Atlas Generation for Segmentation of Head and Neck Computed Tomography Scans

Erin Birkwood

School of Computer Science

McGill University, Montréal

August 10, 2021

A thesis submitted to McGill University in partial fulfillment of
the requirements of the degree of Master of Science

Abstract

The field of atlas-based organ segmentation is multifaceted and has a long history. The selection, or creation, of the atlas is itself a rich area of research; the results of segmentation depend heavily on how representative this reference image is of the subjects to which it corresponds. The first goal of this thesis is to create a head and neck segmentation pipeline that uses an atlas constructed from representative images selected from our dataset of computed tomography (CT) scans, and evaluate segmentation accuracy with the Dice score. Our second goal is to explore whether anatomical subtypes can be automatically discovered through clustering, and subsequently used to create anatomy-specific atlases.

The initial atlas is created by the co-registration of four CT scans, of different subjects, extracted from our larger dataset. For the first objective, all images in our collection and their corresponding organ contours are registered to the atlas. Then, the STAPLE algorithm is used to create a consensus labelling from a training subset of images, which is then used to segment another subset of images. To address the second objective, k-Means clustering is

applied to the entire set of images, to automatically group them into anatomical subtypes. For each cluster, a representative atlas is created, through the co-registration of the three images that are closest to the cluster centre.

While the segmentation pipeline achieved varying results, its performance was strongly correlated with the similarity of the original images to the atlas. This highlighted the relevance of the clustering experiment, which successfully partitioned the images into four groups, based on anatomical similarity. The creation of the atlases for each anatomical subtype was not successful, suffering from poor registration between images. This was in part due to image distortion caused by metallic artifacts; thus, a larger dataset with artifact-free images may resolve this issue.

Résumé

Le domaine de la segmentation d'organes basée sur l'atlas est complexe et a une longue histoire. La sélection, ou la création, de l'atlas est en elle-même un riche domaine de recherche étant donné que les résultats de la segmentation dépendent fortement de la représentativité de l'image de référence envers les individus correspondants. L'objectif de cette thèse est, premièrement, de créer une méthode pour segmenter les images de tomodensitométrie (TDM) de la tête et du cou, en utilisant un atlas construit à partir d'images représentatives. Ensuite, la précision est évaluée avec le coefficient de Dice. Le deuxième objectif est d'explorer la possibilité d'améliorer les segmentations en les basant sur plusieurs atlas provenant de sous-types anatomiques automatiquement découverts grâce au regroupement 'k-Means'.

Pour le premier objectif, un atlas est produit avec le recalage de quatre tomodensitogrammes, chacun provenant d'un individu différent de nos base de données. Toutes les images de notre collection et leurs contours d'organes correspondants sont recalées à l'atlas. L'algorithme STAPLE est ensuite utilisé sur une portion des images réservée pour l'entraînement afin d'en dériver des étiquettes, qui sont par la suite utilisées

pour segmenter le reste. Le deuxième objectif vise à améliorer les résultats de la première en utilisant plusieurs atlas basés sur les regroupement k-Means des sous-types anatomiques. Pour chaque regroupement, un atlas représentatif est créé en se basant sur les trois images les plus proches du centre.

Alors que la première méthode de segmentation a obtenu des résultats variables, celles-ci étant fortement corrélées à la similitude des images originales avec l'atlas, le regroupement k-Means des images en fonction de la similitude anatomique a bien marché et a produit quatre groupes. Par contre, la création des atlas pour chacun des sous-types a échoué dû à un mauvais recalage entre les images. Cela était en partie dû à la distorsion des images causée par des artefacts métalliques. Un ensemble de données plus volumineux avec des images sans artefact devrait résoudre ce problème.

Acknowledgements

I would first like to thank my primary supervisor, Dr. Peter Savadjiev, who has guided me through every part of this thesis. I am grateful for his patience in answering my questions, and his detailed and insightful advice. Dr. Savadjiev has helped shape the way I think and express myself as a scientist, and for that I am truly appreciative.

I would like to express my gratitude to my co-supervisor, Dr. Reza Forghani, for his clinical expertise, and for welcoming me into his lab, where I was able to work on this multidisciplinary project. I would also like to thank the rest of the AIPHL lab, especially Farhad for his guidance, and Nikesh for providing me with the data used in this thesis.

On a personal note, I am thankful to my family for their encouragement to pursue graduate studies, and for providing me with the educational opportunities throughout my life which eventually led me to this degree. I am also grateful to all of my wonderful friends, in Montréal and elsewhere, for lifting my spirits many times during this process. Finally, I'd like to thank my partner, Alex, for always being there to provide a listening ear, a comforting word, or just some good food.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The accurate delineation of the organs in the head and neck is vital in the diagnosis, treatment, and observation of head and neck cancer patients. For example, many such patients undergo radiation therapy; a precise labelling of the affected region is needed to avoid damaging the organs surrounding the tumor area [1–3]. Manual organ segmentation is a time-consuming process subject to variability even amongst experts [1]. Thus, it is extremely advantageous to have automated segmentation tools for medical image analysis.

In recent years, image segmentation research has shifted toward deep learning approaches. While these methods have had many successes, they still suffer from certain limitations, such as the need for large amounts of data, which is often not available in medicine [4]. In the case of supervised deep learning methods, this vast quantity of data must also be labelled. In this thesis, we focus on more traditional atlas-based methods.

To understand atlas-based segmentation, it is first necessary to understand image registration. Image registration is the process of transforming one image into the space of another image such that the maximum number of corresponding points between the images are aligned [5]. For example, one may wish to register two images of a patient taken at different times to monitor disease progression.

In atlas-based segmentation, a novel image is registered to an atlas, which is a labelled anatomical reference template. The new image deforms to align with the atlas, and then the labels from the atlas may be transferred to the new image.

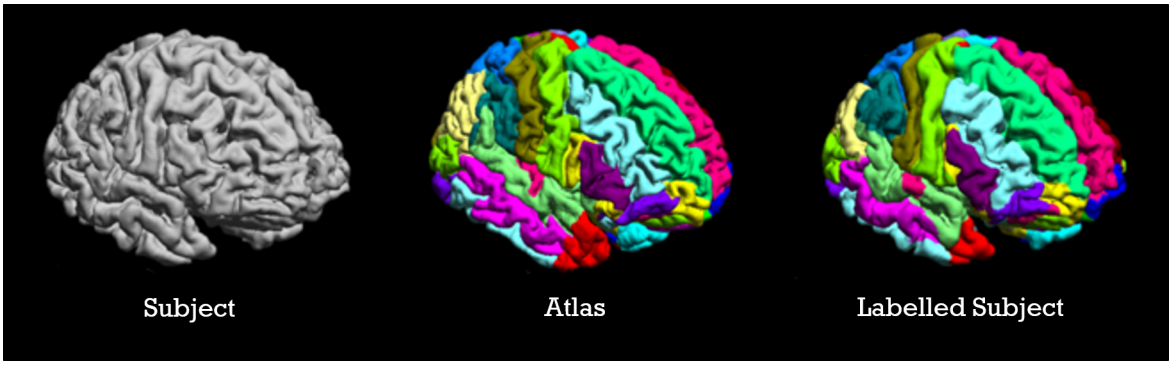


Figure 1.1: Example of atlas-based segmentation, adapted from BrainSuite [6]. Here, the colours represent different anatomical structures in the brain. After registration, the labels are transferred to the subject.

Originally, these methods used only a single atlas, which was either based on a single patient, or a generated "average" image [7, 8]. However, a single reference image is not sufficient to capture anatomical variability across a subject pool, and may lead to inaccurate, or significantly biased results. This led to the development of "multi-atlas" techniques, where

a novel image could be registered to, and segmented using, multiple reference templates. The results of these multiple segmentations can be consolidated in various ways, the most common of which involve a form of majority voting process [9].

Another technique to reduce bias and ensure that segmentation is accurate for a subgroup of people is to use a population-specific atlas. For example, Ridwan et al. [10] created a template for the older adult brain after noticing that most brain studies for this population were using standardized references, such the MNI-ICBM templates [11], which are not optimized for the older brain. The brain changes over time, so certain brain patterns common in older adults may not be present in references based on young or middle-aged adults.

Such atlases can be generated using techniques involving image registration. Several images are co-registered in order to obtain the ”average image” of the group. A common method of atlas creation is 'symmetric group-wise normalization' (SyGN), by Avants et al. [12], which covered in greater detail later in this thesis.

1.1 Motivation and Approach

This thesis is split into two parts. Both require an atlas, or reference template; this will be an ”average” population image constructed from four dataset images.

The first portion of the thesis produces an atlas-based segmentation pipeline. We generate labels for the previously created reference image with a technique inspired by

multi-atlas segmentation. This experiment uses images that have corresponding segmentations from radiologists. We first register all images to the reference template, transforming them into a common space. We then apply these registration transformations to the images' corresponding segmentations, putting those in the space of the reference template as well. The subjects are split into two groups, and then we use an algorithm known as STAPLE to merge each set of contours, creating one consensus segmentation per group. Finally, we segment each group of images by applying their inverse registration transformations to the opposing group's contour. We evaluate the results by comparing the generated contours to the originals.

The second part of the thesis explores the automated clustering of subjects into anatomical subtypes. This would allow the construction of group-specific templates, which we hypothesize should lead to more successful segmentation for members of that group. So, instead of aiming to create an atlas for a particular population, we use unsupervised learning techniques to automatically partition the subjects into anatomical groups. More specifically, this experiment proceeds as follows, using an unlabelled dataset. All images are registered to the previously created reference template. After performing some dimensionality reduction steps, clustering techniques are applied to the images; each group represents an anatomical subtype. Then, a population-specific template is created for each cluster. In future work, these group-specific templates can be used with the segmentation pipeline from the first experiment to obtain more accurate segmentations.

This research could have a few interesting applications. The combination of the segmentation pipeline and the population subgroup clustering is significant in one main way. Once the dataset images are partitioned into anatomical subgroups, one can obtain labels from experts for only the subjects closest to the cluster centres. These labels can then be used to generate the population-specific atlases, which can subsequently be used with our segmentation pipeline to label the remaining subjects. This idea is similar to the LEAP algorithm described by Wolz et al. in 2010 [13].

The clustering of anatomical subgroups could also be useful on its own. For other atlas-based segmentation methods, one could simply compute the similarity between each cluster centre image and a set of pre-existing atlases, using the most closely matched template to label all subjects. This technique may also be applicable to other segmentation pipelines, specifically, those using deep learning. In 2019, a French research group introduced AtlasNet [14], which outperformed other state-of-the-art deep learning frameworks for the segmentation of interstitial lung disease, by first registering its training images to several atlases that represented different anatomical types. In this case, the atlases were selected by radiologists, so it could be interesting to see how such pipelines perform when anatomical subtypes are determined in a data-driven manner, as they are here.

1.2 Thesis Outline

The remainder of this thesis is constructed as follows. Chapter 2 provides background about three relevant topics: image registration, image segmentation, and unsupervised learning methods. Chapter 3 describes the datasets used and outlines the two experiments. Chapter 4 presents the results of those experiments and discusses their significance. Chapter 5 then summarizes the thesis work and suggests future directions of research.

1.3 Contributions

The main contributions of this thesis are summarized here:

- The creation of an atlas-based head and neck segmentation pipeline using selected representative images from our head and neck CT dataset, with an assessment of segmentation accuracy
- An investigation of the automated clustering of head and neck CT scans into subgroups based on anatomical similarity
- The generation of anatomy-specific atlases for each cluster, which can subsequently be used with the segmentation pipeline in order to improve segmentation accuracy in each group

Chapter 2

Background

2.1 Image Registration

2.1.1 Introduction

The goal of image registration is to match points in one image to their corresponding positions in another image. The process of image registration aims to find the optimal transformation to align one image to another, such that the correspondence between homologous points is maximized. The image that undergoes this transformation is often referred to as the 'source', or 'moving', image; the image to which it is aligned is known as the 'target', or 'fixed', image [5]. Further details regarding the steps of registration are discussed in sections 2.1.2 - 2.1.4.

There are numerous applications of image registration, in industrial settings as well as

in the field of medical image analysis. In the industry, techniques often focus on matching geometrically-shaped elements between photos [5]. For example, one may use image registration to align photos of a city skyline taken at different positions, and then stitch these photos together to create a panorama.

Medical Applications

In medicine, however, the process is not quite as straightforward. Anatomical structures are not typically geometrically-shaped, and their location may differ between subjects. In fact, some structures may not even exist in certain people. Thus, more complex techniques are typically required to register medical images [5]. An example of medical image registration is provided in Figure 2.1.

Components of Registration Algorithms

There are three main elements of a registration algorithm: the transformation model, the similarity metric, and the optimization strategy [15,16]. The transformation model describes the method used to align the source image with the target. The similarity metric evaluates how well the transformed source matches the target, and the optimization strategy helps minimize or maximize the similarity metric. All of these components are described in more detail in the following sections.

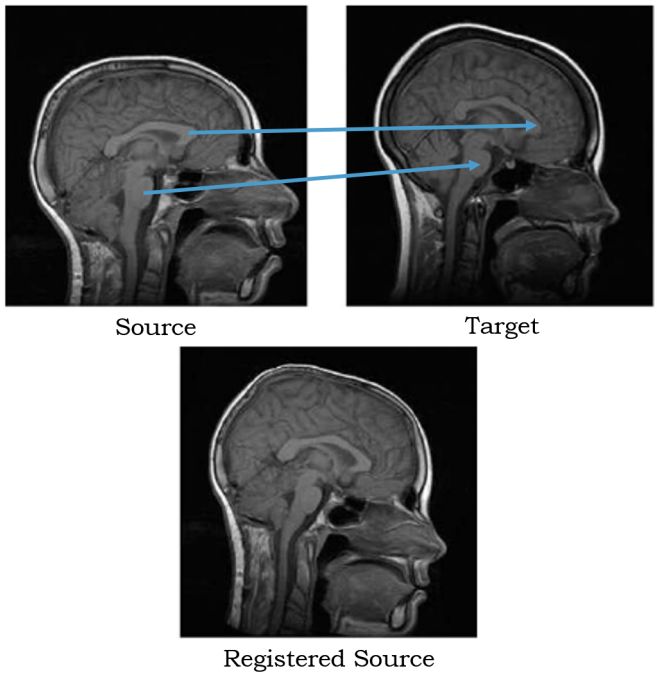


Figure 2.1: Example of medical image registration, adapted from [17]. Points in the source image are aligned with corresponding points in the target image.

2.1.2 Transformation Models

As mentioned above, the transformation model describes the spatial transformation used to map one image to another. There are two main classes of transformations in the context of image registration: linear and non-linear.

Linear Registration

The linear portion of registration includes rigid and affine transformations. Rigid transformations include only rotations and translations. In 3D, this amounts to six degrees of freedom. Affine transformations may include scaling and shearing in addition to rotation

and translation; they encompass 12 degrees of freedom in 3D [16]. A generic representation of a 3D rigid or affine transformation is provided below, where matrix M may represent rotation, scaling, or shearing, and matrix T represents a translation.

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = M_{3x3} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + T_{3x1} \tag{2.1}$$

In medicine, rigid and affine transformations may be performed as a form of preprocessing, or 'pre-registration', before a non-linear, deformable transformation. Non-linear transformations are necessary in this context, as there is not likely to be a uniform and linear relationship between all points in the source and target images [16].

Non-Linear Registration

Non-linear and deformable, or 'free-form', transformations describe a non-uniform mapping between images. When choosing or creating non-linear, deformable registration algorithms, the following four properties are commonly considered [18]:

- *inverse consistency*, which means that the forward and backward transformations must be consistent,
- *symmetry*, which guarantees that the resulting transformation will not be affected by the order of input images,

- *topology preservation*, which ensures that the mapping is one-to-one and continuous, with a continuous inverse, and
- *diffeomorphism*, which describes a property that is met when the transformation function is invertible, and both the forward-mapping function and its inverse are differentiable [18].

Different strategies for non-linear, deformable registration may exhibit some or all of these properties. However, if an algorithm is diffeomorphic, the properties of inverse consistency and topology preservation are already met, by the definition of diffeomorphisms [18,19]. The process of diffeomorphic registration is, in fact, a series of diffeomorphic transformations typically described as occurring over a time t, where $t \in [0, 1]$. If I is a source image being deformed to a target image J, then at $t = 0$, I has not undergone any warping. At $t = 1$, I has been fully warped to image J. If this process is stopped before its completion, we obtain a partially deformed source image [19, 20]. This is a notable property, because it is often interesting to calculate the 'average' of two or more images; the image obtained when I is 'halfway' registered to J, at $t = 0.5$, represents an average between these images. An average image is useful to assess anatomical variability within a population [21]. A visual representation of diffeomorphisms is presented in Figure 2.2.

Diffeomorphism, however, does not guarantee the property of symmetry. Symmetry is also important, because if the registration result is dependent on the order of input images, it is biased, and thus, less meaningful. Many registration algorithms attempt to achieve

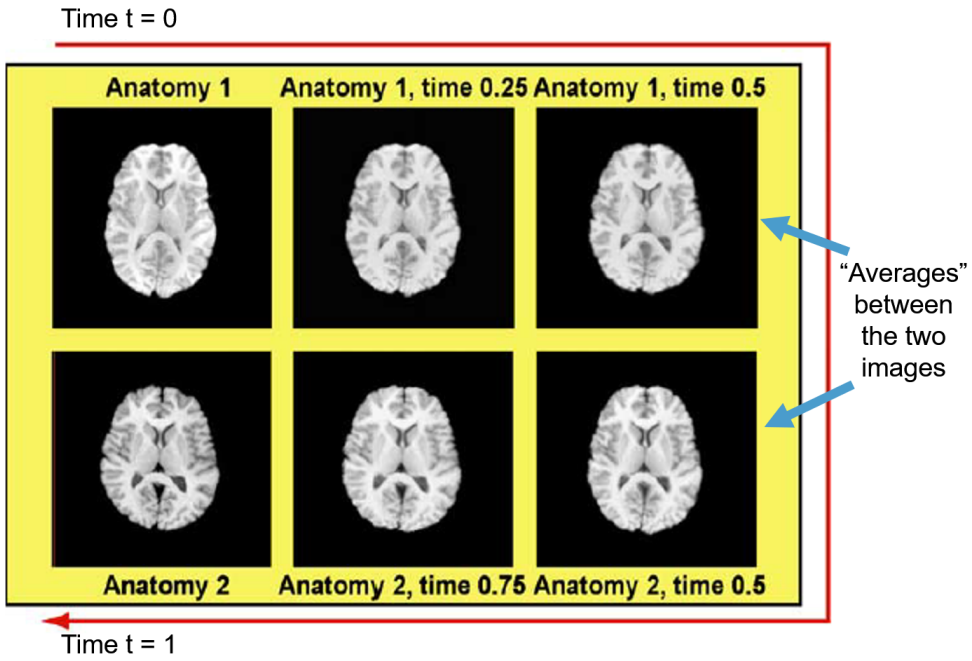


Figure 2.2: A visual representation of diffeomorphisms, adapted from [20].

symmetry by simultaneously estimating the forward and backward transformations, or applying penalties for asymmetry [19].

Below, we discuss the symmetric normalization method (SyN), which is both diffeomorphic and symmetric. While there are numerous strategies for deformable registration, SyN, despite being created over a decade ago, is considered to be state-of-the-art. It remains widely used by researchers, who note that it is still the top performing algorithm of its kind [22, 23]. However, we acknowledge that many similar algorithms exist, including those evaluated in this survey: [24].

Symmetric Normalization (SyN) Method

Here, we discuss the symmetric normalization method (SyN), described in a 2008 paper by Avants et al. [19]. It was developed as an extension to the authors' previous work on a Langrangian diffeomorphic deformable registration method; more details about this work can be found here: [25]. Below, we examine how SyN exploits the properties of diffeomorphisms in order to achieve symmetry.

Consider a diffeomorphism, ϕ. The result of registration for a source image I to some target image can be described as ϕI:

$$\phi I = I(\phi(x, t = 1)) \tag{2.2}$$

where x represents a spatial coordinate in I, and the diffeomorphism is complete at $t = 1$, as described in the previous section [19].

We note that the diffeomorphic transformation ϕ can be split into two parts, ϕ_1 and ϕ_2, where ϕ_1 represents the forward transformation, and ϕ_2 represents the backward transformation. Suppose that $t \in [0, 1]$ and indexes both ϕ_1 and ϕ_2, but in opposing directions. Then, for an image I being registered to another image, J, where y represents a spatial coordinate in J [19]:

$$\phi_1(x,1)I = J \tag{2.3}$$

$$\phi_2^{-1}(\phi_1(x,t),1-t)I = J \tag{2.4}$$

$$\phi_2(\phi_2^{-1}(\phi_1(x,t),1-t),1-t)I = \phi_2(y,1-t)J \tag{2.5}$$

$$\phi_1(x,t)I = \phi_2(y,1-t)J \tag{2.6}$$

Previously, to evaluate the similarity between a transformed source image I and target image J, we would use:

$$|\phi_1(x,1)I - J| \tag{2.7}$$

but now we can instead use:

$$|\phi_1(x,t)I - \phi_2(y,1-t)J| \tag{2.8}$$

which means that the optimization problem can now be solved from both sides, toward the middle, at $t = 0.5$ [19]. This is illustrated in Figure 2.3.

Since I and J are interchangeable, the problem is symmetric by nature, and the result is guaranteed to be symmetrical, regardless of similarity metric used; details about similarity metrics are provided in section 2.1.3. Additionally, in order to avoid any interpolation errors, SyN also includes invertibility constraints in its optimization [19].

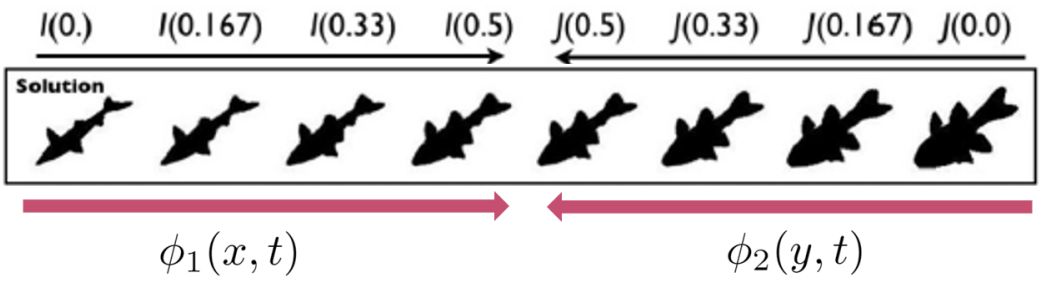


Figure 2.3: An illustration of SyN, adapted from [19].

In the previously mentioned 2009 survey of deformable registration techniques [24], SyN outperformed all other methods. Today, it is still considered to be one of the most advanced registration algorithms, and is widely used [22, 23].

2.1.3 Similarity Metrics

During registration, some measure is required in order to determine which transformation is optimal. The chosen metric will calculate the similarity between the target image and the transformed source image in order to find the best transformation. Depending on the strategy chosen, it will be at its maximum or its minimum when the desired transformation is found.

There are two main categories of similarity metrics; feature-based and intensity-based. Feature-based methods assess similarity based on the placement of particular structures in the images. These structures could be organs, or, they could be predefined patches of the image. Similarity between the images is determined based on the distances between corresponding structures; Euclidean distance may be used for this purpose [16]. A related area is feature-based morphometry, which aims to discover group-specific anatomical

patterns by first detecting scale-invariant features, and then describing them through probabilistic modelling [26]. These detected features can then be incorporated into feature-based similarity measures.

Intensity-based methods, however, only consider the intensities of the voxels between images. Such methods are more commonly seen in medical image registration, and three of the most frequently used similarity functions are described below [16].

Sum of Squared Differences (SSD)

Suppose that the target image is represented by X, the source image is represented by Y, and $T(Y)$ denotes the transformed version of Y that is registered to X. The idea of the sum of squared differences (SSD) approach is that after registration, voxels at the same location in the overlapping domain, Ω, of X and $T(Y)$, should have similar intensities, because they should correspond to the same anatomical structures. SSD can be calculated as follows, where x_i and y_i denote corresponding voxels within X and $T(Y)$, $i \in \Omega$, and N is the size of Ω:

$$SSD = \frac{1}{N} \sum_{i \in \Omega} (x_i - y_i)^2 \tag{2.9}$$

Thus, the lower the sum of squared differences value is, the better the registration has worked; so, the optimal transformation occurs when SSD is minimized [16].

Cross-Correlation (CC)

The cross-correlation (CC) method is motivated by the notion that a linear relationship exists between the intensities of corresponding structures in two images. Suppose again that the target image is represented by X, the transformed source image is represented by $T(Y)$, and that x_i and y_i are corresponding voxels where $i \in \Omega$, the overlapping domain of X and $T(Y)$. Here, $\bar{x}$ and $\bar{y}$ represent the mean voxel values of X and $T(Y)$ in Ω. Then:

$$CC = \frac{\sum_{i\in\Omega}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i\in\Omega}(x_i - \bar{x})^2 \sum_{i\in\Omega}(y_i - \bar{y})^2} \tag{2.10}$$

The more similar the intensities between the two images, the higher the CC value will be; so, the optimal registration transformation is the one which maximizes CC [16].

Mutual Information (MI)

For pairs of images that have been obtained from the same form of medical imaging, also known as 'monomodal' images, SSD and CC are successful because their intensities should be similar, and can be directly compared. However, it is often interesting to register multimodal images, for example, a computed tomography (CT) scan with a magnetic resonance image (MRI) scan from the same patient. In such cases, differing intensity value patterns are present due the different scanning techniques, so methods such as SSD and CC will not be meaningful. However, there should still exist some relationship between the intensities of the two images; this motivates the mutual information metric.

Mutual information (MI), at its core, measures how well one image explains another image. MI uses the concept of entropy, or 'randomness', calculated as follows, where $H(X)$ is the entropy of an image X, and $p(X = i)$ is the probability that a voxel in image X will be equal to i, for all possible values i:

$$H(X) = -\sum_i p(X = i) \log(p(X = i)) \tag{2.11}$$

Then, we can similarly calculate the joint entropy of two images, X and Y, as follows, where j represents all possible values for the voxels in Y:

$$H(X, Y) = -\sum_i \sum_j p(X = i, Y = j) \log(p(X = i, Y = j)) \tag{2.12}$$

Finally, the mutual information metric, MI, is calculated for images X, Y as follows:

$$MI(X, Y) = H(X) + H(Y) - H(X, Y) \tag{2.13}$$

The optimal registration transformation is one that maximizes MI. This is logical because the joint entropy term, $H(X, Y)$ will be minimized when a particular intensity value, a, in X always corresponds to the same intensity value, b, in Y (a need not equal b). Figure 2.4 explains the intuition.

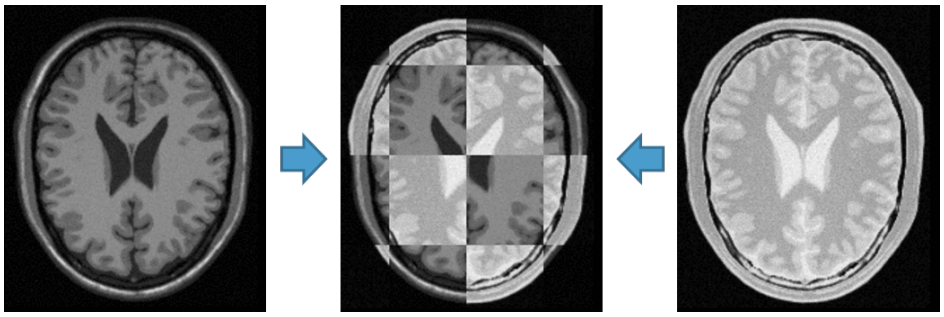


Figure 2.4: Example of mutual information, adapted from [27]. The left and right images are optimally registered when they align as seen in the middle image.

2.1.4 Optimization Strategy

In order to find the transformation which minimizes or maximizes the given similarity measure, an optimization strategy is required. Many such optimization methods take an iterative, coarse-to-fine approach. This means that for a specified number of iterations, a lower resolution version of the source and target images are registered. The resulting transformation is used as the initial transformation for the next step, where higher resolution versions of the images are registered. This proceeds until some convergence criteria is met, or until the maximum number of iterations at each step have been performed [16].

At each iteration, the function parameters must be adjusted in a particular way so that they may proceed toward an optimum. Again, there are various ways to do this, such as those examined in this survey: [28]. Here, we focus on a technique commonly used in registration as well as machine learning, gradient descent (or ascent).

Gradient Descent and Ascent

The goal of gradient descent or ascent is to find a local optimum for a function. In gradient descent, the optimum is a minimum; in gradient ascent, it is a maximum. In either case, the gradient techniques work by adjusting the parameters according to a "learning rate", typically specified by the user. The larger the learning rate, the more quickly the algorithm may proceed toward its optimum; however, if it is too large, it may hover around the target value and never reach it. Conversely, if the learning rate is too small, it could be a very long time before the optimum is found [29].

Consider an example using gradient descent with the sum of squared differences (SSD) function seen in Equation 2.9. We recall that x_i and y_i are values at a corresponding voxel i for target image X and registered source image $T(Y)$. We rewrite the equation as:

$$SSD = \frac{1}{N}\sum_{i\in\Omega}(X(i) - T(Y(i)))^2 \tag{2.14}$$

Let each transformation T have a corresponding set of parameters, Θ. Then, $SSD(\Theta)$ is the SSD corresponding to the transformation of Y with the parameters Θ.

$$SSD(\Theta) = \frac{1}{N}\sum_{i\in\Omega}(X(i) - T(Y(i);\Theta))^2 \tag{2.15}$$

Then, we can calculate the gradient of this function as $\nabla SSD(\Theta)$. If Θ_t represents the parameters at time t, then let the parameters at time $t+1$ be Θ_{t+1}. Gradient descent

calculates Θ_{t+1} in the direction of the negative gradient, as follows:

$$\Theta_{t+1} = \Theta_t - \tau \nabla SSD(\Theta_t) \tag{2.16}$$

Here, τ represents the learning rate discussed earlier. The value of SSD with the new parameters is evaluated, and the process of updating the parameters and evaluating the function continues until it is no longer possible to minimize the SSD.

2.1.5 Template Creation

As an extension to simple pairwise registration, it may be desirable to coregister all images from a particular population in order to create a representative, 'average' image [12]. This technique can be used to generate a reference template that will later be used for atlas-based segmentation, which is described in section 2.2.2. The main template that we created for this thesis appears in Figure 3.3.

The most commonly used method to create such a template is an extension to the SyN method discussed in section 2.1.2. It is known as symmetric group-wise normalization (SyGN), also proposed by Avants et al. [12].

Suppose that we wish to find the optimal template, I^*, for a set of N images, $\{J^i\}$, where i denotes the ith image. Essentially, we are searching for the smallest "parameterization" of the dataset, meaning that the similarity metric for each pairwise registration is optimized, and the length of each path of diffeomorphisms is minimized [12].

Let E be a function representing these properties, to be minimized. Let E_s represent a pairwise registration problem to be solved with SyN. Let ϕ^i represent the path of diffeomorphisms for each image, J^i to register to a template, I. The template begins as an average of all input images, and is updated at each iteration of the SyGN algorithm. The template shape has its own corresponding diffeomorphism, and is represented by ψ. Then [12]:

$$E(I) = \sum_i E_s(I, J^i, \phi^i), \ where \ \forall i, \ \phi^i(x, 0) = \psi(x) \tag{2.17}$$

Similarly to in section 2.1.3 where SyN was discussed, x is a spatial coordinate, and ϕ^i is provided as input x, as well as a time t.

SyGN optimizes each pairwise registration where each image J^i is initialized with the deformation ψ. ψ begins as the identity transformation, but can be updated after $E(I)$ is optimized. Mathematical details regarding the calculation of a new ψ are found here: [12]. The steps are then repeated with a different ψ. Figure 2.5 explains the intuition behind the process of SyGN.

2.1.6 Jacobian Determinant

While not explicitly tied to image registration, we use a "Jacobian determinant" technique on the 3D vector fields that arise from our non-linear, deformable registration in order to calculate volumetric changes at each voxel, so we provide background about our Jacobian-

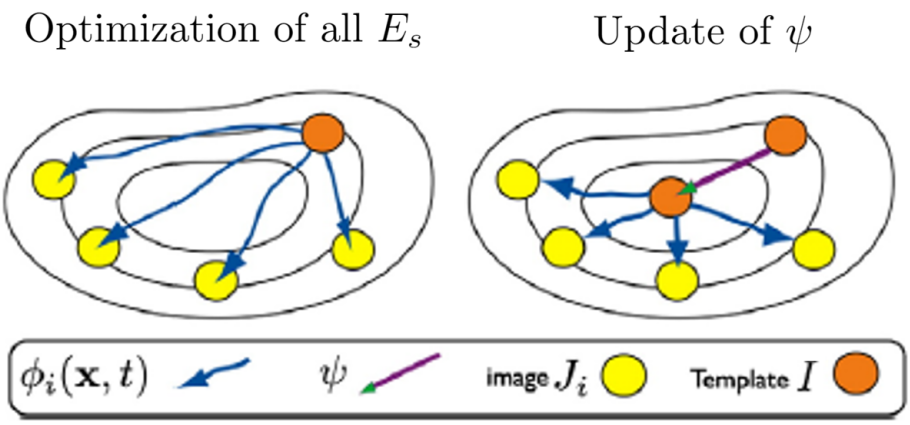


Figure 2.5: A visualization of the SyGN process, adapted from [12].

related work here.

The goal of this technique is to obtain a single value at each voxel describing the magnitude of image deformation that occurred at this particular location during registration to the template. This is a common step in deformation-based morphometry, an analysis technique which is often used to measure differences in brain regions over time, or between patients [30–32]. This is what inspired the use of this method in our work.

Suppose that $D(x,t) = (D_1, D_2, D_3)$ is a 3D vector representing the deformation, or displacement, of an image voxel, $x = (x_1, x_2, x_3)$, at time t. So, this means that $x + D(x,t)$ represents the spatial location of voxel x after being deformed at time t. The local change in volume around voxel x at time t can then be represented by the Jacobian, J, which is the gradient of $D(x,t)$, calculated as follows [32]:

$$\frac{\partial D}{\partial x}(x,t) = \begin{pmatrix} \frac{\partial D_1}{\partial x_1} & \frac{\partial D_1}{\partial x_2} & \frac{\partial D_1}{\partial x_3} \\ \frac{\partial D_2}{\partial x_1} & \frac{\partial D_2}{\partial x_2} & \frac{\partial D_2}{\partial x_3} \\ \frac{\partial D_3}{\partial x_1} & \frac{\partial D_3}{\partial x_2} & \frac{\partial D_3}{\partial x_3} \end{pmatrix} \tag{2.18}$$

For simplicity, we can represent the Jacobian as $J(x,t)$ and rewrite the matrix as such:

$$J(x,t) = \begin{pmatrix} j_{11} & j_{12} & j_{13} \\ j_{21} & j_{22} & j_{23} \\ j_{31} & j_{32} & j_{33} \end{pmatrix} \tag{2.19}$$

Then, we can find the Jacobian determinant, $|J|$, which is a single value representing the local volume change at x. This can be calculated as:

$$|J| = j_{11}(j_{22}j_{33} - j_{23}j_{32}) - j_{21}(j_{12}j_{33} - j_{13}j_{32}) + j_{31}(j_{12}j_{23} - j_{13}j_{22}) \tag{2.20}$$

This will output a single value representing how the voxel has deformed. As seen in Figure 2.6 below, the voxels can be coloured according to the Jacobian determinant value to represent their volumetric changes.

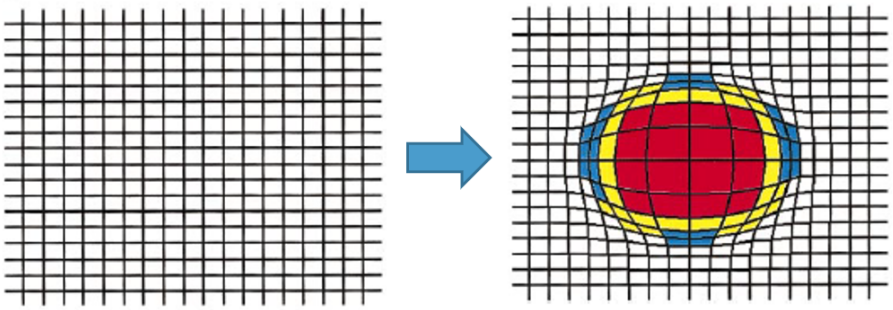


Figure 2.6: Example of Jacobian determinant for deformation field, adapted from [32]. Here, red represents volumetric increase, blue represents volumetric decrease, and yellow represents translation.

2.1.7 Challenges in Registration of Head and Neck Images

One significant challenge affecting medical image analysis is the issue of metallic artifacts. These are distortions in images that occur due to metallic objects in subjects, as demonstrated in Figure 2.7 [33]. In the head and neck area, which is featured in this thesis, items such as dentures, metal fillings, or braces are often present. Since the advent of modern computed tomography (CT) scanners over 40 years ago, methods of removing these artifacts has been a topic of research and development. Their existence not only hinders radiologists from properly assessing an image, but also impacts any subsequent computational analysis [34]. In this thesis, we explain our efforts to manage these artifacts, as well as their potential impacts on our results.

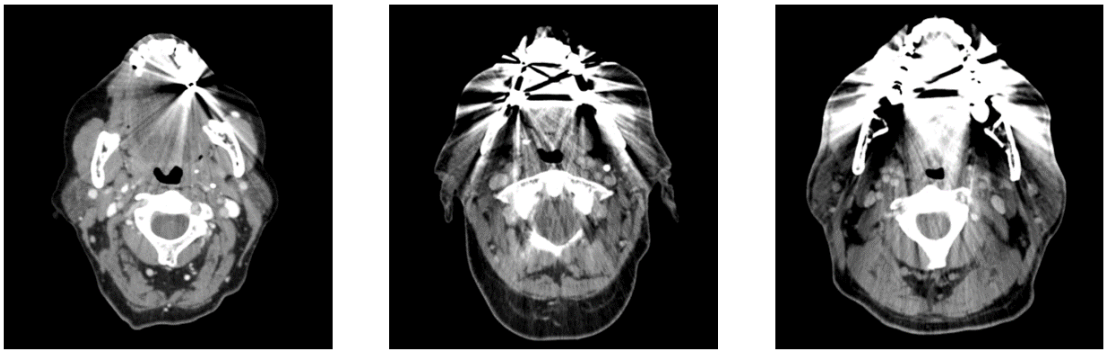


Figure 2.7: Three CT scans affected by metallic artifacts (axial view).

Another difficulty in the registration of head and neck images is anatomical variability. As demonstrated in Figure 2.8, there is a wide range of possible appearances for this area. There is not only a difference in the size and location of certain anatomy, but there are also often differences in the positions of patients. For example, in the third image, the patient's neck appears to have more significant curvature, and they are looking much more upward than the others. All of these factors make registration, the alignment of common features between images, significantly harder.

Figure 2.8: Examples of anatomical variation in head and neck CT scans (sagittal view).

2.2 Image Segmentation

2.2.1 Introduction

In a general sense, the goal of image segmentation is to divide an image into its constituent parts. How this division is performed and how many objects need to be identified depends on the problem under consideration [35]. Segmentation can, of course, be performed manually by an expert; however, this approach is too time-consuming and expensive for many applications [36]. Thus, having computational techniques that can annotate images accurately is extremely beneficial.

In medicine, image segmentation focuses on the automated identification of particular anatomical structures, or abnormalities, such as tumors [9]. It is important to note that the field of medical image segmentation has a rich history beyond the scope of this thesis; here, we discuss some key concepts that relate to our experiments.

2.2.2 Atlas-Based Approaches

Atlas-based approaches involve the registration of a novel image to an 'atlas', or reference image. Using the inverse of the registration transformation, the labels from the atlas can be transferred to the new image.

Originally, an atlas was a single image, labelled by an expert. The image was either a scan of one subject, or an "average" image created from several subjects. However, this

single-atlas approach is not sufficient to capture the potential anatomical variability across subjects [9]. This issue led to the development of 'multi-atlas' techniques, which aim to label a novel image by drawing information from several atlases.

The process of consolidating labels from multiple atlases is known as 'label fusion', and is a vast area of research. The simplest method, 'best atlas selection', calculates the similarity of a novel image to each of the templates in the group, and uses only the most similar template to label the image. However, this ignores potentially useful information from the other atlases. Another approach, majority voting, allows each atlas to 'vote' on which segment each voxel in the target image belongs to. A common extension of majority voting is weighted majority voting, where the labels of certain atlases carry more weight. For example, the algorithm described by Sabuncu et al. in 2010 [37] automatically assigns more weight to atlases that are most similar to the image being labelled. Another category of algorithms based on STAPLE, described in detail in section 2.2.4, utilize weighted voting in a different way. Here, more weight is given to atlases deemed to be reliable; atlas reliability is itself determined by STAPLE [38].

There has been a considerable amount of previous work on head and neck atlas-based segmentation pipelines. All research mentioned here uses the Dice score metric to evaluate segmentation accuracy; this metric is described in more detail in the following section. In 2008, Han et al. [39] evaluated both single and multi-atlas methods of segmenting the head and neck. They found that the multi-atlas pipeline outperformed the single-atlas method

for most of the structures they examined. The partoid glands, which we are interested in, had a corresponding average Dice score of 0.80 using an optimal single atlas, and 0.83 using multiple atlases; the authors found this difference to be statistically significant. A 2012 MIT thesis [40] tested a multi-atlas approach with weighted voting, and achieved a mean Dice score of roughly 0.77 for each parotid gland. In 2014, Fritscher et al. [36] combined multi-atlas techniques with geodesic active contours and statistical appearance models, and obtained a mean Dice score of 0.84 for the left parotid gland, and 0.81 for the right parotid gland.

2.2.3 Evaluation Metrics

To evaluate a generated segmentation against the ground truth, a similarity measure is required. There are various metrics that can be applied, but among the most popular are the Dice score and the Jaccard index. For this thesis, we chose the Dice score due to its widespread use and simplicity [41].

To calculate the Dice score, the number of overlapping pixels from the generated segmentation, G, and the ground truth, T are multiplied by two, and this result is divided by the total number of pixels of both contours; in other words, the score computes twice the intersection over the union of the segmentations [42].

$$Dice(G,T) = \frac{2|G \cap T|}{|G| + |T|} \tag{2.21}$$

2.2.4 Simultaneous Truth and Performance Level Estimation (STAPLE)

While the evaluation of a segmentation against a ground truth can be performed using the Dice score or Jaccard index, it is less clear how to measure the quality of a segmentation when no reliable ground truth exists. While the most reputable sources of reference segmentations are domain experts, there is often significant inter-observer variability. If human experts cannot agree on a correct contouring, it is difficult to establish how well an automated algorithm has performed.

This issue motivated the development of the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm in 2004, by Warfield et al. [43] STAPLE takes as input several segmentations, generated by humans or automatically, and determines a ground truth contouring using a probabilistic model. At the same time, it produces an accuracy score for each of the individual segmentations. The algorithm establishes the hidden ground truth segmentation as well as the performance accuracy of each annotator by determining the maximum likelihood scenario [43].

Suppose that for an image with N voxels, there are M sets of segmentations. Then, let $\mathbf{p} = \{p_1...p_M\}^T$ represent the sensitivity and $\mathbf{q} = \{q_1...q_M\}^T$ represent the specificity for each of the M segmentations. The sensitivity is the proportion of voxels that are correctly identified as part of the structure being segmented ("true positive rate"), and the specificity is the proportion of voxels that are correctly identified as not being part of the segmented

structure ("true negative rate"). Finally, let S be a NxM matrix representing all input binary segmentation decisions, and T be an N-dimensional vector representing the ground truth segmentation. Then, the segmentation sensitivity and specificity values may be estimated as $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ for the $(\mathbf{p}, \mathbf{q})$ which maximize the probability mass function, $f(\mathbf{S}, \mathbf{T}|\mathbf{p}, \mathbf{q})$:

$$(\hat{\mathbf{p}},\hat{\mathbf{q}}) = \underset{p,q}{\operatorname{argmax}} \ln f(\mathbf{S}, \mathbf{T}|\mathbf{p}, \mathbf{q}) \tag{2.22}$$

After the sensitivity and specificity for each segmentation are estimated, the ground truth contouring can be found by assigning greater weight to more reliable segmentations. STAPLE can also be altered to include a priori information, such as a statisical anatomical atlas [43].

2.2.5 Deep Learning Methods

One cannot discuss modern medical image segmentation methods without touching on deep learning. While the details of deep learning segmentation methods are beyond the scope of this thesis, we cover the basic concepts as well as some of the drawbacks of these approaches.

To train such models to perform segmentations, labelled medical image data is fed into a deep neural network. These networks consist of several layers of interconnected nodes that 'learn' the properties of the training dataset. In this case, they learn which image areas correspond to the target structure, and which do not. After the training phase is complete,

a new image can be segmented using the model [4]. The first deep learning model to achieve impressive segmentation results was the Fully Convolutional Network (FCN) [44], however, the results were not sensitive enough to detail for medical images, in part due to the lack of availability of labelled pixel data for training. Ronneberger et al. then proposed the U-Net [45], which is trained on image patches, rather than full images. This approach uses the training data more effectively, and allows for much more precise segmentation of anatomical structures. Another approach trains on bounding boxes that encompass structures, rather than using pixel-level annotations. This method leads to decreased segmentation accuracy, but greatly reduces the cost of obtaining training data, as it is considerably easier to draw a box which roughly encompasses a structure than to precisely label each pixel along the structure border [46].

While deep learning methods have had numerous successes as mentioned above and in [4], they also suffer from certain limitations. Deep learning models require a large amount of data to be trained, and datasets that are sufficient for this purpose are rare in medicine. This means that networks trained on medical images may be subject to overfitting, where they are extremely sensitive to the images already in the dataset, but cannot successfully segment new images. While this issue can be managed with data augmentation, this is obviously not preferable to having enough unaltered images to use for training [4].

Another concern about relying on deep networks for segmentation relates to the potential for a 'combinatorial explosion'. This describes the idea that as time goes on, the problems

we wish to solve will become increasingly complex. As deep learning methods are primarily data-driven, this means that increasingly complex datasets will also be required. However, such datasets may not be feasible to obtain [47].

Though it is of course worthwhile to pursue research in deep learning in medicine, it is important to not to abandon more traditional image analysis methods that rely more heavily on prior knowledge. In fact, combining these two approaches could help resolve certain deep learning limitations, such as the lack of available data [21].

2.3 Unsupervised Learning Methods

2.3.1 Introduction

In this thesis, we are interested in identifying anatomical subtypes within our dataset through the grouping of the deformations that arise from the registration of our images to a common template. To achieve this, we require automated methods that can partition the deformation fields based on similarity, without any prior information.

The idea of utilizing population subgroups is touched upon in a 2019 publication by Vakalopoulou et al. [14], which presents their deep learning segmentation framework, AtlasNet, tested on interstitial lung disease. They registered all of their initial images to six pre-selected atlases representing different anatomies, and trained one network per atlas group. The labels from each network were then merged into a consensus segmentation.

AtlasNet achieved higher accuracy in the segmentation of interstitial lung disease compared to other state-of-the-art techniques, such as SegNet, even with a smaller training dataset [14].

In the case of AtlasNet, the representative atlases were selected by radiologists. This serves as partial motivation for our research into the automated detection of population subgroups; it would not only be useful in combination with our segmentation pipeline, but could be relevant for other applications as well.

The field of unsupervised learning aims to find patterns in unlabelled data, and use this information to construct a representative model. When this model receives new data, it should be able to make decisions or predictions that are consistent with the properties of the original dataset. There are many applications of unsupervised learning algorithms, for example, anomaly detection, clustering, and dimensionality reduction [48]. Here, we focus on the last two areas.

2.3.2 Dimensionality Reduction

Independent Components Analysis (ICA)

Independent Components Analysis (ICA) is a variant of 'blind source separation' algorithm which finds the user-specified number of independent vectors in a set of data with many components [49]. In simpler terms, the idea of ICA is that a set of data was generated by the mixing of several independent factors, which we would like to recover.

This technique is commonly used to reduce the dimensions of a dataset prior to the application of other machine learning techniques, to reduce computational complexity [50,51]. It has also previously been used to separate components of functional magnetic resonance imaging (fMRI) images, which motivated our use of ICA for CT scan data [49, 52]. In this thesis, we use ICA to help cluster deformation patterns in images to detect anatomical subtypes in an unsupervised manner, as detailed in Chapter 3.

ICA works by assuming that an M-dimensional collection of data, $x = [x_1...x_M]^T$, was generated by an N-dimensional vector of independent components, $s = [s_1...s_N]^T$, combined with some mixing matrix A, of dimension MxN [52]:

$$\mathbf{x} = \mathbf{As} \tag{2.23}$$

The goal of ICA is to estimate an unmixing matrix, W, of dimension NxM, such that:

$$\mathbf{y} = \mathbf{Wx} \tag{2.24}$$

where y is a vector which estimates the true independent components, s [52].

t-Distributed Stochastic Neighbor Embedding (t-SNE)

In this thesis, ICA is used to reduce the dimensionality of our data before we perform our key operations, such as clustering, which is covered in the next section. However, we

still require a method of visualizing results for data in higher dimensions.

For this, we use t-Distributed Stochastic Neighbor Embedding (t-SNE). Essentially, t-SNE maps higher-dimensional data into two or three dimensions based on the distances between the datapoints in their original space. While the specifics of this algorithm are beyond the scope of this thesis, details are available in this paper: [53].

2.3.3 Clustering

While there are many types of clustering, some of the most commonly seen categories of clustering algorithms include hierarchical, distribution-based, and partitional [54].

In hierarchical clustering, each data point normally begins in its own cluster. Then, neighbouring clusters are progressively combined until all points are part of the same cluster. This algorithm may also work in the opposite way, where all points begin in one cluster and are then progressively divided into smaller groups based on their characteristics. Examples of these algorithms include BIRCH and CURE [54].

In distribution-based clustering, points are grouped together based on the distribution from which they were generated. The parameters of the distributions are not known in advance, but are discovered by using an expectation-maximization (EM) strategy. Here, the expectation (E) step calculates the probability that each point belongs to each cluster, and the maximization (M) step updates the relevant parameters of the model; this may include values such as mean, variance, and density. This is repeated until the parameters

that maximize likelihood are found. The most well known example of this is the Gaussian mixture model (GMM) [54, 55].

In partitional clustering, each data point may only belong to one cluster. The group to which it belongs is determined based on the distance from the point to the cluster centres; it will join the group whose centre point, or 'centroid', is the closest. Partitional clustering is perhaps the most frequently used type of clustering, and the most popular algorithm to perform this is k-Means [54]. k-Means is, in fact, a special case of Gaussian mixture model, where the Gaussian functions are spherical, and points may only belong to one group. We employ k-Means in our own approach, so we describe it in more detail below.

2.3.4 K-Means Algorithm

In k-means, there are two main steps performed at each iteration of the algorithm: defining the centroids of each cluster, and assigning the remaining points to the appropriate cluster. These can be thought of as the expectation (E), and maximization (M) steps, respectively. In the context of this thesis, the cluster centres will represent an anatomical subtype, and other subjects will join the cluster of the nearest centroid to them.

Initially, the cluster centres may be selected randomly. The number of clusters is defined by the user. Next, the distances between the other points and each of the centroids is determined. For a 2D dataset, this may be calculated as follows, where d is the distance of some point $p(x, y)$ from a cluster centroid C_k [56]:

$$d = ||p(x, y) - C_k|| \tag{2.25}$$

Each of the non-centroids is then 'attached' to the nearest cluster. Then, for all k clusters, a new cluster centre is then calculated according to the new group. If $\{C_k\}$ represents the set of points associated with centroid C_k, then [56]:

$$C_k = \frac{1}{|\{C_k\}|} \sum_{y \in \{C_k\}} \sum_{x \in \{C_k\}} p(x, y) \tag{2.26}$$

These steps are then repeated until the cluster centres no longer change, or a maximum number of iterations is reached [56].

Determination of Cluster Number

As the number of clusters, k, is determined by the user, the question of how to determine the optimal number of clusters is raised. While other techniques, such as the average silhouette method, exist, the most widely used strategy is the 'elbow method'. To find the optimal number of clusters, one should plot the number of clusters against the sum of squared errors (SSE), calculated as follows, for k clusters, where p represents a point in the cluster [56]:

$$SSE = \sum_{k=1}^{k} \sum_{p \in \{C_k\}} ||p - C_k||_2^2 \tag{2.27}$$

The optimal value of k is where the SSE value drops dramatically, and the curve appears to plateau. This is determined by visual inspection; the slope of the curve will suddenly become much less steep after a particular number of clusters, and the SSE value will gradually decrease after this point.

If this pattern is not observed, such as in the case where the slope undergoes a slow and gradual decrease, then a meaningful clustering does not exist, since the SSE remains similar despite the number of groups.

Chapter 3

Methodology

In this thesis, we discuss two main experiments, both of which require the creation of a reference template, as explained in section 3.3. The first experiment, described in section 3.4, involves the use of registration to the reference template as a tool to perform organ segmentation. The second experiment, detailed in section 3.5, explores the automated clustering of subjects into anatomical subtypes, with the goal of creating anatomy-specific reference templates. The grouping is performed based on the similarity of the images to the reference template from section 3.3.

3.1 Datasets

The activities described in sections 3.3-3.5 are performed using the following two datasets. Both were obtained from the Augmented Intelligence and Precision Health

Laboratory (AIPHL), which is part of McGill University's Department of Diagnostic Radiology, in the Faculty of Medicine and Health Sciences [57].

Both datasets consist of computed tomography (CT) scans of the head and neck. A CT scan is a common medical imaging technique that uses a beam of x-rays to create several cross-sectional images of a patient. These cross-sectional images are known as "slices". An intravenous "contrast agent", like iodine, is sometimes given to patients in order to better visualize particular structures on the scan. CT scanners may also produce images at multiple energy levels, measured in kiloelectron volts (keV). This is because some tissues or anatomical features are easier to see at certain energy levels [58].

All images were collected with the same 64-slice dual-energy computed tomography (CT) scanner, the Discovery CT750 HD from GE Healthcare. Patients were scanned after 80 mL injection of the contrast agent iopamidol, dispensed at a rate of 2 mL/s. There was a delay of 65s before scanning. While scans from 40-140 keV were obtained, here, we only use the 65 keV energy level, as it is optimal for tissues of the head and neck [59].

The datasets are aggregations of scans used in previous publications by the AIPHL lab, examples here: [3, 59, 60].

3.1.1 SRG Dataset

The SRG ('spectral radiogenomic') dataset consists of 3D CT scans of the head and neck regions of 56 subjects. All patients have tumors in this region, either due to HNSCC

(head and neck squamous cell carcinoma) or lymphoma. Each subject has corresponding contours, which are labelled regions corresponding to the relevant anatomical structures and the background, for nine normal organs. Here, however, we focus on only two organs: the left and right parotid glands, seen in Figure 3.1. We chose the parotid glands simply because they were the largest available contoured organs, and it is easier to visualize segmentation results for larger structures.

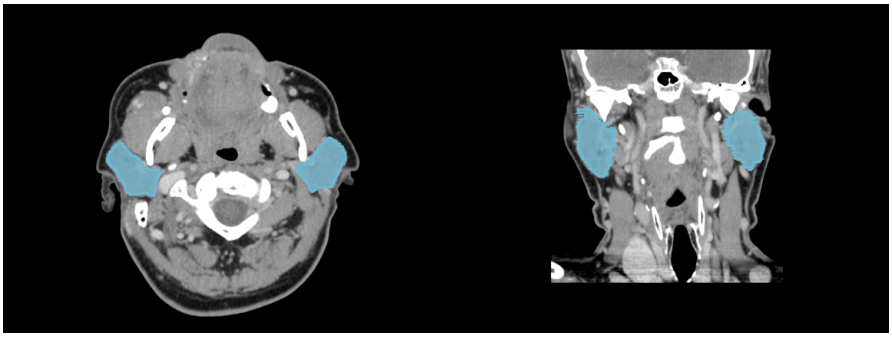


Figure 3.1: The left and right parotid glands (left: axial view, right: coronal view).

Of the 56 subjects, two only have scans in which their mouths were open, so these are excluded from the experiments, as the rest have closed-mouth scans. An additional four subjects have contours that use different labelling schema, so they are excluded as well. Thus, we consider the remaining 50 subjects.

3.1.2 HNSCC Dataset

The HNSCC ('head and neck squamous cell carcinoma') dataset also consists of 3D CT scans of the head and neck, for 95 subjects. It was collected from patients who have

HNSCC tumors at different stages. There are four subjects included who do not have tumors, nor do they have metallic artifacts. These artifact-free subjects are used for the reference template creation described in section 3.3. During the group template creation portion of the experiment described in section 3.5, subjects with severe artifacts are filtered out (detailed in section 4.2.2). This dataset does not have any corresponding contours.

3.2 Image Preprocessing

3.2.1 Masking

Many of the SRG and HNSCC images contain parts of the CT scan table. To avoid negatively affecting the registration process, we must remove these from the images. The subjects' shoulders must also be cropped, for a similar reason; we are only interested in the registration of the head and neck region, and inclusion of the shoulders will make registration more difficult.

Both of these modifications are done via a 'masking' process using MATLAB, in which every subject has a customized, elliptic cylindrical bounding box to remove the undesired elements. The cylinder dimensions are determined manually by observing the masked result in 3D Slicer, an open source software for medical image visualization [61]. The cylinder begins aligned to the centre of the image, and is shifted and adjusted in size until the shoulders are removed and any parts of the CT scan table are cropped. This process is partially inspired

by a 2012 Master of Engineering thesis from MIT [40].

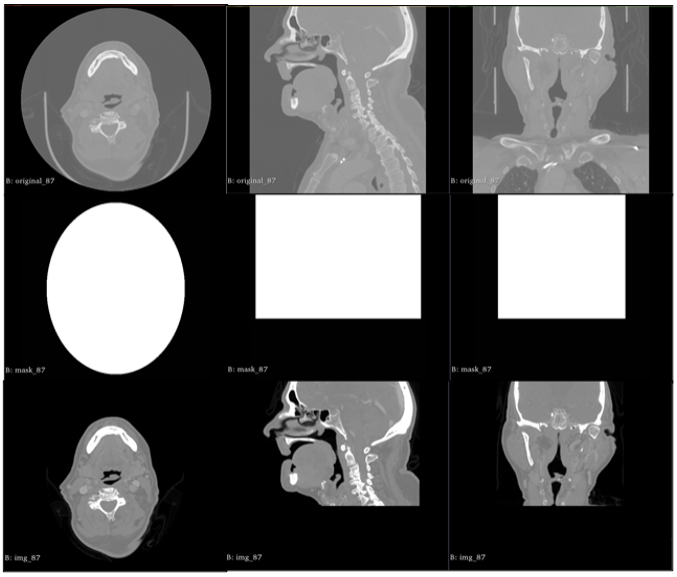


Figure 3.2: An example of the masking process. Top: original image, middle: mask, bottom: masked image.

3.2.2 Image File Formatting

For most of our image registration and segmentation related work, we use the Advanced Normalization Tools (ANTs) toolkit, implemented on the command line. ANTs is widely considered to be a state-of-the-art medical image registration and segmentation toolkit, popularly used in the research community. One of its creators is Dr. Brian B. Avants, who pioneered SyN and its related algorithms, discussed in section 2.1.2 [62].

The ANTS toolkit notes a preference for the NIfTI file format. Since the HNSCC dataset was in DICOM format, and the SRG dataset was in NRRD format, they are both converted

to NIfTI using Plastimatch, another medical image analysis command line tool [63].

3.3 Reference Template Creation

Both of our experiments require the creation of an initial, single custom reference template. To do this, we require images which are free of metallic artifacts and tumors, so that the template is representative of a standard anatomy. If the reference image contains abnormalities, the quality of subject registration may be diminished.

In the HNSCC set, we have four such images. We include all of these in our template, as they represent reasonably different anatomical types. While we acknowledge that it would be ideal to include even more images, template creation is a very time-consuming process that requires the registration of all subjects to one another, so each added image greatly lengthens the process. This is especially true for the head and neck region, which can vary quite significantly between subjects. It is necessary to consider the trade-off between the amount of anatomical variation that another image adds versus the time to create the template. In our case, a smaller amount of images is sufficient, as our goal is only to create a proof-of-concept for our pipeline.

To create the template, all four input images are registered to each other, resulting in the final ”average” image. The process is illustrated in Figure 3.3. Further details about template creation concepts were provided in section 2.1.5.

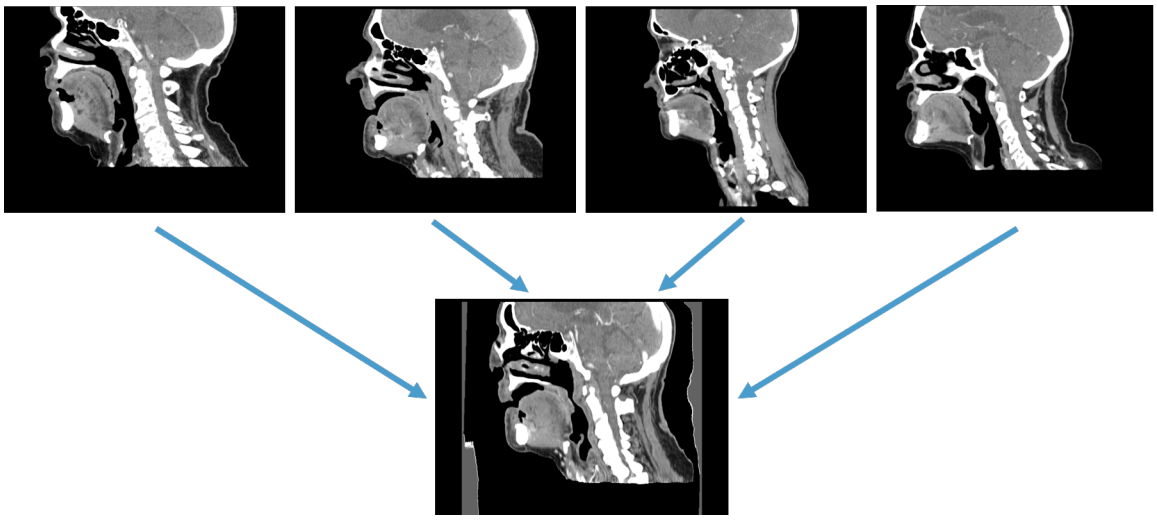


Figure 3.3: The four input images with the result of template creation (sagittal view).

Our implementation uses the ANTs script 'antsMultivariateTemplateConstruction2', for which relevant parameters are described in more detail in Appendix A.1.

We first perform an initial rigid body registration of inputs before template creation; this is recommended by the ANTs documentation if there is no initial template, as is the case here. During the creation of the template, we choose not to include the 'full affine transformation' in our template updates, meaning that the rigid portion is excluded. While this may seem unusual, we obtained final results that were strangely warped and unrealistic when using the full affine transformation. For the non-linear, deformable step, we use the SyN algorithm with a cross-correlation similarity metric, described in sections 2.1.2 and 2.1.3, respectively. The registration takes a coarse-to-fine approach (described in section 2.1.4), with four levels of registration. Each level has corresponding shrink factors, smoothing factors, and levels of iteration (parameters f, s, and q).

Template Construction

```
antsMultivariateTemplateConstruction2.sh -d 3 -f 8x4x2x1 -s 3x2x1x0 -m CC
-r 1 -y 0 -t SyN[0.5,2,0] -g 0.4 -q 100x70x50x10
-o ${outputPath}[image 1] [image 2] [image 3] [image 4]
```

3.4 Segmentation using Reference Template Registrations

In this experiment, we aim to use the registrations of the SRG dataset to the reference template to create a consensus contouring for the template created with the HNSCC dataset. We then use registration to map the template contours to new subjects.

3.4.1 Architecture

An overview of this pipeline is presented below in Figure 3.4:

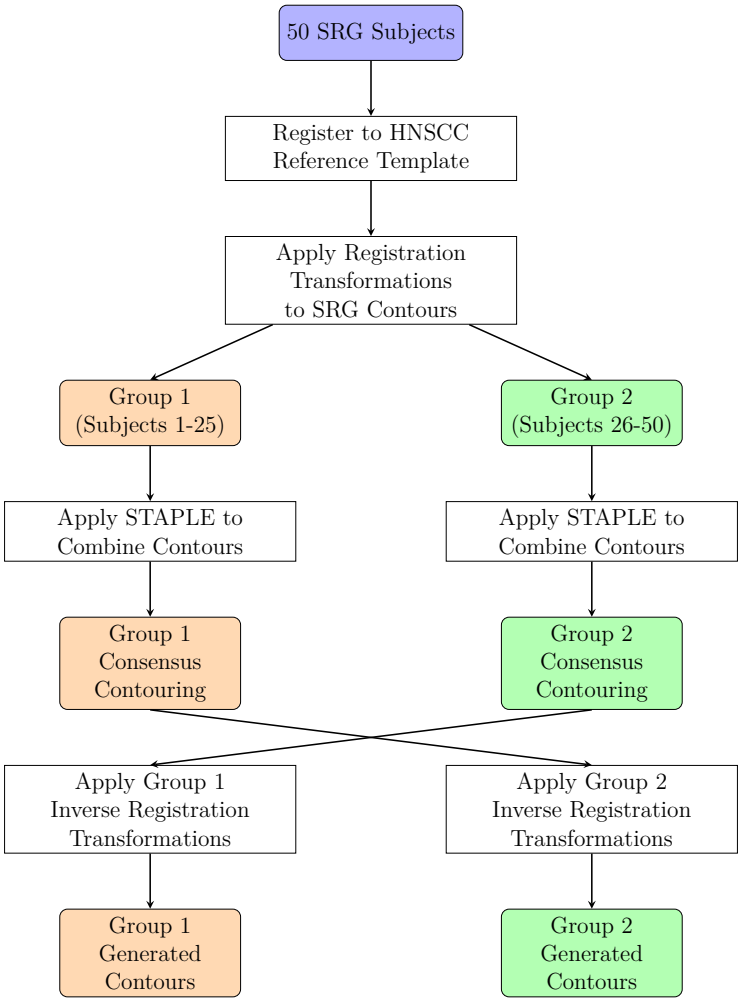


Figure 3.4: Architecture of first experiment.

3.4.2 Registration to Reference Template (SRG)

Image Registrations

For the initial step of registering the SRG dataset images to the reference template from section 3.3, we use the ANTs toolkit script 'antsRegistration'. This process is done in two

parts; the rigid step and the non-linear, deformable step. While it is unusual, we exclude the affine step here, after empirical observations that its inclusion resulted in strangely warped and unrealistic images at the end of registration. The other parameters are chosen according to recommendations in the ANTs documentation [62].

In the rigid registration, the 'source' is the input image, and the 'target' is the reference template. In the deformable registration, the 'source' is the result obtained after rigid registration, and the 'target' is still the reference template. For both steps, we use the cross-correlation similarity metric. In the rigid step, the value in square brackets denotes the gradient step size. In the SyN step, the values in brackets are: gradient step size, update field variance in voxel space, and total field variance in voxel space. For the SyN step, we chose to only have three levels of registration, excluding the last (finest) level. This is because the difference between our results including and excluding the last level were minimal, but the process took twice as long to run.

Relevant parameters of the 'antsRegistration' script are described in more detail in Appendix A.2.

Rigid Registration

```
antsRegistration -d 3 -o ${rigidResult} -n Linear -w [0.005,0.995] -u 1
-r [${referenceTemplate},${inputImage},1] -t Rigid[0.1]
-m CC[${fixedPath},${movingPath},1,8] -c [1000x500x250x100,1e-6,10]
-f 8x4x2x1-s 3x2x1x0vox
```

Deformable Registration

```
antsRegistration -d 3 -o ${finalResult} -n Linear -w [0.005,0.995]
-u 1 -t SyN[0.5,2,0] -m CC[${referenceTemplate},${rigidResult},1,8]
-c [100x70x50,1e-6,10] -f 8x4x2 -s 3x2x1vox
```

Application of Transformations to Contours

We then apply the rigid and deformable transformations to the SRG set's corresponding normal organ contours, so that these contours will also be in the space of the reference template. We perform this step using the ANTs script 'antsApplyTransforms'. There are two steps here, as well, for rigid and then deformable registration.

Relevant parameters of the 'antsApplyTransforms' script are described in more detail in Appendix A.3.

Rigid Registration

```
antsApplyTransforms -d 3 -i {originalContour} -o {resultDirectory}
-t {rigidTransform} -r {rigidResult}
```

Deformable Registration

```
antsApplyTransforms -d 3 -i {rigidResult} -o {resultDirectory}
-t {SyNTransform} -r {finalResult}
```

3.4.3 Consensus Contouring

Again referring to Figure 3.4, we split the collection of 50 template-aligned contours into two groups of 25. For each group, we create a 'consensus contouring', in the space of the reference template.

This is performed using Insight Toolkit (ITK). ITK is an open-source library that provides extensive scientific image analysis functionality. The creators have also produced the 'SimpleITK' library for easy integration of ITK features into other code. It is available for several programming languages, including Python, which we use in our implementation [27]. Here, we execute the STAPLE algorithm described in section 2.2.4.

The 'STAPLEImageFilter' function used in the script below requires a specific foreground value that references the object to be segmented, while the rest of the image is considered as the background. We set the numerical foreground value as 'STAPLE_FOREGROUND_VAL'. We then execute the function on our directory of subject contours ('IMAGE_CONTOURS'), which are labelled regions of the structure area and the background, that have been previously transformed according to section 3.4.2. As output, we get a singular image representing the consensus contouring created from all of the original contours that have been registered to our reference template.

We obtain one consensus contouring per group, for each desired normal organ. Here, we compute the contours for the left and right parotid glands, giving us two consensus contours per group, resulting in four total contours.

STAPLE Execution

```
import SimpleITK as sitk

# Set up STAPLE filter

staple_filter = sitk.STAPLEImageFilter()

staple_filter.SetForegroundValue(STAPLE_FOREGROUND_VAL)

# Execute STAPLE

staple_img = staple_filter.Execute(IMAGE_CONTOURS)
```

3.4.4 Generation of Contours

As seen in Figure 3.4, for each of the groups, we use the opposing group's consensus contours to generate labels in the space of the original images. That is to say, we use the first group's consensus contours to generate automatic labels for the second group's images, and follow a similar procedure to get the first group's labels from the second group's consensus contours.

To achieve this, we apply the inverse of the registration transformations generated in section 3.4.2. We again use 'antsApplyTransforms' here, in two steps. Since we wish to transform in the reverse direction, that is to say, back to the space of the original images, we now perform the deformable step before the rigid step.

While the commands are the same as the forward transformations in section 3.4.2, the 'rigidTransform' and 'synTransform' parameters refer to inverted versions of the

transformations. The 'antsRegistration' script already outputs the inverted deformable transformation in a separate file, so we simply use this. For the rigid step, 'antsApplyTransforms' allows us to specify that the transformation should be inverted by writing it as such: -t [rigidTransform, 1].

After this step, we have generated contours for all of the original SRG images, in the original image space.

3.4.5 Evaluation

Dice Score

We then use the Dice score metric described in section 2.2.3 to evaluate these segmentation results. The score will tell us the accuracy of the generated contours compared to the original labels, which were obtained from radiologists.

Similarity to Template

After calculating the Dice scores of the segmentation results on a per patient basis, we will examine a possible relationship between Dice score and the similarity of images to the template.

To do this, we can use a sum of squared differences approach. For voxels in the registered images that also exist within template area, we calculate the absolute value of the difference in intensities between the template and the registered images. We add the square of these

absolute value differences to a running total. This leaves us with a single numerical value describing the similarity of a particular image to the template. The larger this value is, the less similar an image is to the template.

3.5 Clustering Subjects into Anatomical Subtypes

In this experiment, we use the k-Means clustering technique described in section 2.3.4 to group subjects by anatomical similarity. We evaluate subject similarity to the reference template using the deformable transformation that arose during its registration to that template. In other words, the images will be clustered based on the magnitude of deformation that they underwent at each voxel to align with the template. We then create representative reference templates for each of the discovered anatomical subtypes, with the goal of implementing the pipeline from section 3.4 within each group to improve segmentation accuracy.

3.5.1 Architecture

To illustrate this process, we refer to Figure 3.5:

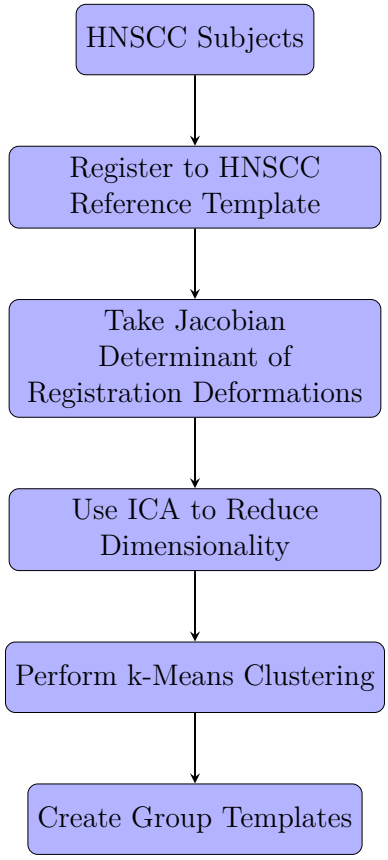


Figure 3.5: Architecture of second experiment.

3.5.2 Registration to Reference Template (HNSCC)

Following the steps outlined in section 3.4.2, we register all of the HNSCC dataset images to the reference template from section 3.3. Again, we exclude the affine registration step, after empirical observations that doing so improved the quality of registrations for our particular datasets.

For all HNSCC images, we now have the final warped-to-template registration results as

well as the rigid and deformable transformations. In this case, we do not have any contours, so there is nothing further to do in this step.

3.5.3 Jacobian Determinant

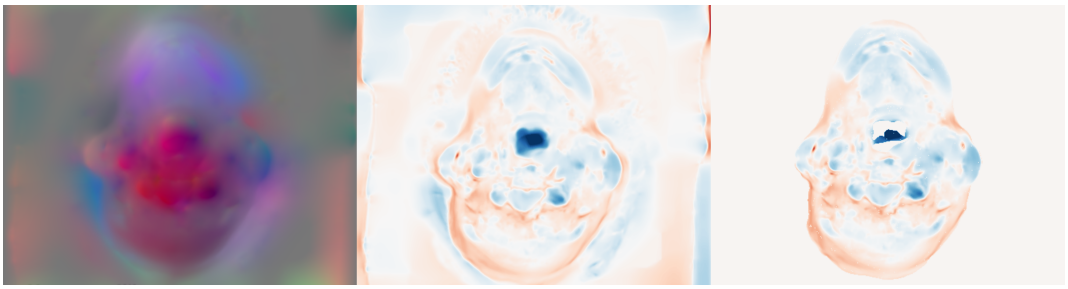
We then take the Jacobian determinant of the deformable registration transformation. The goal of this step is to obtain a single value at each voxel describing the magnitude of image deformation that occurred at this particular location during the registration to the template, as described in section 2.1.6. A visualization of this is shown in Figure 3.6.

This step uses another ANTs script, 'CreateJacobianDeterminantImage', which simply computes the Jacobian determinant of the image that it is passed as a parameter; the usage of this command is described below. In our case, we have three dimensional images, and we simply pass each image's corresponding deformation field (transformation) to the script to obtain the corresponding Jacobian determinant images.

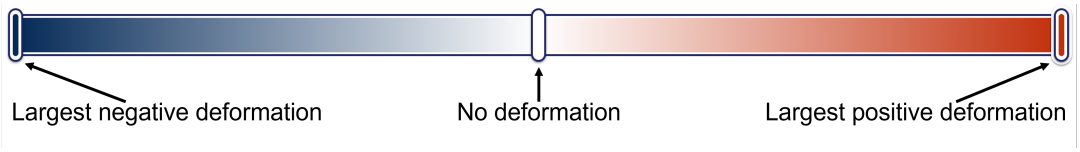
Usage of 'CreateJacobianDeterminantImage'

```
CreateJacobianDeterminantImage <imageDimension> <deformationField>
    <outputImage>
```

We then apply a mask of the reference template to the resulting image, as we wish to exclude any voxels that are not part of the template.



Left to right (axial view): deformation field, Jacobian determinant image, masked Jacobian determinant image.



Colour bar for determinant images. The leftmost colour represents the largest 'negative' deformation, the rightmost colour represents the largest 'positive' deformation, and the middle colour represents no deformation. Note that 'negative' and 'positive' represent opposite directions.

Figure 3.6: Masked Jacobian determinant image

3.5.4 Independent Components Analysis

Our datasets consist of 3D CT scan images, of size 512x512x233. If we flatten one of these images into 1D, we have 61,079,552 individual points. With this amount of data for each of the 95 images, any operations or analysis will be very computationally heavy. This is what motivates our use of Independent Components Analysis (ICA).

To do this, we use the Scikit-learn Python library, which is popular for computational techniques related to machine learning [64]. We use Scikit-learn's 'FastICA' function. We tested a few different numbers of ICA components, and found that five components provided a reasonable clustering result (described in the next step); so, we had FastICA extract five

components. We set the maximum number of iterations to 500, and the tolerance to 0.05. The defaults for these parameters were 200 and 1e-4, respectively, but they did not lead to convergence for our data. Before adjusting the tolerance, FastICA did not converge, even after thousands of iterations. After updating the tolerance to 0.05, FastICA was able to converge in fewer than 500, but greater than 200, iterations.

We execute this using the 'fit_transform' function of FastICA on the flattened representations of our images ('FLATTENED_IMAGES'), the 1D arrays of size 61,079,552.

Scikit-learn FastICA Implementation

```
from sklearn.decomposition import FastICA

# Set up & run ICA

ica = FastICA(n_components=5, max_iter=500, tol=0.05)

ica_result = ica.fit_transform(FLATTENED_IMAGES)
```

3.5.5 k-Means Clustering

Now that each image is only represented by five components, we perform k-Means clustering of the images, as described in section 2.3.4. To do this, we again use the Scikit-learn library.

We import the 'KMeans' function and provide it with our desired number of clusters ('NUM_CLUSTERS'). We then pass KMeans the ICA representations of all of our HNSCC

images ('ICA_IMAGES') for clustering.

All cluster numbers from 1 to the number of images will be tested. We test all of these values not only to find the ideal number of clusters, but to determine whether a meaningful clustering exists. Further details about cluster number selection are provided in the following section.

Scikit-learn k-Means Implementation

```
from sklearn.cluster import KMeans

# Set up the KMeans clustering function

kmeans = KMeans(n_clusters=NUM_CLUSTERS)

# Cluster ICA representations of images

labels = kmeans.fit_predict(ICA_IMAGES)
```

3.5.6 Evaluation of Clustering

To find the optimal number of clusters, we perform the 'Elbow Method', as described in section 2.3.4. For practical purposes, we do not want to have more than four or five clusters to create our group-specific templates. However, we still wish to test each cluster size from 1 to the number of images for two reasons. First, we would like to learn what the optimal number of clusters is; while it is impractical to create more than 4-5 group templates, if the ideal cluster size is considerably higher than five, then we may not obtain noticeable

anatomical similarity between cluster subjects in a fewer number of groups. The second reason is that, if the ideal cluster size is equivalent to the number of images, then we can conclude that no meaningful clustering exists.

3.5.7 Creation of Group Templates

The creation of the group-specific templates is the same as in section 3.3. However, here, we use the three most representative members of the cluster to create the template. We use the centroid image, which is the image closest to the centre of the cluster, and the two other images that are closest to the centroid. By using the cluster members closest to the centre, we are using the most representative images of that anatomical subtype. This allows us to account for a significant amount of intra-group variance.

Since some of the clusters are not very large, we would like to use as few images as possible that still capture adequate variance. As explained further in the next section, we would eventually like to test the segmentation pipeline from section 3.4 in each cluster; so, if all images are used in the group template, there will be no non-template images with which to test segmentation.

Similar to the template creation discussed in section 3.3, it would be ideal to use a greater number of images for the group templates; however, we are limited by our small dataset. Again, in our case, we are developing a proof-of-concept, so we have a sufficient number of images for this purpose.

3.5.8 Evaluation of Group Templates

One way to assess the quality of each group template is by visual inspection. If we see areas where it appears that registration did not complete (for example, if there are two noses visible), then the template creation was not successful. If the resulting template looks like a realistic CT scan, then it was successful.

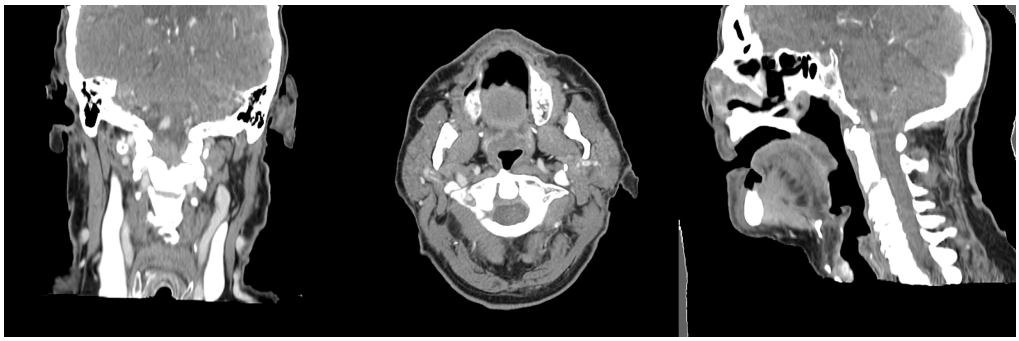
We can also evaluate the effectiveness of the created templates by using them to segment the images in their respective clusters with the pipeline from section 3.4. Theoretically, this should lead to increased accuracy compared to the segmentation of the same images using a generic template, as was used in the previous experiment.

Chapter 4

Experimental Results

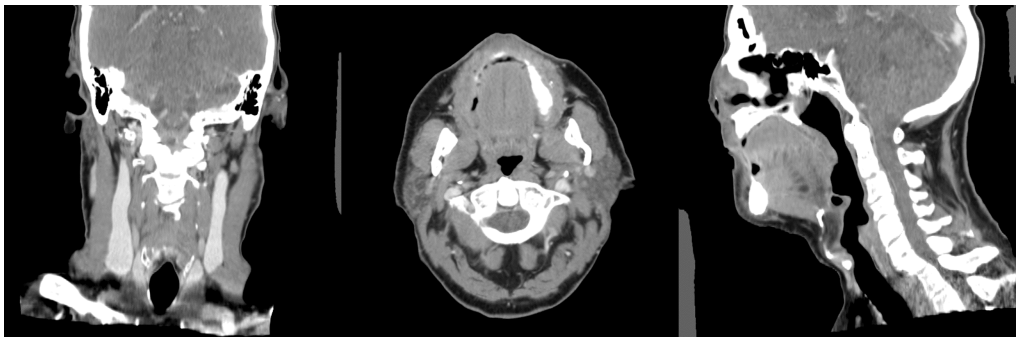
4.1 Segmentation using Reference Template Registrations

Here, we examine the results of the segmentation pipeline described in section 3.4 for the left and right parotid glands. We first observe some examples of registration; we consider one subject where registration was successful, and another subject where it was unsuccessful. We then evaluate the success of parotid gland segmentation by calculating the Dice score of the generated contours against the original contours, for each desired organ.



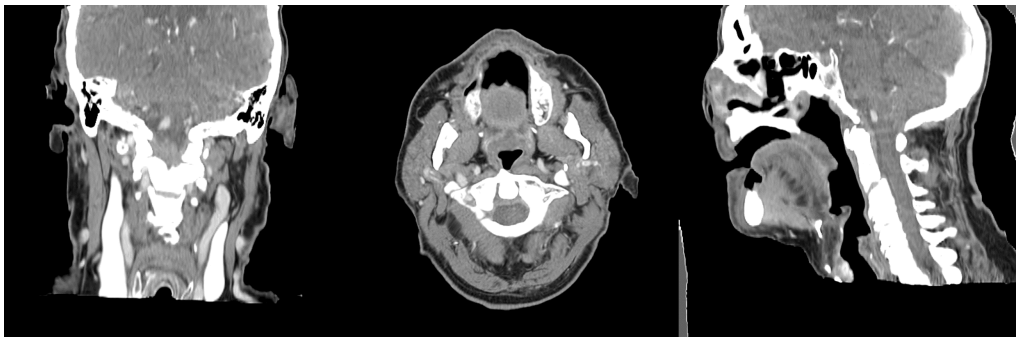
(a) The template to which subjects were registered

(b) The original subject image

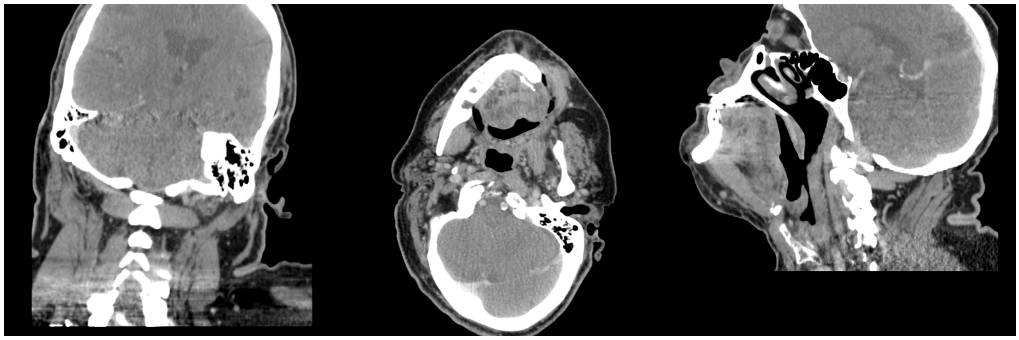


(c) The result of subject registration to the template

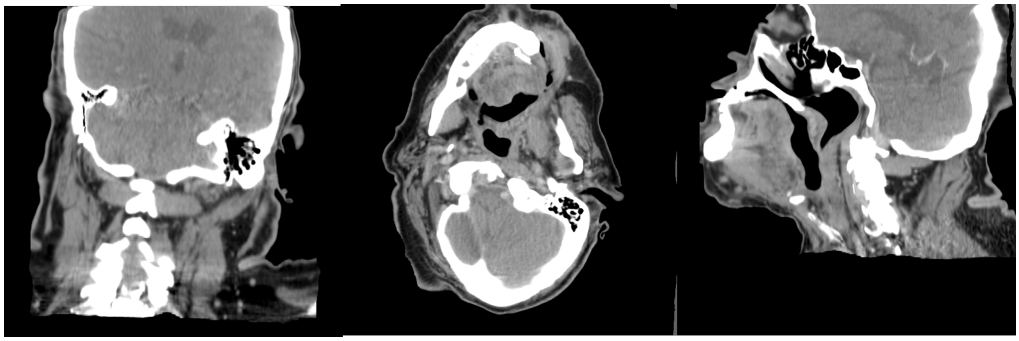
Figure 4.1: An example of a successful registration. This particular subject was able to align quite well to the custom template that we created. (Left to right: coronal, axial, sagittal views)



(a) The template to which subjects were registered



(b) The original subject image



(c) The result of subject registration to the template

Figure 4.2: An example of an unsuccessful registration. This subject had considerable difficulty aligning to the custom template that we created. (Left to right: coronal, axial, sagittal views)

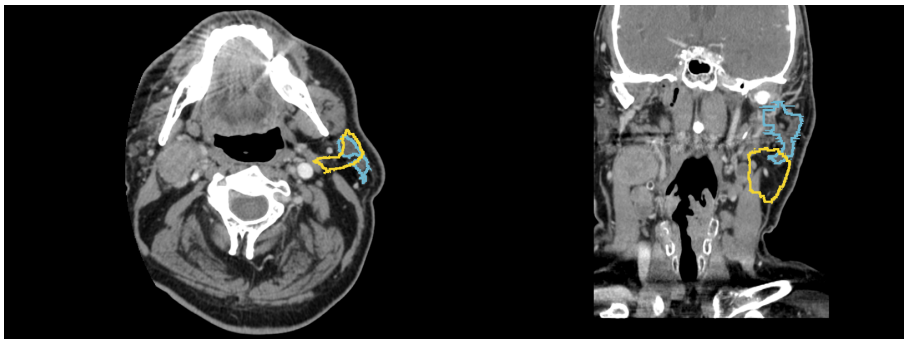
We note that five SRG subjects had a Dice score of zero for the left parotid gland (LPG) segmentation, and six SRG subjects had a score of zero for the right parotid gland (RPG). Upon further inspection of these subjects, we discovered that their original contouring excluded these particular glands. Thus, we consider the scores of 45 total subjects for the LPG and 44 total subjects for the RPG. Table 4.1 shows the average Dice score for each group and organ, and the combined average score for each organ.

	Left Parotid Gland (LPG)	**Right Parotid Gland (RPG)**
Group 1	0.574 *(Range: 0.107 - 0.764)*	0.597 *(Range: 0.007 - 0.820)*
Group 2	0.559 *(Range: 0.286 - 0.804)*	0.621 *(Range: 0.217 - 0.775)*
All Subjects	0.567 *(Range: 0.107 - 0.804)*	0.608 *(Range: 0.007 - 0.820)*

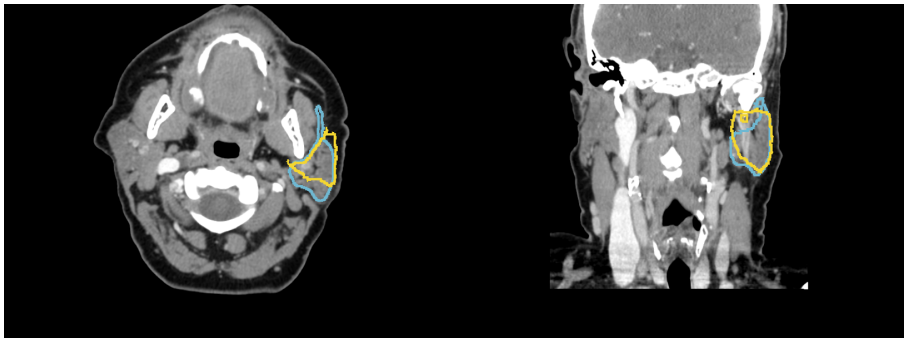
Table 4.1: Average Dice score, including the range of values (minimum - maximum), for LPG and RPG segmentations, calculated per group, and then across all subjects.

Upon inspection of our results, we observe that there are some instances where our pipeline performs especially well and some where it performs much more poorly than average. Figures 4.3 and 4.4 show examples of poor, average, and good segmentation outcomes.

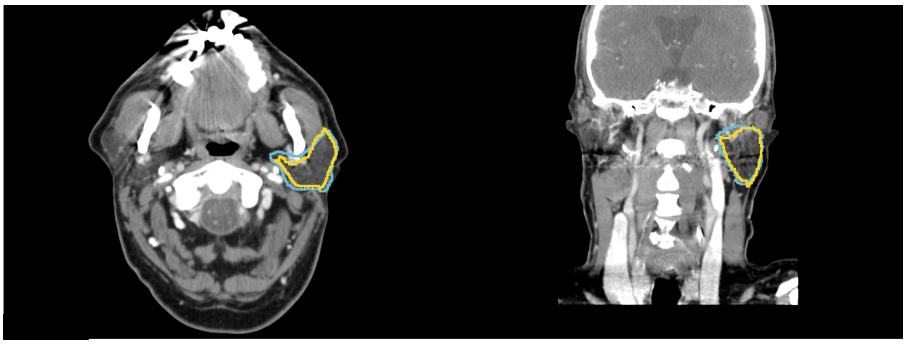
We next examine whether the Dice score is related to the similarity of an image to the template. Template similarity is determined by the sum of squared differences approach described in section 3.4.5. Figures 4.5 and 4.6 show image distances from the template and the corresponding Dice scores for each parotid gland.



(a) Poor; Dice score = 0.107

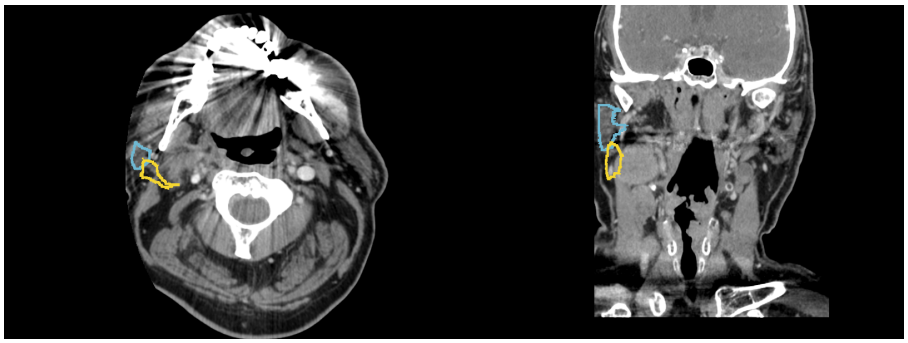


(b) Average; Dice score = 0.566

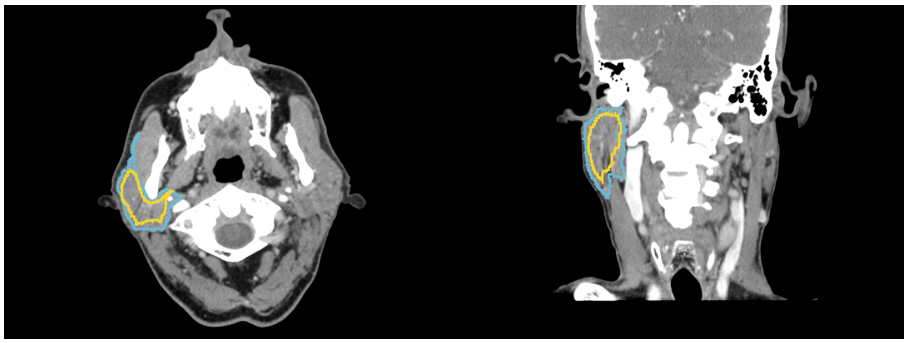


(c) Good; Dice score = 0.804

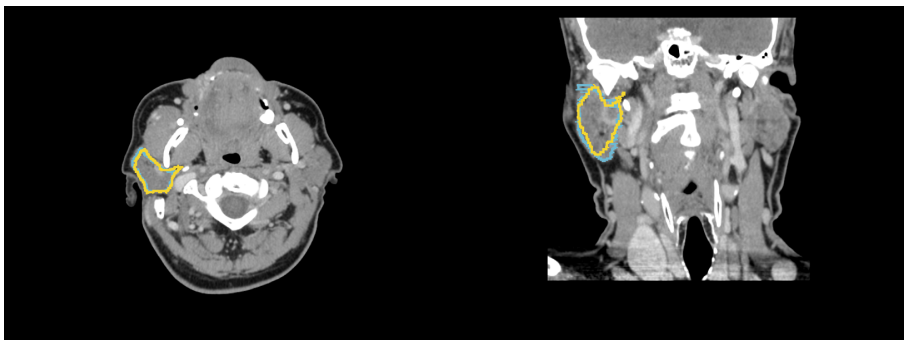
Figure 4.3: Sample segmentation results for left parotid gland (LPG). Original contour is blue, generated contour is yellow. (Left: axial view, right: coronal view)



(a) Poor; Dice score = 0.007



(b) Average; Dice score = 0.613



(c) Good; Dice score = 0.820

Figure 4.4: Sample segmentation results for right parotid gland (RPG). Original contour is blue, generated contour is yellow. (Left: axial view, right: coronal view)

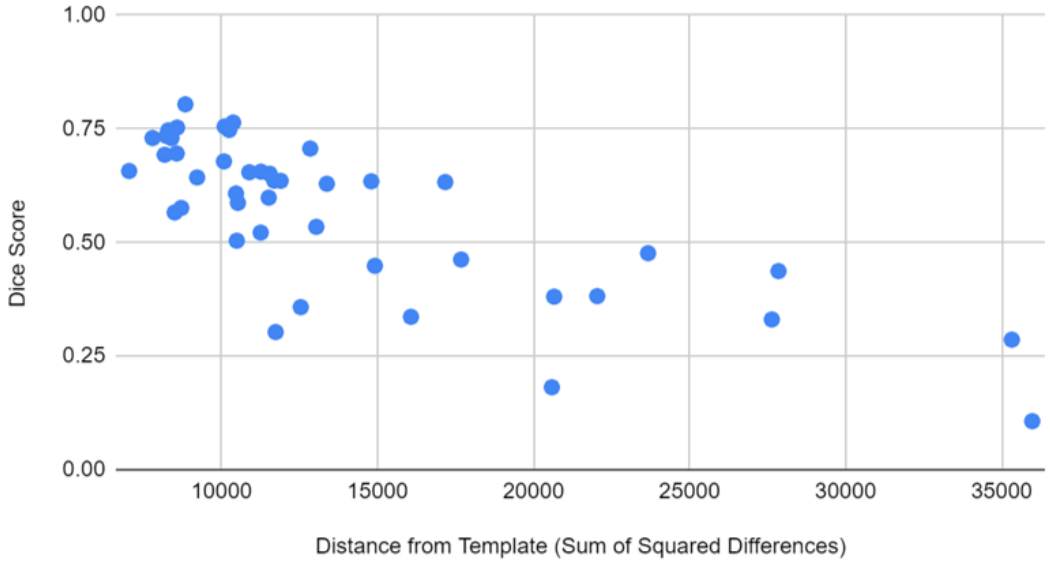


Figure 4.5: Dice Score vs. Distance from Template (LPG). Template distance is calculated using sum of squared intensity differences.

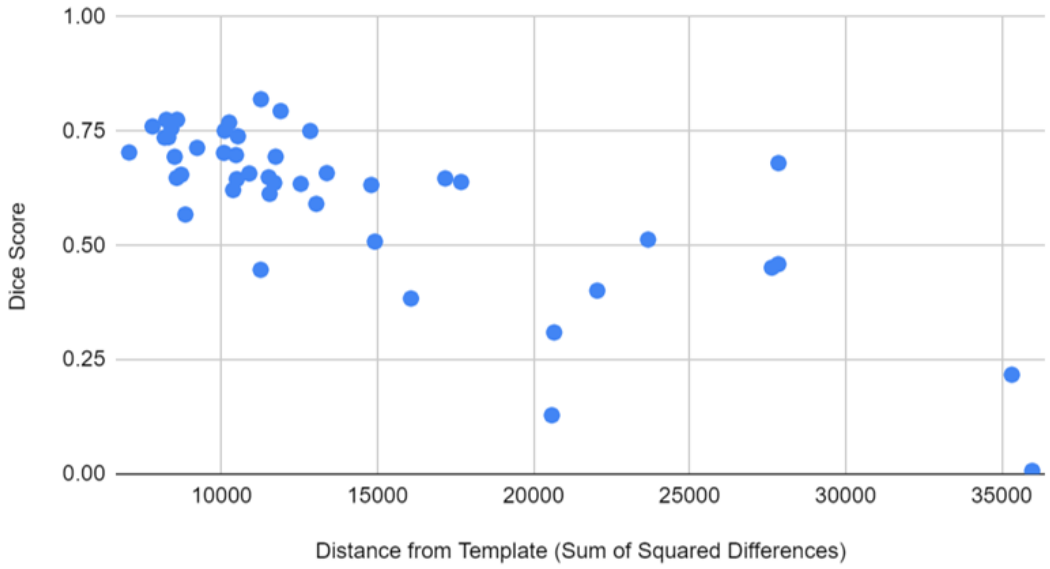


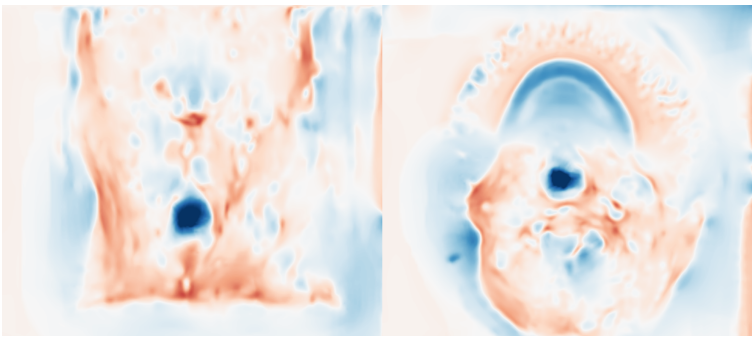
Figure 4.6: Dice Score vs. Distance from Template (RPG). Template distance is calculated using sum of squared intensity differences.

For each parotid gland, we calculate the Spearman coefficient for the Dice score against the sum of squared difference from template, shown in Table 4.2. We use the Spearman coefficient as it does not assume an underlying distribution of the data, and our data does not follow a normal distribution.

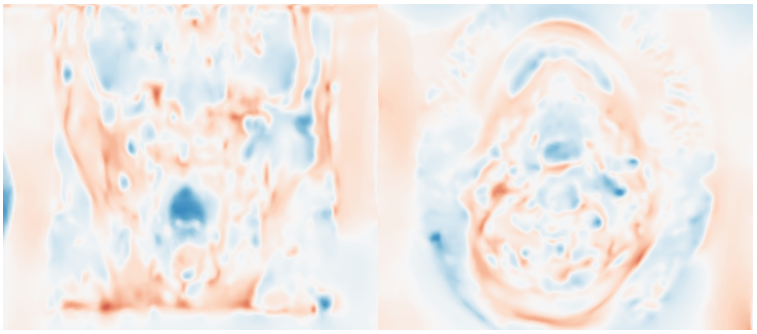
	Spearman Correlation Coefficient
Left Parotid Gland	-0.760
Right Parotid Gland	-0.676

Table 4.2: Spearman correlation coefficient for LPG and RPG for Dice score vs. distance from template.

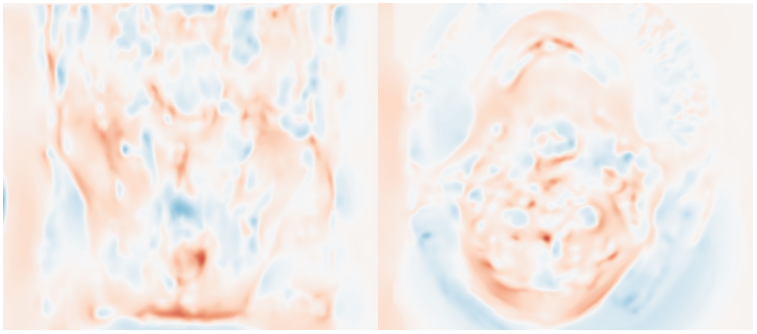
We recall the subjects from Figure 4.5, where we presented results for the segmentation of the left parotid gland. In order to further examine the relationship between distance from template and Dice score, we observe the Jacobian determinant images created from the deformation fields that arose during the registration of these subjects to the template.



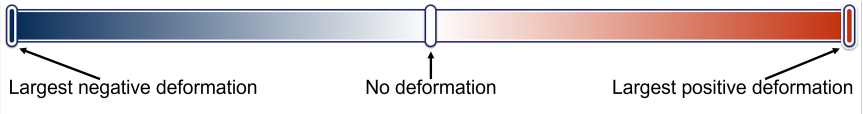
(a) Subject with poor segmentation outcome (Dice score = 0.107, image L2 norm: 942)



(b) Subject with average segmentation outcome (Dice score = 0.566, image L2 norm: 862)



(c) Subject with good segmentation outcome (Dice score = 0.804, image L2 norm: 700)



Colour bar for determinant images, as seen in Figure 3.6.

Figure 4.7: Jacobian determinant images corresponding to the subjects in Figure 4.3. These images represent the Jacobian determinants of the deformation fields that arose from their registrations. Areas with more intense colour correspond to deformations of greater magnitude. Again, 'positive' and 'negative' are opposite directions relative to the template. Dice scores and L2 norms calculated over full 3D (512x512x233) image. L2 norms calculated only over template mask seen in Figure 3.6. (Left: coronal view, right: axial view)

4.2 Clustering Subjects into Anatomical Subtypes

We now examine the results for the steps described in section 3.5, where we clustered our images based on anatomical similarity, and attempted to generate one representative template per group. After some experimentation, we chose to use five components for our ICA, as it gave the most reasonable clustering result. This means that the k-Means clustering was performed on five-dimensional representations of image data.

4.2.1 Results of Clustering

As described in section 3.5.6, we perform the elbow method to evaluate whether a meaningful clustering exists. These results are presented below. However, as noted previously, we do not wish to use more than four or five clusters, for practical purposes.

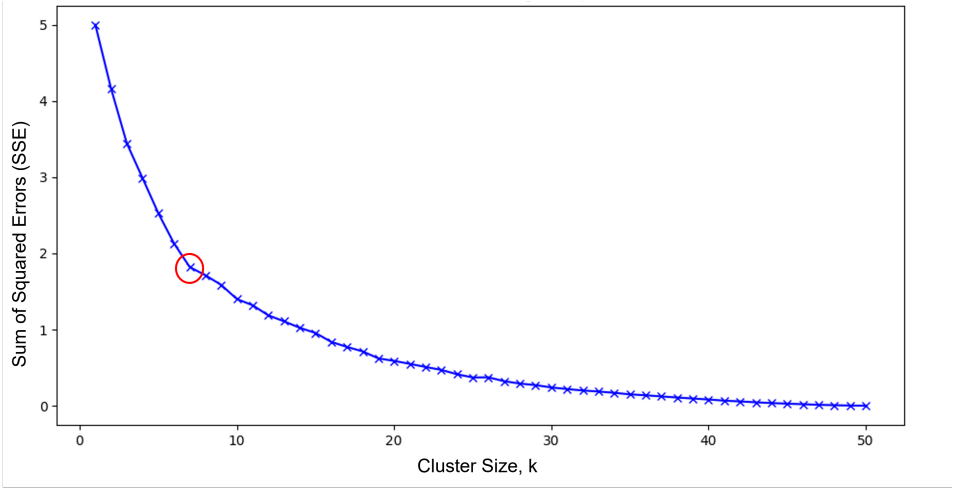


Figure 4.8: Elbow method result for our subjects. The ideal number of clusters is around seven, showing that a meaningful clustering exists.

After testing with four and five clusters, we determined empirically that four clusters gave a more natural-seeming grouping. A representation of the clusters is presented in Figure 4.9, visualized in two dimensions using Scikit-learn's t-SNE function [64].

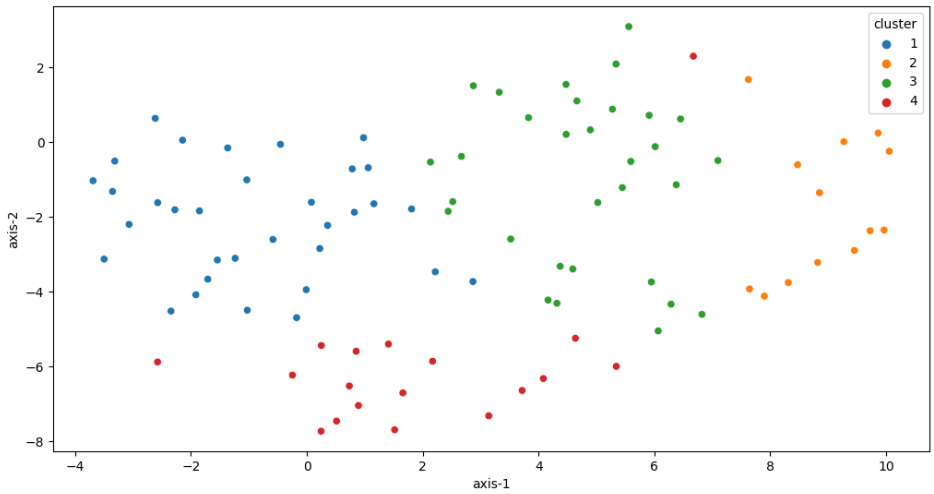
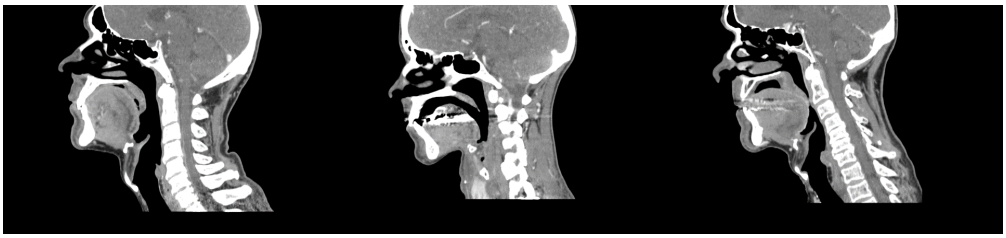


Figure 4.9: k-Means clustering of subjects, clustered in 5D, visualized in 2D with t-SNE.

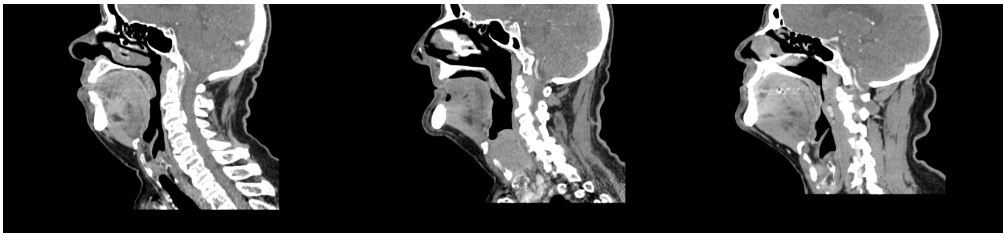
In the following figure, we observe a selection of members from each of these clusters, in order to visually compare their anatomies.

(a) Three members of the blue cluster



(b) Three members of the red cluster

(c) Three members of the green cluster



(d) Three members of the orange cluster

Figure 4.10: Examples of cluster members. The red cluster members appear to have more elongated necks, whereas the blue cluster members have shorter necks. The orange members are facing upward, and the green members are facing straight ahead or slightly downward and have similar nose shapes. A sagittal view is used to more easily see anatomical features. Note that the original dataset images are shown (i.e. those before template registration).

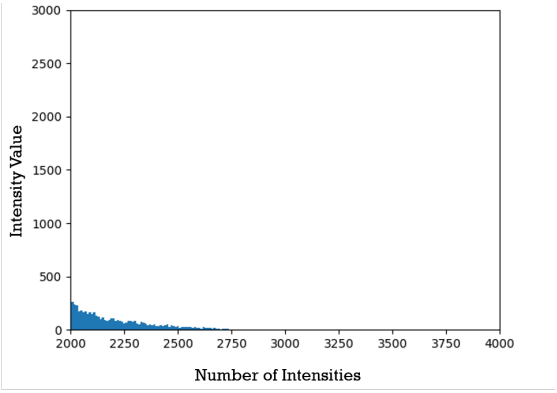
4.2.2 Filtering of Subjects with Artifacts

The step after clustering is the creation of group templates; however, many of the subjects closest to the cluster centroids were discovered to contain significant metallic artifacts. "Severe" artifacts were filtered from the data, and "mild" artifacts were allowed to remain. See Figure 4.11 for examples of the two categories.

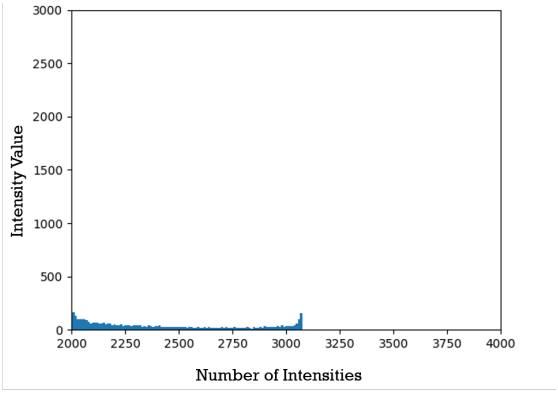
(a) Mild artifacts **(b)** Severe artifacts

Figure 4.11: Mild vs. severe artifacts (axial view)

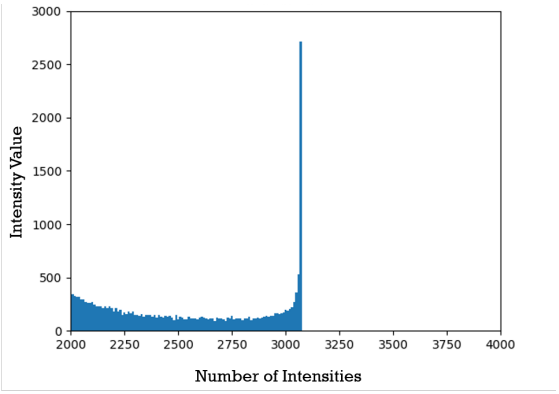
In Figure 4.12, we see examples of histograms demonstrating sample intensity patterns for subjects with no, mild, and severe artifacts. In subjects without artifacts, there are very few, if any, image voxels with intensities above 2000. Thus, to judge which subjects have severe artifacts and which do not, we examine only the tail end of the intensities; that is to say, we only consider the intensity values above 2000. Through trial and error with filtering criteria, we determined that severe artifact images were those that had over 15% of tail intensities above 3000, *and* over 65% of the intensities above 3000 were also above 3050.



(a) No artifacts



(b) Mild artifacts



(c) Severe artifacts

Figure 4.12: Histograms of image intensities for subjects with no, mild, and severe artifacts.

Artifact-free Clustering

After the filtering, only 49 of 95 subjects remain. Figure 4.13 shows the clustering with the filtered dataset. "Artifact-free" images are those without severe artifacts.

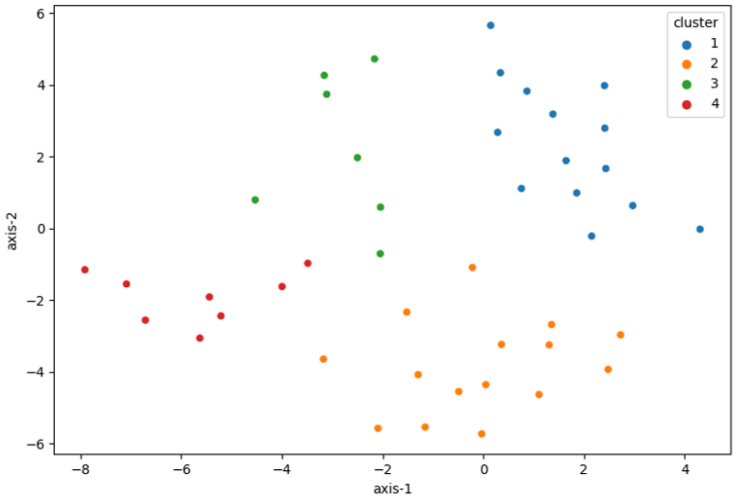
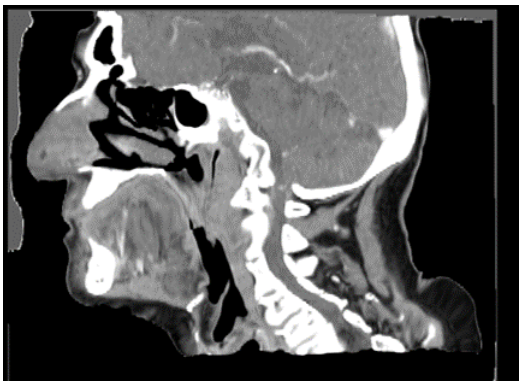


Figure 4.13: k-Means clustering of artifact-free subjects, clustered in 5D, visualized in 2D using Scikit-learn's t-SNE function.

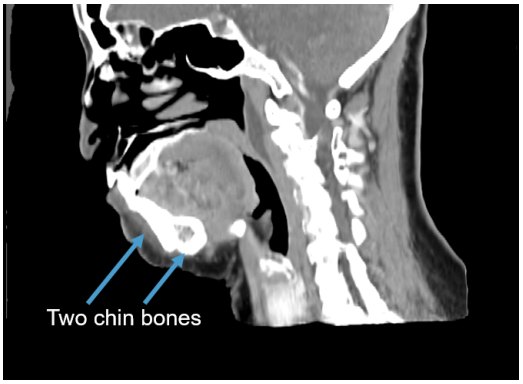
4.2.3 Results of Group Template Creation

Finally, we attempt to create the group-specific templates with the artifact-free images. An example of this effort for each cluster is presented on the following page.

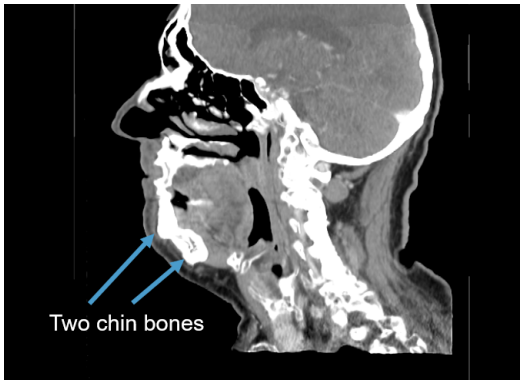


(a) Blue Cluster Template

(b) Red Cluster Template



(c) Green Cluster Template



(d) Orange Cluster Template

Figure 4.14: Examples of group template results, corresponding to the clusters in Figure 4.13. In each of the bottom two images, two chin bones are observed, which signals poor registration. A sagittal view of the images is presented.

Chapter 5

Discussion and Conclusions

In this thesis, we created a proof-of-concept for our atlas-based segmentation pipeline, in which the atlas was constructed from representative dataset subjects, and labelled with consensus segmentations generated using the STAPLE algorithm. We also investigated whether it was possible to use clustering techniques to partition subjects into anatomical subtypes, and create population-specific reference templates for these groups.

This work has three important contributions. The first is that, with the partitioned dataset, one could obtain labels for only a subset of each group in order to make population-specific templates. Then, one could contour the remaining images by implementing our segmentation pipeline in each group. The second is that one could simply select, for each cluster, the pre-existing atlas that is most similar to the centroid image, and then label the group members using a standard atlas-based strategy. In future

work, both of these techniques could lead to improved atlas-based segmentation accuracy for anatomical subgroups that may not be well-represented by a standardized reference template. Finally, our anatomical subtype clustering and template creation methods could be integrated with other pipelines, particularly those based on deep learning. One such example is AtlasNet [14], discussed in sections 1.1 and 2.3.1, which used the registration to population-specific atlases to improve segmentation performance. In this case, the atlases were selected by radiologists, but it would be interesting to test their framework with anatomical subtypes that were discovered automatically.

5.1 Analysis of Segmentation Pipeline Results

5.1.1 Relationship between Template Similarity and Dice Score

After observing the results of our segmentations, such as the examples provided in Figures 4.3 and 4.4, we were motivated to investigate whether a relationship existed between the distance of a registered image from the reference template, and the Dice score. This correlation would be logical, given that images that were originally less similar to the reference template would have more difficulty with registration, and the final segmentations are performed using inverse registration transformations. This is consistent with the sample registration results that were presented in Figures 4.1 and 4.2; the subject who had very poor registration is vastly different from our template.

The Spearman coefficient results in Table 4.2 support this notion. The left parotid gland showed a moderate to strong negative correlation between the distance from the template and the Dice score, and the right parotid gland showed a moderate negative correlation. This supports our hypothesis that the two values are related, and shows that the greater the distance between an image and the template, the lower the Dice score, and thus, segmentation accuracy, will be.

This relationship can also be demonstrated with the Jacobian determinant images seen in Figure 4.7. In 4.7a, the subject with the poor segmentation outcome has sizeable areas where the colour is very intense or dark; this indicates that these portions of the image underwent a very large deformation. If such considerable deformations were necessary, we can conclude that the original image must have been very different from the template. In 4.7c, which corresponds to a subject with a high Dice score, we do not observe any areas with very intense or dark colour; the intensities across the image are much more uniform. This means that this subject did not experience any immense deformations, so they must have already been quite similar to the reference template; however, we know that minor deformations were still required, since the image is not completely uniform.

5.1.2 Comparison with Previous Work

While other research efforts on multi-atlas segmentation for the head and neck, such as those discussed in section 2.2.2, had achieved equivalent or higher Dice score accuracy

compared to our pipeline, they had not explored the idea of population subgroups. As observed in Figures 4.5 and 4.6, the Dice scores for images that are more similar to the template are closer to 0.70-0.80, which is comparable to previous results. If our segmentation pipeline is implemented in each of the clusters, we should be able to achieve that level of accuracy or higher, for all subjects; however, this will need to be verified in future work.

Differences between Left and Right Parotid Glands

We note that the segmentation of the right parotid gland was slightly more successful than the left parotid gland, across our subject pool, as seen in Table 4.1. The average Dice score for the RPG was roughly 0.04 higher than for the LPG. This difference was found to be statistically significant at the 0.05 level by the paired t-test; however, our sample size is relatively small, so it would be beneficial to confirm this with additional data.

A possible explanation for these results could be that the LPG is more variable in shape and size than the RPG, making it more difficult to construct an accurate consensus contouring. Across other atlas-based segmentation efforts discussed in section 2.2.2, the differences between the glands is not consistent. In the 2012 MIT thesis [40], the average scores for the glands appear to be the same, and in Fritscher et al.'s work [36], the Dice score for the LPG was 0.03 higher than for the RPG. A particular group of images, such as our dataset and those used in the aforementioned papers, is typically labelled by one, or a small group of, experts, so perhaps inter-observer variability plays a factor here.

Experimentation on additional data would be required to see if this trend persists.

5.2 Analysis of Anatomical Subtype Clustering and Group Template Creation

The anatomical subgroup clustering showed promise, as the subjects within groups appeared to share common characteristics when inspected visually. However, we were not able to successfully create all the population-specific atlases.

5.2.1 Success of Clustering

As discussed in section 2.3.4, the ideal number of clusters for k-Means clustering can be found through the ”elbow method”. However, here, we did not want more than four or five clusters for two practical reasons: 1) template creation is a time-consuming process, and 2) we wanted to have several cluster members who are not in the template, to test registration.

Despite this fact, it was still interesting to use the elbow method results to confirm that a meaningful clustering exists; these results were presented in Figure 4.8. In this case, the optimal number of clusters was around seven; seven is much smaller than the number of images, so this meant that a significant grouping did indeed exist.

We determined empirically that four clusters provided a grouping that appeared most natural; a representation of these clusters was presented in Figure 4.9. We observed that the

group members did bear clear anatomical resemblances, as seen in Figure 4.10.

5.2.2 Impact of Registration Method on Clustering

As discussed in sections 3.3, 3.4.2, and 3.5.2, we were not able to include the full affine step in our registrations, as its inclusion led to strangely warped registration outcomes. Since no affine registration step was incorporated into our experimental pipeline, the Jacobian determinant, as calculated in section 3.5.3, will include the effects of scaling as well as the effects of non-linear deformations. In other words, the Jacobian determinant quantifies overall changes in the image (except for rotation and translation), and not strictly those that are non-linear.

The exclusion of an affine registration step is certainly a limitation of our research, which will need to be addressed in future work to improve our overall results. For example, the registration of the subject seen in Figure 4.2 clearly suffered due to the lack of an affine transformation. Thus, resolving this issue should improve experimental results for this subject, as well as others that experienced similar issues.

5.2.3 Issues with Group Template Creation

Unfortunately, certain areas of the group templates suffered from poor registration. Examples of this issue can be observed in Figures 4.14c and 4.14d, where two chin bones are visible. There are at least two factors potentially contributing to the lack of template

creation success that require further investigation.

Effects of Metallic Artifacts

As discussed in section 2.1.7, artifacts negatively impact registration; this is an issue for two reasons. The first is that the registration during template creation will be inhibited, and the second is that if these distortions exist in the group templates, then the registration of novel images to the templates will be affected. Thus, it was necessary to filter out subjects with significant artifacts from the dataset. As metallic objects were present in most subjects, only those with severe artifacts (for example, Figure 4.11b) were excluded, and those with mild artifacts (for example, Figure 4.11a) were allowed to remain. The filtering method described in section 4.2.2 was derived based on the intensity patterns in subject images with no, mild, and severe artifacts, represented in Figure 4.12.

After the filtering, only 49 of 95 subjects remained; the updated clustering results were presented in Figure 4.13. The clusters were considerably smaller than when the full dataset was used, and the subjects were further apart from each other. We performed tests using three clusters instead of four, however, with three groups, the anatomical similarities amongst members were not as clear. Four clusters gave groups with the same characteristics as prior to the filtering, so it made the most sense to keep $k = 4$.

The more spread out the images within the clusters are, the less similar they will be. This is certainly a limitation of our work; if this experiment was repeated on a larger artifact-free

dataset, there may be more representative images closer to the cluster centre, allowing for more successful registration during template creation.

Anatomical Differences

A second, but related, element that may be inhibiting template creation is simply anatomical differences. While subjects who clustered together had features in common, they were different enough from each other to make registration difficult. For example, in the group seen in Figure 4.10d, the members are all looking upward, and the back of their necks are less smooth compared to the other groups. However, they still have some differences in anatomy between them, such as spine curvature, which may make it hard to build the average image amongst them. Thus, it is possible that even with more images, it would be difficult to create representative templates for the groups, because they may still be too different from each other to register well. Again, a larger dataset of artifact-free images is needed to further investigate the severity of this issue.

5.3 Current Limitations and Future Work

Our main limitations concern the small size of our dataset. For example, in sections 3.3 and 3.5.7, we note that the initial and group templates would benefit from being created using more images. Furthermore, as detailed in section 5.2.3, one of the main issues affecting our research is the lack of available dataset images without severe artifacts. If additional

data, or a method to remove artifacts from our current data, allows the creation of the population-specific templates, then we can pursue the following endeavours.

As mentioned in the beginning of this chapter, we would like to implement our segmentation pipeline in each of the clusters, using the group-specific templates. We could then compare, for each subject, the resulting segmentation accuracy to the accuracy obtained when we tested our pipeline with the initial custom template.

Additionally, we are interested in combining our clustering work with deep learning-based approaches to segmentation. As discussed earlier, we could test AtlasNet [14] with population-specific templates generated based on automatically discovered anatomical subgroups. However, the idea of registering training images to anatomically representative templates for the purpose of data augmentation could certainly be useful for many other deep learning medical segmentation frameworks.

5.4 Conclusion

To summarize, the first portion of experimentation tested whether a segmentation pipeline with a custom atlas that used inverse registration transformations to label novel images would be successful; the second part explored how this segmentation technique could be improved if it was performed on groups of subjects who already shared anatomical similarities. The clustering of dataset images into anatomical subgroups was successful, while the subsequent creation of group-specific templates was not. This was possibly

related to the lack of available dataset images without severe metallic artifacts; access to a greater number of artifact-free images may lead to a different outcome. Thus, while our techniques showed encouraging results, further investigation is necessary to fully realize their potential.

Bibliography

[1] N. Tong, S. Gou, S. Yang, D. Ruan, and K. Sheng, "Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks," *Medical Physics*, vol. 45, no. 10, pp. 4558–4567, 2018.

[2] D. Močnik, B. Ibragimov, L. Xing, P. Strojan, B. Likar, F. Pernuš, and T. Vrtovec, "Segmentation of parotid glands from registered CT and MR images," *Physica Medica*, vol. 52, no. August, pp. 33–41, 2018.

[3] R. Forghani, A. Chatterjee, C. Reinhold, A. Pérez-Lara, G. Romero-Sanchez, Y. Ueno, M. Bayat, J. W. Alexander, L. Kadi, J. Chankowsky, J. Seuntjens, and B. Forghani, "Head and neck squamous cell carcinoma: prediction of cervical lymph node metastasis by dual-energy CT texture analysis with machine learning," *European Radiology*, vol. 29, no. 11, pp. 6172–6181, 2019.

[4] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability (Switzerland)*, vol. 13, no. 3, pp. 1–29, 2021.

[5] H. Lester and S. R. Arridge, "A survey of hierarchical non-linear medical image registration," *Pattern Recognition*, vol. 32, no. 1, pp. 129–149, 1999.

[6] "BrainSuite — Magnetic Resonance Image Analysis Tools." [Online]. Available: http://brainsuite.org/

[7] D. L. Pham, C. Xu, and J. L. Prince, "Current Methods in Medical Image Segmentation," *Annual Review of Biomedical Engineering*, vol. 2, pp. 315–337, 2000.

[8] J. L. Lancaster, L. H. Rainey, J. L. Summerlin, C. S. Freitas, P. T. Fox, A. C. Evans, A. W. Toga, and J. C. Mazziotta, "Automated Labeling of the Human Brain: A Preliminary Report on the Development and Evaluation of a Forward-Transform Method," *Hum. Brain Mapping*, vol. 5, pp. 238–242, 1997.

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

[10] A. R. Ridwan, M. R. Niaz, Y. Wu, X. Qi, S. Zhang, M. Kontzialis, C. Javierre-Petit, M. Tazwar, D. A. Bennett, Y. Yang, and K. Arfanakis, "Development and evaluation of a high performance T1-weighted brain template for use in studies on older adults," *Human Brain Mapping*, vol. 42, no. 6, pp. 1758–1776, 2021.

[11] "Atlases — McConnell Brain Imaging Centre - McGill University." [Online]. Available: https://www.mcgill.ca/bic/software/tools-data-analysis/anatomical-mri/atlases

[12] B. B. Avants, P. Yushkevich, J. Pluta, D. Minkoff, M. Korczykowski, J. Detre, and J. C. Gee, "The optimal template effect in hippocampus studies of diseased populations," *NeuroImage*, vol. 49, no. 3, pp. 2457–2466, 2010.

[13] R. Wolz, P. Aljabar, J. V. Hajnal, A. Hammers, and D. Rueckert, "LEAP: Learning embeddings for atlas propagation," *NeuroImage*, vol. 49, no. 2, pp. 1316–1325, 1 2010.

[14] M. Vakalopoulou, G. Chassagnon, N. Bus, R. Marini, E. Zacharaki, M.-p. Revel, N. Paragios, M. Vakalopoulou, G. Chassagnon, N. Bus, R. M. Silva, and E. Zacharaki, "AtlasNet : Multi-atlas Non-linear Deep Networks for Medical Image Segmentation," 2018.

[15] B. Avants, N. Tustison, G. Song, P. Cook, A. Klein, and J. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," *NeuroImage*, vol. 54, no. 3, pp. 2033–2044, 2 2011.

[16] F. P. Oliveira and J. M. R. Tavares, "Medical image registration: A review," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 17, no. 2, pp. 73–93, 2014.

[17] S. Periaswamy and H. Farid, "Medical image registration with partial data," *Medical Image Analysis*, vol. 10, no. 3 SPEC. ISS., pp. 452–464, 2006.

[18] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[19] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[20] B. Avants and J. C. Gee, "Geodesic estimation for large deformation anatomical shape averaging and interpolation," *NeuroImage*, vol. 23, no. SUPPL. 1, pp. 139–150, 2004.

[21] P. Savadjiev, C. Reinhold, D. Martin, and R. Forghani, "Knowledge Based Versus Data Based: A Historical Perspective on a Continuum of Methodologies for Medical Image Analysis," *Neuroimaging Clinics of North America*, vol. 30, no. 4, pp. 401–415, 2020.

[22] A. V. Dvorak, T. Swift-LaPointe, I. M. Vavasour, L. E. Lee, S. Abel, B. Russell-Schulz, C. Graf, A. Wurl, H. Liu, C. Laule, D. K. Li, A. Traboulsee, R. Tam, L. A. Boyd, A. L. MacKay, and S. H. Kolind, "An atlas for human brain myelin content throughout the adult life span," *Scientific Reports*, vol. 11, no. 1, pp. 1–13, 2021.

[23] Y. Zou, W. Zhu, H. C. Yang, I. Jang, N. L. Vike, D. O. Svaldi, T. E. Shenk, V. N. Poole, E. L. Breedlove, G. G. Tamer, L. J. Leverenz, U. Dydak, E. A. Nauman, Y. Tong, T. M. Talavage, and J. V. Rispoli, "Development of brain atlases for early-to-middle adolescent collision-sport athletes," *Scientific Reports*, vol. 11, no. 1, pp. 1–15, 2021.

[24] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M. C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, and R. V. Parsey, "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *NeuroImage*, vol. 46, no. 3, pp. 786–802, 2009.

[25] B. B. Avants, P. T. Schoenemann, and J. C. Gee, "Lagrangian frame diffeomorphic image registration: Morphometric comparison of human and chimpanzee cortex," *Medical Image Analysis*, vol. 10, no. 3 SPEC. ISS., pp. 397–412, 2006.

[26] M. Toews, W. Wells, D. L. Collins, and T. Arbel, "Feature-based morphometry: Discovering group-related anatomical patterns," *NeuroImage*, vol. 49, no. 3, pp. 2318–2327, 2 2010.

[27] "ITK — Insight Toolkit." [Online]. Available: https://itk.org/

[28] S. Klein, M. Staring, and J. P. Pluim, "Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2879–2890, 2007.

[29] S. Ruder, "An overview of gradient descent optimization algorithms," pp. 1–14, 2016. [Online]. Available: http://arxiv.org/abs/1609.04747

[30] N. Leporé, C. Brun, X. Pennec, Y.-Y. Chou, O. L. Lopez, H. J. Aizenstein, J. T. Becker, A. W. Toga, and P. M. Thompson, "Mean Template for Tensor-Based Morphometry Using Deformation Tensors BT - Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007," N. Ayache, S. Ourselin, and A. Maeder, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 826–833.

[31] E. M. Meintjes, K. L. Narr, A. J. Van Der Kouwe, C. D. Molteno, T. Pirnia, B. Gutman, R. P. Woods, P. M. Thompson, J. L. Jacobson, and S. W. Jacobson, "A tensor-based morphometry analysis of regional differences in brain volume in relation to prenatal alcohol exposure," *NeuroImage: Clinical*, vol. 5, pp. 152–160, 2014.

[32] M. K. Chung, K. J. Worsley, T. Paus, C. Cherif, D. L. Collins, J. N. Giedd, J. L. Rapoport, and A. C. Evans, "A unified statistical approach to deformation-based morphometry," *NeuroImage*, vol. 14, no. 3, pp. 595–606, 2001.

[33] M. Katsura, J. Sato, M. Akahane, A. Kunimatsu, and O. Abe, "Current and novel techniques for metal artifact reduction at CT: Practical guide for radiologists," *Radiographics*, vol. 38, no. 2, pp. 450–461, 2018.

[34] L. Wei, B. Rosen, M. Vallières, T. Chotchutipan, M. Mierzwa, A. Eisbruch, and I. El Naqa, "Automatic recognition and analysis of metal streak artifacts in head and neck computed tomography for radiomics modeling," *Physics and Imaging in Radiation Oncology*, vol. 10, no. January, pp. 49–54, 2019.

[35] N. M. Zaitoun and M. J. Aqel, "Survey on Image Segmentation Techniques," *Procedia Computer Science*, vol. 65, no. Iccmit, pp. 797–806, 2015.

[36] K. D. Fritscher, M. Peroni, P. Zaffino, M. F. Spadea, R. Schubert, and G. Sharp, "Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours," *Medical Physics*, vol. 41, no. 5, pp. 1–11, 2014.

[37] M. R. Sabuncu, B. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," *IEEE Transactions on Medical Imaging*, vol. 29, no. 10, pp. 1714–1729, 2010.

[38] J. E. Iglesias, M. R. Sabuncu, I. Aganj, P. Bhatt, C. Casillas, D. Salat, A. Boxer, B. Fischl, and K. Van Leemput, "An algorithm for optimal fusion of atlases with different labeling protocols," *NeuroImage*, vol. 106, pp. 451–463, 2015.

[39] X. Han, M. S. Hoogeman, P. C. Levendag, L. S. Hibbard, D. N. Teguh, P. Voet, A. C. Cowen, and T. K. Wolf, "Atlas-based auto-segmentation of head and neck CT images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5242 LNCS, no. PART 2, pp. 434–441, 2008.

[40] A. M. Arbisser, "Multi-atlas Segmentation in Head and Neck CT Scans," Ph.D. dissertation, Massachusetts Institute of Technology, 2012.

[41] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimization for Medical Image Segmentation: Theory and Practice When Evaluating With Dice Score or Jaccard Index," *IEEE transactions on medical imaging*, vol. 39, no. 11, pp. 3679–3690, 2020.

[42] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[43] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[44] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 4 2017.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351. Springer Verlag, 2015, pp. 234–241.

[46] S. Bonechi, P. Andreini, M. Bianchini, and F. Scarselli, "Generating Bounding Box Supervision for Semantic Segmentation with Deep Learning," in *Lecture Notes in*

Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11081 LNAI. Springer Verlag, 2018, pp. 190–200.

[47] A. L. Yuille and C. Liu, "Deep Nets: What have They Ever Done for Vision?" *International Journal of Computer Vision*, vol. 129, no. 3, pp. 781–802, 2021.

[48] Z. Ghahramani, "Unsupervised learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3176, pp. 72–112, 2004.

[49] M. J. McKeown, L. K. Hansen, and T. J. Sejnowsk, "Independent component analysis of functional MRI: What is signal and what is noise?" *Current Opinion in Neurobiology*, vol. 13, no. 5, pp. 620–629, 2003.

[50] A. X. Stewart, A. Nuthmann, and G. Sanguinetti, "Single-trial classification of EEG in a visual object task using ICA and machine learning," *Journal of Neuroscience Methods*, vol. 228, pp. 1–14, 2014.

[51] T. Radüntz, J. Scouten, O. Hochmuth, and B. Meffert, "Automated EEG artifact elimination by applying machine learning algorithms to ICA-based features," *Journal of Neural Engineering*, vol. 14, no. 4, 2017.

[52] V. D. Calhoun, J. Liu, and T. Adali, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data." *NeuroImage*, vol. 45, no. 1 Suppl, pp. 163–172, 2009.

[53] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2625, 11 2008.

[54] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015.

[55] M. S. Yang, C. Y. Lai, and C. Y. Lin, "A robust EM clustering algorithm for Gaussian mixture models," *Pattern Recognition*, vol. 45, no. 11, pp. 3950–3961, 2012.

[56] A. Et-Taleby, M. Boussetta, and M. Benslimane, "Faults detection for photovoltaic field based on k-means, elbow, and average silhouette techniques through the segmentation of a thermal image," *International Journal of Photoenergy*, vol. 2020, 2020.

[57] "AIPHL — Augmented Intelligence & Precision Laboratory." [Online]. Available: http://aiphl.lab.mcgill.ca/

[58] "National Institute of Biomedical Imaging and Bioengineering —." [Online]. Available: https://www.nibib.nih.gov/

[59] S. Lam, R. Gupta, M. Levental, E. Yu, H. D. Curtin, and R. Forghani, "Optimal virtual monochromatic images for evaluation of normal tissues and head and neck cancer using

dual-energy CT," *American Journal of Neuroradiology*, vol. 36, no. 8, pp. 1518–1524, 2015.

[60] M. Seidler, B. Forghani, C. Reinhold, A. Pérez-Lara, G. Romero-Sanchez, N. Muthukrishnan, J. L. Wichmann, G. Melki, E. Yu, and R. Forghani, "Dual-Energy CT Texture Analysis With Machine Learning for the Evaluation and Characterization of Cervical Lymphadenopathy," *Computational and Structural Biotechnology Journal*, vol. 17, pp. 1009–1015, 2019.

[61] "3D Slicer." [Online]. Available: https://www.slicer.org/

[62] "ANTs (Advanced Normalization Tools)." [Online]. Available: http://stnava.github.io/ANTs/

[63] "Plastimatch." [Online]. Available: https://plastimatch.org/index.html

[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

Appendix A

Relevant Parameters of ANTs Scripts

A.1 antsMultivariateTemplateConstruction2

Parameter	Description
d	Dimension of images
r	Perform rigid registration of inputs
y	Update template with full affine transform
t	Transformation model for registration (For 'SyN': [gradientStep, updateFieldVarianceInVoxelSpace, totalFieldVarianceInVoxelSpace])
m	Similarity metric for registration
f	Shrink factors for each level
s	(Gaussian) smoothing sigmas for each level
g	Gradient step size
q	Max iterations for registration at each level
o	Output file path

A.2 antsRegistration

Parameter	Description
d	Dimension of images
n	Interpolation option, from ITK
w	Winsorize image intensities [lowerQuantile, upperQuantile]
u	Histogram match images before registration
m	Similarity metric for registration; CC[fixedImage, movingImage, metricWeight, radius]
r	Set initial moving transform, [fixedImage, movingImage, initializationFeature]
t	Transformation model for registration: Rigid[gradientStep] SyN[gradientStep, updateFieldVarianceInVoxelSpace, totalFieldVarianceInVoxelSpace]
c	'Convergence', iterations per level; [levelIterations, convergenceThreshold, convergenceWindowSize]
f	Shrink factors for each level
s	(Gaussian) smoothing sigmas for each level
o	Output file path

A.3 antsApplyTransforms

Parameter	Description
d	Dimension of images
i	Path to image to be transformed
t	Transformation file
r	Name for result file (transformed image)
o	Output directory