developed by [67], RANSAC

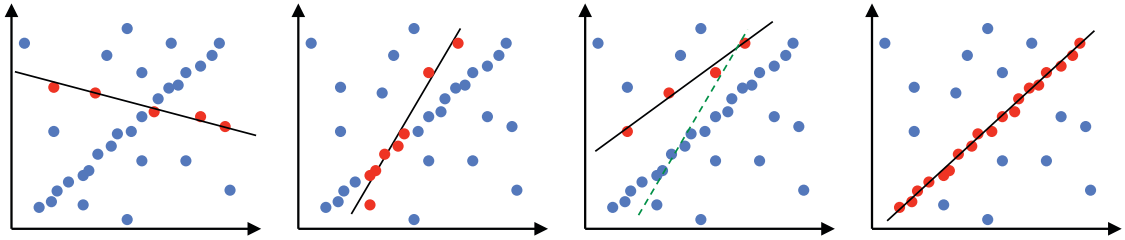


Figure 5.2: From left to right, four successive iterations of RANSAC. Models are randomly initialized, the best model is kept until convergence or exit criteria is met.

is a non-deterministic algorithm that iteratively estimates parameters of a model in the presence of large amounts of outliers. Rather than using the full set of points to estimate a model, RANSAC uses a minimal random subset to estimate an initial model. This model is then tested on the remainder of the data points and if any other point is well represented by the model (up to an error tolerance) it is added to the subset. This process is repeated on different subsets until the rank of the model is above a certain predefined threshold. The rank is determined by the number of points contained in the subset. The higher the number of points, the higher the rank and the better the model explains the data. Figure 5.2 shows a basic example of RANSAC.

5.3 Data and Results

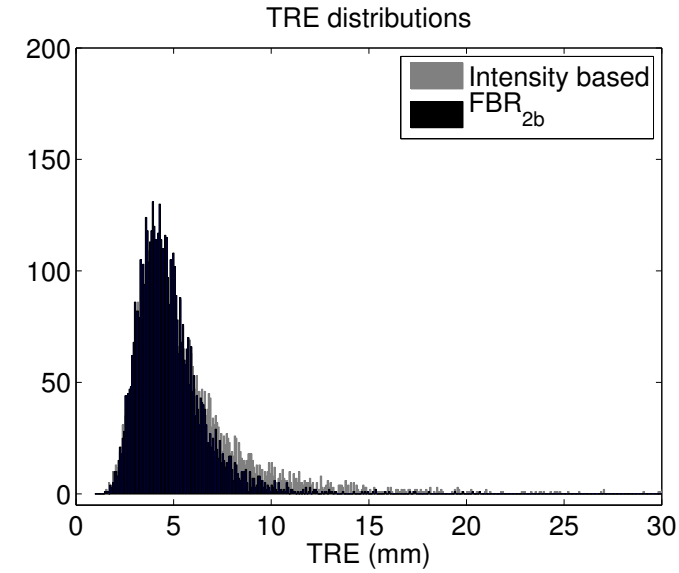
Images used to evaluate the proposed method where obtained from the OAI public use dataset (groups 1.C.0 and 1.E.0, available at http://www.oai.ucsf.edu). A subset of 75 images were randomly selected and manually annotated by an expert using three orthogonal views by placing four landmark points on the anterior collateral ligament (ACL) and posterior collateral ligament (PCL) insertions on the femur and the tibia (see Appendix C for a description on how the landmarks are defined). The central voxel of each ligament insertion is selected at the bone interface. The OAI dataset consists of multiple image sequences for each subject. In the following, the fat-suppressed, sagittal 3D double echo steady state (DESS) sequence with selective water excitation was used. The images have an in-plane resolution of 0.36 x 0.36mm and slice thickness of 0.7mm [158] (see Appendix B for more details).

We followed the work flow outlined in Figure 5.1: Using down-sampled images, we

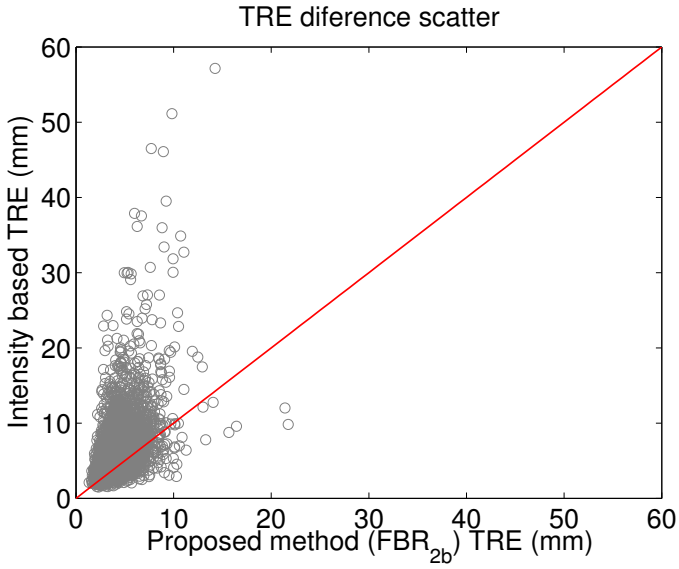
filtered out smooth regions of the image using the structure tensor (Section 5.2.2). After this, we calculated 3D local SS features in the remaining regions for a regular grid for every second voxel in the in-plane direction and for each slice using a correlation window and a patch size of radius of 5 voxels. We empirically found these values to be adequate for representing the structures at hand. We then reduced the number of features using sparsity and SS thresholds of 0.25 and 0.9 respectively (Section 5.2.2). For a each pair of images we used a forward-backward matching algorithm to find stable points, using a search window of ± 30 voxels.

Further to this pre-processing two different variants for the matching were explored: In the first variant we use the matching features as input for a point-based affine registration in order to obtain a affine transformation between the images, which we call feature based registration (FBR). In the second variant we used RANSAC to estimate the parameters of the affine transformation model. We refer to this as feature based registration with RANSAC (FBR^{+}). Since the 3D local SS descriptors are not completely rotationally invariant, an iterative process could prove to be beneficial, as feature matches would become more accurate as the images are increasingly better aligned. Using the output from FBR and FBR^{+}, the process was repeated using the transformed image as input for a second iteration. At this point, two different search window sizes, ± 15 and ± 40, in the forward-backward matching algorithm where tested. The method referred to as FBR_{2a} used a ± 15 voxel window, while the methods FBR_{2b} and $\text{FBR}^{+}_{2b^{+}}$ (using RANSAC on both iterations) used ± 40 voxel window. To assess the performance of the proposed method, the FRE of the previously defined landmarks was compared to an intensity based affine registration (AfR), that minimizes the normalized mutual information using gradient descend optimization [198]. FRE values where calculated for all the possible pairwise registration (n=5550).

Table 5.3 shows FRE for the proposed methods. Our method shows a very substantial improvement in terms of mean FRE and standard deviation, 2.13 and 2.47 mm respectively, $\sim$36% over AfR. In Figure 5.3 (a) and Figure 5.4 (a) the FRE distributions of the proposed method and AfR are shown. In both cases the proposed method distribution is more skewed towards zero (specially in Figure 5.4) than AfR, indicating an overall improvement. Figure

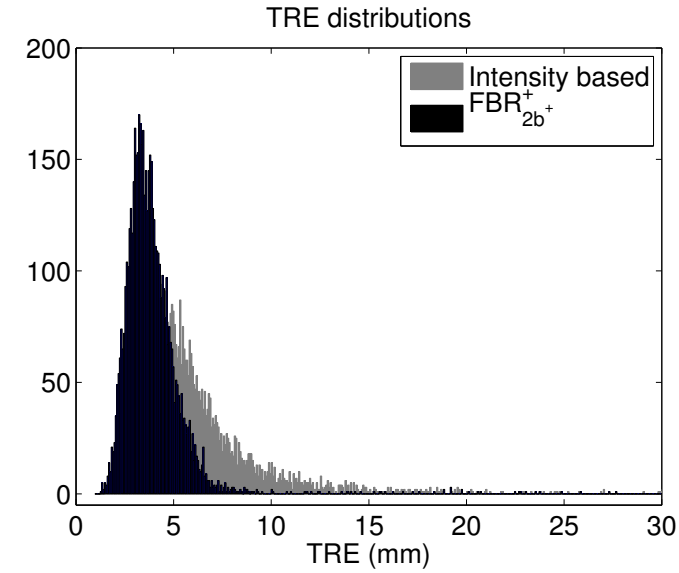


(a)

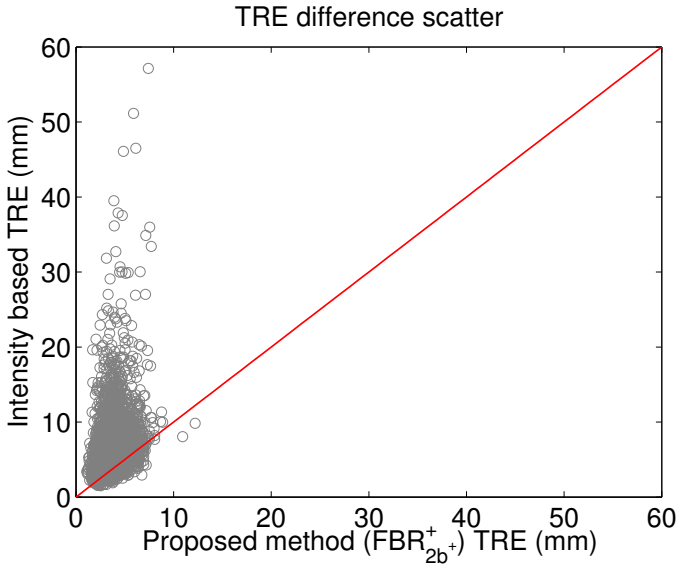


(b)

Figure 5.3: FRE comparison between the proposed method without using RANSAC and intensity based affine registration.



(a)



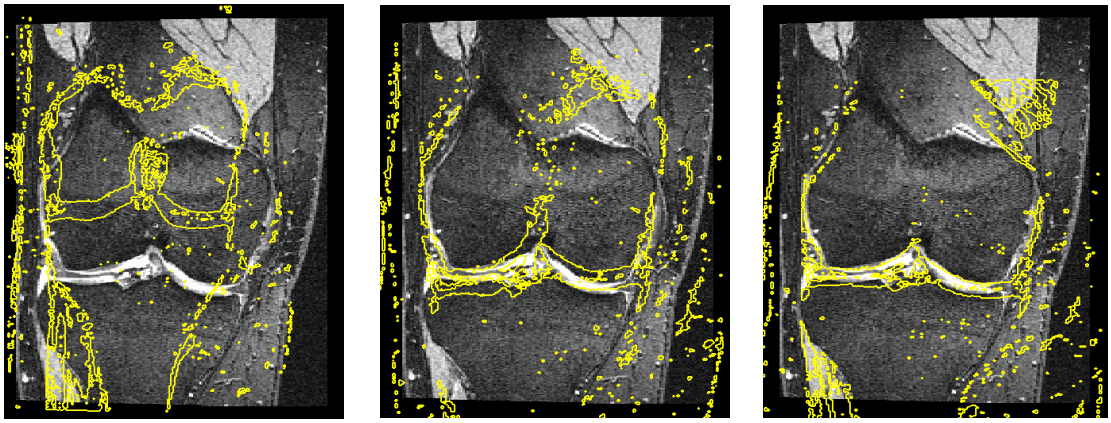
(b)

Figure 5.4: FRE comparison between our method using RANSAC as an estimator and intensity based affine registration.

Landmark	AfR	FBR	FBR^{+}	FBR_{2a}	FBR_{2b}	$FBR^{+}_{2b^{+}}$
ACL femur	5.62(3.88)	4.43(2.75)	3.15(1.94)	3.52(2.11)	3.66(1.74)	**3.05(1.41)**
ACL tibia	6.17(4.10)	5.70(3.72)	4.50(2.73)	4.75(3.03)	5.03(2.62)	**4.12(2.04)**
PCL femur	6.08(3.97)	5.04(2.97)	**3.68**(2.11)	4.19(2.38)	4.42(2.25)	**3.68(1.79)**
PCL tibia	5.56(3.82)	4.86(3.03)	4.36(2.58)	4.38(2.51)	4.12(2.17)	**4.07(2.00)**
All mean error	5.86(3.53)	5.00(2.58)	3.92(1.77)	4.21(1.93)	4.38(1.41)	**3.73(1.06)**

Table 5.1: Accuracy of the proposed methods. Errors in mm with standard deviation in brackets.

5.3 (b) and 5.4 (b) show scatter plots of the FRE obtained by the proposed method FBR_{2b} vs AfR and $FBR^{+}_{2b^{+}}$ vs AfR, respectively. Values above the diagonal line indicate and improvement of the proposed method over AfR. Furthermore, outliers (in regards to the FRE of the four landmarks) are almost completely removed in the cases of FBR_{2b} and $FBR^{+}_{2b^{+}}$, which shows the robustness of the algorithm. Figure 5.5 shows a comparison between AfR, FBR_{2b} and $FBR^{+}_{2b^{+}}$. In the background is the target image, while the overlaid contours correspond to the source image after registration, a clear improvement is shown.



(a) (b) (c)

Figure 5.5: Registration results after using (a) AfR, (b) FBR_{2b} and (c) $FBR^{+}_{2b^{+}}$. Average FRE of 25.19, 7.08 and 3.15mm, respectively.)

5.4 Conclusions

We have proposed a method that uses SS features instead of image intensities in order to establish feature correspondences between a pair of images. This enables us to register images in a more robust and accurate way. We have shown quantitative results that demonstrate the improvements obtained over intensity-based registration. This has been demonstrated using

a subset of images from the OAI database consisting of 75 randomly sampled MR images. Using exhaustive pairwise registration we obtained a FRE improvement in $\sim$82% of the cases, while virtually eliminating any outliers. It was observed that a two-iteration approach in which two consecutive feature matching/transformation are done, that is the output of the first iteration (after transformation) is the input of the next one, helped to eliminate outliers. We also observed that adding a further intensity-based registration only degrades the quality of the registration. This confirms that the use of the 3D local SS features play a key role in the robust registration. However, a week point of the evaluation done is that it relies in the calculation of the FRE for only four landmarks, which is far from ideal to measure true registration error. Additionally, the four landmarks used do not cover the whole volume of the image and focus on a rather specific part knee (the ligament insertions). An interesting avenue for further work would be to use the proposed model for non-rigid registration. However, non-rigid transformation models usually have a high degree of parameters and RANSAC might not be able to estimate a meaningful model from the relatively low number of matching features. This could be addressed using a piecewise affine model. The following chapter retakes the brain MR image landmark localization task introduced in Chapter 4. Here, a machine learning framework that, leveraging on manifold learning and regression, estimates the direction and distance to the landmark at every location. This is in contrast to the use of a sliding window approach that detects (or not) the location of the landmark.

Chapter 6

Laplacian eigenmaps for automatic landmark localization

This Chapter is based on:

- R. Guerrero, R. Wolz, D. Rueckert. Laplacian eigenmaps manifold learning for landmark localization in brain MR images. Medical image computing and computer-assisted intervention (MICCAI), volume part II, pages 566-573, 2011.

Abstract

In this chapter we propose to address the problem of identifying anatomical landmarks in medical images. A manifold learning approach based on Laplacian eigenmaps that learns an embedding from patches drawn from a training set of annotated images is introduced. The position of the patches in the manifold can be used to predict the location of the landmarks via regression. New image patches are embedded in the manifold and the resulting coordinates are used to predict the landmark position in the new image. The output of multiple regressors is fused in a weighted fashion to boost accuracy and robustness. We demonstrate this framework localizing 20 anatomical landmarks in 3D brain MR images from the ADNI database. In addition we locate 7 landmarks in a database of face images in order to demonstrate the method's ability to generalize beyond medical images. We compare the proposed method to two alternative approaches, a Sliding window detector with Haar features and non-rigid registration-based landmark localization. The proposed approach has an average landmark localization accuracy of ∼1.24mm for brain MR images and ∼1.75 pixels for facial images. This demonstrates improved performance compared to sliding-window and registration-based approaches.

6.1 Introduction

The localization of anatomical landmarks is a crucial step in many medical imaging applications. In registration, landmarks can be used to define corresponding anatomical points in different images. Matching the landmarks across images and interpolating the correspondences between landmarks, e.g. using thin-plate splines [20, 89, 19], yields a registration that represents faithfully the anatomy of the structures that the landmarks belong to. Similarly, several segmentation algorithms require seed point initialization and anatomical landmarks can be used to initialize such algorithms [159, 237]. In morphometry, landmarks can be used to obtain quantitative measures from anatomical structures and compare them across different images; e.g. in [75], PCA of the distribution of 24 brain landmarks across different subjects is analyzed and a statistically significant difference between left and right

hemispheres is revealed (the landmarks are located on a 3D reconstruction of the brain's cortical surface).

In this work we propose a manifold learning approach that is capable of learning a low-dimensional embedding of image patches. The assumption is that the local anatomy around a particular landmark is well-represented in this embedding. We can then learn a regression model that predicts the displacements between the patch location and the landmark. Image patches from unseen images are mapped to the learned manifold using an out-of-sample approach [18] and the regression model is then used to obtain an estimate of the landmark's position. Finally, a consensus of the predictions made by several patches (belonging to the same image) is formed by computing a weighted average of all the estimates. The approach has been evaluated by training on a large dataset of 100 brain MR images from CN subjects, MCI and AD patients from the ADNI[1] study. A different set of 100 images from ADNI is used for testing the proposed approach. Additionally, in order to demonstrate the method's ability to generalize to different types of images, we also evaluate it using 400 face images from a publicly available database for facial expression analysis [154]. The face data was randomly split into two independent subsets of 300 images for training and 100 for testing.

6.2 Method

6.2.1 Manifold Learning

Manifold learning in general refers to a set of machine learning techniques that aim at finding a low dimensional representation of high dimensional data while trying to faithfully represent the intrinsic geometry of the data. Some of the most popular manifold learning techniques include MDS [209], Isomap [203], LLE [180] and Laplacian eigenmaps [17]. A review and comparison of different manifold learning techniques can be found in Chapter 3.

In our framework, given a set of N_I images, we extract N_P equally sized patches from each image in a ROI around a landmark. Each of these patches, consisting of D voxels, is stored as a vector of intensities $\mathbf{x}_n = \{\mathrm{x}_1, ..., \mathrm{x}_D\} \in \mathbb{R}^D$. The set of patches is denoted

[1] www.loni.ucla.edu/ADNI

as $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, where $N = N_I \cdot N_P$. Our aim is to learn the underlying manifold in $\mathbb{R}^d$ ($d \ll D$) that represents the relationship between patches in the vicinity of a given landmark. Specifically, we intend to learn a manifold that can be used to predict the displacement $\Delta_n = \{\delta_x, \delta_y, \delta_z\}$ between the center of the patch and the landmark in question. Manifold learning techniques offer a powerful approach to find a representation of images that facilitates the application of statistical machine learning techniques such as regression. Since the patches are expected to lie on or near to a non-linear manifold, the Euclidean distance between patches in the original space is not necessarily meaningful and cannot be used for regression. After uncovering the manifold structure in the data, the Euclidean distance in the embedded space provides a more meaningful approximation of the geodesic distance in the original space and is thus more suitable for regression.

Laplacian eigenmaps

Laplacian eigenmaps can be used to find a low-dimensional representation of the data $f : \mathbf{X} \rightarrow \mathbf{Y}$, $\mathbf{y}_i = f(\mathbf{x}_i)$ while preserving the local geometric properties of the manifold [17]. Laplacian eigenmaps use a local neighborhood graph to approximate geodesic distances between data points. In this work we use the Euclidean norm as a distance (similarity) metric to identify the k-neighborhood around each point. From these distances a sparse neighborhood graph G is constructed. Furthermore, a weight matrix $\mathbf{W}$ assigns a value to each edge in G, and is computed using a Gaussian heat kernel

$$w_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2}\right) , \tag{6.1}$$

with standard deviation σ.

Laplacian eigenmaps aim to place points $\mathbf{x}_i$ and $\mathbf{x}_j$ close together in the low-dimensional space if their weight $w_{i,j}$ is high, e.g. if they are close in the original, high-dimensional space. This is done by means of minimizing the cost function given by

$$\phi(\mathbf{Y}) = \operatorname{argmin} \sum_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 w_{i,j} , \tag{6.2}$$

under the constraint that $\mathbf{y}^T\mathbf{D}\mathbf{y} = 1$ which removes an arbitrary scaling factor in the embedding and prevents the trivial solution where all $\mathbf{y}_i$ are zero. The minimization of Equation (6.2) can be formulated as an eigenproblem [10], through the computation of the degree matrix $\mathbf{M}$, which is a diagonal matrix that contains information about the degree of each vertex of $\mathbf{W}$, and the Laplacian $\mathbf{L}$, where $\mathbf{L} = \mathbf{M} - \mathbf{W}$ and $m_{i,i} = \sum_j w_{i,j}$. Hence, the low-dimensional manifold $\mathbf{Y}$ that represents all data points can be obtained via solving a generalized eigenproblem

$$\mathbf{L}\nu = \lambda\mathbf{M}\nu \,, \tag{6.3}$$

where ν and λ are the eigenvectors and eigenvalues, and in turn the d eigenvectors ν corresponding to the smallest (non-zero) eigenvalues λ represent the new coordinate system.

Approximate nearest neighbors (ANN)

Since we are learning a manifold from a large number of examples (in our case 133100 examples, see Section 6.4), the similarity matrix $\mathbf{W}$ that needs to be calculated in the Laplacian eigenmaps algorithm is very large ($\sim$18 billion elements). Even though it is strictly k-sparse, calculating exact nearest neighbors would mean that a non-sparse matrix would need to be calculated first, in order to find the k nearest neighbors and then sparsify $\mathbf{W}$; making the calculation of all the exact pairwise distances computationally unfeasible. We therefore, instead calculate approximate nearest neighbors using a hierarchical k-means tree, in order to speedup the nearest neighbor search. A k-means tree is constructed by splitting all the data points into k_m distinct regions using the k-means clustering algorithm (of complexity $O(N^{Dk_m+1}\log N)$, where D is the dimensionality) for a given number of iterations where the k_m seed points are chosen randomly. This is repeated recursively (on each of the newly formed clusters) until the number of data points in each region falls below k_m [150]. This is implemented in the fast library for approximate nearest neighbors (FLANN) [149]. Queries are computed by exploring the tree in a best-bin-first manner, as this has been shown to improve the exploration of kd-trees by up to two orders of magnitude. This search has a complexity of $O(N^{1-1/d} + k)$, where d is the dimension of the tree.

Out of Sample Extension

For the application considered in this work, it is necessary to map new patches into the manifold in order to use the embedded coordinates to make a prediction via regression. For linear dimensionality reduction techniques like PCA this is straightforward, as they provide a projection matrix for exact transformation between the original and the embedded space. Unfortunately, this is not the case for most non-linear methods. Therefore, approximation techniques must be used. We address this problem by using an out of sample technique that employs the Nyström approximation [163], which approximates the eigenvectors of a large matrix based on the eigendecomposition of a submatrix of the large matrix, to formulate a training set dependent normalized kernel. Using this kernel $\tilde{K}$, an approximate mapping from the high dimensional space to the low dimensional manifold is obtained. To embed a point into a manifold first we find the nearest neighbors of the new point, belonging to the test set $\mathbf{X}'$, in the training set $\mathbf{X}$. Then, the kernel $\tilde{K}$ is used to assign a weight to each of its nearest neighbors. Finally an approximate mapping to the manifold is calculated using the weighted average of the low dimensional coordinates of its high dimensional nearest neighbors. The equivalent, training set dependent normalized kernel, is given by:

$$\tilde{K}(\mathrm{x}'_i, \mathrm{x}_j) = \frac{1}{N} \frac{K(\mathrm{x}'_i, \mathrm{x}_j)}{\sqrt{E_{\mathbf{X}'}[K(\mathrm{x}'_i, \mathbf{X})] E_{\mathbf{X}}[K(\mathrm{x}_j, \mathbf{X})]}} \,, \tag{6.4}$$

where K is a Gaussian heat kernel (Equation (6.1)), x_j and x'_i are points from the training $\mathbf{X}$ and test $\mathbf{X}'$ datasets, respectively. The expectations are taken over the empirical data and N is the number of training samples (see [18] for full analysis). For every new patch we embed in the manifold we need to find its k nearest neighbors. This is done using either exact nearest neighbors or approximate nearest neighbors using the search tree defined in Section 6.2.1.

6.2.2 Spatial Prior Probabilities

Assuming that the structure of interest is in some approximately known orientation and position, a landmark's spatial location is bounded, to a certain extent, to a particular volume

or area within the image. That is, once the images have been affinely registered, the possible locations of each of the n landmarks is bounded within this space. Thus, we can restrict the search for each landmark to those locations which have a non-zero probability for the location of the landmark. We model the spatial prior probabilities of each landmark, based on the position of the landmark in the training set, using kernel (or parzen window) density estimation. This can be formulated as:

(6.5)

where p are either the 3D voxel coordinates (x,y,z) or 2D pixel coordinates (x,y), $\iota_i \in \{\iota_1,...,\iota_n\}$ are all the landmarks in the training set, $\mathbf{K}_p(\cdot)$ is the the kernel function in a d-dimensional space such that $\int_{\Re^d} \mathbf{K}_p(\mathrm{x})d\mathrm{x} = 1$, and $h_n > 0$ is the width of the kernel. The kernel function $\mathbf{K}_p(\cdot)$ is modeled as a Gaussian function. Figure 6.1 shows an example of the prior probability map of 20 landmarks in a brain MR image.

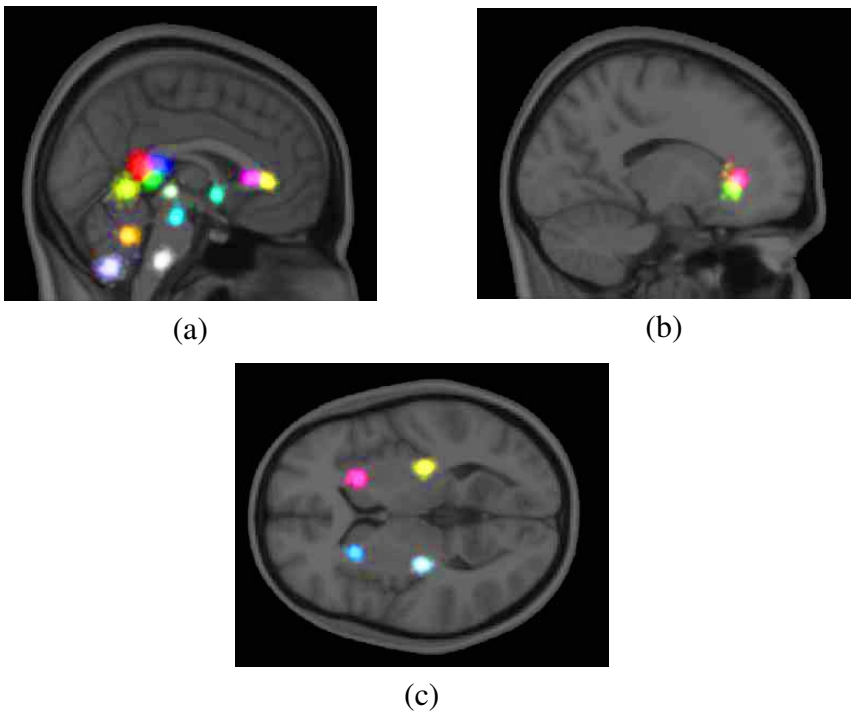


Figure 6.1: Estimated prior probability distribution for all the landmarks. (a) Splenium and genu of corpus callosum, superior and inferior tip of the cerebellum, fourth ventricle, anterior and posterior commisure, and superior and inferior aspect of the pons. (b) Anterior and inferior tip lateral ventricle (only left side shown). (c) Superior and inferior tip of the putamen (left and right). Contrast in probability maps has been enhanced to facilitate visualization.

6.2.3 Landmark Prediction

Using the low-dimensional coordinates $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ of each patch (that are estimated from the training set) and their corresponding displacements Δ_n (between the center of the patch in image space and the position of the landmark), we fit a linear regressor, using $\mathbf{Y}$ as independent variables and Δ_n as dependent variables, to obtain an estimated displacement $\Delta'_n = \{\delta'_x, \delta'_y, \delta'_z\}$,

$$\Delta'_n = \mathbf{Y}'\mathbf{b} + \varepsilon \simeq \begin{pmatrix} 1 & \mathrm{y}_{11} & \cdots & \mathrm{y}_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathrm{y}_{N'1} & \cdots & \mathrm{y}_{N'd} \end{pmatrix} \begin{pmatrix} b_{x0} & b_{y0} & b_{z0} \\ \vdots & \vdots & \vdots \\ b_{xd} & b_{yd} & b_{zd} \end{pmatrix}, \tag{6.6}$$

where the error term ε has been neglected.

A test dataset $\mathbf{X}'$ is built by extracting patches from an unseen images at random locations within its specified non-zero prior probability region. These patches are embedded in the landmark specific-manifold (e.g. a separate manifold is constructed for each landmark) using the out-of-sample approach, described in Section 6.2.1, to obtain their low-dimensional representation $\mathbf{Y}'$. Using the learned regressor coefficients β, an estimate of the displacement from patch n to the landmark is obtained. Since patches that predict small displacements tend to have a higher accuracy than those that make large displacement predictions, a weighted average is calculated, where the weights of individual predictions' are based on the magnitude of the displacement Δ'_n. We use a Gaussian kernel, $\mathbf{K}_w(\Delta'_n) = \exp(-\Delta'_n/2{\sigma_w}^2)$, to calculate the weight of each prediction.

6.3 Comparison to other landmark detection approaches

In this section we briefly describe two other different methods commonly used for landmark localization: (a) Sliding window detector with Haar features (SW) and (b) nonrigid image registration (REG). This two other methods were used as a comparison to the method proposed here (Section 6.5).

6.3.1 Sliding window detector with Haar features (SW)

A sliding window detector consists of testing a classifier in a subregion of an image (window) to see whether or not the object of interest is in this subregion, storing the result, moving the window to another location and testing again, until the whole image (or ROI) has been tested. In this work, a detector for each landmark was built using a variation of the Viola-Jones face detector [218]. In this approach, Haar-like features as shown in Figure 4.2 have been used due to their simplicity and fast computational speed. See Section 4.3.1 for further details on this type of landmark detector. The 3D sliding window detector, was trained using 400 positive examples, from randomly chosen images from the ADNI dataset (96 AD, 192 MCI and 112 CN patients) and 4000 negative ones, taken from the vicinity (within the ROI of the landmark and outside ± 4 voxels from it) of the positives to improve robustness. The classifiers were built as a single monolith (instead of a cascade of classifiers as in [218]) 100 feature (3D Haar features) classifier.

6.3.2 Nonrigid image registration (REG)

Registration can be used to propagate the annotation (e.g. a set of points or labels) present in a reference image to a new image. In this case the annotated reference image acts as a template (or atlas). Hence, landmarks that are annotated in the template can be propagated to new images. Figure 4.3 in Chapter 4 illustrates the landmark propagation procedure. An initial set of landmark annotations are carried out on the MNI152 template and once the images have been registered to this template, the locations of the landmarks can be determined by applying the non-rigid transformation to their position. In this work we used the method proposed in [182], where a FFD model based on B-splines is used to deform the underlying mesh of control points until an intensity similarity measure is optimized (in this case normalized mutual information). We used a hierarchical FFD approach in which the control point mesh is refined from a 20 mm spacing to 10 mm and then 5 mm. Preliminary experiments using an even finer control point spacing of 2.5 mm showed little or no improvement.

6.4 Data

Two different datasets were used to evaluate the proposed method: A 3D brain MR image dataset and, in order to show the generalization capabilities of the method, a 2D facial dataset.

6.4.1 Brain MR images

The images that were used to evaluate the proposed method were obtained from the ADNI database [148], see Appendix A. In this work, we used a subset of 1.5T T1-weighted images of 100 randomly chosen subjects for training and another 100 randomly chosen subjects for testing. In both (training and testing) datasets there are 24 AD, 48 MCI and 28 healthy subjects, to truthfully represent the full ADNI dataset. All images were acquired at baseline. Brain images were skull-stripped, affinely aligned to the MNI152 brain template and normalized using linear intensity rescaling, prior to landmark localization.

In total 20 landmarks were used to learn 20 different landmark specific manifolds. Figure 4.5 in Chapter 4 shows the location of the 20 anatomical landmarks used in this chapter. Landmarks were manually annotated by an expert using three orthogonal views, Appendix C.1 gives their description. The high-dimensional training set $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, is obtained by collecting 3D image cubic patches of 21^3 voxels around a regular grid that is centered at the landmark, from 100 different brain MR images. The regular grid has a spacing of 3 voxels and a displacement of ± 18 voxels in each axis from the landmark. This volume is chosen so that it includes the non-zero probability volume obtained from the PDF estimation. For each image in the training set we sample 11^3 (1,331) patches from this grid, as this amount was deemed sufficient to learn the subspace. Doing this for the 100 images in the training set and rearranging them so that each patch is represented as a column vector (with 21^3 intensity values), yields a 133,100 by 9,161 (N, D) matrix that contains all the patches, from all the training images, around the landmark in question.

6.4.2 Facial images

The database is comprised of 400 color images of faces (taken from the web-based database for facial expression analysis [154]), with various expressions and different styles of head and facial hair. Seven landmarks (see Table 6.5.2 for listing) have been manually annotated by an expert. Based on this annotations we first rigidly aligned the faces, crop them, convert to grayscale and normalized both intensity and size. Each of the preprocessed images has a size of 128^2 pixels. The 400 images where randomly split in to two independent sets, 300 for training and 100 for testing.

The high-dimensional training set, $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, used in this work, is obtained by collecting image patches of 27^2 pixels around a regular grid that is centered at the landmark, from the 300 training facial images. The grid has a spacing of 2 pixels and a displacement of ± 10 pixels in each direction from the landmark. This area is chosen so that it includes the non-zero probability area obtained from the PDF estimation. For each image in the training set we sample 11^2 (121) patches from this grid. Doing this for the 300 images in the training set yields a 36,300 by 729 (N, D) matrix that contains all the patches, from all the training images, around the landmark in question.

6.5 Results

6.5.1 Brain dataset

From the training set $\mathbf{X}$ we learn the underlying low-dimensional manifold, using Laplacian eigenmaps (Section 6.2.1). The parameters k (nearest neighbors in the neighborhood graph), d (the output dimensionality of the data) and σ from the Gaussian heat kernel, were empirically set to 50 (except for the case of the inferior aspect of pons, which required 130 nearest neighbors), 80 and 1, respectively. The parameter k was chosen to yield a fully connected neighborhood graph. This ensures that all the displacements, Δ_n, from the landmark were equally represented. That is, any patch left out would mean an under representation of its associated displacement in final graph (largest connect component). The coefficient

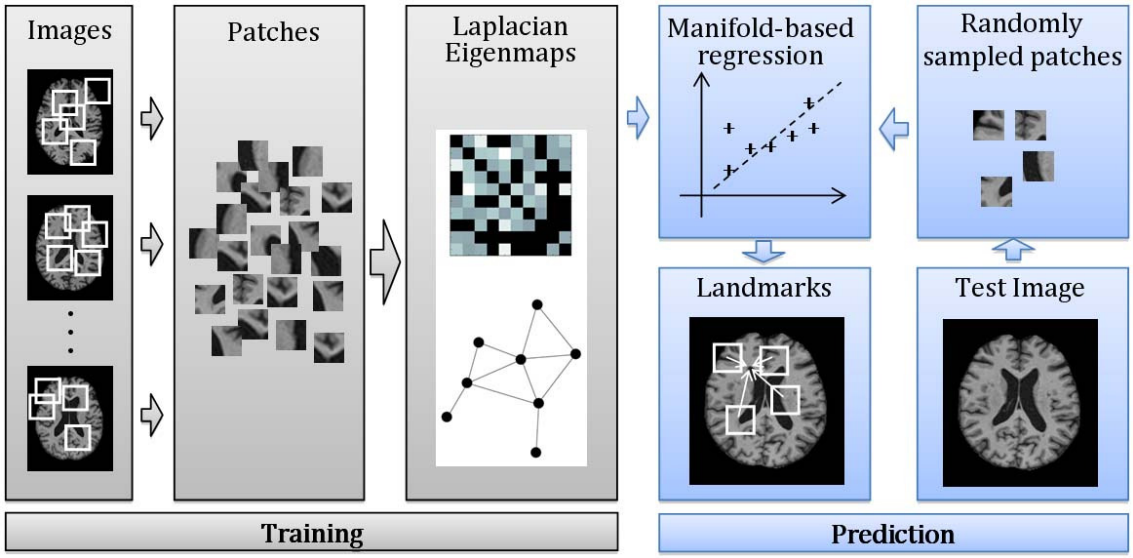


Figure 6.2: Diagram of method's training and testing steps.

of determination, R^2, for the linear regression was used as an indicative to determine the final dimensionality d of $\mathbf{Y}$, with values of around 0.9 obtained for 80 dimensions, with stable behavior of the algorithm observed for final dimensionalities between 30-400. Finally, varying the parameter σ showed little improvement.

For each new test image, we sampled 100 patches at random locations, within a non-zero probability ROI obtained from the estimated PDF and not necessarily belonging to the grid used in the training. We then embedded the new points (patches) into the learned low-dimensional manifold using the out-of-sample technique described in Section 6.2.1 and used the learned regression model to obtain a prediction from each point. A final landmark prediction for each image is obtained using a weighted average, as described in Section 6.2.3, where σ_w from the weighting Gaussian kernel $\mathbf{K}_w(\Delta'_n)$ was tuned to each landmark-specific case (with σ_w 1-1.35). An overview of the whole process is shown in Figure 6.2.

Table 6.1 shows results comparing the proposed method, landmark specific manifold (LM), with the two other possible approaches described in Section 6.3: SW and REG. A five random subsample cross-validation of the method was carried out in order to asses the results and ensure reproducibility (the average of the five tests is shown, with a variability among tests of $\sim$0.1mm). Stable behavior was observed in the five tests. To embed new patches in the manifold for testing, we used the search tree to find approximate nearest

neighbors learned in the training phase. Our methods shows relatively consistent result with all landmarks throughout the tested dataset.

A statistical comparison of the three classes of patients (AD, MCI and CN) was performed on the training set. It was observed that, obtaining the average landmark position (for each landmark) and comparing intra-class distance variation, shows no statistical differences between classes. Also, the average intra-class landmark prediction error shows no statistically significant variation across the groups. In an attempt to achieve higher accuracy we explored using exact nearest neighbors in the testing stage, as the similarity matrix calculated here is substantially smaller, thus making its computational cost more feasible. However, a lower accuracy was obtained with this method (as well as higher computation times).

As can be seen in Figure 6.3, landmarks with higher variability, that is, their location in the brain is less constrained, e.g. the superior aspect of cerebellum (L8), tend to have a higher prediction error than those landmarks with a more stable location, e.g. anterior and posterior commissure (L14, L15), in terms of the prediction accuracy by the proposed algorithm. A 2D visualization of the splenium of corpus callosum's (outer aspect) manifold is shown in Figure 6.4. In order to facilitate visualization, only 1000 points where plotted (instead of the 133,100). The corresponding MR image (a sagital slice of the patch) of four pairs of points are displayed in order to show that local neighborhoods in the manifold represent patch-similarity (structural significance) in the input space.

The proposed method was implemented in Matlab, using a combination of the FLANN [149] and the Matlab Toolbox for Dimensionality Reduction [212], as well as some code optimization of the latter, mainly the out-of-sample implementation. Training a landmark-specific manifold requires around 3 hours on a 2.67GHz 12-core machine with 64Gb of RAM. Although only the eigensolver takes advantage of more than one core, most of the process runs on a single core. For testing, if exact nearest neighbors are found for the out-of-sample extension, embedding new points in the learned manifold takes about 0.16 seconds, meaning that each landmark is located in about 16 seconds (100 patches are used to predict each landmark). If instead of exact nearest neighbors, we use the learned search tree used

Anatomical landmark	LM	SW	REG
Splenium of corpus callosum (outer aspect)	**1.27(0.63)**+*	1.75(1.04)	3.95(1.43)
Splenium of corpus callosum (inferior tip)	**1.13(0.43)**+*	1.46(0.75)	2.10(0.89)
Splenium of corpus callosum (inner aspect)	**1.30(0.57)**+*	1.81(0.99)	2.31(1.31)
Genu of corpus callosum (outer aspect)	**1.13(0.50)**+*	1.58(1.09)	1.73(1.01)
Genu of corpus callosum (inner aspect)	**1.03(0.47)**+*	1.28(0.67)	1.47(0.64)
Superior aspect of pons	**1.15(0.54)***	1.22(0.63)	2.79(1.26)
Inferior aspect of pons	**1.16(0.49)**+*	1.86(0.92)	1.70(0.85)
Superior aspect cerebellum	**1.68(0.72)**+*	2.27(1.71)	2.99(1.64)
Fourth ventricle	1.33(**0.54**)+*	**1.09**(0.65)	5.57(2.70)
Putamen posterior (left)	**1.14(0.47)**+*	2.21(1.22)	4.36(1.81)
Putamen anterior (left)	**1.24(0.48)**+*	1.86(1.13)	2.48(1.29)
Putamen posterior (right)	**1.08(0.48)**+*	2.20(1.22)	3.53(1.78)
Putamen anterior (right)	**1.28(0.52)**+*	2.31(1.61)	2.79(1.43)
Anterior commissure	**1.00**(0.79)+*	1.27(0.72)	1.05(**0.42**)
Posterior commissure	0.88(**0.36**)*	**0.79**(0.60)	1.85(0.48)
Inferior aspect cerebellum	**1.35(0.64)**+*	2.13(1.69)	3.71(1.68)
Anterior tip of lateral ventricle (left)	**1.28(0.58)**+*	1.86(1.14)	3.67(1.72)
Anterior tip of lateral ventricle (right)	**1.50(0.85)**+*	1.84(1.20)	3.65(1.73)
Inferior tip of lateral ventricle (left)	**1.47(0.85)**+*	2.28(1.42)	4.44(2.07)
Inferior tip of lateral ventricle (right)	**1.11(0.54)**+*	2.18(1.18)	4.01(1.79)

Table 6.1: Accuracy of the proposed method (LM) on the ADNI database, for the specified landmarks, and comparison SW and REG approaches. Errors in mm with standard deviation in brackets. Statistical significance (to %5, results not corrected for multiple comparisons) is indicated by + and *, for comparisons between LM and SW or REG, respectively.

to find the approximate nearest neighbors in the training stage, each landmark is detected in about 5.8 seconds.

A direct comparison between the method presented in this chapter and Chapter 4 (landmark localization using SS features) is not possible due to the fact that slightly different datasets where used. Nonetheless, on average manifold learning landmark localization clearly outperforms 3D local SS. The main disadvantage of manifold learning landmark localization is the memory requirements of the manifold.

6.5.2 Face dataset

Experiments where also carried out on a face database covering 400 subjects, as detailed in Section 6.4. The parameters k, d, σ and σ_w empirically set to 50, 20 (except for the left and right side of the mouth, which required 100 to ensure a fully connected graph), 1 and 1, respectively. The dataset was split into 300 images for training and 100 testing. For each landmark we predict, 100 patches are drawn from the non-zero probability area and embed

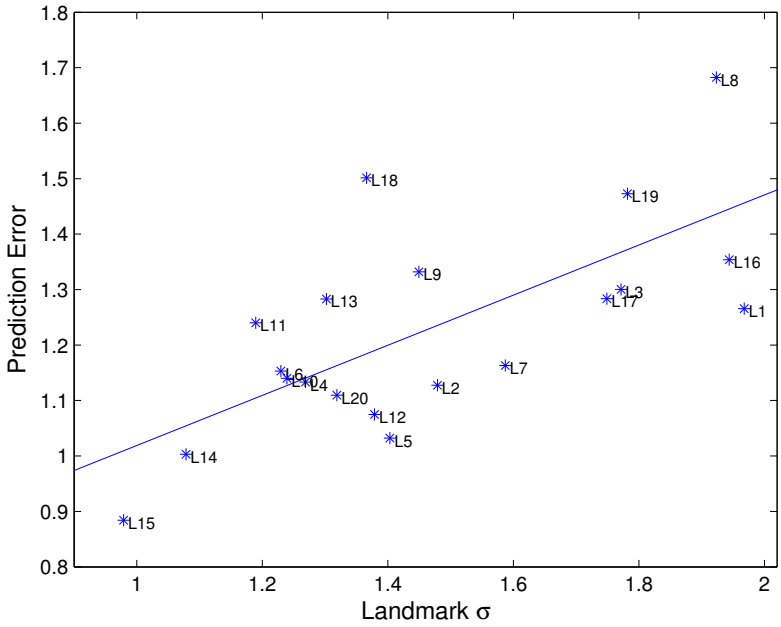


Figure 6.3: Average landmark predicted error vs. standard deviation of distance from average landmark position. More variable landmarks tend to have a higher prediction error.

them in the manifold to obtain a prediction from the learned regressors. The results of using the proposed method on this database are shown in Table 6.5.2, along with a comparison of localization accuracies of the same points by Martinez et al. [142]. A direct comparison to the work presented by Martinez et al. is not entirely possible, as the dataset used in their work is a combination of several datasets (including the one used for the evaluation of the proposed method). However, results are added for illustration purposes. Results are given in an interocular normalized distance in order to ease comparison between the methods. Figures 6.5 show some comparisons between the predicted landmarks and the ground truth.

Landmark	**LM**	**Martinez et al. [142]**
Lower lip	**0.034(0.021)**	0.061(0.098)
Upper lip	**0.033(0.023)**	0.037(0.041)
Left side of mouth	**0.031(0.018)**	0.041(0.079)
Right side of mouth	**0.036(0.022)**	**0.036**(0.040)
Nose	**0.030(0.018)**	0.035(0.36)
Left eye (pupil)	0.025(0.015)	—
Right eye (pupil)	0.025(0.014)	—

Table 6.2: Intermolecularly normalized accuracy of the proposed method on the facial images database and a comparison the the accuracies obtained by Martinez et al [142]. Best results in bold and standard deviation in brackets.

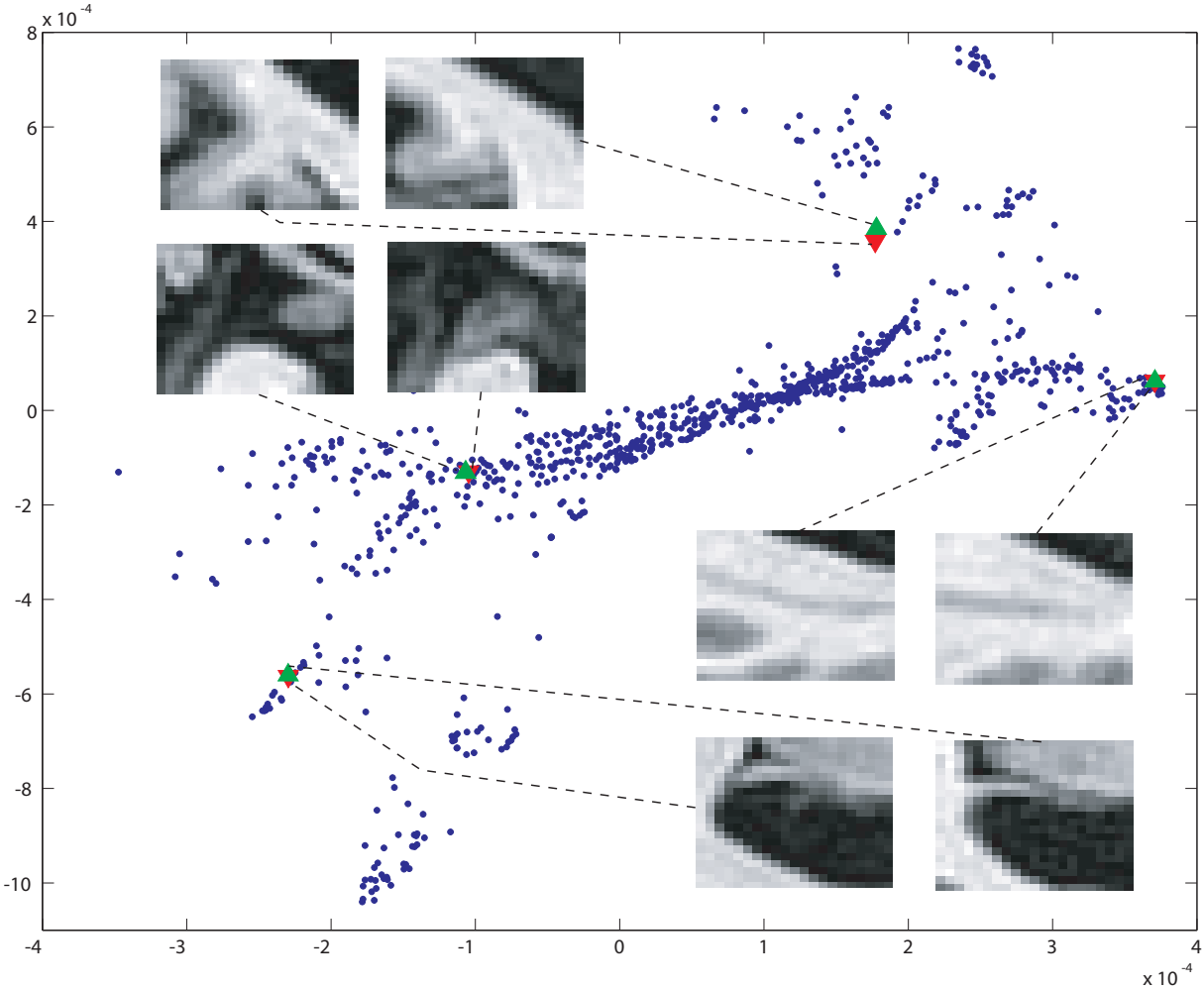


Figure 6.4: 2D visualization of the splenium of corpus callosum's (outer aspect) manifold. 1000 points used for visualization.

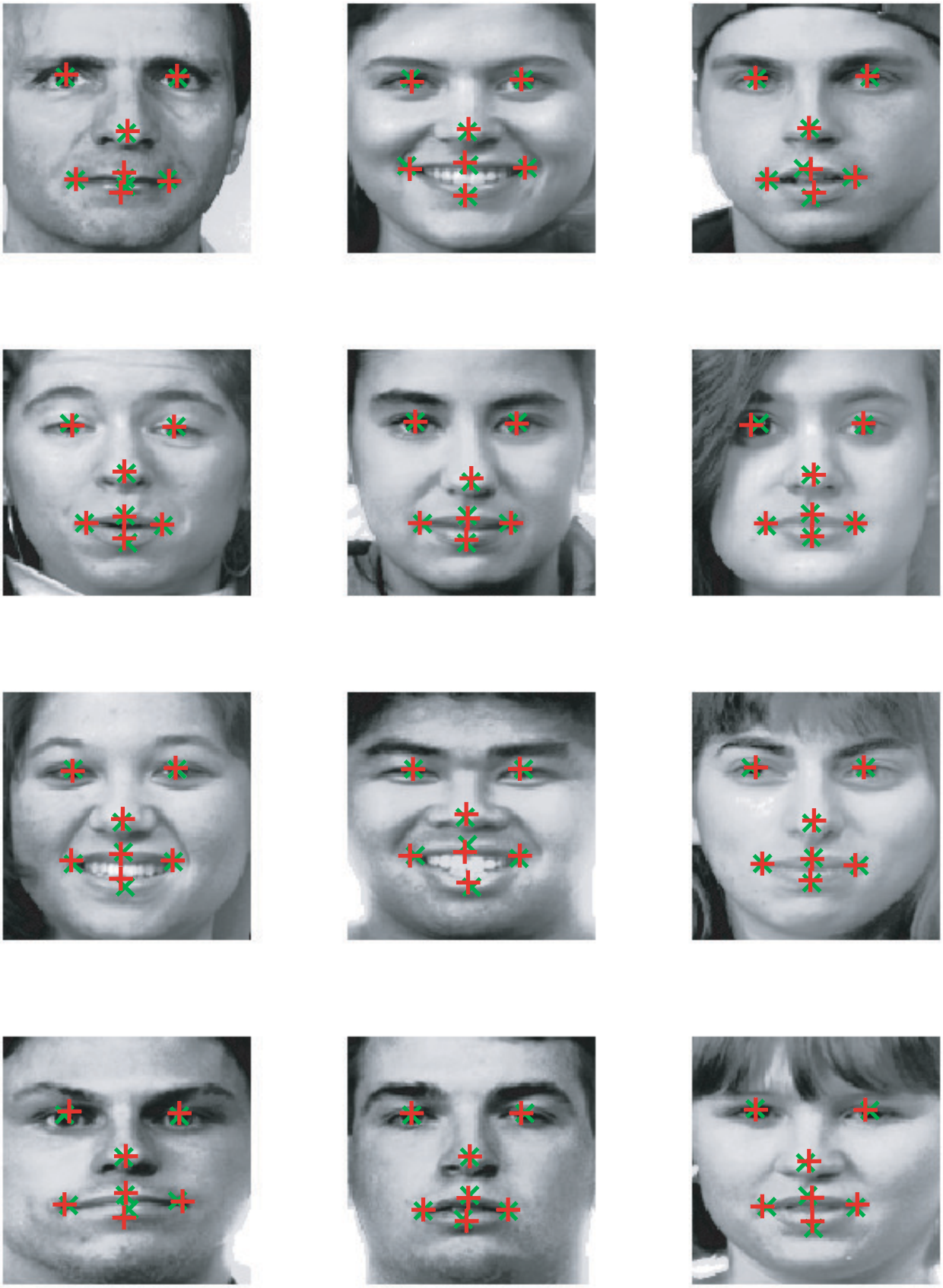


Figure 6.5: Results on face dataset. In red the landmarks predicted with the proposed method and in green the manually annotated landmarks. Best seen in color.

6.6 Conclusions

We have proposed a method that uses Laplacian eigenmaps to learn a low-dimensional manifold that represents local anatomy around a specific landmark in brain MR images and human faces with different styles of facial and head hair. The landmark specific low-dimensional manifolds were learned using image patches (around the vicinity of the landmark) from 100 brain MR images belonging to the ADNI dataset and also from 400 facial images from the web-based database for facial expression analysis. Prior knowledge of the spatial distribution of the landmarks was used to reduce the search space. Our results show that the proposed method significantly outperforms the 3D sliding-window and non-rigid registration approaches mentioned in Section 6.3 in the MR image landmark localization task. Additionally, the method was shown to compare favorably to another state-of-the-art method in the face database. A key drawback of the presented approach is its high memory requirements as each manifold, along with its high dimensional data points, need to be preserved in order to embedded new patches to estimate a landmarks position.

Further improvement to the framework could be achieved using a more powerful regression techniques such as support vector regression or a multiple output regression. An interesting avenue to explore for future research is to extend the manifold learning approach from ROIs to the whole image. This would allow the fast identification of an arbitrary, dense set of landmarks. The located landmarks could then be used to establish correspondences and provide a fast, initial registration. However due to the need large amount of expert knowledge required to annotate training images, this seems unlikely to be done. If the approach is extended to pseudo-landmarks (e.g. control points as they are used in free-form deformation-based registration), this would enable learning the correlation between the appearance of image patches appearances and the associated deformations, similar to the work proposed in [121]. This may significantly reduce the computational cost and improve the robustness to local minima by finding subject-specific patch correspondences.

In this chapter we have described a very accurate way of locating landmarks in both brain MR and face images via linear regression on a learned subspace manifold. In Chapter 4 another method to locate landmarks in brain MR images was presented, where 3D local

SS features where used as image point descriptors. However, the implicit assumption that brain MR landmarks can be well represented using generic feature descriptors does not necessarily hold. In the following chapter we will explore the idea of learning descriptors specially suited to represent the data at hand, rather than making assumptions on how the data should be best represented.

Chapter 7

Data-specific feature point descriptor matching

This Chapter is based on:

- Ricardo Guerrero, Daniel Rueckert. ”Data-specific Feature Point Descriptor Matching Using Dictionary Learning and Graphical Models”. SPIE Medical Imaging, pages 866921-8, 2013.

Abstract

Matching landmarks in a pair of images is a challenging task. Although off-the-shelf feature point descriptors are powerful at describing points in an image, they are generic by nature, as they have been usually developed for applications in a general computer vision setting where there is little prior knowledge about the images. This chapter describes a general framework that leverages recent developments in the machine learning community with an aim of building feature point descriptors that are data-specific. The proposed approach describes landmarks as feature descriptors based on a sparse coding reconstruction of a patch surrounding the landmark (or any point of interest), using a data-specific learned dictionary. Since strong spatial constraints exist in medical images, we also combine spatial information of surrounding point descriptors in an online built graphical model. We demonstrate accurate results in matching one-to-one anatomical landmarks in brain MR images. This is in contrast to the methods for landmark localization developed in Chapters 4 and 6, where several annotated target images are used to estimate the landmark position in a source image (several-to-one).

7.1 Introduction

The detection of landmarks is a crucial step in many medical imaging applications, including registration, shape modeling and morphometry. Approaches to landmark detection can be roughly classified into three main categories: geometric-, classification- and regression-based techniques.

Recently dictionary learning and sparse coding have emerged as powerful tools in image analysis and machine learning. They have been successfully used in denoising [2], inpainting [139], classification [233], segmentation [171] and reconstruction [173]. Sparse coding aims to represent a given signal as a sparse mixture model of some basis, while dictionary learning tries to find which set of basis best represents the signals to be reconstructed. In this work we propose to leverage ideas from dictionary learning and sparse coding to learn data-specific feature point descriptors that are based on sparse reconstructions of the dictionary

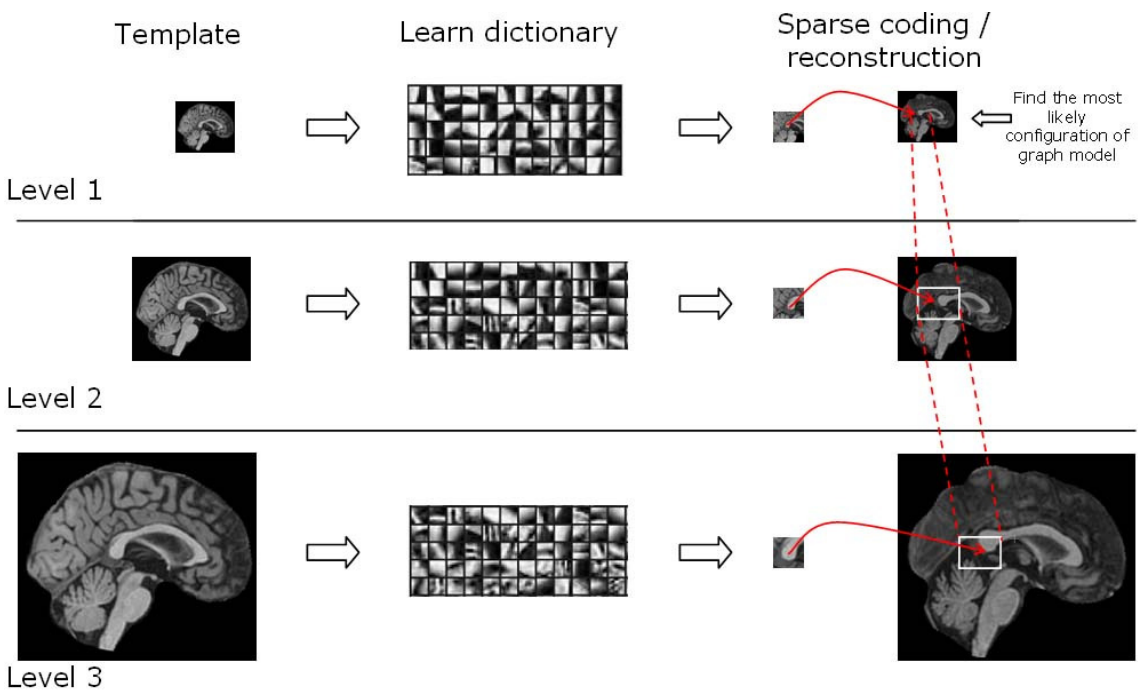


Figure 7.1: Overview of the proposed method.

elements.

7.2 Methods

Given an image with known anatomical landmark positions, we can use the sparse coding coefficients as feature point descriptor of its known landmark positions. Then, using these descriptors their matching counterparts can be found in another unseen source image. To do this we adopt a multi-resolution pyramid approach (see Figure 7.1).

A dictionary for each level in the resolution pyramid is learned either on-line using the pair of images or off-line using a set of training images. Feature vectors are obtained from the sparse coding coefficients of: A patch around the known landmark in the target image, a fixed number of randomly sampled support points from the target image and the complete source image. Initially this is carried out for images at the coarsest level of the resolution pyramid. A similarity map between the reconstruction coefficients from the target image (landmark plus support points) and all the coefficients belonging to each voxel in the source image is calculated; we expect that similarly looking patches should have similarly sparse coding coefficients. Using a graphical model, built from the configuration of the landmark

and support points in the target image (see Figure 7.2), we then find the most likely location of this model (all points) in the source image. The most likely location of the landmark in the source image defines a search window for the next level in the resolution pyramid and the process is repeated until the finest level is reached.

7.2.1 Sparse coding

Sparse coding seeks to represent a signal $\mathbf{x} \in \mathbb{R}^n$ as a sparse linear combination of basis signals that belong to an over-complete dictionary $\mathbf{D} \in \mathbb{R}^{n \times \kappa}$ that contains κ basis signals or atoms as column vectors, $\{\mathbf{d}_j\}_{j=1}^{\kappa}$. Typically this representation is an approximation $\mathbf{x} \approx \mathbf{D}\alpha$, subject to $\|\mathbf{x} - \mathbf{D}\alpha\|_2 \leq \varepsilon$, where ε is the error tolerance and the vector $\alpha \in \mathbb{R}^{\kappa}$ are the sparse coefficients that act on $\mathbf{D}$ to reconstruct signal $\mathbf{x}$. In the case where $n < \kappa$ the number of possible solutions is infinite. A common constraint to the problem is to look for the sparsest solution by minimizing

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad s.t. \quad \|\mathbf{x} - \mathbf{D}\alpha\|_2 \leq \varepsilon \tag{7.1}$$

where $\|\cdot\|_0$ is the l_0 norm that counts the number of non-zero entries of a vector. Given a 3D image I of size $N = N_x \cdot N_y \cdot N_z$, we can break up image I into N small patches centered around each voxel $\mathbf{x}_i \in \mathbb{R}^n$. The sparse representation of each patch in $\mathbf{x}$ can then be found solving Equation (7.1) independently

$$\min_{\mathbf{x}_i} \|\mathbf{x}_i\|_0 \quad s.t. \quad \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2 \leq \varepsilon \tag{7.2}$$

where the coefficients α_i provide a linear combination of atoms for each voxel and its surrounding patch. Since α_i represents a basis mixture model, one can expect that patches that have a similar appearance will have similar mixture models and hence one can regard α_i as a feature point descriptor.

7.2.2 Dictionary learning

State-of-the-art results indicate that in general is better to learn a dictionary from the data itself [171], rather than using an off-the-shelf dictionary, e.g. wavelets or discrete cosine transform. Given a set of signals $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, the assumption is that there exist a dictionary $\mathbf{D}$, from which a mixture of its elements can represent all signals in $\mathbf{X}$ via a sparse linear combination of them. Dictionary learning aims at finding the best possible set of basis to sparsely represent the signals in $\mathbf{X}$, by minimizing

$$\min_{\mathbf{D},\boldsymbol{\alpha}} \sum_i \|\alpha_i\|_0 \quad s.t. \quad \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_F^2 \leq \varepsilon\,, \tag{7.3}$$

or similarly

$$\min_{\mathbf{D},\boldsymbol{\alpha}} \left\{ \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_F^2 \right\} \quad s.t. \quad \forall_i, \|\alpha_i\|_0 \leq T_0 \tag{7.4}$$

for a fixed error tolerance, ε, or sparsity, T_0.

The K-SVD [2] algorithm aims to solve Equation (7.4) iteratively. First, using an of-the-shelf dictionary (discrete cosine transform, wavelets, curvelets, etc.) as an initial estimate of the dictionary, $\mathbf{D}$ is fixed and the best coefficient matrix $\boldsymbol{\alpha}$ is found using the orthogonal matching pursuit [32] algorithm. Then a dictionary update step is performed, by fixing all columns of $\mathbf{D}$ except one, $\mathbf{d}_k$, and finding a new best representation for $\mathbf{d}_k$. The process is repeated for every column in $\mathbf{D}$. If the algorithm has not converged or an early exit criteria met, then the whole process is repeated (find α, then $\mathbf{D}$).

7.2.3 Graphical model

In order to incorporate spatial constraints into the landmark matching problem, we define a pictorial structure [64] for the image. The graphical model is built as a tree of depth two with the root node being a patch around the anatomical landmark and additional support point patches drawn at random from the image acting as the leaves on the tree. Following the notation of Felzenszwalb et al. [64], the graphical model is defined as a undirected tree-shaped graph $G = (V, E)$, where the vertices $V = \{v_r, v_1, v_2, ...v_n\}$ correspond to the parts

(nodes in the graphical model) defined as image patches around the landmark v_r and support points v_i, and the edges E connect the root patch v_r to every other patch v_i. An instantiation of the model is given by the configuration $L = (l_r, l_1, l_2, ..., l_n)$, where l_r and l_i specify the location of the landmark and support points respectively (Figure 7.2). Since our aim is to build a graphical model on-line and from just one example, we cannot learn the parameters of the model from the data. Instead we simply model the relationship between patches $\Psi_{i,j}$ as a Gaussian distribution (represented by the spring in Figure 7.2 right).

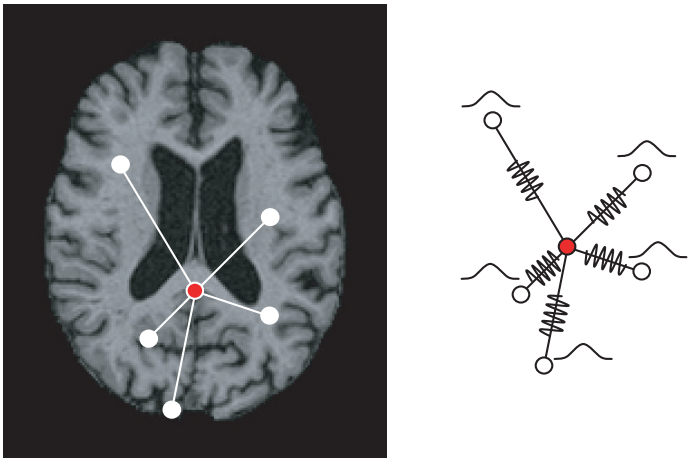


Figure 7.2: Graphical model: The red and white points indicate the landmark and support points respectively.

7.2.4 Model matching

The problem of finding the best match across images, not for an individual point, but rather an ensemble of points in the form of a graphical model can be defined by the following optimization problem:

$$L^* = \arg\min_L \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_r, v_i) \in E} \Psi_{r,i}(l_r, l_i) \right) \tag{7.5}$$

Here $\Psi_{r,i}(l_r, l_i)$ is a function that models the relationship between patch v_r and patch v_i, and $m_i(l_i)$ is a function that associates a cost of placing patch v_i at location l_i. In our case we model $m_i(l_i)$ as the $L2$-norm between the target patch descriptor α_T and the descriptor α_S at every point in the source image, $m_i(l_i) = \|\alpha_T - \alpha_S(l_i)\|_2$. As stated before, $\Psi_{r,i}$ follows a Gaussian distribution.

In general, solving Equation (7.5) for a graph of arbitrary shape is an NP-hard problem [22]. However, if the graph $G = (V, E)$ is restricted to a tree-shaped topology it can be solved using an algorithm based on the Viterbi recurrence [221]. Defining a root vertex $v_r \in V$ in the graph (in our case is the landmark sought), every other vertex in the graph $v_i \in V$ has a depth k_i. In our case $k_i = 1$ which implies that only the root and leaves are nodes in our tree. For leave vertices v_i, the only edge incident to them is (v_r, v_i), thus the only contribution of l_i to the energy in Equation (7.5) is $m_i(l_i) + \Psi_{r,i}(l_r, l_i)$ and quality of the best location is given by

$$B_i(l_r) = \min_{l_i} \{m_i(l_i) + \Psi_{r,i}(l_r, l_i)\} \ , \tag{7.6}$$

Here the best location l_i^* for patch v_i, as a function of its parent's location l_r, can be found by replacing "min" by "arg min" in Equation (7.6):

$$l_i^* = \arg\min_{l_i} (m_i(l_i) + \Psi_{r,i}(l_r, l_i)) \ . \tag{7.7}$$

For the root node v_r the value $B_c(l_r)$ is known for each of its children, hence the best location can be found as

$$l_r^* = \arg\min_{l_r} \left(m_r(l_r) + \sum_{v_i \in C_j} B_i(l_j) \right) \ . \tag{7.8}$$

Equation (7.6) describes the quality of the optimal location l_i^* for patch v_i as a function of its parent patch v_r. Using the known parent's location and Equation (7.6), we can back-trace from the root to the leaves to obtain the optimal location L^* for the whole model.

7.3 Comparison to other landmark detection approaches

In this section we briefly describe two other different methods that can be used for landmark localization in a one-to-one matching approach: (a) 3D local SS descriptors and (b) REG. These two methods were used as a comparison to the method proposed.

7.3.1 3D local SS descriptors

3D local SS descriptors define a landmark based on the structural pattern of its neighborhood. In 3D this is done by computing the similarity between a small spherical patch around the landmark and every other point (another small sphere patch) in a larger surrounding spherical image region, which results in an internal similarity map. This similarity map is then binned into a log-spherical representation. Each bin is filled with the highest similarity that falls within its supported range. A detailed description of SS descriptors can be found in Chapter 4.

7.3.2 Non-rigid image registration

Intensity-based registration can be used to propagate a set of annotated landmarks from a source image to a target image. Figure 4.3 in Chapter 4 illustrates the landmark propagation procedure. In this work we used the method proposed in [182], where a FFD model based on B-splines is used to deform the underlying mesh of control points until an intensity similarity measure is optimized (in this case normalized mutual information). We used a hierarchical approach with a control point mesh that is refined from a 20 mm spacing to 10 mm and then 5 mm.

7.4 Data and Results

The proposed method has been tested using brain MR images from the ADNI database [148] (see Appendix A). In particular, we used a subset of 1.5T T1-weighted baseline images of 100 randomly chosen subjects to learn a dictionary for each of the three resolution levels and another 100 randomly chosen subjects for testing. Both (training and testing) datasets include 24 AD, 48 MCI and 28 healthy subjects, to faithfully represent the full spectrum of subjects in the ADNI dataset. All brain MR images were skull stripped, affinely aligned to MNI space and normalized using linear intensity rescaling. To validate the proposed method 20 landmarks (Table 7.2) were manually annotated by an expert observer using three orthogonal views. The landmarks defined in the MNI template can be seen on Figure

4.5 (see Appendix C for a detailed description).

We have evaluated four different approaches: In the first approach, we search for the best matches for the manually selected landmarks' descriptors at the coarsest level of the resolution pyramid. In this scenario (called *one path* (OP)) we discarded the feature vectors from previous levels. Also, an experiment using graph-based spatial constraints was carried out (called *one path plus graphical model* (OPGM)). In another approach we follow the same structure as before, except that we concatenated the descriptors from previous levels with the current ones and subsequently find the best matches based on these descriptors (called *keeping previous vectors* (KPV)). As before, we also tested using graph-based spatial constraints (called *keeping previous vectors plus graphical model* (KPVGM)). Table 7.1 illustrates the main characteristics of the four different approaches.

	OP	OPGM	KVP	KVPGM
Graphical model	✗	✓	✗	✓
Combined level descriptors	✗	✗	✓	✓

Table 7.1: Different approaches and their associated characteristics.

An experiment was carried out were one of the test images was used as source and its landmarks were matched to the remaining 99 (source) images, the process was repeated 100 times. The three dictionaries used in all experiments (one for each level) were learned from patches extracted from down-sampled versions of the training images, and each dictionary consisted of 130 atoms. The average accuracy results for the four mentioned approaches are presented in Table 7.2.

It is clear that the approach that discards descriptors from previous levels and incorporates spatial knowledge (OPGM) outperforms the other approaches for every landmark both in terms of accuracy and standard deviation. Results of a comparison to other landmark detection approaches are summarized in Table 7.4. It should be noted that the SS approach used in this chapter follows the same validation strategy as OPGM (one-to-one matching), this is in contrast to how it was implemented in Chapter 4 (several-to-one matching). The proposed method achieves on average a better landmark localization error. However accu-

Anatomical landmark	OP	OPGM	KPV	KPVGM
Splenium of corpus callosum (outer aspect)	8.57 (6.90)	**3.57** (2.10)	6.40 (3.73)	5.12 (2.91)
Splenium of corpus callosum (inferior tip)	7.21 (7.44)	**2.80** (1.18)	4.67 (2.65)	3.36 (1.72)
Splenium of corpus callosum (inner aspect)	7.28 (7.58)	**3.10** (1.17)	4.83 (2.70)	3.51 (1.70)
Genu of corpus callosum (outer aspect)	6.41 (6.84)	**2.39** (1.64)	5.98 (5.83)	2.58 (1.79)
Genu of corpus callosum (inner aspect)	4.80 (6.83)	**1.81** (1.03)	4.01 (4.34)	1.97 (1.20)
Superior aspect of pons	1.63 (1.56)	**1.60** (1.70)	2.27 (2.09)	2.27 (2.15)
Inferior aspect of pons	2.29 (0.84)	**2.32** (1.05)	3.03 (1.77)	3.07 (1.78)
Superior aspect cerebellum	11.89 (7.80)	**4.90** (1.41)	7.43 (3.16)	6.13 (2.62)
Fourth ventricle	2.82 (2.58)	**2.04** (1.45)	2.89 (1.81)	2.85 (1.71)
Putamen posterior (left)	8.24 (9.43)	**4.02** (1.38)	6.21 (2.81)	4.96 (1.23)
Putamen anterior (left)	4.89 (5.27)	**3.02** (0.97)	5.86 (1.63)	4.93 (1.26)
Putamen posterior (right)	8.59 (8.79)	**3.62** (1.03)	5.79 (2.892)	4.19 (1.31)
Putamen anterior (right)	4.26 (3.56)	**3.04** (1.21)	4.66 (1.58)	4.33 (1.38)
Anterior commissure	4.18 (7.92)	**1.48** (0.90)	3.29 (3.11)	2.69 (2.12)
Posterior commissure	3.49 (5.31)	**1.25** (0.53)	2.74 (2.64)	1.97 (1.86)
Inferior aspect cerebellum	5.52 (1.93)	**5.42** (1.85)	6.59 (2.40)	6.44 (2.35)
Anterior tip of lateral ventricle (left)	5.28 (6.42)	**2.85** (2.14)	5.09 (4.71)	3.73 (2.69)
Anterior tip of lateral ventricle (right)	4.25 (5.22)	**2.45** (1.34)	3.32 (5.89)	2.62 (1.44)
Inferior tip of lateral ventricle (left)	5.89 (5.88)	**3.76** (1.55)	6.03 (4.18)	4.84 (2.64)
Inferior tip of lateral ventricle (right)	4.01 (3.05)	**2.82** (0.88)	3.37 (1.47)	3.01 (0.90)
Mean	5.58 (5.71)	**2.91** (1.33)	4.72 (3.07)	3.73 (1.84)

Table 7.2: Accuracy of some variations of the proposed method. Mean error in mm with standard deviations in brackets.

racy results on individual landmarks are more evenly distributed, with OPGM (proposed method) being the most accurate in 35% of the cases, while SS and REG being the most accurate in 25% and 40% of the cases, respectively.

7.5 Conclusions

In this work we have proposed a framework that combines dictionary learning, sparse coding and graphical models in order to develop a data-specific feature point descriptor matching algorithm that is spatially consistent between target and source images. To our knowledge, this is the first time that dictionary learning along with sparse coding have been used to address the problem of learning feature point descriptors. By using the learned dictionary basis to reconstruct a patch, feature descriptors are specially tailored to represent the imaging data at hand (brain MR images in our case).

In addition, using on-the-fly constructed graphical models, we have shown that the proposed method produces accurate results in a one-to-one landmark matching setting. This is in contrast to learning appearance from several training examples, as it is generally the case

Anatomical landmark	OPGM	SS	REG
Splenium of corpus callosum (outer aspect)	2.38 (2.65)*	**2.12** (1.97)	4.18 (1.52)
Splenium of corpus callosum (inferior tip)	2.40 (1.33)+*	3.50 (2.43)	**2.08** (1.15)
Splenium of corpus callosum (inner aspect)	3.21 (2.07)+*	5.79 (4.30)	**2.52** (1.40)
Genu of corpus callosum (outer aspect)	**1.59** (0.97)*	1.79 (1.62)	2.04 (1.37)
Genu of corpus callosum (inner aspect)	1.50 (0.89)+	4.34 (3.34)	**1.46** (0.93)
Superior aspect of pons	**1.60** (0.75)*	1.67 (1.54)	2.80 (1.24)
Inferior aspect of pons	5.10 (4.59)+*	2.57 (2.27)	**1.71** (1.07)
Superior aspect cerebellum	3.67 (3.24)+*	8.46 (6.20)	**3.03** (1.40)
Fourth ventricle	1.52 (0.83)*	**1.34** (1.24)	5.97 (2.98)
Putamen posterior (left)	4.46 (2.19)+	**3.74** (2.47)	4.62 (2.15)
Putamen anterior (left)	3.79 (3.04)+*	4.50 (3.33)	**2.58** (1.55)
Putamen posterior (right)	4.20 (2.31)+*	4.84 (3.30)	**3.31** (1.51)
Putamen anterior (right)	**2.30** (1.67)+*	4.17 (2.76)	2.87 (1.58)
Anterior commissure	**1.07** (1.18)+	2.37 (1.78)	1.10 (0.41)
Posterior commissure	4.71 (2.79)+*	**1.39** (1.16)	1.91(0.51)
Inferior aspect cerebellum	4.81 (3.86)+*	6.33 (5.04)	**3.98** (1.96)
Anterior tip of lateral ventricle (left)	**1.44** (0.95)+*	2.39 (2.16)	3.83 (1.58)
Anterior tip of lateral ventricle (right)	2.39 (1.35)*	**2.22** (1.51)	3.83 (1.76)
Inferior tip of lateral ventricle (left)	**3.74** (2.68)+	5.68 (5.26)	4.13 (1.79)
Inferior tip of lateral ventricle (right)	**2.44** (1.37)+*	3.83 (2.83)	3.75 (1.84)
Mean	**2.91** (2.04)	3.65 (2.83)	3.08 (1.48)

Table 7.3: Accuracy of some variations of the proposed method. Mean error in mm with standard deviations in brackets. Statistical significance (to %5, results not corrected for multiple comparisons) is indicated by + and *, for comparisons between OPGM and SS or REG, respectively.

when using a sliding window classifier, manifold learning or a voting scheme using other descriptors (e.g. self similarities). One of the main disadvantages of the proposed approach is high computational cost of the sparse coding of the source image. Due to the highly parallelizable nature of the sparse coding we believe this could greatly be ameliorated using a graphics processing unit.

A direct comparison between the methods presented in this chapter and Chapter 6 (manifold learning landmark localizations) is not strictly speaking possible due to the fact that slightly different datasets were used. Additionally, the methods presented in this chapter find landmarks doing a one-to-one matching, while manifold landmark localization (Chapter 6) uses information from several-to-one images. Nevertheless, it can be observed that on average manifold learning landmark localization clearly outperforms the OPGM, 3D local SS and REG methods. As stated before, the main disadvantage of manifold learning landmark localization is the memory requirements of the manifold and in comparison to the methods presented here, the prior higher expert input requirements.

Chapters 4, 6 and 7 have presented methods for brain landmark localization, while Chap-

ter 5 presented an automatic feature point matching framework for knee MR image affine registration. In the following chapter a framework, that leverages the machine learning expertise developed in previous chapters, to extract MR brain imaging AD biomarkers is presented.

Chapter 8

Manifold population modelling as an imaging biomarker

This chapter is based on:

- Ricardo Guerrero, Robin Wolz, Amil Rao, Daniel Rueckert. ”Manifold population modelling as a neuro-imaging biomarker: Application to ADNI and ADNI-GO”. *NeuroImage 2013*, submited.

Abstract

We propose a framework for feature extraction from learned low-dimensional manifold subspaces that represent anatomical inter-subject variability. The manifold subspace is built from data-driven regions of interest rather than regions that are specified using a-priori knowledge. In this chapter we present an application of this framework in the context of AD. Specifically, regions are learned via sparse regression using the mini-mental state examination (MMSE) score as an independent variable which correlates better with the actual disease stage than a discrete class label. Sparse regression is used to perform variable selection and we use a re-sampling scheme to reduce sampling bias. We then use the learned manifold coordinates to perform classification of the images. Results of the proposed approach are shown using the ADNI and ADNIGO datasets. Two types of classifier, including a new MRI-DSS classifier, are tested in conjunction with two learning strategies: In the first case, subjects with AD and progressive mild cognitive impairment (pMCI) are grouped together, while subjects that are CN or have stable mild cognitive impairment (sMCI) are also grouped together. In the second approach, the classifiers are learned using the original class labels (without grouping). We show that the results obtained using the ADNI database are comparable to other state-of-the-art methods. A classification rate of 71%, of arguably the most clinically relevant subjects, sMCI and pMCI, is obtained. Additionally, we present classification results for CN and early MCI (eMCI) subjects using the ADNIGO database and show a classification accuracy of 65%. To our knowledge this is the first time that results for CN/eMCI classification have been reported.

8.1 Introduction

AD is the most common form of dementia, usually associated with the elderly population (over 65 years of age). AD has a world-wide prevalence of around 26.6 million cases reported in 2006 and predictions suggest that this figure will increase four-fold to over 100 million by the year 2050 [28]. If intervention or treatment could achieve even a modest one year delay of both disease onset and progression, there would be nearly nine million fewer

cases of the disease by 2050 [28]. One of the challenges in the management of AD is that postulated interventions are more likely to be effective in the very early stages of the disease. These figures underline the huge impact advances in early diagnosis might have on the overall well being of the population, the burden to caregivers and family members, as well as the associated financial costs to the world's health systems. Several studies over the recent years have concluded and confirmed that AD can be diagnosed by clinical assessment alone accurately in 90% of the cases when validated against neuropathological standards [172]. However, by the time a patient is diagnosed he/she may already suffer from substantial loss of quality-of-life and chances for improvement, or even deceleration disease progression, may have deteriorated. Hence, the diagnosis of very early onset dementia is crucial.

Several medications are currently approved by the U.S. Food and Drug Administration (FDA) to treat people who have been diagnosed with AD. Treating the symptoms of AD can provide patients with comfort, dignity, and independence for a longer period of time and can encourage and assist their caregivers as well. Disease modifying treatments are more likely to have a significant impact in earlier stages of the disease. Population stratification is important to allow the recruitment of appropriate subjects for clinical trials, and explore the effects of novel treatments in subjects where results are expected to be most effective, hence, reducing overall costs of the trial by removing unsuitable subjects in an earlier stage. Of special interest are subjects with MCI, which is a prodromal form of AD. Existing studies have suggested that about 10-12% of subjects with MCI progress to probable AD per year [160]. However, individual patients can remain in a stable MCI (sMCI) condition for years. From a clinical perspective it is therefore particularly interesting to identify those subjects that are at immediate or medium-term risk of progressing from MCI to AD (pMCI).

ADNI is a study with the primary goal of testing whether longitudinal MR imaging, PET, other biological markers and clinical and neuropsychological assessment can be combined to measure the progression of MCI and AD. Recent studies focus on identifying subjects at risk at a much earlier stage. In the ADNIGO and ADNI2 studies, a group of eMCI patients is included [3] which represents individuals with milder degrees of cognitive and functional impairment than the MCI subjects included in ADNI. With a slower rate of progression, they

form an especially interesting subject group as biomarker manifestation could potentially be different at such an early stage of the disease.

Imaging biomarkers play an increasingly important role in the early diagnosis of neurodegenerative diseases like AD. MR imaging often forms part of clinical assessment for patients with MCI or suspected AD. Biomarkers based on MR imaging are considered to be more sensitive to change after symptoms from amyloid-based biomarkers start to appear (e.g. accumulation of amyloid-β) [79]. Figure 8.1 illustrates the rate of change of several types of AD biomarkers at different disease stages. Imaging biomarker measurements can be key in the development of disease-modifying therapies and interventions. They can be used to explore the modifying effects these therapies may have on the disease progression across time, and also as a screening tool to select a more homogeneous prodromal patient population that is expected to have higher risk for rapid imminent clinical progression, thus increasing the efficiency of clinical trials [97]. Recently, there have been many studies with a main focus on automatically identifying such imaging biomarkers. Many of the well-established and well-known biomarkers used in dementia that are derived from MR imaging are based on morphological measurements of specific a-priori defined brain structures (e.g. hippocampus, amygdale, cortex, enthorhinal cortex) and include features such as volumes or shapes ([43, 229, 130, 37, 232, 124, 226, 169, 34, 48]). However, patterns of neurodegeneration may not necessarily follow strict standard definitions of anatomical structures or functional regions. Hence, limiting the analysis to predefined regions could potentially reduce the power of the biomarker to detect subtle differences or changes over time.

More recently, the problem of learning clinically useful biomarkers has been cast as a machine learning problem. Models that derive from developments in the machine learning community have been put forward as alternatives which seek for discriminative features that could act as AD biomarkers independent from a predefined parcellation of structures ([63, 62, 84, 216, 235, 230, 146, 50, 61]). This independence could potentially lead to a better modelling of a disease trajectory for the whole brain and across time. Furthermore, this would account for the fact that the disease trajectory manifests itself at different regions at different times.

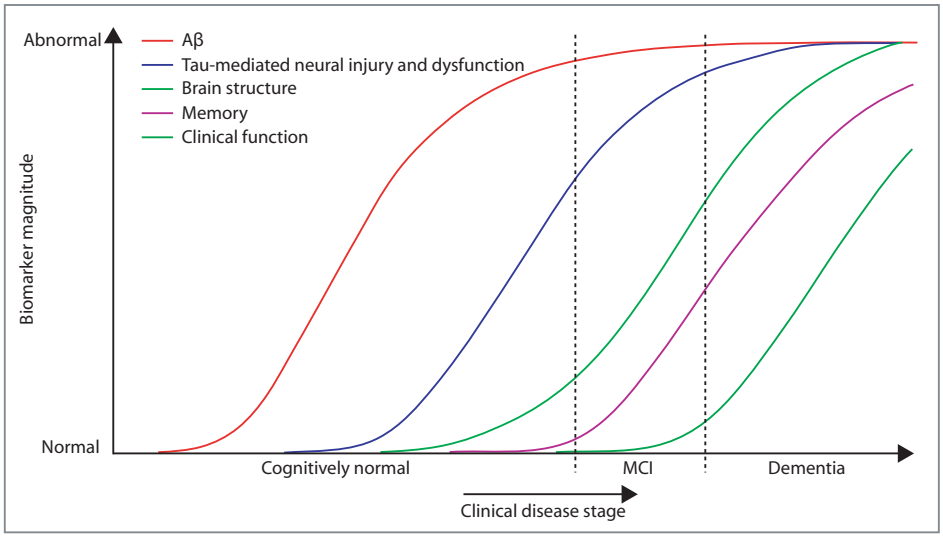


Figure 8.1: Different biomarkers of the Alzheimers pathological cascade (from [109]).

Some of the potential pitfalls when working with high-dimensional data, such as medical images, can be associated to the curse of dimensionality. This describes a general paradox that occurs in high-dimensional space, where if a neighborhood is considered "local", then it will be most likely "empty", while a "non-empty" neighborhood will probably be "non-local". This implies that in high-dimensional space the variance-bias trade-off cannot be accomplished very well, unless there is a very large amount of samples available. That is, to keep variance low the neighborhood has to be made large enough to include enough samples, but then a very large bias is introduced due to the large neighborhood, and vice versa [186].

Learning a low-dimensional subspace representation of complex, high-dimensional objects (e.g. images) is a central problem of machine learning and pattern recognition. Several methods have been proposed to find the underlying low-dimensional space of intrinsically low-dimensional data that is embedded in a high-dimensional space. A low-dimensional representation of the data allows the use of modelling techniques that suffer from the small sample size problem in high-dimensional spaces. There is a long history in the use of linear dimensionality reduction techniques, like PCA and MDS [44], across different domains. Recently, nonlinear techniques like principal curves [98], ISOMAP [203], LLE [180] or Laplacian eigenmaps [17] have been proposed to better capture the variation of highly nonlinear data. For a comprehensive and recent review of dimensionality reduction techniques

see [213].

In recent years these techniques have been applied increasingly to medical images: Working with brain MR images and using concepts from dimensionality reduction, Aljabar et al. [6] applied spectral analysis to pairwise label overlaps obtained from a structural segmentation to discriminate AD patients from CN subjects. Klein et al. [122] used vectors defined by the similarities between a given test subject and a set of training images as features from which to learn a classifier. Some dimensionality reduction techniques aim to model global variabilities over the whole dataset, which could potentially limit their generalization power of the learned subspace when dealing with complex datasets. In recent work, it has been suggested that this is indeed the case when dealing with brain MR images and that nonlinear methods better capture the natural variabilities of such images [85, 96]. Wolz et al. [230] propose to classify a subjects state in a manifold that is learned from image similarities measured over an a-priori defined ROI and (clinical) meta-information related to the subject. However, as stated before, patterns of neurodegeneration may not necessarily be best observed in the predefined ROI, since useful information could potentially be ignored. On the other hand the ROI will most likely contain regions that are not associated with the patterns of neurodegeneration and this could confound the learned subspace. Furthermore, subject classification is performed by applying a SVM approach to the manifold coordinates. SVM finds a separation hyperplane defined by only a subset of subjects (support vectors) that lie close to the hyperplane in the learned subspace.

There are two fundamental problems when dealing with high dimensional data such as 3D brain MR images: First, there is a large amount of variables (voxels) available in images and not all contribute equally (or at all) to the modelling of the disease status and trajectory. Relevant variable selection from this large pool of potential predictor variables is a way to tackle this problem. We assume that the underlying disease trajectory manifests itself in a small subset of variables in an image and so it can be modelled using a sparse set of voxels. In this context, the L_1 norm has been proposed as an effective solution to the variable selection problem [206, 239]. Secondly, this variable selection process often is an ill-posed problem, where the sample size is much smaller than the number of variables and variables

are highly correlated. That is, the L_1 norm can only select up to N uncorrelated variables, where N is the number of samples. Although the dimensionality reduction techniques mentioned before can deal this issue, all variables contribute to the manifold estimation process.

We propose to use sparse regression to learn a ROI in which a distance metric allows to define a manifold that better describes the different stages of AD. The resulting ROI defines the brain regions where the disease trajectory can be best observed and quantified. The compact manifold representation allows us to model different populations directly from the learned manifold coordinates. Population distribution models of the observed data can be used to infer the disease state of a new patient by embedding it in the manifold and obtaining a probabilistic score for the class correspondence as opposed to a discrete label as in classification approaches. This probabilistic estimation allows us to move away from a discrete decision based on hyperplanes to a continuous characterization or modelling of disease progression. The proposed MRI-DSS formulation aims at modelling the disease progression while fully taking advantage of manifold coordinates. The MRI-DSS metric yields a continuous variable on the disease trajectory.

8.1.1 Biomarkers for AD

The methods used to assess the possibility of a given individual to be affected by dementia can be broadly divided into two categories: (I) psychological tests and (II) quantitative measurements. Psychological tests such as MMSE [72] or clinical dementia rating (CDR) [147] are used in most memory clinics to assess the cognitive state of a patient. They typically involve several questions testing the short-term memory of the patient. While an existing impairment can be identified in most cases, a much earlier identification of people at risk is necessary to enable a potentially successful treatment.

AD is caused by neurofibrillary tangles and neuritic plaques [23]. In the later stages of the disease these degenerative changes in the human neurotransmitter system lead to atrophy in selected brain regions [225]. The study of the generation of tangles and plaques has emerged as a promising approach for detecting the disease while at its earlier stage. Another commonly associated AD risk factor is the concentration of the tau-protein and the

amyloid-beta-peptide $A\beta_{42}$ in the cerebrospinal fluid (CSF) [210]. Although obtaining a patient's CSF sample is invasive, measurements of this biomarker can give a good insight of the patient's state.

PET, in combination with the use of a Fludeoxyglucose ^{18}F tracer, can detect the decrease in brain metabolism of glucose and oxygen caused by AD [33]. An alternative tracer that has shown promising results as an AD biomarker is the Pittsburgh Compound B (PiB) tracer [107], which selectively binds to Aβ deposits and thereby can be used to visualize beta-amyloid deposits in the brain. Structural images acquired with MR on the other hand allow the analysis of the current morphology of brain degeneration. The volume of brain structures and their change over time are widely accepted as biomarkers for AD, e.g., [112]. A more detailed introduction to biomarkers for AD can be found in, e.g., [210].

8.2 Material and Methods

8.2.1 Data

In this work, we used the subset of 523 subjects for which T1-weighted 1.5T MR images were available at baseline, 12 and 24 month follow-up, as of June 2012. Experiments in this work were performed using baseline images. 12 of those subjects where discarded due to label ambiguities (subjects whose labels changed from MCI to CN or from AD to MCI). The remaining 511 subjects consisted of 106 patients diagnosed as probable AD, 230 as MCI (114 sMCI and 116 pMCI) and 175 CN (see Table 8.1 for a description of the demographics). Patients considered as pMCI where those that had converted from MCI to AD as of June 2012. The remaining 315 out of the 838 baseline images (CN = 56, sMCI = 119, pMCI = 49 and AD = 91) that did not contain all time points where used as training data in the variable selection scheme (Section 8.2.2).

Additionally, experiments where carried out using the ADNIGO dataset [3]. From this dataset, all the available baseline MR images labeled as CN or eMCI, as of June 2012, were selected and preprocessed in the same way as with ADNI (see Table 8.2 for a description of the demographics).

	N	Age	MMSE	Men	CDR	Weight
CN	175	76.34±5.11	29.17±0.97	52% (91)	0±0.1	74.43±15.57
sMCI	114	75.12±6.67	27.29±2.25	66% (75)	0.49±0.05	77.02±12.83
pMCI	116	74.73±6.93	26.64±1.7	63% (73)	0.5±0.05	74.56±14.41
AD	106	75.4±7.39	23.25±1.97	53% (56)	0.77±0.25	72.58±13.81

Table 8.1: Subject group's mean age, sample size, MMSE scores, gender distribution, CDR scores and weight data (with standard deviation in brackets) from the ADNI database.

	N	Age	MMSE	Men	Weight
CN	134	73.77±10.85	28.99±1.23	51% (68)	75.68±15.08
eMCI	229	67.42±18.61	28.29±1.53	54% (124)	81.47±15.89

Table 8.2: Subject group's mean age, sample size, MMSE scores, gender distribution and weight data (with standard deviation in brackets) from the ADNIGO database.

Image preprocessing

In this study, all the images used were skull stripped using multi-atlas segmentation [132] and intensity normalized at a global scale using a piecewise linear function [153]. Intensity normalization was carried out following an iterative scheme, where all images are normalized to a common template/subject, then the template was changed and all the images were re-normalized to the new template. This was repeated N times, where N is the number of subjects to aid in removing normalization bias [43]. Also, all images were transformed to a common space, the MNI152 template, and hence re-sliced and re-sampled to isotropic voxel size of 1mm. A coarse free-form-deformation [182], using a control point spacing of 10mm, was carried out to remove gross anatomical variability while aligning anatomical structures in order to focus on more local variation. In order to account for disease manifestation and progression in left-handed and right-handed populations, and hence find more generalizable regions, the selected variables from Section 8.2.2 are mirrored along left-right hemisphere prior to the subsequent steps.

8.2.2 Relevant variable selection

Regression techniques allow the modelling and analysis of several variables, where the focus is on modelling the relationship between a dependent variable and one or more independent

variables. Over the years, several regression methods have been proposed [207, 239, 206], with arguably the simplest method being OLSR. generalize well beyond the training data. In ridge regression, this is achieved by incorporating an L_2 penalty into the OLSR objective function, which leads to a unique solution in which correlated predictors are given similar regression weights. LASSO regression, on the other hand, uses an L_1 penalty which regularizes the problem by encouraging a sparse solution in which most of the estimated regression weights are zero. This is a highly desirable trait when dealing with high dimensional data because it allows for variable selection. Two of the main problems with LASSO are that it does not perform well in the presence of highly correlated variables (e.g. neighboring voxels in an image would probably be very well correlated) and that it can only select a number of variables that is up to the number of samples (a significant problem for high dimensional data). Elastic net regression [239] seeks to address the drawbacks of the LASSO [206], e.g. it allows to select a number of variables that is greater than the number of samples. This is done by adding an additional L_2 penalty term on the model's coefficients to the L_1 penalty term of LASSO.

Elastic net

Viewed as an image regression problem, elastic net regression identifies regions of interest (predictor variables) within the images $\mathbf{X}$ that are useful to regress against variable $\mathbf{l}$ associated with each image. This could be the clinical label or the MMSE score associated with a patient. The elastic net objective function in Equation (3.38) can be solved efficiently using the LARS-EN algorithm [239], which allows for the number of steps or number of variables selected to be easily incorporated as an early termination criteria. Is worth noting that Equation (3.38), in the special cases where λ_R and λ_L are set to zero, becomes the ordinary least square regression. If λ_L is set to zero then it describes a ridge regression and if λ_R is set to zero we obtain a LASSO regression. Another special case arises when $\lambda_R \to \infty$: It can be shown ([239]) that for each predictor variable $\mathbf{x}_i$, minimizing Equation (3.38) has a closed-form solution that can be written as:

$$\underset{\lambda_R \to \infty}{\hat{\boldsymbol{\beta}}_i} = \left(\left| \mathbf{l}^{\mathrm{T}} \mathbf{x}_i \right| - \frac{\lambda_L}{2} \right)_+ \operatorname{sign}\left(\mathbf{l}^{\mathrm{T}} \mathbf{x}_i \right), \qquad i = 1, 2, \ldots, p, \tag{8.1}$$

where $(\cdot)_+$ denotes the positive part.

This can be solved very efficiently, since $\mathbf{l}^{\mathrm{T}} \mathbf{x}_i$ is the univariate regression coefficient of the ith predictor, the estimates $\hat{\beta}_i$ are obtained by applying a soft threshold to the univariate regression coefficients. Equation (8.1) is also known as univariate soft thresholding.

As stated before, the L_2 regularization term (ridge) encourages the selection of correlated variables. In medical images it can be expected that voxels (variables) that belong to the same anatomical structure will have a high degree of correlation within each other. Choosing $\lambda_R \to \infty$ imposes a maximal grouping condition on Equation (3.38). In this setting elastic net regression can be used as a ROI learning algorithm.

As $\lambda_R \to \infty$, we are left with only one free parameter λ_L, from which we will drop the subindex and refer to it only as λ from now on. Equation (8.1) can be solved for a range of regularization parameters λ when we find the full regularization path, $\lambda_{\min} \leq \lambda \leq \lambda_{\max}$ up to the desired stopping criteria in the same way as one would using the LARS-EN algorithm. In our case we limit the step size on the path such that we ensure that at each step we add only one variable.

Training re-sampling

In order to increase robustness against sampling errors from the dataset, we adopt a re-sampling scheme. In this approach, the regularization path is found on B random subsets, solving Equation (8.1) over a range of values $\lambda \in [\lambda_{max}, \lambda_{min}]$, such that zero variables are included at λ_{max}, K variables are included at λ_{min} and with each step only one variable is added. At each step k on a subset b we obtain a set of regression coefficients $\hat{\boldsymbol{\beta}}_{b,k}(\lambda_{b,k})$, where $b = 1, 2, \ldots, B$ and $k = 1, 2, \ldots, K$. We define an indicator variable $\Psi_{bk}(\lambda_{b,k})$ which is set to one if the coefficient corresponding to variable x_j is non-zero and to zero otherwise. The relevance of each variable is measured by defining the probability of it being selected

by the regressor as

$$P_{v_j}(\lambda_{B,K}) = \frac{1}{B \cdot K} \sum_{b=1}^{B} \sum_{k=1}^{K} \Psi_{b,k}(\lambda_{b,k}), \quad j = 1, 2, ..., p. \tag{8.2}$$

Thresholding the probabilities P_v at τ to keep the most relevant voxels yields a mask that defines a ROI that correlates with the disease progression.

8.2.3 Manifold Learning

Manifold learning in general refers to a set of machine learning techniques that aim at finding a low dimensional representation of high dimensional data while trying to faithfully represent the intrinsic geometry of the data. For example, if we have an image dataset and each image is considered a single point in a very high dimensional space, then this high dimensional space is probably overcomplete in the sense that a sub-manifold (most likely to be non-linear) of far fewer dimensions may represent most of the variation in the dataset. In manifold learning, a distance matrix is typically used to represent the relations between pairs of data items, which can be assumed to be either the original images or some set of features derived from the images. This matrix can be interpreted as a graph in which each node corresponds to an image and the weight of each edge encodes the distance between images or derived features.

In our framework, given a set of N vectors of length D that define the most relevant voxels (variables) $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N\} \in \mathbb{R}^D$ from a set of images, the aim is to learn the underlying manifold in $\mathbb{R}^d$ ($d \ll D$) that best represents the relationship of images in the population using $\mathbf{V}$. Here $\mathbf{v}_i = \{v_1, v_2, ..., v_D\}$ are the weighted, most relevant voxels in image i.

Laplacian eigenmaps

Let us recall from Chapter 3 how Laplacian eigenmaps can be used to find a low-dimensional representation of the data $f : \mathbf{X} \to \mathbf{Y}$, $\mathbf{y}_i = f(\mathbf{x}_i)$ while preserving the local geometric properties of the manifold [17]. This can be formulated as an eigenproblem [10]. Hence, the

low-dimensional manifold $\mathbf{Y}$ that represents all the data points can be obtained via solving a generalized eigenproblem

$$\mathbf{L}\nu = \lambda \mathbf{M}\nu \ , \tag{8.3}$$

where $\mathbf{M}$ is the degree matrix, $\mathbf{L}$ is the Laplacian, ν and λ are the eigenvectors and eigenvalues. In turn the d eigenvectors ν corresponding to the smallest (non-zero) eigenvalues λ represent the new coordinate system.

8.2.4 Population distribution modelling

It is now widely accepted that both pathological processes and clinical decline occur gradually over time, with AD being the end stage of the accumulation and progression of these pathological changes. Additionally, the current consensus on AD is that these changes begin years before the earliest clinical symptoms occur [109]. Hence, AD biomarkers need to reflect this temporal progression, and imaging biomarkers are not an exception.

The aim in this work is to model the different discrete stages using continuous probabilistic Gaussian mixture models in order to make predictions of group assignment or disease evolution of unseen samples. This modeling is done using the coordinates of the low-dimensional representation found using Laplacian eigenmaps in order to avoid the curse of dimensionality associated with the high-dimensional space. In Parzen kernel density estimation (KDE), each observation sample is treated as a component in a mixture model. Treating each sample as a single Dirac delta function, which can be written as a Gaussian with zero covariance, with its probability concentrated at the point itself, we can define a multivariate and N-component Gaussian mixture model of the sample distribution as [127]:

$$P_s(\mathrm{y}) = \sum_{i=1}^{N} = \alpha_i \phi_{\Sigma_i}(\mathrm{y} - \mathrm{y}_i) \ , \tag{8.4}$$

where $\phi_{\Sigma_{si}}$ defines a Gaussian of mean y_i and covariance Σ_i that belongs to the sample mixture model distribution.

Defining the KDE, $\hat{P}_{KDE}(\mathrm{y})$, as the convolution between the sample distribution $P_s(\mathrm{y})$ and a Gaussian kernel with a covariance matrix (also known as the bandwidth) $\mathbf{H}$, we get:

$$\hat{P}_{KDE}(\mathrm{y}) = \phi_{\mathbf{H}}(\mathrm{y}) * P_s(\mathrm{y}) = \sum_{i=1}^{N} \alpha_i \phi_{\mathbf{H}+\Sigma_i}(\mathrm{y} - \mathrm{y}_i) \ , \tag{8.5}$$

where * denotes a convolution.

Considering the case where the sample distribution P_s is a Gaussian mixture model, with $\Sigma_i = 0$ (Dirac delta functions), Equation (8.5) can be rewritten as

$$\hat{P}_{KDE}(\mathrm{y}) = \sum_{i=1}^{N} \alpha_i \phi_{\mathbf{H}}(\mathrm{y} - \mathrm{y}_i) \ . \tag{8.6}$$

The asymptotic mean integrated squared error (AMISE) allows us to measure the fit of the estimated distribution $\hat{P}_{KDE}(\mathrm{y})$ to the unknown underlying distribution $P_u(\mathrm{y})$ and is defined as

$$\mathrm{AMISE} = (4\pi)^{-d/2} |\mathbf{H}|^{-1/2} N_\alpha^{-1} + \frac{1}{4} d^2 \int \mathrm{tr}^2 \{\mathbf{H} \mathcal{G}_{P_u}(\mathrm{y})\} \, d\mathrm{y} \tag{8.7}$$

where $\mathrm{tr}\{\cdot\}$ is the trace, $\mathcal{G}_{P_u}(\mathrm{y})$ is the Hessian of the unknown probability $P_u(\mathrm{y})$ and $N_\alpha = (\sum_{i=1}^{N} \alpha_i^2)^{-1}$

We can use AMISE to determine the optimal bandwidth $\mathbf{H}$ (according to the observable data) of the kernel used in $\hat{P}_{KDE}(\mathrm{y})$ to estimate $P_u(\mathrm{y})$. Defining $\mathbf{H}$ in terms of scale ξ and structure $\mathbf{F}$ as $\mathbf{H} = \xi^2 \mathbf{F}$, the AMISE measure is minimized by

$$\xi_{opt} = \left[d(4\pi)^{d/2} N_\alpha |\mathbf{F}|^{1/2} R(P_u, \mathbf{F}) \right]^{-1/(d+4)} \ , \tag{8.8}$$

with

$$R(P_u, \mathbf{F}) = \int \mathrm{tr}^2 \{\mathbf{F} \mathcal{G}_{P_u}(\mathrm{y})\} \, d\mathrm{y} \, . \tag{8.9}$$

Since P_u is unknown, $R(P_u, \mathbf{F})$ is approximated as

$$\hat{R}(P_u, \mathbf{F}, G) = \int \mathrm{tr}\{\mathbf{F} \mathcal{G}_{P_G}(\mathrm{y})\} \, \mathrm{tr}\{\mathbf{F} \mathcal{G}_{P_s}(\mathrm{y})\} \ . \tag{8.10}$$

Here P_s is the sample and P_G is the so-called pilot distribution with covariance matrix $\Sigma_{gj} = G + \Sigma_{sj}$ and G is the pilot bandwidth estimated using the multivariate normal scale

rule as

$$G = \hat{\Sigma}_{smp} \left(\frac{4}{(d+2)N_\alpha} \right)^{2/(d+4)} . \tag{8.11}$$

Here $\hat{\Sigma}_{smp}$ is the estimated covariance from all available samples and the structure $\mathbf{F}$ of the bandwidth $\mathbf{H}$ is approximated using the covariance matrix of the samples as $\mathbf{F} = \hat{\Sigma}_{smp}$ [222].

If the number of samples N is large and is made available to the population density estimation procedure described above, then the Gaussian mixture model defined by $\hat{P}_{KDE}(\mathrm{y})$ can be unnecessarily complex and may over fit the data. Hence a model compression step can be used to reduce the model components [127] from N to M, where $M < N$, as long as the compressed distribution $\hat{P'}_{KDE}(\mathrm{y})$ is within a certain Helliger distance [165]. This means that if $K \in N$ sample points are close to each other, then their corresponding Gaussians in the mixture model can be combined into a single Gaussian with a weight $\hat{\alpha}_i = \sum_{j=1}^{K} \alpha_j$.

8.2.5 MRI Disease-State-Score

We propose to model different stages in the disease trajectory using the probabilistic distribution of different classes that can be estimated from different class populations (Section 8.2.4) and from the samples' coordinates in the low-dimensional manifold (Section 8.2.3). We then construct a probabilistic scoring function that determines the class likelihood in the low-dimensional space, and hence, model the disease trajectory as a continuous variable. Thus

$$\begin{aligned} f(y) &= P_B(\mathrm{y}) - P_A(\mathrm{y}) \\ f(y) &= \sum_{i=1}^{N_B} \alpha_{Bi} \phi_{\Sigma_{Bi}}(\mathrm{y}) - \sum_{i=1}^{N_A} \alpha_{Ai} \phi_{\Sigma_{Ai}}(\mathrm{y}) \ , \end{aligned} \tag{8.12}$$

where P_A and P_B are the estimated probability distributions of classes A and B, respectively, α_{*i} and Σ_{*i} are the weights and covariances associated to the i-th element in the Gaussian mixture model and N_A and N_B are the number of components in each model.

The difference between class probability functions (mixture of Gaussians) evaluated at

a test point y (the unseen or test subject embedded in the manifold), can be written as the logarithm of their division. Normalizing the difference (logarithmic division) and rewriting this using a sigmoid (logistic) function we obtain the following scoring function:

$$S(y) = \left(\frac{2}{1 + \frac{P_A(y)}{P_B(y)}} \right) - 1 \ . \tag{8.13}$$

Here $S(y)$ ranges from -1 to 1, and the sign represents the class while the absolute value indicates the class likelihood probability. The continuous nature of the proposed metric provides a richer variable that can be used to define "heat" maps in the manifold associated with a particular class, e.g. AD, CN, sMCI or pMCI. This could be used to define high certainty regions in the manifold where predictions can be made with a high degree of confidence. Additionally, the "heat" maps provide a color coded visualization tool of a patients current "state" for clinicians. Restricting classification/prediction to high confidence areas can be used for patient enrollment in clinical trials, e.g. it might be of particular interest to find subjects that with a certain (high) degree of confidence will convert from MCI to AD in a certain amount of time.

8.3 Results

Using sparse regression (elastic net) as described in Section 8.2.2, we obtain a probabilistic mask of the relevance of each variable or voxel in the image. This mask relates the importance of each voxel in a regression that models the MMSE score. We have used MMSE scores instead of the disease label since they offer a more continuous representation of disease progression. An example of the probabilistic mask obtained from the voxel selection using elastic net regression can be seen in Figure 8.2: This shows three orthogonal 2D views of the MNI template and the voxel importances. In this image, it can be observed that the variables with higher probability cluster around the hippocampus, which is a well known marker of AD. Thresholding this mask at a certain probability of a voxel being "picked" by the sparse regression, produces a ROI. In our experiments we found empirically that a 10% threshold produces the best results, which yields a mask of 1,331 voxels. This parameter

can be also tuned using cross validation.

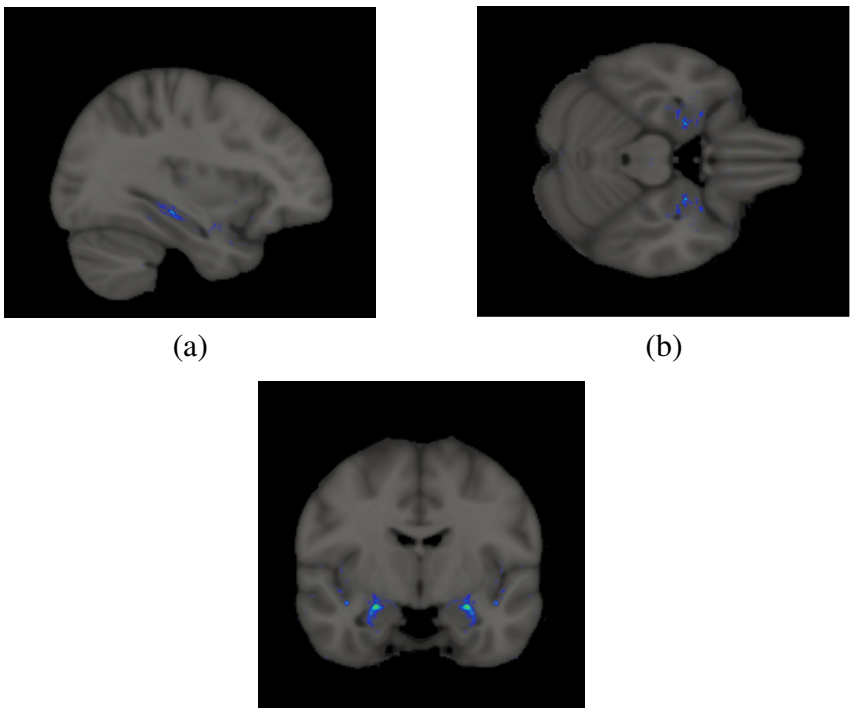


Figure 8.2: (a) Sagittal, (b) axial and (c) coronal orthogonal views of MMSE probabilistic variable selection mask in MNI space (best seen in color). Brighter colors (light blue) indicate a higher probability of the voxel being selected by Equation (8.1).

Using the obtained mask to define the ROI in the images, we learn a low-dimensional representation of the ROI using Laplacian eigenmaps (see Section 8.2.3). Within the ROI we compute cross correlation as a similarity metric between subjects' ROIs. Finally, the population distribution modelling is carried out directly on the learned subspace using the methodology described in Section 8.2.4. An overview diagram of the methods main steps is shown in Figure 8.3.

In order to measure the different aspects of the proposed methodology, a number of experiments were designed. Although the proposed MRI-DSS metric allows for a continuous disease modelling, experiments based on classification tasks are presented in order to allow easy comparison to previous work. In the following sections we report the classification performance for the clinically relevant class separations of the ADNI and ADNIGO datasets. Additionally, the value of performing variable selection as well as manifold learning are illustrated by showing an overall improved classification accuracy. We also show accurate

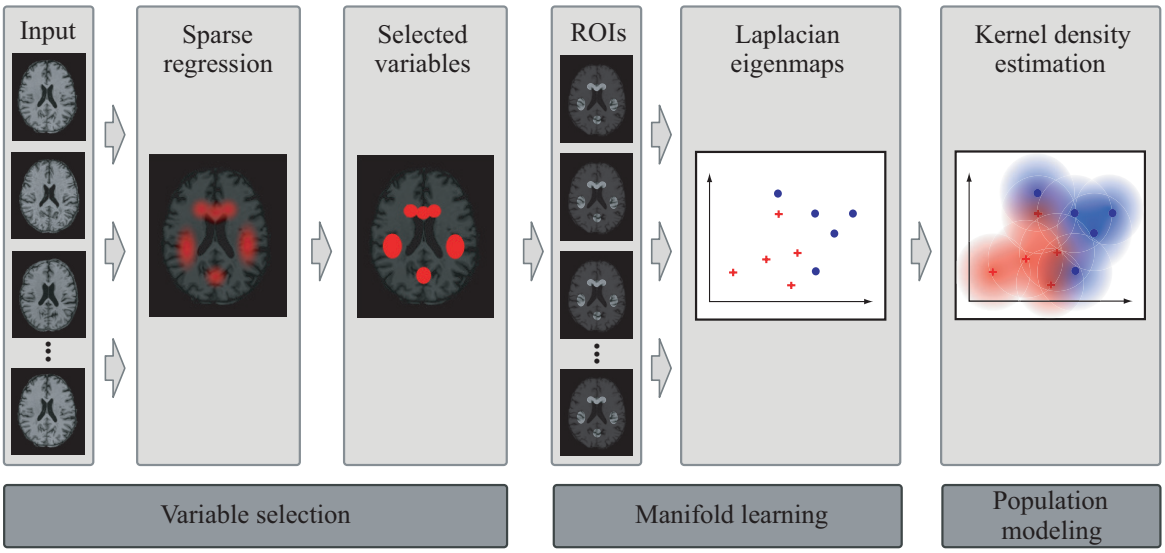


Figure 8.3: Diagram showing the three main stages of the method: variable selection, manifold learning and population modelling (best seen in color).

MMSE score prediction using the proposed MRI-DSS.

8.3.1 ADNI Classification

All classification tasks were carried out in the manifold subspace learned from the selected variables according to Section 8.2.2. In order to incorporate new subjects, the manifold subspace must be re-learned using all available subjects. Two types of classifiers were explored in this work: A SVM approach [42] and the proposed probabilistic distribution threshold, MRI-DSS.

SVM uses training data to find an optimal separating hyperplane between two classes in an n-dimensional feature space. Using this n-dimensional hyperplane, test subjects are classified according to their relative position in the manifold. We used a SVM with a linear kernel function as well as Matlab's default settings.

The probabilistic distribution threshold was obtained by combining the estimated distribution from both classes and normalizing values to form a sigmoid shaped MRI-DSS function. Values range from -1 to 1 and the absolute value indicates class likelihood probability. Thresholding this scoring function (Equation (8.12)) at zero allows us to binarize the scores in order to obtain a classification.

We used both methods to measure classification ACC, SEN and SPE. The results for

the comparisons CN vs AD, sMCI vs pMCI and CN vs pMCI, using the ADNI dataset (see Table 8.1) are presented in Table 8.4. Results for CN vs eMCI, using the ADNIGO dataset (see Table 8.2) are shown in Table 8.6. In all experiments we used a leave 10% out cross-validation strategy and the results presented reflect the average over 1,000 runs.

Two classifier learning paradigms were explored in order to obtain classification accuracy: One, which is referred to as classifier A, where the classifier is learned on single classes (CN, sMCI, pMCI and AD), and another, called classifier B, where we group together similar classes and treat them as one (CN-sMCI and AD-pMCI). Table 8.3 highlights how the classifier paradigms are employed.

	Train \ Test	AD vs CN	pMCI vs sMCI	pMCI vs CN
A	AD, CN	✓	✗	✗
	pMCI, CN	✗	✓	✗
	pMCI, sMCI	✗	✗	✓
B	pMCI-AD, sMCI-CN	✓	✓	✓

Table 8.3: Classifier paradigms A and B with their associated testing and training classes.

Considering a disease progression that follows a trajectory from CN to MCI to AD, and assuming that sMCI subjects tend to be more CN like, while at the same time pMCI subjects tend to be more AD like, grouping them together in order to boost the classifier training data can be justified. By doing so we can train a classifier or probability distribution as a class that includes CN and sMCI, and pMCI and AD in another group. From this point these two classification paradigms will be referred to as classifier A and classifier B, respectively.

In this work we used SVM with a linear kernel function, soft-margin constant $C = 1$ and quadratic programing optimization. Fine tunning the soft-margin constant provides slightly better results, however results are generally robust for a very large range of values ($1e^{-6} < C < 1e^{5}$). Figure 8.4 shows a boxplot of a grid search of C for sMCI vs pMCI classification using a type B classifier. Here, instances are an average of the accuracies obtained in manifolds with 1-50 dimensions, while the red line, box, whiskers and crosses represent the median, the 75th percentile, the extremes and the outliers, respectively, of 100 runs were done for every value of C. Preliminary experiments showed that using non-linear kernels provided little to no improvement, while adding more tuning parameters to the

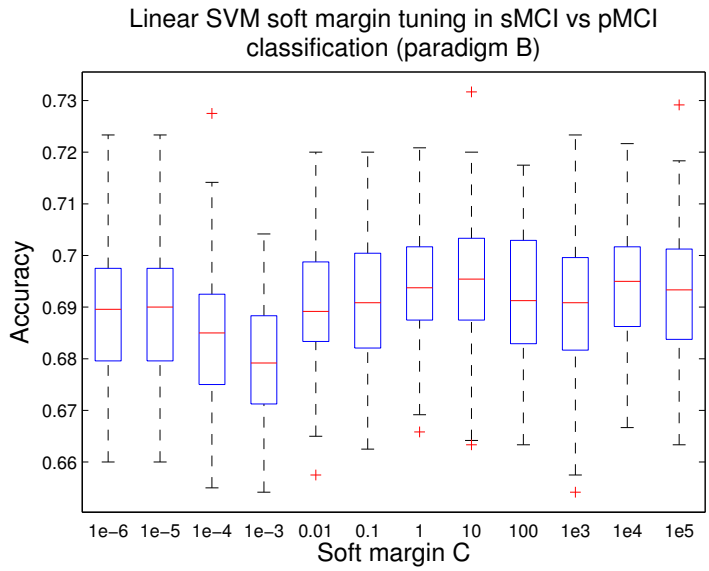


Figure 8.4: Boxplot of results from a grid search of the soft margin parameter C in linear SVM. Instances are an average of the accuracies obtained across 50 manifolds (with 1-50 dimensions). The middle red line, box, whiskers and crosses represent the median, the 75th percentile, the extremes and the outliers, respectively. 100 runs done for every value of C.

framework. Results for type A classifiers as well as other classification tasks show similar robustest to the setting of C.

The parameter for the Laplacian eigenmaps algorithm were set empirically based on previous experience [232, 230, 92]. The number of nearest neighbors used to build the similarity graph was set to 20, although similar results are obtained for values between 10-25. Finally, the dimensionality of the manifold was explored systematically from 1-50 dimensions and the best values are reported. We also found that the results are robust against the choice of these parameters, with stable SVM classification results in manifold dimensionalities from 10-30 while for the case of MRI-DSS the best performing dimensionalities are consistently in the 1-10 range. This is due to the relatively low number of samples used to learn the higher dimension probabilistic models.

Table 8.4 shows classification results on the manifold learned from the selected variables, which in the table are referred to as learned mask SVM and learned mask MRI-DSS. It can be seen that the results are comparable to the state-of-the-art (see Table 8.9). In general, results indicate that SVM performs on average better than MRI-DSS, however, it must be noted that the proposed metric tries to model a more complex variable (the whole population

distribution) with the added benefit of providing good visualization capabilities of the results that can potentially be used to show progression from a “low-risk” zone to a “mild“ or ”high-risk“ zone in the manifold. In order to asses the value of doing variable selection, as opposed to using a predefined structural mask, we repeated the experiments using the same structural mask used in [232], which defines a ROI of around 30,000 voxels around the hippocampus. The results of classification on the manifold learned based on this ROI and in the same manner as before are presented in Table 8.4 (Hippocampal mask SVM and Hippocampal mask MRI-DSS). It can be seen that, in every case, using the learned mask provides more accurate results. Furthermore, we can notice that classifier paradigm B on average produces a slightly higher accuracy than paradigm A in the AD vs CN and pMCI vs sMCI classification tasks. We believe this due to the added training samples which should provide a more robust classifier. However, this trend seems to reverse for the pMCI vs CN classification task. A paired t-test reveals mixed results on the statistical significance between classifier paradigm A and B (see Table 8.3.1), with the highest significance seen in the pMCI vs sMCI classification. We also note that the testing data belongs only to the groups specified, regardless of classifier paradigm. Figure 8.5, shows a 2D visualization of the subjects in the learned subspace manifold, based on the learned ROI, and Figure 8.6 shows the probability distribution mixture of classes.

Another important part of the proposed methodology is the use of manifold learning. We have evaluated the importance of this by performing a comparison of classifiers trained and tested on the subjects without manifold learning. The results for this experiment are presented in the bottom line of Table 8.4. Again it can be observed that for every case learning classifiers on the manifold space outperforms classifiers learned in their original space. Note that only a classifier like SVM that is able to deal with relatively high dimensional data can be used for comparison.

8.3.2 ADNIGO classification

Experiments were carried out using the learned ROI from the ADNI database to classify the ADNIGO subjects performing manifold learning and population modelling, see Table 8.2.

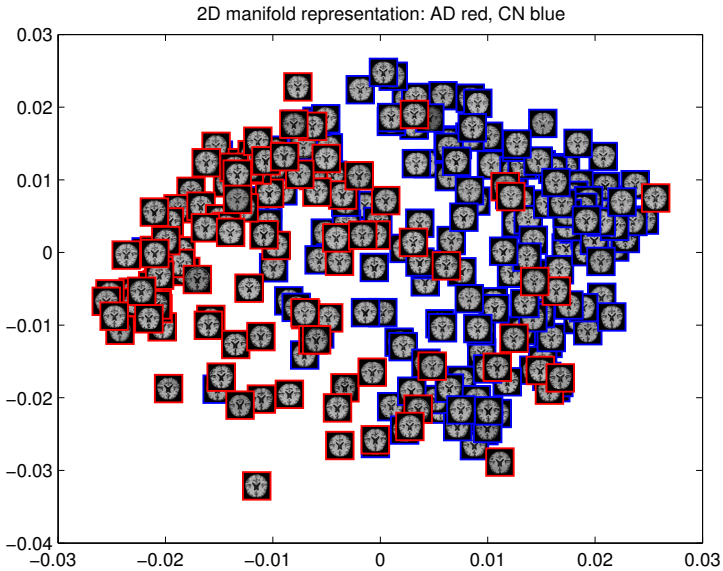


Figure 8.5: 2D visualization of the manifold of AD (red outline) and CN (blue outline) subjects for the ADNI dataset (best seen in color).

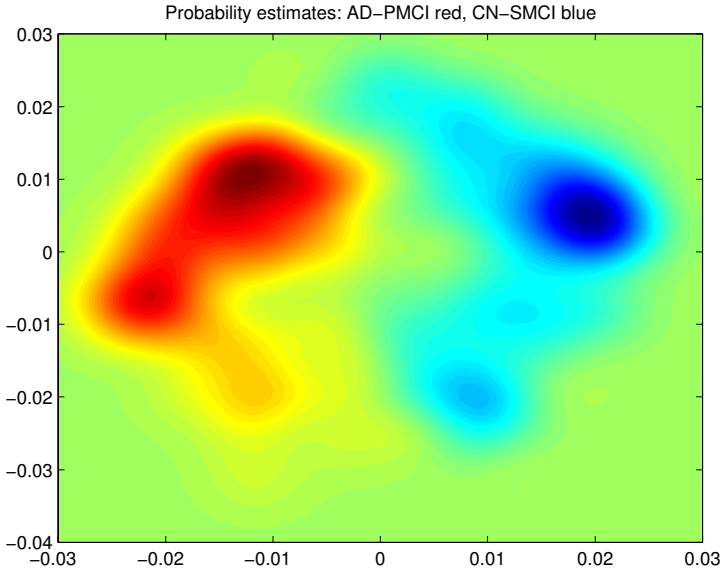


Figure 8.6: 2D visualization of the probability estimates in the manifold for the ADNI dataset (best seen in color). Dark red indicates high AD probability and dark blue very low AD probability.

	AD vs CN			pMCI vs sMCI		
	ACC	SEN	SPE	ACC	SEN	SPE
Learned mask SVM	84/**86**	84/86	**85/85**	69/**71**	**77**/75	60/**67**
Learned mask MRI-DSS	81/81	80/83	82/82	66/67	72/71	59/64
Hippocampal mask SVM	81/81	79/83	82/79	60/66	67/70	53/61
Hippocampal mask MRI-DSS	76/78	82/87	71/69	58/61	53/60	63/62
No manifold learning SVM	84/75	**91**/76	77/73	61/62	61/69	61/55

	pMCI vs CN		
	ACC	SEN	SPE
Learned mask SVM	**82**/81	81/**86**	**83**/76
Learned mask MRI-DSS	77/78	72/85	82/70
Hippocampal mask SVM	76/75	75/71	77/79
Hippocampal mask MRI-DSS	70/69	63/55	77/82
No manifold learning SVM	68/66	77/77	59/55

Table 8.4: Classification results in percentage on the manifold built using the learned ROI (Learned mask SVM and Learned mask MRI-DSS) and on a manifold built from the hippocampal mask used in [232] (Hippocampal mask SVM and Hippocampal mask MRI-DSS). In all cases results for classifiers A and B are presented separated by "/". Best results shown in bold numbers.

	AD vs CN	pMCI vs sMCI	pMCI vs CN
Learned mask SVM	p=0.006	p<0.001	p=0.367
Learned mask MRI-DSS	p=0.593	p=0.745	p=0.415
Hippocampal mask SVM	p=0.255	p<0.001	p<0.001
Hippocampal mask MRI-DSS	p<0.001	p<0.001	p<0.001
No manifold learning SVM	p=0.041	p<0.001	p=0.180

Table 8.5: p-values from paired t-tests between classifier paradigms A and B.

Preprocessing was carried out in the same manner as for ADNI. The results are presented in Tables 8.6, 8.7 and 8.8. Table 8.6 presents the results obtained using the same ROI as for the experiments using ADNI. The p-values indicate the probability that the manifold coordinates from both classes belong to the same distribution. Two permutation tests were used to assess this, multivariate analysis of variance (MANOVA) and the Cramer test [14]. The former assumes a normal distribution of the data, while the latter does not make such an assumption. As can be seen in Figure 8.7, the normality assumption of the data distribution does not necessarily hold. Nevertheless results from both tests are presented. Table 8.7 presents the results of using the hippocampal mask used in [232], again to show the added value of the variable selection step. The results shown in Table 8.8 use a ROI obtained from the variable selection procedure with a threshold of 1% on the probabilistic soft mask. This thresholding yielded 17,428 voxels that includes more varied areas of the brain other

than the hippocampus and its vicinity. The improvement in the results as a result of using the learned ROI is hypothesized to be due to the subtle contributions of areas of the brain other than the hippocampus. Figure 8.7 shows the population in the manifold and Figure 8.8 shows the class probability distributions. As expected, we see that the classes' probability distributions pose more challenging questions, hence, accounting for the relatively low classification accuracy.

	eMCI vs CN			
	ACC	SEN	SPE	p-value (MANOVA/Cramer)
SVM	**61**	**76**	46	0.0003/**0.001**
MRI-DSS	**61**	66	**56**	**0.0002**/0.004

Table 8.6: Classification results using selected variable mask from ADNI thresholded at 10% to learn a manifold for ADNIGO. Best results shown in bold numbers.

	eMCI vs CN			
	ACC	SEN	SPE	p-value (MANOVA/Cramer)
SVM	**57**	**59**	54	0.0041/0.004
MRI-DSS	**57**	54	**59**	**0.0019/<0.0001**

Table 8.7: Classification results using hippocampal mask to learn a manifold for ADNIGO. Best results shown in bold numbers.

	eMCI vs CN			
	ACC	SEN	SPE	p-value (MANOVA/Cramer)
SVM	**65**	**61**	69	**<0.0001/<0.0001**
MRI-DSS	61	50	**72**	**<0.0001/<0.0001**

Table 8.8: Classification using selected variable mask from ADNI thresholded at 1% to learn a manifold for ADNIGO. Best results shown in bold numbers.

8.3.3 MMSE prediction

An additional experiment was carried out to estimate MMSE scores from the manifold. Using a linear regression model built from the MRI-DSS obtained from the learned low dimensional population distributions yielded an average error of 1.5 points. From Table 8.1 we can see that in ADNI class mean MMSE values are separated by 2.2 points for AD-MCI, 3.72 points for MCI-CN, and a smaller separation of 0.65 points exists between pMCI-sMCI

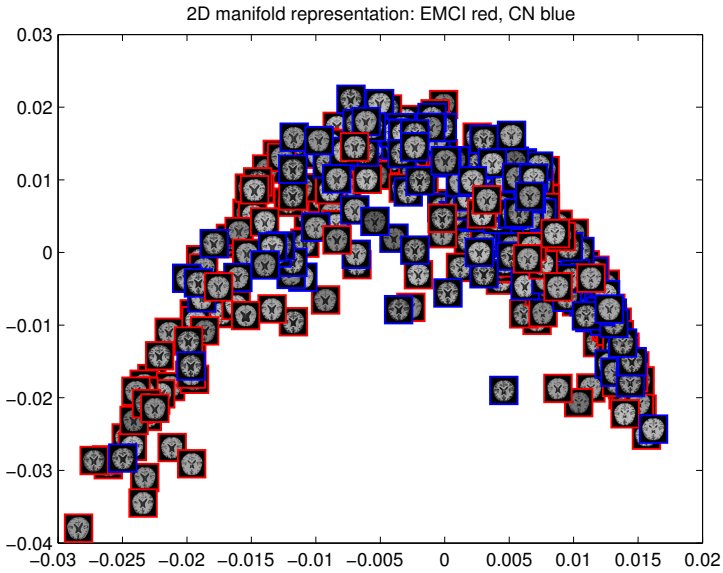


Figure 8.7: 2D visualization of the estimated manifold of eMCI (red outline) and CN (blue outline) subjects for the ADNIGO dataset (best seen in color).

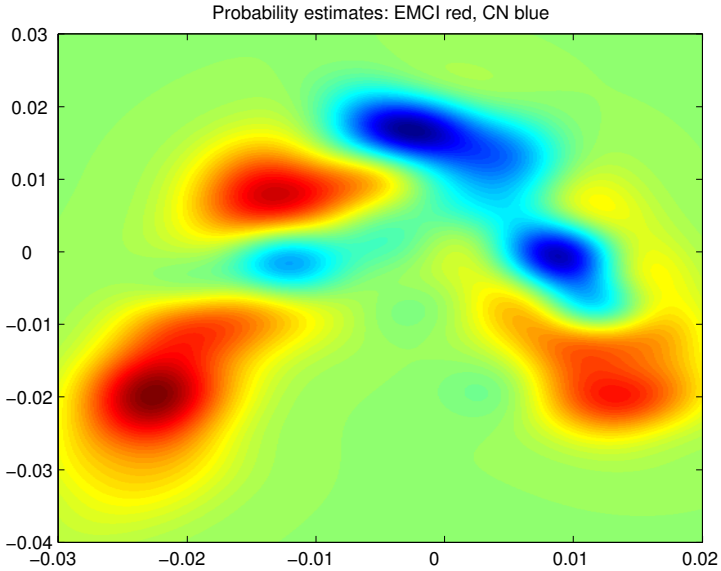


Figure 8.8: 2D visualization of the probability estimates in the manifold for the ADNIGO dataset (best seen in color). Dark red indicates high AD probability and dark blue very low AD probability.

MMSE values. Furthermore, in ADNIGO (Table 8.2), a separation between CN-eMCI mean MMSE scores of 0.7 points can be observed. When originally proposed, the MMSE [72] was shown to have test-re-test mean variation of 1.1 points when the same tester performs both examinations within a 24 hour period on the same patients while a slightly higher mean variation of 1.3 can be observed when one tester performs the first examination and another the second. Thus, the prediction accuracy of the presented method is comparable to the variability of the test itself.

8.4 Conclusions

Recently the task of predicting conversion to AD has received a lot of attention, particularly for subjects in the MCI group. Several approaches that seek to classify the data in order to carry out this prediction task have been proposed in the literature. The proposed method learns a ROI using elastic net regression with a richer response variable (MMSE scores) rather than what could be considered over-simplistic class labels that do not fully explain the disease stage. In a database such as ADNI the MMSE scores should be highly correlated with the class labels since MMSE scores form part of the inclusion and diagnostic criteria of the study. Another important point to note is that the proposed MRI-DSS metric parameterizes the class likelihood as a continuous score. This could potentially be used to define areas of high or low diagnostic certainty. An added benefit of the proposed MRI-DSS is the intuitive visualization of the probability maps in lower dimensional spaces (1-3 dimensions). Classification results reported for other methods are shown in Table 8.9. A direct comparison between methods is difficult due to differences in the datasets, preprocessing steps, validation techniques, etc. However, some observations can be made about the advantages and disadvantages of the different methods. Here we focus the discussion on studies that report results on the sMCI/pMCI classification task, as this is arguably the most clinically relevant.

Cho et al. [34] classified subjects based on cortical thickness features using the same dataset as used in Cuingnet et al. [48], obtaining an accuracy of 71% but with relatively

low sensitivity of 63%. Chupin et al. [37] obtain a classification accuracy of 64% using hippocampal volumes as features, but also report a low sensitivity of 60%. Coupé et al. [43] use patch-based segmentation to segment the hippocampus while at the same time scoring voxels as AD-like or CN-like. To our knowledge their results represent the best results achieved so far using all available images from the ADNI cohort. They have reported an accuracy of 74% although they have a more complex preprocessing pipeline. Cuingnet et al. [48] evaluated various structural methods, for which the obtained accuracies range from 57% to 71% with relatively low sensitivities. Davatzikos et al. [50] used voxel-based morphometry (VBM) to classify subjects. They achieved an accuracy of 56% with a high sensitivity of 95% but at the cost of a very low specificity of 38%. Eskildsen et al. [61] used tensor-based morphometry (TBM) along with cortical ROIs: Subjects with similar time to conversion were pooled together and tested independently achieving high accuracies ($\sim$79%), however when the features selected were used on the whole dataset the accuracy obtained was 68%. Koikkalainen et al. [124] used TBM with a combination of classifiers to achieve an accuracy of 72%. However it is suggested by [61, 43] that this high accuracy might be biased since the ROI used are obtained using the training and testing dataset. Misra et al. [146] use VBM to find discriminative ROI in the images, the highest accuracy reported is of 82%, however the low number of subjects included in the study makes it hard to compare it to other methods. Querbes et al. [169] used cortical thickness features within ROI to achieve an accuracy of 73%. As in [124], the ROI is learned from both training and testing dataset. Westman et al. [226] used predefined cortical ROIs and subcortical structure volumes to predict conversion, achieving and accuracy of 59%. Wolz et al. [232] used a combination of methods and features to obtain classification accuracies between 64-68%. Zhang et al. [235] used longitudinal images to learn ROIs within the whole brain. Their highest accuracy reported is 78%. However, as in [146, 169] the small number of subjects used in this study makes it hard to compare with other methods.

As it can be seen, the proposed method offers comparable classification and prediction results to other state-of-the-art techniques. One of the main strengths of the proposed method is the ability to model the entire population. This provides good visualization prop-

erties in the learned manifold, which can also be used to define ”hot“ spots where there is a high degree of confidence in the classification/prediction made. However, as with any other method it has some disadvantages, mainly the fact that the manifold an distribution have to be relearned every time a new subject is added to the cohort.

Article	Feature(s)	Method	Conversion period	N(sMCI, pMCI)	ACC (SEN/SPE) %
Cho et al. [34]	Cortex	Cortical thickness	0-18 months	131, 72	71 (63/76)
Chupin et al. [37]	Hip. and amygdale	Atlas based	0-18 months	134, 76	64 (60/65)
Coupé et al. [43]	Hip. and entorhinal cortex	Atlas based (LNOCV)	0-48 months	238, 167	74 (73/74)
	-	Atlas based (LOOCV)	-	-	71 (70/71)
Cuingnet et al. [48]	Hippocampus	Atlas based	0-18 months	134, 76	67 (62/69)
	Whole brain	VBM (gray matter)	-	-	71 (57/78)
	Cortex	Cortical thickness	-	-	70 (32/91)
Davatzikos et al. [50]	Whole brain	VBM	0-36 months	170, 69	56 (95/38)
Eskildsen et al. [61]	Cortex	TBM, Cortical ROIs	0-48 months	227, 161	68 (68/69)
Koikkalainen et. al. [124]	Whole brain	TBM, combination of classifiers	0-36 months	215, 164	72 (77/71)
Misra et al. [146]	Whole brain	VBM, ROIs	0-36 months	76, 27	82 (-/-)
Querbes et al. [169]	Cortex	Cortical thickness	0-24 months	50, 72	73 (75/69)
Westman et al. [226]	Cortical and subcortical	Thickness and volume	0-12 months	256, 62	59 (74/56)
Wolz et al. [232]	Hippocampus	Atlas based	0-48 months	238, 167	65 (63/67)
	Whole brain	TBM	-	-	64 (65/62)
	Hip. and amygdale ROI	Manifold learning	-	-	65 (64/66)
	Cortex	Cortical thickness	-	-	56 (63/45)
	Combination	Combination	-	-	68 (67/69)
Zhang et al. (2012) [235]	Whole brain	Whole brain ROIs	0-24 months	50, 38	78 (79/78)

Table 8.9: Previous state-of-the-art work results on classification of sMCI vs pMCI.

Chapter 9

Conclusions and future work

This thesis has proposed several new learning-based methods for landmark localization, feature matching and biomarker extraction. We have described the methodological aspects of the novel approaches and have compared them to other well-established state-of-the-art techniques. We have demonstrated the applicability of the methods to different clinical problems. The evaluation of the proposed methods has been carried out on several large and diverse datasets including 3D MR images of the brain, 3D MR images of the knee and 2D facial images.

One of the drawbacks of classical approaches to medical image analysis, is that they tend to require large amounts of expert knowledge. Machine learning-based techniques have the ability to learn complex patterns from the data. This enables methods to require less expert input. The three main areas of contribution of this thesis, presented in Chapters 4, 5, 6, 7 and 8, are as follows:

- Anatomical landmark localization in brain and knee MR images, as well as in facial images, using techniques such as boosting, manifold learning, regression, 3D local SS descriptors, dictionary learning and sparse coding.

- Feature matching and affine transformation estimation in knee MR images based on a combination of 3D local SS descriptors, forward-backward matching and RANSAC as robust estimator.

- AD biomarker discovery in brain MR images using machine learning: Sparse regression for data-driven variable selection, manifold learning for dimensionality reduction, and non-parametric density estimation for population modeling.

The validation of the proposed methods has been carried out using large datasets in order to show not only accuracy, but also robustness, which is considered an important factor throughout this work. The ADNI (see Appendix A) study is the biggest study on MR imaging in dementia so far [148]. With around 830 participants and dozens of imaging sites, using several types of scanners, ADNI provides a very rich dataset that comes close to real clinical practice. Similar to ADNI in its scope, the ADNIGO dataset provides more brain MR images and has approximately 360 participants. The OAI (see Appendix B) is an observational study of knee osteoarthritis (OA), with aim to facilitate the scientific evaluation of OA biomarkers. The OAI established a database that included clinical, radiological (x-ray and MR) and a biospecimen repository from 4796 men and women ages 45-79.

The work presented in Chapters 4, 6 and 7 mainly dealt with the localization of landmarks in brain MR images that were manually annotated by an expert. The most accurate method was presented in Chapter 6: This technique is based on Laplacian Eigenmaps, with prior knowledge of the spatial distribution of the landmarks regression. The results showed that the proposed method significantly outperforms a 3D SW classifier, data-specific feature descriptors, SS feature descriptors and non-rigid registration in localizing the landmarks. The main drawback of this approach is that is not very scalable, as it requires a landmark-specific manifold to be learned. This would require very large amounts of memory to be stored, while at the same time requiring a high degree of expert knowledge.

Chapter 5 presented a method that uses self-similarity features, rather than raw intensities, to establish feature point matches between a pair of images. This, in combination with RANSAC, enables the registration of images in a more robust and accurate way. Using a subset of 75 randomly selected 3D baseline MR images from the OAI public dataset, quantitative results about the improvements made over raw intensity based registration, were presented. Landmark alignment accuracy improvements are reported in $\sim$82% of the cases, while virtually eliminating all misregistrations. However, the main disadvantage of the pro-

posed framework is that it can currently only handle affine registrations, which can limit its applicability.

The biomarker presented in Chapter 8, which is based on sparse regression, manifold learning and non-parametric density estimation, provides an AD classification methodology that is more data-driven than traditional biomarkers. The accuracy of the developed biomarker in classifying the different stages of AD is comparable to other well established biomarkers. Another important point to note is that the proposed biomarker parameterizes the class likelihood as a continuous score. This could potentially be used to define areas of high or low diagnostic certainty. An added benefit of the proposed method is the intuitive visualization of manifold coordinates' probability maps associated to each class. A drawback of the proposed framework is the fact that the manifold has to be recalculated every time a new subject is added and, hence, the class probabilities also need to be re-estimated. Additionally, even though the use of the estimated probability distribution provides a good visualization tool, lower classification accuracies are observed using this metric over a SVM classifier.

9.1 Future work

Landmark localization methods presented in this thesis provide accuracies in the range of 1.2-3.1 mm for the 20 subcortical brain landmarks described in Appendix C.1. According to Rueckert et al. [181], the average intra-observer variability for the same 20 brain landmarks in 25 schizophrenic subjects is 0.84 mm while the inter-observer variability is 1.05 mm. Although caution must be taken as this measures are derived from a different dataset than the one used in this thesis, it is at least an indicative that there might be room for improvement in terms of accuracy in most of the landmark localization methods presented in this thesis. Taking this into account, we could consider the fact that landmarks that belong to specific structures, e.g. the brain, knee or the face, are not only constrained in their position within the structure at hand, but they are also constrained by the position of other landmarks. For example, the inner and outer aspect, and the inferior tip of the splenium of the corpus

callosum, are part of the same anatomical structure, so their locations with respect to each other are strongly correlated. A graphical model of the joint spatial distribution probabilities of related landmarks, e.g. representing the fact they belong to the same anatomical structure, in the form of a Markov random field, could be used to introduce spatial awareness to the model. The joint spatial probabilities between related landmarks could be modelled as a multivariate Gaussian. Choosing the most probable configuration of all the landmarks according to the graphical model would allow the estimation of landmarks with sub-pixel accuracy. The marginal spatial probabilities of the landmarks, e.g. a Gaussian fit to the output from all the predictions made from the one of the proposed method, could act as a weight on the joint spatial probabilities. For the method presented in Chapter 7, this would only require using other landmarks as support points, rather than performing a random selection. Another interesting avenue to explore for future research is to extend some of the proposed landmark localization techniques to work with arbitrary points or features. This would allow the fast identification of an arbitrary, dense set of landmarks. The located landmarks could then be used to establish correspondences and provide a fast, initial registration. If the approaches are extended to pseudo-landmarks (e.g. control points as used in free-form deformation registration) which are arranged as a dense regular grid over the image voxels, this would enable its use as a similarity metric for a dense image registration algorithm.

There are several fundamental open questions that need to be answered in AD classification/analysis. In the ADNI database, there is no definitive ground truth about subject labeling, only gold standards provided by experts. Diagnosis by clinical assessment alone is accurate in 90% of the cases when validated against neuropathological standards [172]. Hence, there is always a level of uncertainty associated to a subject's label, e.g. it could be that although the subject suffers from dementia, it might not be necessarily AD but instead another form of dementia, e.g. fronto-temporal dementia (FTD) or dementia with Lewes bodies. In terms of classification accuracy this must be taken into account, as any improvement over the intrinsic label error in the data might be due to over fitting rather than an overall improvement in accuracy. A method that could detect outliers in the data could potentially point to possible mislabeling of specific data instances. It would be interesting if

such data driven approach could reach an agreement with clinical experts. Additionally, in realtion to the work on biomarker extraction detailed in Chapter 8, there are several ways to extend the proposed methodology. For instance, in the work presented a 10 mm FFD [182] grid was used to align images to a common space. The justification of doing so is the removal of coarse non-linear inter-subject anatomical variations, while aligning smaller structures. There is no guarantee that a 10 mm control point spacing of a FFD is optimal, or furthermore, there is no guarantee that there exists any optimal one. Future work could include a multilevel variable selection step, where each observation can be a concatenation of the same image with different levels of alignment (e.g. from 20 to 5mm FFD control point spacing). Sparse regression could be used to select variables from this extended observations set. Another avenue to explore would be incorporation of longitudinal features, variable selection could be done also on longitudinal images in similar fashion, that is, concatenating longitudinal images.

Chapter 10

Publications

Journal Publications

(1) Ricardo Guerrero, Robin Wolz, Anil Rao, Daniel Rueckert. ”Manifold population modelling as a neuro-imaging biomarker: Application to ADNI and ADNI-GO”. NeuroImage, submited, 2013.

Conference Proceedings

(2) Ricardo Guerrero, Daniel Rueckert. ”Data-specific Feature Point Descriptor Matching Using Dictionary Learning and Graphical Models”. SPIE Medical Imaging, pages 866921-8, 2013.

(3) Ricardo Guerrero, Claire Donoghou, Luis Pizarro, Daniel Rueckert. ”Learning correspondences in knee MR images from the Osteoarthritis Initiative”. Machine learning in medical imaging - Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 7588, pages 218-225, 2012.

(4) R. Guerrero, L. Pizarro, R. Wolz, and D. Rueckert. Landmark localisation in brain MR images using feature point descriptors based on 3D local self-similarities. In IEEE International Symposium on Biomedical Imaging (ISBI), pages 1535-1538, 2012.

(5) S. Pszczolkowski, L. Pizarro, R. Guerrero, D. Rueckert. Nonrigid free-form registration using landmark-based statistical deformation models. In Proceedings of SPIE, 2012,

Volume 8314.

(6) K.P. Tung, W.Z. Shi, L. Pizarro, H. Tsujioka, H.Y. Wang, R. Guerrero, R. De Silva, P.E. Edwards, D. Rueckert. Automatic detection of coronary stent struts in intravascular OCT imaging. Proceedings of SPIE, San Diego, USA, February 2012, Volume 8315.

(7) R. Guerrero, R. Wolz, D. Rueckert. Laplacian eigenmaps manifold learning for landmark localization in brain MR images. Medical image computing and computer-assisted intervention (MICCAI), volume part II, pages 566-573, 2011.

Appendix A

ADNI and ADNIGO

ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the FDA, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MR imaging, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the United States and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 CN older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

The purpose of the ADNIGO study is to build upon the information obtained in the original ADNI study and examine how brain imaging can be used with other tests to measure the progression of MCI and early AD. ADNIGO seeks to define and characterize the mildest

symptomatic phase of AD, referred to in this study as early amnestic MCI (eMCI). However, generally no formal sub-categorization between eMCI and MCI (or late MCI) exists. The eMCI subjects represent individuals with milder degrees of cognitive and functional impairment than the MCI subjects, and their rate of progression is slower [3].

A.1 MR image acquisition

In the ADNI study, image acquisition was carried out at multiple sites based on a standardized MRI protocol [110] using 1.5T scanners manufactured by General Electric Healthcare (GE), Siemens Medical Solutions, and Philips Medical Systems. Out of two available 1.5T T1-weighted MR images based on a 3D MPRAGE sequence, we used the image that has been designated as "best" by the ADNI quality assurance team [110]. Acquisition parameters on the SIEMENS scanner (parameters for other manufacturers differ slightly) are echo time of 3.924 ms, repetition time of 8.916 ms, inversion time of 1000 ms, flip angle 8°, to obtain 166 slices of 1.2-mm thickness with a 256 $\times$ 256 matrix.

All images were preprocessed by the ADNI consortium using the following pipeline:

1. *GradWarp*: A system-specific correction of image geometry distortion due to gradient non-linearity [118].

2. *B1 non-uniformity correction*: Correction for image intensity non-uniformity [110].

3. *N3*: A histogram peak sharpening algorithm for bias field correction [193].

Since the Philips systems used in the study were equipped with B1 correction and their gradient systems tend to be linear [110], the preprocessing steps 1 and 2 were applied by ADNI only to images acquired with GE and Siemens scanners.

Appendix B

The Osteoarthritis Initiative (OAI)

The OAI is a multi-center, longitudinal, prospective observational study of knee OA comprised of five public-private partnership (N01-AR-2-2258; N01-AR-2-2259; N01- AR-2-2260; N01-AR-2-2261; N01-AR-2-2262), funded by the National Institutes of Health (NIH), a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. The overall aim of the OAI is to develop a public domain research resource to facilitate the scientific evaluation of biomarkers for OA as potential surrogate endpoints for disease onset and progression.

OA is a joint debilitating pathology characterized by erosion of the articular cartilage. It is a widespread disease which causes joint pain, tenderness and stiffness in patients. Around 35 million people in the United States (13 % of the population) are 65 or older, and more than half of them have radiological evidence of osteoarthritis in at least one joint. By 2030, 20 % of Americans (about 70 million) will be over 65 and will be at risk for OA. Hence, in addition to the impact on individuals it also represents a significant financial cost to society.

At present, therapies available to treat osteoarthritis are limited. Most current treatments are designed only to relieve pain and reduce or prevent the disability caused by bone and cartilage degeneration. Drug therapies target the symptoms but not the cause of this disease; no treatment inhibits the degenerative structural changes that are responsible for its progres-

sion. Furthermore, clinical testing of new therapies is complicated by the highly variable way in which OA is manifested in individual patients.

Four clinical centers and a data coordinating center conducted the OAI, a public-private partnership that bring together new resources to help find biochemical, genetic and imaging biomarkers for development and progression of OA. The OAI established and maintained a natural history database for osteoarthritis that included clinical evaluation data, radiological (x-ray and magnetic resonance) images, and a biospecimen repository from 4796 men and women ages 45-79 enrolled between February 2004 and May 2006. Three 3.0 Tesla MR imaging scanners, one at each clinical center (with one shared between locations), where dedicated to imaging the knees of OAI participants annually over four years. The project recruited participants who had, and those who where at high risk for developing symptomatic knee osteoarthritis. All data and images collected are be available to researchers worldwide to help quicken the pace of biomarker identification, scientific investigation and OA drug development. Access to biospecimens is given through application to the National Institute of Arthritis, Musculoskeletal and Skin Diseases (NIAMS).

This manuscript uses an OAI public use dataset and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or private funding partners.

B.1 MR image acquisition

In the OAI study acquisition was carried out at multiple sites using dedicated 3T Siemens Trio MR scanners. In this study we used only Sagittal 3D DESS with water excitation sequences. Parameters on these scanners where echo time of 4.7 ms, repetition time of 16.3 ms, inversion time of 4.7 ms, flip angle 25°, to obtain 160 slices of 1.2-mm thickness with a 384×384 matrix.

Sagittal 3D DESS with water excitation enables quantization of cartilage volume over the entire knee (patellofemoral and femorotibial joints). Another primary use of the 3D DESS acquisition is to identify osteophytes in both the original sagittal (superior-inferior patella, anterior-posterior femur and tibia) as well as in the coronal (medial / lateral femur

and tibia) and axial medial-lateral patella (MPR). Secondarily, it also potentially provides assessment of sub-articular marrow edema and cysts both in the original sagittal plane as well as in the coronal (central femur and tibia) and axial (patella) MPR. This latter marrow assessment does not have proven sensitivity and specificity, but is presumed to be less sensitive than a fat suppressed IW or T2W.

Appendix C

Landmark definition

C.1 Brain landmarks

Landmarks where manually selected by an expert observer using 3 orthogonal views. They are defined in the MNI space as follows:

From a sagittal view, along the mid-sagittal plane (MSP). The outer aspect, inferior tip and inner aspect of the splenium of the corpus callosum are the most anterior, lowest and inflection point of this structure. The outer aspect of the genu of the corpus is the most posterior point of the outer wall of the structure and the inner aspect is the most posterior point of the inner wall. The superior aspect of the pons is located by following the superior edge of the pons until the recess at the juncture of the pons with the tegmentum of the mesencephalon is found, the inferior aspect of the pons is found similarly, but following the inferior edge. The superior and inferior aspect of the cerebellum are defined as the most superior and inferior points of the structure. The fourth ventricle extends from the aqueduct of the midbrain to the central canal of the upper end of the spinal cord, the most inner point in the cerebellum is chosen. The anterior commissure can generally be found at the tip of the fornix. The posterior commissure is located at top of the superior colliculus. The anterior tips of the lateral ventricles (left and right) are located at the most anterior point of the lateral ventricles. The inferior tips of the lateral ventricles (left and right) are defined as the lowest points of the lateral ventricle while in the same sagittal plane as the inferior tips.

From an axial view. The putamen posterior and anterior (left and right) are defined as the most frontal and most posterior points of the putamen, which can be easily found using an axial view.

C.2 Knee landmarks

75 MR images were randomly selected from the OAI dataset for testing the methodology. These images were manually annotated by an expert using three orthogonal views by placing four landmark points on the ACL and PCL insertions on the femur and the tibia. These where chosen since they are clearly identifiable in the subject's joint space. The middle voxel of each ligament insertion is selected on the bone interface.

Bibliography

[1] Cortical surface-based analysis: I. segmentation and surface reconstruction.

[2] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Medical Imaging*, 54(11):4311–4322, 2006.

[3] P. S. Aisen, R. C. Petersen, M. C. Donohue, A. Gamst, R. Raman, R. G. Thomas, S. Walter, J. Q. Trojanowski, L. M. Shaw, L. A. Beckett, R. J. Clifford, Jr., W. Jagust, A. W. Toga, A. J. Saykin, J. C. Morris, R. C. Green, and M. W. Weiner. Clinical core of the Alzheimer's disease neuroimaging initiative: Progress and plans. *Alzheimer's & Dementia*, 6(3):239–246, 2010.

[4] A. Aizerman, E. M. Braverman, and L. I. Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[5] P. Aljabar. *Tracking longitudinal change using MR image data*. PhD thesis, University of London, 2007.

[6] P. Aljabar, R. Heckemann, A. Hammers, J. Hajnal, and D. Rueckert. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage*, 46(3):726 – 738, 2009.

[7] P. Aljabar, D. Rueckert, and W. Crum. Automated morphological analysis of magnetic resonance brain imaging using spectral analysis. *NeuroImage*, 43(2):225–235, 2008.

[8] P. Aljabar, R. Wolz, and D. Rueckert. *Machine Learning in Computer-Aided Diagnosis: Medical Imaging Intelligence and Analysis*, chapter Manifold learning for medical image registration, segmentation and classication. IGI Global, 2012.

[9] Y. Amit and D. G. Y. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

[10] W. N. Anderson and T. D. Morley. Eigenvalues of the Laplacian of a graph. *Linear and Multilinear Algebra*, 18:141–145, 1985.

[11] J. Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95 – 113, 2007.

[12] R. Bajcsy and S. Kovacic. Multiresolution elastic matching. *Comput. Vision Graph. Image Process.*, 46(1):1–21, 1989.

[13] R. Bajcsy, R. Lieberson, and M. Reivich. A computerized system for the elastic matching of deformed radiographic images to idealized atlas images. *J. of Computer Assisted Tomography*, 7(4):618–625, 1983.

[14] L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190 – 206, 2004.

[15] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.

[16] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, 2005.

[17] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*, volume 14, pages 585–591, 2002.

[18] Y. Bengio, J. F. Paiement, P. Vincent, O. Delalleau, N. L. Roux, and M. Ouimet. Out-of-sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering. In *In Advances in Neural Information Processing Systems 16*, volume 16, pages 177–184, 2004.

[19] F. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:567–585, 1989.

[20] F. Bookstein. Thin-plate splines and the atlas problem for biomedical images. In *International Conference on Information Processing in Medical Imaging (IPMI)*, pages 326–342, 1991.

[21] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.

[22] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:2001, 2001.

[23] H. Braak and E. Braak. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, 82(4):239 – 259, 1991.

[24] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[25] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[26] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. 1 edition, 1984.

[27] M. Bro-Nielsen and C. Gramkow. Fast fluid registration of medical images. In *Visualization in Biomedical Computing*, pages 267–276. Springer-Verlag, 1996.

[28] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi. Forecasting the global burden of Alzheimers disease. *Alzheimer's & Dementia*, 3(3):186–191, 2007.

[29] A. Buades, B. Coll, and J.-M. Morel. A review of image denoising algorithms, with a new one. *SIAM Multiscale Modeling and Simulation*, 4(2):490–530, 2005.

[30] Y. Cao, M. Miller, R. Winslow, and L. Younes. Large deformation diffeomorphic metric mapping of vector fields. *IEEE Transactions on Medical Imaging*, 24(9):1216–1230, 2005.

[31] J. Carballido-Gamio and S. Majumdar. Atlas-based knee cartilage assessment. *Magnetic Resonance in Medicine*, 66(2):574–83, 2011.

[32] S. Chen, S. A. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5):1873–1896, 1989.

[33] G. Chetelat, B. Desgranges, V. De La Sayette, F. Viader, F. Eustache, and J. Baron. Mild cognitive impairment: Can FDG-PET predict who is to rapidly convert to Alzheimer's disease? *Neurology*, 60(8):1374–7, 2003.

[34] Y. Cho, J.-K. Seong, Y. Jeong, and S. Y. Shin. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. *NeuroImage*, 59(3):2217–2230, 2012.

[35] G. Christensen. *Deformable Shape Models for Anatomy*. PhD thesis, Washington University, 1994.

[36] G. Christensen, R. Rabbitt, and M. Miller. Deformable templates using large deformation kinematics. *IEEE Transactions on Medical Imaging*, 5(10):1435–1447, 1996.

[37] M. Chupin, E. Gérardin, R. Cuingnet, C. Boutet, L. Lemieux, S. Lehéricy, H. Benali, L. Garnero, and O. Colliot. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. *Hippocampus*, 19(6):579–587, 2009.

[38] A. Colin Cameron and F. A. G. Windmeijer. An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2):329–342, 1997.

[39] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal. Automated multi-modality image registration based on information theory. *International Conference on Information Processing in Medical Imaging (IPMI)*, pages 263–274, 1995.

[40] T. Cootes, A. Hill, C. Taylor, and J. Haslam. The use of active shape models for locating structures in medical images. In *International Conference on Information Processing in Medical Imaging (IPMI)*, pages 33–47, 1993.

[41] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 484–498, 1998.

[42] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[43] P. Coupé, S. F. Eskildsen, J. V. Manjn, V. S. Fonov, and D. L. Collins. Simultaneous segmentation and grading of anatomical structures for patient's classification: Application to Alzheimer's disease. *NeuroImage*, 59(4):3736 – 3747, 2012.

[44] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 1994.

[45] M. Craene, O. Camara, B. H. Bijnens, and A. F. Frangi. Large diffeomorphic FFD registration for motion and strain quantification from 3D-US sequences. In *Proceedings of the 5th International Conference on Functional Imaging and Modeling of the Heart*, FIMH '09, pages 437–446, 2009.

[46] W. Crum, L. Griffin, D. Hill, and D. Hawkes. Zen and the art of medical image registration: correspondence, homology, and quality. 20(3):1425 – 1437, 2003.

[47] J. G. Csernansky, S. Joshi, L. Wang, J. W. Haller, M. Gado, J. P. Miller, U. Grenander, and M. I. Miller. Hippocampal morphometry in schizophrenia by high dimensional

brain mapping. *Proceedings of the National Academy of Sciences*, 95(19):11406–11411, 1998.

[48] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehricy, M.-O. Habert, M. Chupin, H. Benali, and O. Colliot. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, 56(2):766–781, 2011.

[49] C. Davatzikos. Spatial transformation and registration of brain images using elastically deformable models. *Comput. Vis. Image Underst.*, 66(2):207–222, 1997.

[50] C. Davatzikos, P. Bhatt, L. M. Shaw, K. N. Batmanghelich, and J. Q. Trojanowski. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging*, 32(12):2322.e19–2322.e27, 2011.

[51] C. Davatzikos, J. Prince, and R. Bryan. Image registration based on boundary mapping. *IEEE Transactions on Medical Imaging*, 15(1):112–115, 1996.

[52] C. Davatzikos, M. Vaillant, S. M. Resnick, J. L. Prince, S. Letovsky, and R. N. Bryan. A Computerized Approach for Morphological Analysis of the Corpus Callosum. *Journal of Computed Assisted Tomography*, 20(1):88–97, 1996.

[53] M. Davis, A. Khotanzad, D. Flamig, and S. Harms. A physics-based coordinate transformation for 3-D image matching. *Medical Imaging, IEEE Transactions on*, 16(3):317–328, 1997.

[54] V. de Silva and J. B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *Advances in Neural Information Processing Systems 15*, pages 705–712, 2002.

[55] C. Donoghue, A. Rao, A. M. J. Bull, and D. Rueckert. Manifold learning for automatically predicting articular cartilage morphology in the knee with data from the osteoarthritis initiative (OAI). In *In Proceedings of SPIE*, page 7962, 2011.

[56] C. R. Donoghue, A. Rao, L. Pizarro, A. M. J. Bull, and D. Rueckert. Fast and accurate global geodesic registrations using knee MRI from the Osteoarthritis Initiative. In *Computer Vision and Pattern Recognition Workshops*, pages 50–57, 2012.

[57] D. L. Donoho and C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. Technical report, 2003.

[58] G. Dougherty. *Medical Image Processing*. Biological and Medical Physics. Springer New York, 2011.

[59] J. Duchon. Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *Revue d'automation, informatique et recherche operationelle*, 10(12):5–12, 1976.

[60] F. Eckstein, F. Cicuttini, J.-P. Raynauld, J. Waterton, and C. Peterfy. Magnetic resonance imaging (mri) of articular cartilage in knee osteoarthritis (oa): morphological assessment. *Osteoarthritis and Cartilage*, 14, Supplement 1(0):46 – 75, 2006.

[61] S. F. Eskildsen, P. Coupé', D. García-Lorenzo, V. Fonov, J. C. Pruessner, and D. L. Collins. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *NeuroImage*, 2012.

[62] Y. Fan, N. Batmanghelich, C. M. Clark, and C. Davatzikos. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage*, 39(4):1731 – 1743, 2008.

[63] Y. Fan, D. Shen, R. Gur, R. Gur, and C. Davatzikos. COMPARE: Classification of Morphological Patterns Using Adaptive Regional Elements. *IEEE Transactions on Medical Imaging*, 26(1):93–105, 2007.

[64] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int. J. Comput. Vision*, 61(1):55–79, 2005.

[65] B. Fischer and J. Modersitzki. Fast diffusion registration. *Contemporary Mathematics*, 313:117–129, 2002.

[66] B. Fischl, N. Rajendran, E. Busa, J. Augustinack, O. Hinds, B. T. Yeo, H. Mohlberg, K. Amunts, and K. Zilles. Cortical folding patterns and predicting cytoarchitecture. *Cerebral Cortex*, 18(8):1973–1980, 2008.

[67] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[68] J. Fitzpatrick, J. West, and J. Maurer, C.R. Predicting error in rigid-body point-based registration. *IEEE Transactions on Medical Imaging*, 17(5):694–702, 1998.

[69] M. Fitzpatrick, D. Hill, and J. C.R. Maurer. *Medical Image Processing*, chapter 8. SPIE, 2000.

[70] Flach, Peter and Hernández-Orallo, Jose and Ferri, Cèsar. A Coherent Interpretation of AUC as a Measure of Aggregated Classication Performance. In *ICML 2011, The 28th International Conference on Machine Learning*, 2011.

[71] P. Fletcher, C. Lu, S. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005, 2004.

[72] M. F. Folstein, S. E. Folstein, and P. R. McHugh. "Mini-Mental State": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, 1975.

[73] L. Ford and D. Fulkerson. *Flows in Networks*. Princeston University Press, 1962.

[74] N. C. Fox, E. K. Warrington, P. A. Freeborough, P. Hartikainen, A. M. Kennedy, J. M. Stevens, and M. N. Rossor. Presymptomatic hippocampal atrophy in Alzheimer's disease: A longitudinal MRI study. *Brain*, 119(6):2001–2007, 1996.

[75] S. L. Free, P. O'Higgins, D. D. Maudgil, I. L. Dryden, L. Lemieux, D. R. Fish, and S. D. Shorvon. Landmark-based morphometrics of the normal adult brain using MRI. *Neuroimage*, 13(5):801–13, 2001.

[76] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[77] B. J. Frey and D. J. C. MacKay. A revolution: Belief propagation in graphs with cycles. In *In Neural Information Processing Systems*, pages 479–485. MIT Press, 1997.

[78] J. Fripp, S. Crozier, S. K. Warfield, and S. Ourselin. Automatic segmentation and quantitative analysis of the articular cartilages from magnetic resonance images of the knee. *IEEE Transactions on Medical Imaging*, 29(1):55–64, 2010.

[79] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6:67–77, 2010.

[80] W. J. Fu. Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

[81] P. Fua. A parallel stereo algorithm that produces dense depth maps and preserves image features. *Machine Vision and Applications*, 6(1):35–49, 1993.

[82] A. J. Gandhe, D. L. Hill, C. Studholme, D. J. Hawkes, C. F. Ruff, T. C. Cox, M. J. Gleeson, and A. J. Strong. Combined and three-dimensional rendered multimodal data for planning cranial base surgery: A prospective evaluation. *Neurosurgery*, 35(3):463–471, 1994.

[83] J. C. Gee, C. Barillot, L. L. Briquer, D. R. Haynor, and R. Bajcsy. Matching structural images of the human brain using statistical and geometrical image features. 1994.

[84] E. Gerardin, G. Chtelat, M. Chupin, R. Cuingnet, B. Desgranges, H.-S. Kim, M. Niethammer, B. Dubois, S. Lehricy, L. Garnero, F. Eustache, and O. Colliot. Multidimensional classification of hippocampal shape features discriminates Alzheimer's

disease and mild cognitive impairment from normal aging. *NeuroImage*, 47(4):1476–1486, 2009.

[85] S. Gerber, T. Tasdizen, P. T. Fletcher, S. Joshi, and R. Whitaker. Manifold modeling for brain population analysis. *Medical Image Analysis*, 14(5):643–653, 2010.

[86] Glaunès, Joan and Vaillant, Marc and Miller, Michael I. Landmark matching via large deformation diffeomorphisms on the sphere. *Journal of Mathematical Imaging and Vision*, 20(1-2):179–200, 2004.

[87] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios. Dense image registration through MRFs and efficient linear programming. *Medical Image Analysis*, 12(6):731 – 741, 2008.

[88] B. Glocker, A. Sotiras, N. Komodakis, and N. Paragios. Deformable medical image registration: Setting the state of the art with discrete methods*. *Annual Review of Biomedical Engineering*, 13(1):219–244, 2011.

[89] A. Goshtasby. Registration of images with geometric distortions. *IEEE Transactions on Geoscience and Remote Sensing*, 26:60–64, 1988.

[90] D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 271–279, 1989.

[91] R. Guerrero, L. Pizarro, R. Wolz, and D. Rueckert. Landmark localisation in brain MR images using feature point descriptors based on 3D local self-similarities. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1535–1538, 2012.

[92] R. Guerrero, R. Wolz, and D. Rueckert. Laplacian eigenmaps manifold learning for landmark localization in brain mr images. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 6892 of *Lecture Notes in Computer Science*, pages 566–573, 2011.

[93] A. Guimond, J. Meunier, and J.-P. Thirion. Average brain models: A convergence study. *Computer Vision and Image Understanding*, 77(2):192 – 210, 2000.

[94] V. Gulshan. http://www.robots.ox.ac.uk/~vgg/software/SelfSimilarity/.

[95] J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes. *Medical Image Registration*. CRC press, 2001.

[96] J. Hamm, D. H. Ye, R. Verma, and C. Davatzikos. GRAM: A framework for Geodesic Registration on Anatomical Manifolds. *Medical Image Analysis*, 14(5):633–642, 2010.

[97] H. Hampel, R. Frank, K. Broich, S. J. Teipel, R. G. Katz, J. Hardy, K. Herholz, A. L. W. Bokde, F. Jessen, Y. C. Hoessler, W. R. Sanhai, H. Zetterberg, J. Woodcock, and K. Blennow. Biomarkers for Alzheimer's disease: Academic, industry and regulatory perspectives. *Nature Reviews Drug Discovery*, 9:560–574, 2010.

[98] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[99] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer New York, second edition, 2009.

[100] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers. Automatic anatomical brain {MRI} segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115 – 126, 2006.

[101] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, J. M. Brady, and J. a. Schnabel. Non-local shape descriptor: a new similarity metric for deformable multi-modal registration. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 14, pages 541–8, 2011.

[102] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832 –844, 1998.

[103] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[104] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[105] P. Hoyer and P. Dayan. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[106] J. R. Hurley and R. B. Cattell. The procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, 7(2):258–262, 1962.

[107] M. D. Ikonomovic, W. E. Klunk, E. E. Abrahamson, C. A. Mathis, J. C. Price, N. D. Tsopelas, B. J. Lopresti, S. Ziolko, W. Bi, W. R. Paljug, M. L. Debnath, C. E. Hope, B. A. Isanski, R. L. Hamilton, and S. T. DeKosky. Post-mortem correlates of in vivo PiB-PET amyloid imaging in a typical case of Alzheimer's disease. *Brain*, 131(6):1630–1645.

[108] C. Izard, B. Jedynak, and C. E. L. Stark. Spline-based probabilistic model for anatomical landmark detection. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2006.

[109] C. R. Jack, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

[110] C. R. Jack Jr., M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, A. M. Dale, J. P. Felmlee, J. L. Gunter, D. L. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C. S. DeCarli, G. Krueger, H. A. Ward, G. J. Metzger, K. T. Scott, R. Mallozzi, D. Blezek, J. Levy, J. P. Debbins, A. S. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, and M. W. Weiner. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

[111] C. R. Jack Jr., R. C. Petersen, Y. C. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, B. F. Boeve, S. C. Waring, E. G. Tangalos, and E. Kokmen. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology*, 52(7):1397–1403, 1999.

[112] C. R. Jack Jr., R. C. Petersen, Y. C. Xu, S. C. Waring, P. C. O'Brien, E. G. Tangalos, G. E. Smith, R. J. Ivnik, and E. Kokmen. Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology*, 49(3):786–794, 1997.

[113] J. Jaremko, R. Cheng, R. Lambert, a. Habib, and J. Ronsky. Reliability of an efficient MRI-based method for estimation of knee cartilage volume using surface registration1. *Osteoarthritis and Cartilage*, 14(9):914–922, 2006.

[114] H. Jia, G. Wu, Q. Wang, Y. Wang, M. Kim, and D. Shen. Directed graph based image registration. 36(2):139–151, 2012.

[115] D. H. J.M. Fitzpatrick and J. C.R. Maurer. *Handbook of Medical Imaging*, volume 2, chapter Image registration, page 447513. SPIE Press, 2000.

[116] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[117] S. Joshi and M. Miller. Landmark matching via large deformation diffeomorphisms. *Image Processing, IEEE Transactions on*, 9(8):1357–1370, 2000.

[118] J. Jovicich, S. Czanner, D. Greve, E. Haley, A. van der Kouwe, R. Gollub, D. Kennedy, F. Schmitt, G. Brown, J. MacFall, B. Fischl, and A. Dale. Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, 30(2):436–443, 2006.

[119] M. Kessler, S. Pitluck, P. Petti, and J. Castro. Integration of multimodality imaging data for radiotherapy treatment planning. *International Journal of Radiation Oncology*Biology*Physics*, 21(6):1653 – 1667, 1991.

[120] R. Keys. Cubic convolution interpolation for digital image processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(6):1153–1160, 1981.

[121] M. Kim, G. Wu, P.-T. Yap, and D. Shen. A generalized learning based framework for fast brain image registration. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 306–314, 2010.

[122] S. Klein, M. Loog, F. van der Lijn, T. den Heijer, A. Hammers, M. de Bruijne, A. van der Lugt, R. P. W. Duin, M. M. B. Breteler, and W. J. Niessen. Early diagnosis of dementia based on intersubject whole-brain dissimilarities. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 249–252, 2010.

[123] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *14th International Joint Conference on Artificial Intelligence*, volume 2, pages 1137–1143, 1995.

[124] J. Koikkalainen, J. Ltjnen, L. Thurfjell, D. Rueckert, G. Waldemar, and H. Soininen. Multi-template tensor-based morphometry: Application to analysis of Alzheimer's disease. *NeuroImage*, 56(3):1134–1144, 2011.

[125] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1568–1583, 2006.

[126] N. Komodakis and G. Tziritas. Approximate labeling via graph cuts based on linear programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1436–1453, 2007.

[127] M. Kristan, A. Leonardis, and D. Skocaj. Multivariate online kernel density estimation with Gaussian kernels. *Pattern Recognition*, 44(10-11):2630–2642, 2011.

[128] J. Kybic and M. Unser. Fast parametric elastic image registration. *IEEE Transactions on Medical Imaging*, 12(11):1427–1442, 2003.

[129] T. Lehmann, C. Gonner, and K. Spitzer. Survey: interpolation methods in medical image processing. *IEEE Transactions on Medical Imaging*, 18(11):1049 –1075, 1999.

[130] J. P. Lerch, J. Pruessner, A. P. Zijdenbos, D. L. Collins, S. J. Teipel, H. Hampel, and A. C. Evans. Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. *Neurobiology of Aging*, 29(1):23–30, 2008.

[131] H. Lester, S. Arridge, K. Jansons, L. Lemieux, J. Hajnal, and A. Oatridge. Non-linear registration with the variable viscosity fluid algorithm. In *International Conference on Information Processing in Medical Imaging (IPMI)*, volume 1613, pages 238–251. 1999.

[132] K. Leung, J. Barnes, M. Modat, G. Ridgway, J. Bartlett, N. Fox, and S. Ourselin. Automated brain extraction using Multi-Atlas Propagation and Segmentation (MAPS). In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 2053 –2056, 2011.

[133] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011.

[134] D. Liu, K. S. Zhou, D. Bernhardt, and D. Comaniciu. Search strategies for multiple landmark detection by submodular maximization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2831–2838, 2010.

[135] D. Loeckx, P. Slagmolen, F. Maes, D. Vandermeulen, and P. Suetens. Nonrigid image registration using conditional mutual information. *IEEE Transactions on Medical Imaging*, 29(1):19–29, 2010.

[136] X. Lu, B. Georgescu, A. Littmann, E. Mueller, and D. Comaniciu. Discriminative joint context for automatic landmark set detection from a single cardiac MR long axis slice. In *International Conference on Functional Imaging and Modeling of the Heart (FIMH)*, volume 5528, pages 457–465, 2009.

[137] M. Vaillant and M.I. Miller and L. Younes and A. Trouvé. Statistics on diffeomorphisms via tangent space representations. *NeuroImage*, 23, Supplement 1(0):S161 – S169, 2004.

[138] G. Maddala. *Limited Dependent and Qualitative Variables in Econometrics*. Econometric Society monographs in quantitative economics. CAMBRIDGE University Press, 1983.

[139] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[140] P. Markelj, D. Tomazevic, F. Pernus, and B. T. Likar. Robust gradient-based 3-D/2-D registration of CT and MR to X-ray images. *IEEE Transactions on Medical Imaging*, 27(12):1704–14, 2008.

[141] S. Marsland and C. J. Twining. Constructing diffeomorphic representations for the groupwise analysis of nonrigid registrations of medical images. *IEEE Transactions on Medical Imaging*, 23(8):1006–1020, 2004.

[142] B. Martinez, M. F. Valstar, X. Binefa, and M. Pantic. Local evidence aggregation for regression-based facial point detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1149–1163, 2013.

[143] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, 5(4):115–133, 1943.

[144] J. Meinguet. Multivariate interpolation at arbitrary points made simple. *Zeitschrift fr angewandte Mathematik und Physik ZAMP*, 30(2):292–304, 1979.

[145] M. I. Miller and L. Younes. Group actions, homeomorphisms, and matching: A general framework. *International Journal of Computer Vision*, 41:61–84, 2001.

[146] C. Misra, Y. Fan, and C. Davatzikos. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. *NeuroImage*, 44(4):1415–1422, 2009.

[147] J. Morris. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology*, 43(11):2412 – 2414, 1993.

[148] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging Clinics of North America*, 15(4):869–877, 2005. Alzheimer's disease: 100 years of progress.

[149] M. Muja. http://www.cs.ubc.ca/~mariusm/index.php/FLANN/FLANN.

[150] M. Muja and D. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Application (VISSAPP)*, pages 331–340, 2009.

[151] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *In Proceedings of Uncertainty in AI*, volume 9, pages 467–475, 1999.

[152] R. Nicolson, T. J. DeVito, C. N. Vidal, Y. Sui, K. M. Hayashi, D. J. Drost, P. C. Williamson, N. Rajakumar, A. W. Toga, and P. M. Thompson. Detection and mapping of hippocampal abnormalities in autism. *Psychiatry Research: Neuroimaging*, 148(1):11–21, 2006.

[153] L. G. Nyl and J. K. Udupa. On standardizing the mr image intensity scale. *Magnetic Resonance in Medicine*, 42(6):1072–1081, 1999.

[154] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proceedings of IEEE Int'l Conf. Multimedia and Expo (ICME'05)*, pages 317–321, July 2005.

[155] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Representation and Reasoning Series. Morgan Kaufmann, 1988.

[156] X. Pennec, P. Cachier, and N. Ayache. Understanding the demons algorithm: 3d non-rigid registration by gradient descent. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 1999.

[157] X. Pennec, C. Guttmann, and J.-P. Thirion. Feature-based registration of medical images: Estimation and validation of the pose accuracy. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 1496, pages 1107–1114, 1998.

[158] C. Peterfy and E. Schneider. The osteoarthritis initiative: report on the design rationale for the magnetic resosnace imaging protocol for the knee. *Osteoarthritis and Cartilage*, 16(12):1433–1441, 2008.

[159] J. Peters, O. Ecabert, C. Meyer, H. Schramm, R. Kneser, A. Groth, and J. Weese. Automatic whole heart segmentation in static magnetic resonance image volumes. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 4792, pages 402–410. 2007.

[160] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen. Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology*, 56(3):303–308, 1999.

[161] R. R. Picard and R. D. Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984.

[162] L. Pizarro, P. Mrázek, S. Didas, S. Grewenig, and J. Weickert. Generalised nonlocal image smoothing. *International Journal of Computer Vision*, 90(1):62–87, 2010.

[163] J. C. Platt. Fastmap, metricmap, and landmark mds are all nyström algorithms. In *International Workshop on Artificial Intelligence and Statistics*, pages 261–268, 2005.

[164] J. P. W. Pluim, J. Maintz, and M. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004, 2003.

[165] D. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2002.

[166] G. Postelnicu, L. Zollei, and B. Fischl. Combined volumetric and surface registration. *IEEE Transactions on Medical Imaging*, 28(4):508–522, 2009.

[167] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 3 edition, 2007.

[168] J. Purdy. 3D treatment planning and intensity-modulated radiation therapy. *Oncology*, 13:155–168, 1999.

[169] O. Querbes, F. Aubry, J. Pariente, J.-A. Lotterie, J.-F. Dmonet, V. Duret, M. Puel, I. Berry, J.-C. Fort, P. Celsis, and T. A. D. N. Initiative. Early diagnosis of Alzheimer's disease using cortical thickness: Impact of cognitive reserve. *Brain*, 132(8):2036–2047, 2009.

[170] S. Rajeev and C. Krishnamoorthy. Discrete optimization of structures using genetic algorithms. *Journal of Structural Engineering*, 118(5):1233–1250, 1992.

[171] I. Ramírez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3508, 2010.

[172] N. A. Ranginwala, L. S. Hynan, M. F. Weiner, and C. L. White, 3rd. Clinical criteria for the diagnosis of Alzheimer disease: Still good after all these years. *The American Journal of Geriatric Psychiatry*, 16:384–388, 2008.

[173] S. Ravishankar and Y. Bresler. Mr image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE Transactions on Medical Imaging*, 30(5):1028–1041, 2011.

[174] T. Rohlfing, D. Russakoff, and J. Maurer, C.R. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. *IEEE Transactions on Medical Imaging*, 23(8):983–994, 2004.

[175] R. Rohling, A. Gee, and L. Berman. Three-dimensional spatial compounding of ultrasound images. *Medical Image Analysis*, 1(3):177–93, 1997.

[176] K. Rohr. On 3D differential operators for detecting point landmarks. *Image and Vision Computing*, 15(3):219–233, 1997.

[177] K. Rohr, M. Fornefett, and H. Stiehl. Spline-based elastic image registration: integration of landmark errors and orientation attributes. *Computer Vision and Image Understanding*, 90(2):153 – 168, 2003.

[178] K. Rohr, H. Stiehl, R. Sprengel, T. Buzug, J. Weese, and M. Kuhn. Landmark-based elastic registration using approximating thin-plate splines. *Medical Imaging, IEEE Transactions on*, 20(6):526–534, 2001.

[179] K. Rohr and S. Worz. An extension of thin-plate splines for image registration with radial basis functions. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 442–445, 2012.

[180] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[181] D. Rueckert, A. F. Frangi, and J. A. Schnabel. Automatic construction of 3-D statistical deformation model of the brain using nonrigid registration. *IEEE Transactions on Medical Imaging*, 22(8):1014–1025, 2003.

[182] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid registration using free-form deformations: Application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[183] D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, editors. *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*. MIT Press, 1986.

[184] N. Ryan, C. Heneghan, and P. de Chazal. Registration of digital retinal images using landmark correspondence by expectation maximization. *Image and Vision Computing*, 22(11):883–898, 2004.

[185] B. Schoelkopf, A. J. Smola, and K.-R. Mueller. Kernel principal component analysis. *Lecture Notes in Computer Science*, 1327:583–591, 1997.

[186] D. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley, 1992.

[187] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th International Conference on Multimedia*, pages 357–360, 2007.

[188] T. W. Sederberg and S. R. Parry. Free-form deformation of solid geometric models. *SIGGRAPH Comput. Graph.*, 20(4):151–160, 1986.

[189] C. E. Shannon and W. Weaver. A mathematical theory of communication, 1948.

[190] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[191] A. Shekhovtsov, I. Kovtun, and V. Hlavac. Efficient MRF deformation model for non-rigid image matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, 2007.

[192] A. Simonsen, J. Mattila, K. Hejl, A.-M.and Frederiksen, S.-K. Herukka, M. Hallikainen, M. van Gils, J. Lötjönen, H. Soininen, and G. Waldemar. Application of the PredictAD software tool to predict progression in patients with mild cognitive impairment. *Dementia and Geriatric Cognitive Disorders*, 34:344–350, 2012.

[193] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17(1):87–97, Feb. 1998.

[194] R. W. So, T. W. Tang, and A. C. Chung. Non-rigid image registration of brain magnetic resonance images using graph-cuts. *Pattern Recognition*, 44(10-11):2450 – 2467, 2011.

[195] A. Sotiras, D. Christos, and N. Paragios. Deformable Medical Image Registration: A Survey. Technical Report RR-7919, INRIA, 2012.

[196] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36(2):111–147, 1974.

[197] C. Studholme. Simultaneous population based image alignment for template free spatial normalisation of brain anatomy. In *Biomedical Image Registration*, volume 2717, pages 81–90. 2003.

[198] C. Studholme, D. Hill, and D. Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern Recognition*, 32(1):71–86, 1999.

[199] C. Studholme, D. L. G. Hill, and D. J. Hawkes. Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures. *Medical Physics*, 24(1):25–35, 1997.

[200] M. Styner, J. A. Lieberman, D. Pantazis, and G. Gerig. Boundary and medial shape analysis of the hippocampus in schizophrenia. *Medical Image Analysis*, 8(3):197 – 203, 2004.

[201] H. Tang, E. Wu, Q. Ma, D. Gallagher, G. Perera, and T. Zhuang. {MRI} brain image segmentation by multi-resolution edge detection and region selection. *Computerized Medical Imaging and Graphics*, 24(6):349 – 357, 2000.

[202] T. Tang and A. Chung. Non-rigid image registration using graph-cuts. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 4791, pages 916–924. 2007.

[203] J. B. Tenenbaum, V. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[204] J.-P. Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998.

[205] P. Thompson and A. Toga. A surface-based technique for warping three-dimensional images of the brain. *IEEE Transactions on Medical Imaging*, 15(4):402–417, 1996.

[206] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[207] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed problems*. 1977.

[208] K. Toennies. *Guide to Medical Image Analysis: Methods and Algorithms*. Advances in Computer Vision and Pattern Recognition. Springer, 2012.

[209] W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17:401–419, 1952.

[210] J. Q. Trojanowski. Searching for the Biomarkers of Alzheimers. *Practical Neurology*, 3:30–34, 2004.

[211] M. Unser. Splines: a perfect fit for signal and image processing. *Signal Processing Magazine, IEEE*, 16(6):22–38, 1999.

[212] L. van der Maaten, E. Postma, and J. van den Herik. Dimensionality reduction: A comparative review. *Tilburg University Technical Report*, 2009.

[213] L. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. Technical Report TiCC-TR 2009-005, Tilburg University, 2009.

[214] J. Vandemeulebroucke, S. Rit, J. Kybic, P. Clarysse, and D. Sarrut. Spatio-temporal motion estimation for respiratory-correlated imaging of the lungs. *Medical Physics*, 38:166–178, 2011.

[215] V. Vapnik and A. Lerner. Pattern Recognition using Generalized Portrait Method. *Automation and Remote Control*, 24, 1963.

[216] P. Vemuri, J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S. Knopman, B. F. Boeve, R. C. Petersen, and J. Clifford R., Jr. Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies. *NeuroImage*, 39(3):1186–1197, 2008.

[217] L. Verard, P. Allain, J. Travere, J. Baron, and D. Bloyet. Fully automatic identification of AC and PC landmarks on brain MRI using scene analysis. *IEEE Transactions on Medical Imaging*, 16(5):610–616, 1997.

[218] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[219] P. Viola and W. I. Wells. Alignment by maximization of mutual information. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 16–23, 1995.

[220] P. A. Viola. *Alignment by maximization of mutual information*. PhD thesis, Massachusetts Institute of Technology, 1995.

[221] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[222] M. P. Wand and M. C. Jones. *Kernel Smoothing (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1 edition, 1994.

[223] C. Watson. Anatomic basis of amygdaloid and hippocampal volume measurement by magnetic resonance imaging. *Neurology*, 42(9):1743–1750, 1992.

[224] I. Wells, W. M., W. E. L. Grimson, R. Kikinis, and F. A. Jolesz. Adaptive segmentation of MRI data. *IEEE Transactions on Medical Imaging*, 15(4):429–442, 1996.

[225] G. Wenk. Neuropathologic changes in alzheimer's disease. *Journal of Clinical Psychiatry*, 64 Suppl 9, 2003.

[226] E. Westman, A. Simmons, J. S. Muehlboeck, P. Mecocci, B. Vellas, M. Tsolaki, I. Koszewska, H. Soininen, M. W. Weiner, S. Lovestone, C. Spenger, and L.-O. Wahlund. AddNeuroMed and ADNI: Similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America. *NeuroImage*, 58(3):818–828, 2011.

[227] F. Windmeijer. Goodness-of-fit measures in binary choice models. *Econometric Reviews*, 14(1):101–116., 1995.

[228] A. E. Wluka, S. Stuckey, J. Snaddon, and F. M. Cicuttini. The determinants of change in tibial cartilage volume in osteoarthritic knees. *Arthritis & Rheumatism*, 46(8):2065–2072, 2002.

[229] R. Wolz, P. Aljabar, J. V. Hajnal, A. Hammers, and D. Rueckert. LEAP: Learning Embeddings for Atlas Propagation. *NeuroImage*, 49(2):1316–1325, 2010.

[230] R. Wolz, P. Aljabar, J. V. Hajnal, J. Ltjnen, and D. Rueckert. Nonlinear dimensionality reduction combining MR imaging with non-imaging information. *Medical Image Analysis*, 16(4):819–830, 2012.

[231] R. Wolz, R. A. Heckemann, P. Aljabar, J. V. Hajnal, A. Hammers, J. Lötjönen, and D. Rueckert. Measurement of hippocampal atrophy using 4D graph-cut segmentation: Application to ADNI. *NeuroImage*, 52(1):109 – 118, 2010.

[232] R. Wolz, V. Julkunen, J. Koikkalainen, E. Niskanen, D. P. Zhang, D. Rueckert, H. Soininen, J. Ltjnen, and the Alzheimer's Disease Neuroimaging Initiative. Multimethod analysis of MRI images in early diagnostics of Alzheimer's disease. *PLoS ONE*, 6(10):e25446, 10 2011.

[233] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, 2009.

[234] Y.-T. Wu, T. Kanade, C.-C. Li, and J. Cohn. Image registration using wavelet-based motion model. *International Journal of Computer Vision*, 38(2):129–152, 2000.

[235] D. Zhang, D. Shen, and the Alzheimer's Disease Neuroimaging Initiative. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PLoS ONE*, 7(3):e33182, 2012.

[236] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.

[237] Y. Zheng, X. Lu, B. Georgescu, A. Littmann, E. Mueller, and D. Comaniciu. Robust object detection using marginal space learning and ranking-based multi-detector aggregation: Application to left ventricle detection in 2D MRI images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1343–1350, 2009.

[238] B. Zitov and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21:977–1000, 2003.

[239] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.