Multi-atlas Segmentation in Head and Neck CT Scans

by

Amelia M. Arbisser

B.S., Computer Science and Engineering, M.I.T., 2011

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering
in Electrical Engineering and Computer Science
at the Massachusetts Institute of Technology

May 2012
[June 2012]

Signature of Author: ______________________________

Amelia Arbisser
Department of Electrical Engineering and Computer Science
May 21, 2012

Certified by: ______________________________

Prof. Polina Golland
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: ______________________________

Prof. Dennis Freeman
Chairman, Masters of Engineering Thesis Committee

Multi-atlas Segmentation in Head and Neck CT Scans

by Amelia M. Arbisser

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Master of Engineering

Abstract

We investigate automating the task of segmenting structures in head and neck CT scans, to minimize time spent on manual contouring of structures of interest. We focus on the brainstem and left and right parotids. To generate contours for an unlabeled image, we employ an atlas of labeled training images. We register each of these images to the unlabeled target image, transform their structures, and then use a weighted voting method for label fusion. Our registration method starts with multi-resolution translational alignment, then applies a relatively higher resolution affine alignment. We then employ a diffeomorphic demons registration to deform each atlas to the space of the target image. Our weighted voting method considers one structure at a time to determine for each voxel whether or not it belongs to the structure. The weight for a voxel's vote from each atlas depends on the intensity difference of the target and the transformed gray scale atlas image at that voxel, in addition to the distance of that voxel from the boundary of the structure. We evaluate the method on a dataset of sixteen labeled images, generating automatic segmentations for each using the other fifteen images as the atlas. We evaluated the weighted voting method and a majority voting method by comparing the resulting segmentations to the manual segmentations using a volume overlap metric and the distances between contours. Both methods produce accurate segmentations, our method producing contours with boundaries usually only a few millimeters away from the manual contour. This could save physicians considerable time, because they only have to make small modifications to the outline instead of contouring the entire structure.

Thesis Supervisor: Polina Golland
Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgments

I would first like to thank my wonderful thesis supervisor, Polina Golland, for introducing me to this project, and for offering me guidance and motivation throughout the research process. This thesis would not have been possible without the support of my advisor at MGH, Gregory Sharp. Our weekly discussions along with Nadya Shusharina allowed me to fully understand the concepts relevant to my research topic, and eventually fully define the direction of my project. It is also a pleasure to thank all of my lab mates, especially Adrian Dalca, Ramesh Sridharan, and Michal Depa, for their eagerness to help and answer any questions I had. A special thanks goes to my academic advisor, Boris Katz, for giving me encouragement and advice throughout my MIT experience.

I owe my deepest gratitude to my parents for their love and support. In particular, I truly appreciate my mom's patience in allowing me to talk out my problems throughout my thesis project, and my dad's useful critiques in the writing process. My grandparents also offered me infinite love and support, and it makes me happy to make them proud. I would like to express my appreciation for my sister, for her constant companionship, and our mutual motivation of one another. I am also grateful to all of my extended family and friends, especially RJ Ryan, Skye Wanderman-Milne, and Piotr Fidkowski for helping me talk through obstacles I encountered in my research. Finally, I would like to thank Carmel Dudley and Samuel Wang, for their critical eyes and endless patience throughout the thesis process.

Contents

List of Figures

Chapter 1

Introduction

In this thesis we explore a method for automatically segmenting anatomical structures on CT scans of the head and neck. We begin by explaining the radiation therapy motivation behind the task, followed by a brief overview of our approach. We discuss the scope of our project, and finally outline the remainder of this document.

■ 1.1 Motivation

Radiation therapy is an effective treatment for cancerous tumors, but because the radiation is also harmful to healthy tissue, detailed treatment plans are necessary before irradiation can begin. To begin planning a course of treatment, a 3-dimensional model

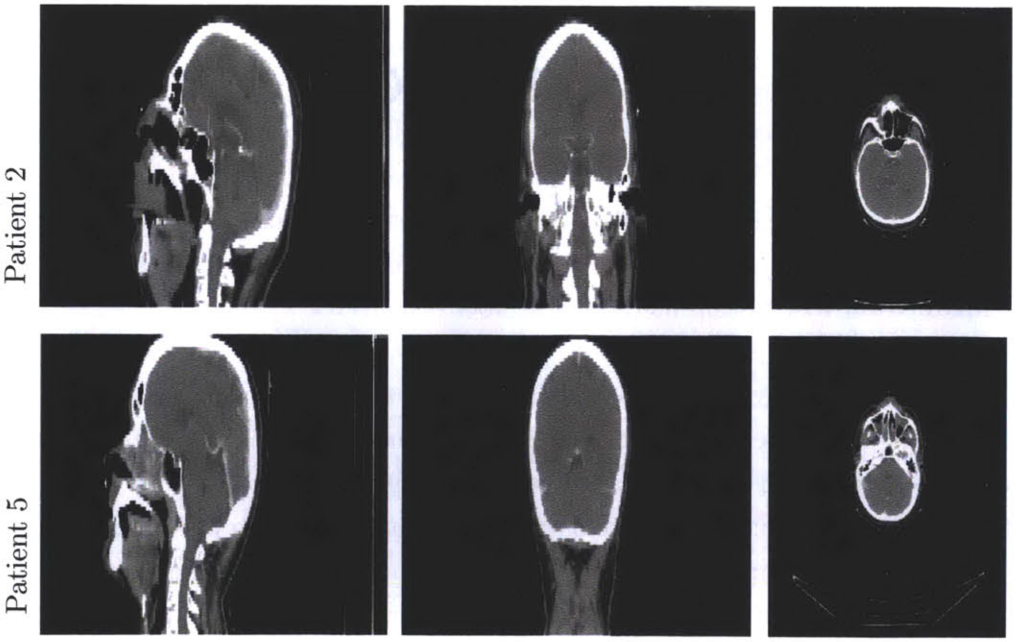


Figure 1.1: Two example CT scans.

is constructed using a CT scan of the patient. On this 3D model, the tumor is labeled, along with any sensitive anatomy in the surrounding area. A complex optimization algorithm uses these labels to compute a radiation plan that delivers the appropriate dosage of radiation to the tumor without incurring too much damage to the surrounding structures. Treatment can take anywhere from a few days to several weeks.

Currently, the 3D models are created manually. A trained technician or doctor must sit down at a computer with the CT scan, and for each axial slice of the scan, mark which voxels comprise each structure. This can take many hours. Figure 1.1 shows cross sections of two example CT scans, central sagittal, coronal, and axial slices.

The goal of this project is to automate this labeling process, at least partially, to save valuable hours of doctors' time, and potentially also reduce cost to patients. We focus on labeling the parotid glands and brainstem in CT scans of the head and neck, but the methods are applicable to other anatomical structures. Even if an automatic labeling of a scan is incomplete or imperfect, it can take less time to correct a labeling than to produce one manually from scratch.

■ 1.2 Scope

In this work, we explore a multi-atlas registration technique most similar to Depa '10 [4]. The atlas is composed of several manually labeled CT scans which we register to the target image. For deformable registration, we employ the diffeomorphic demons registration algorithm [16] after affine alignment. In addition, we experiment with a few different preprocessing steps involving cropping and intensity modification. Finally, we compare voxel-wise voting methods for label fusion.

Specifically, we apply this multi-atlas method to segment the left and right parotid glands because they are particularly challenging due to the high variability in their shape and size between patients, as shown in Figure 1.2. The contour of the right parotid is shown in green. In addition, we applied our method to the brainstem. This structure has less anatomical variation from patient to patient, however, its tissue density is more similar to that of the tissue surrounding it. It therefore poses a different challenge for our segmentation method.

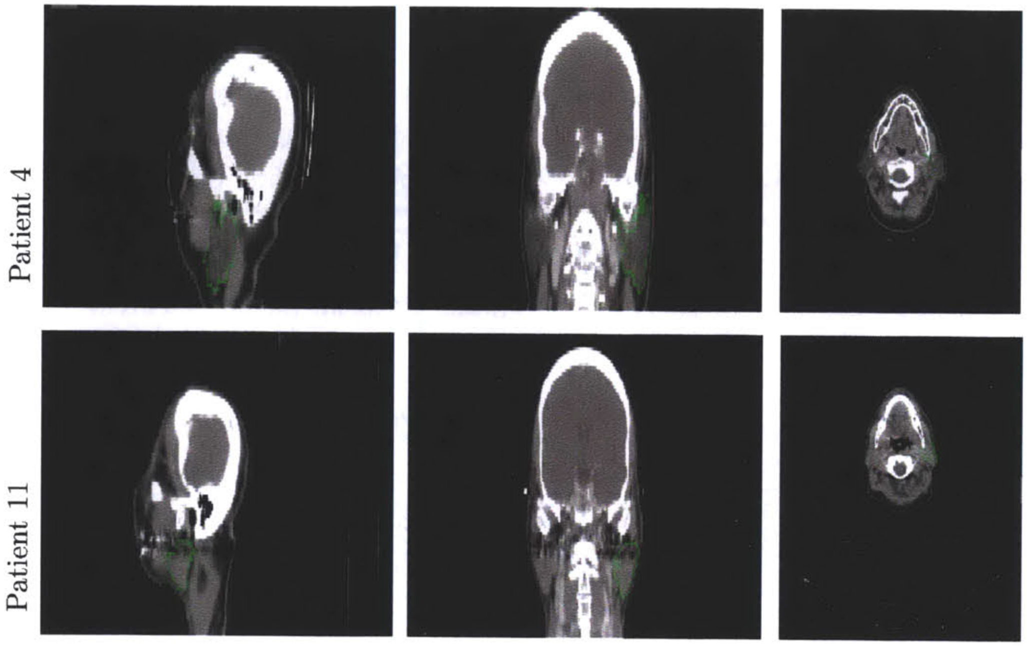
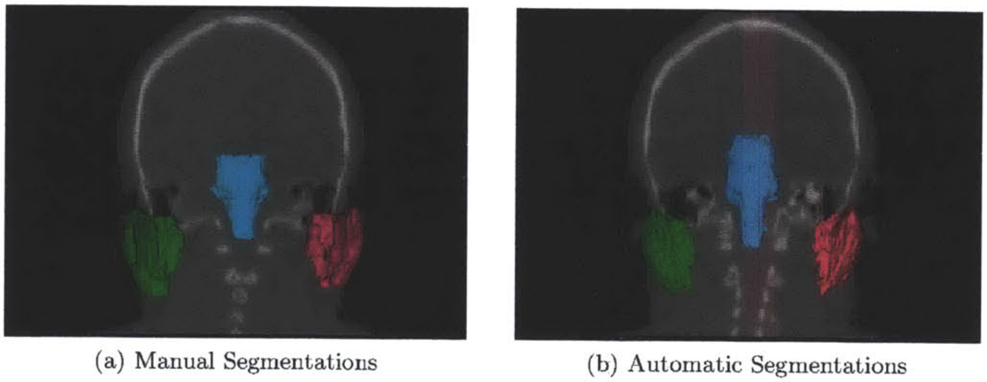


Figure 1.2: The left parotid.


(a) Manual Segmentations

(b) Automatic Segmentations

Figure 1.3: 3D renderings of our segmentation results.

We perform experiments on a dataset of sixteen manually segmented images. To evaluate the results, we compare the automatically estimated segmentations of the three structures to the manual labels using metrics for volume overlap and distances between the contours. An example segmentation can be seen in Figure 1.3.

■ 1.3 Overview

In the remainder of this document, we first discuss prior work in medical image segmentation. Chapter 3 explains the details of our method, focusing primarily on technique and only briefly describing implementation. In Chapter 4, we first describe the dataset and our experimental design. Then we present our results with plots and images, explaining their significance. In the last chapter, we discuss our contributions and draw some conclusions. Lastly, we suggest some specific areas for further research.

Chapter 2

Background

Image segmentation is the task of labeling a particular region in an image, such as an object in the foreground of a 2D image, or in our application, 3D anatomical structures. Image segmentation is a well studied problem in computer vision, with a variety of applicable methods. Some of the most naive image segmentation techniques involve simple intensity thresholding [11]. These techniques partition the image by selecting an intensity range, and determining that everything within that range should have the same label. Slightly more sophisticated methods take into account location data, enforcing contiguity of contours. Other methods employ clustering algorithms, using features such as intensity and location, to create similarity graphs. Unfortunately, these methods are not particularly effective for our purposes, because many structures consist of similar tissues and thus exhibit the same intensity in the images.

Another approach for segmenting anatomical images uses a canonical model for what the target structures are known to look like and where they are located in the average subject [10]. These models are then aligned or registered to their respective structures in the target image. While these methods are often effective for localizing the position of the target structures, they do not always do a good job of discovering the boundaries of the structures. High variability in the shapes of anatomical features between patients makes using a single anatomical model less feasible, because it is very likely that the target structure will have a substantially differently shape from that of the canonical structure.

To account for this inter-patient variability, multi-atlas-based segmentation methods can be used. Instead of assuming a single general model for all patients' structures, multi-atlas-based methods take a set of previously segmented scans of subjects other than the target subject, and use the intensity information from the images along with their labels to construct the label for the target structure. Multi-atlas techniques can be broken down into two primary steps. First, the atlas images must be aligned with the target image. This often involves both moving the subjects into the right position and orientation, and non-rigidly deforming the images. The resulting registration transform is then applied to the structure labels, moving them into the space of the target image. Secondly, the multiple labels resulting from the registration step must be reconciled to form a single structure label for each pixel in the target image. This step is commonly referred to as label fusion.

■ 2.1 Registration Method

One of the most important parts of atlas-based segmentation techniques is which registration method to use. Almost all methods begin with an initial non-deformable transformation step, to align the subjects at the same location in the same orientation. This can be accomplished using a well known optimization method such as gradient descent. The choice of metric to be minimized can have a substantial impact on the effectiveness of this registration step. Most techniques employ some combination of Mutual Information (MI) [18] and Sum of Squared Differences (SSD) [14] of the intensities of the moving and target images. For example, Han '10 [7] registers each atlas image to the target by iteratively optimizing a weighted sum of the mutual information metric and a normalized-sum-of-squared-differences metric.

Another variation in registration method is the direction of registration. Most often atlas images are registered directly to the space the target image, however, each of these registrations can be computationally expensive. To compensate for this, some methods register all images to an average common space [13] [8]. Thus when a new target image is received, only the registration of the target to the common space must be computed, because all the atlas registrations could be done ahead of time. In addition to saving time, this method has the benefit that most images are likely more similar to the average image that they are registering to than they would be to any single image. Registering two more similar images can yield a more accurate registration. However, if the target image is particularly far from the average image and there were atlas images that were more similar to the target image, we cannot take advantage of that similarity like we could if we were registering the atlas images directly to the target.

Arguably the most important part of registration is the non-rigid deformation method. There are many techniques to choose from, such as cubic B-spline [9], locally affine [3], and contour-based methods [1]. In this work we use a diffeomorphic demons registration [15] for non-rigid registration, after performing affine registration first. It is also possible to use a combination of these methods, for example, beginning with a locally affine approach and refining with a contour based method [7].

■ 2.2 Label Fusion

The next step in multi-atlas-based segmentation methods after registration is label fusion, which determines a single label for each structure in the target image from the potentially multiple labels resulting from registering the atlas images. One of the simplest ways to do this is to select a single atlas whose transformed structure label will act as the label for the target image. There are many ways to select this single atlas image. For example, mean squared error (MSE) of intensity values in the registered atlas image and the target image can be a good indicator of how well the two images registered. We would then select the atlas image with the lowest MSE because the structure is probably most similar to the target's structure in the atlas image that is close to the target image in intensity. Any number of complex metrics to select a single atlas image may be devised. Selecting only one atlas image is based on the assumption that some atlas image has a structure very similar to the target structure.

By selecting which atlas label to use locally instead of globally, we can take advantage of local similarity between the target image and each atlas image. One way to do this is a voxel-wise vote, where we look at the problem as selecting the correct label (i.e., structure or background) for each voxel separately. Each transformed atlas structure can then be seen as casting a vote for each voxel. A simple way to reconcile these votes would be a simple majority vote, where the final label for a voxel is determined by which label has more votes from the transformed atlas structures.

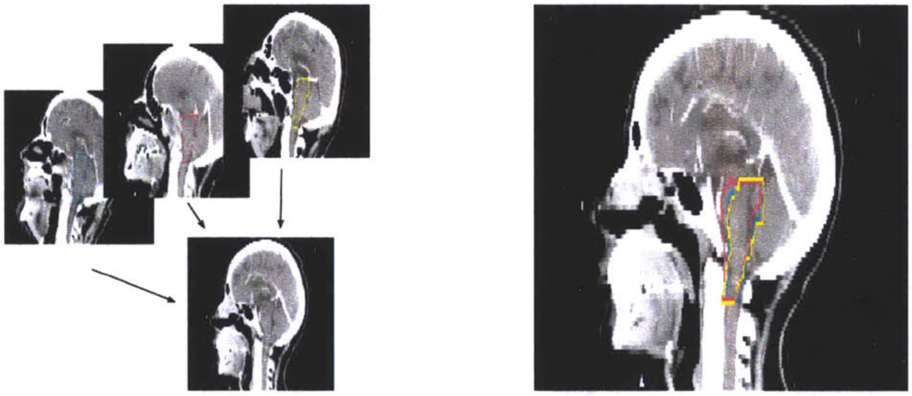
The next logical extension would be to employ weighted voting. This can be done at the level of the entire atlas image, assigning the same weight to every voxel of the atlas image based on the similarity between the atlas image and the target image. Alternatively, it can be done locally, for example, separately at each voxel. One popular technique for doing this is the STAPLE algorithm [7, 17]. STAPLE was initially developed to assess the accuracy of several different segmentations for the same structure in an image by effectively inferring a ground truth segmentation for the image. This is done using expectation maximization to jointly estimate a specificity and sensitivity parameter for each individual segmentation along with the ground truth segmentation. Other methods calculate weights for each atlas structure at the voxel level directly [4, 13]. These methods look at each voxel and assign a weight from each atlas based on features like local intensity similarity.

Chapter 3

Methods

■ 3.1 Overview

Our multi-atlas segmentation method starts with an atlas of N images, $\{I_1, I_2...I_N\}$ and their labels for the relevant structure $\{L_1, L_2...L_N\}$. $I_n(x)$ is the intensity of voxel x in atlas image n; $L_n(x) = 1$ if the voxel is part of the structure and $L_n(x) = -1$ indicates that it is not. Given an unlabeled target image I, we register every atlas image to the target image. Thus we calculate the transform ϕ_n that describes how to deform atlas image I_n to the target image I. We then apply the transforms ϕ_n to the label image L_n, resulting in N transformed labels for the target structure in the space of the target image. This process is illustrated in Figure 3.1, where we see three atlas subjects aligning to the target subject in (a), and then the transformed atlas labels superimposed on the target in (b), ready for label fusion.



(a) Registration

(b) Label Fusion

Figure 3.1: An illustration of the multi-atlas segmentation method.

To fuse these N labels for the structure into a single label for the target image, we use a voxel-wise weighted voting algorithm. The algorithm takes into account both the local similarity in intensity of the two images, and the distance from the structure boundary in the transformed atlas image.

■ 3.2 Preprocessing

Because the location of the cancerous region is different in each patient, the acquired CT images often have very different fields of view. Not all contain the same regions of the patient. In addition, some have artifacts like the couch of the scanner, which can mislead the registration algorithm and are irrelevant to the information of interest, the anatomical structures. Thus, we apply preprocessing to ensure that the assumptions the registration method is making about the similarities of the images are correct.

■ 3.2.1 Cropping

The first preprocessing step that we explored was cropping the images. Head and neck CT images usually include a significant portion of the shoulders. Because all the structures of interest are within the skull and only slightly below, the shoulders were unnecessary. In addition, they sometimes cause a misalignment of the skull, because not all patients have the same angle between their head and shoulders. This is still somewhat of an issue in the neck, but is much less so because the neck is so much smaller than the shoulders, and the intensity differences are not quite as pronounced.

■ 3.2.2 Masking

To eliminate irrelevant artifacts and objects around the edge of the image, we applied a cylindrical mask inscribed in the volume of the image. This mask set the value of all voxels outside the inscribed cylinder to the same intensity as air, about -1000 HU. Figure 3.2 illustrates this process. In the first CT scan you can see the edges of the couch at the right of the sagittal slice and the bottom of the axial slice. The yellow part of the cylindrical mask leaves the image intensities as they are, while all voxels in the black region are set to the HU of air, which comprises most of the background already. In the third image the couch has been masked out and can no longer be seen.

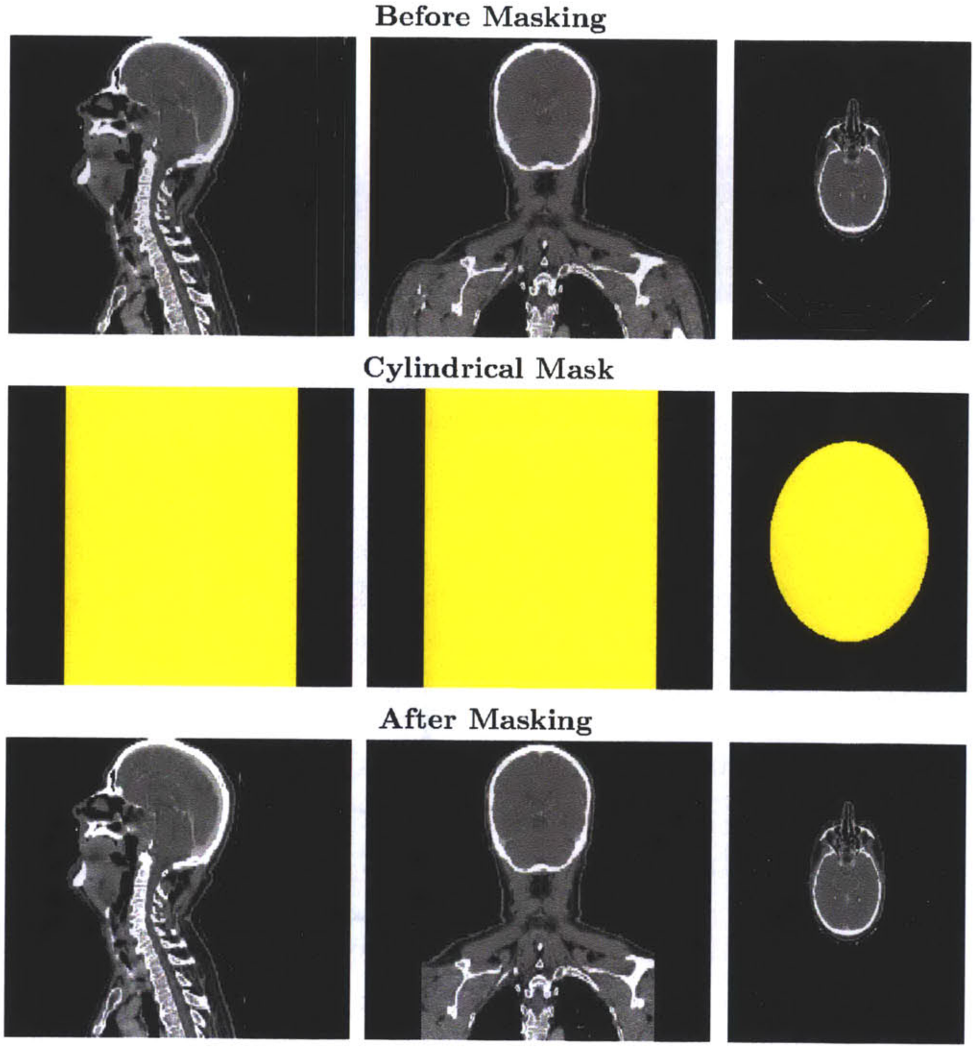
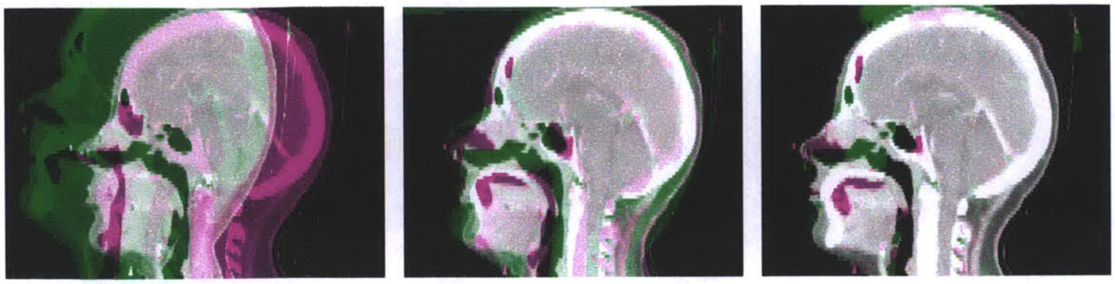


Figure 3.2: Cylindrical mask application.

■ 3.3 Registration

After preprocessing, each atlas image I_n must be separately registered to the target image I. First we appy a non-deformable registration that moves each atlas subject into the same location and position as the target subject. This step transforms the image as a whole, translating, rotating, and stretching the image. Next we apply the diffeomorphic demons registration to get the final deformation field ϕ_n which allows the atlas image to move more freely so that it can better conform to the shape of the target



(a) Before registration. (b) After affine alignment. (c) After demons registration.

Figure 3.3: Registration process.

subject. In Figure 3.3, an example atlas image (in green) is overlaid on the target image (in magenta) at three points during registration. Gray indicates where the intensities of the two images are similar, green shows areas where the atlas image intensity is higher, and magenta indicates where the target image intensity is higher.

3.3.1 Non-deformable Registration

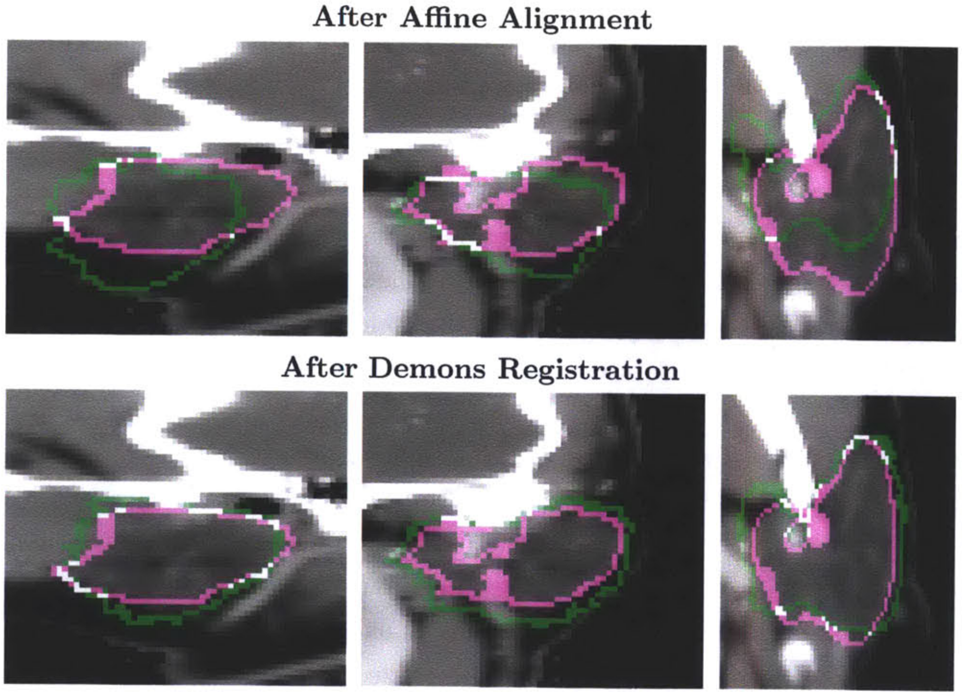
The aim of the non-deformable registration step is to move the atlas image so that the subject is in the same basic position as the target subject. We first apply simple translational alignment, shifting the image in the x, y, and z directions until the subjects are maximally overlapping. We use the sum of squared differences between the intensities in the atlas image and the target image as a metric for this overlap. The actual amount that the atlas image must be shifted in each direction is found using gradient descent.

We experimented with starting with rigid alignment, which allows for rotations in addition to translations. However, this was too unstable and would often result in extreme rotations of the patient. When we first translate the subjects into the correct position, gradient descent for affine alignment is more effective. Affine registration allows rotations, scaling, and shearing. This permits us to compensate for differences in the angle of the subject's head and in the overall size of the patient. In practice, very little shearing occurs.

Both of these registration steps are performed at multiple resolutions. That is, before performing gradient descent at the full resolution of the image, the atlas image and target image are subsampled at a lower resolution and the registrations are computed on those smaller images. When the registration is computed at the next resolution level, the atlas image starts from its transformed position estimated in the previous step. This helps prevent gradient descent from getting stuck in local minima.

These registration steps are performed using Plastimatch's [12] implementation of multi-resolution registration via gradient descent.

■ 3.3.2 Deformable Demons Registration



After Affine Alignment

After Demons Registration

Figure 3.4: The demons algorithm applied to the left parotid.

Once we have the subjects in approximately the same position, we can start to deform them to improve the alignment of individual parts of their anatomy [15]. These deformations are necessary because there is a lot of variation in the shape and relative size of anatomical features. For example, some people's eyes are farther apart than others. This deformable registration step allows us to account for all the inter-subject variability in the shape of the target structures such as the parotid. The diffeomorphic demons registration allows us to correct these differences in proportion while maintaining smoothness and structure in the image.

Figure 3.4 illustrates the effect of the demons algorithm on a left parotid. The grayscale images are the sagittal, coronal, and axial slices of the target image surrounding the parotid, outlined in pink. The green outline on the upper three images shows a transformed atlas structure after affine alignment. In the bottom images, the result of the non-rigid demons transform applied to the affinely aligned structure is shown in green.

Demons registration works iteratively, updating a velocity field that deforms the moving atlas image along the intensity gradient of the target image. This field is calculated by finding the velocity field u that optimizes the following energy function:

$$E(I, I_n, \phi, u) = ||I - I_n \circ \phi \circ \exp(u)||^2 + ||u||^2 \tag{3.1}$$

were I and I_n are the target and atlas images respectively, ϕ is the transformation estimated at the current iteration, and $\exp(u)$ is the deformation field that corresponds to the velocity u. We then smooth the resulting velocity field u by convolving it with a regularization kernel, and iterate.

We employ a multi-resolution scheme for demons registration as well. The registration algorithm will push too far along inconsequential gradients if allowed to run for too long. That is, it will pick up on intensity differences that are not indicative of any anatomical structure in the image. To allow for the most flexible, smooth registrations we first perform alignment at lower resolutions to move larger regions of the patient to overlap, and then increase the resolution to localize the boundaries with better precision. This way we can deform the image far from its initial position without introducing too much non-linearity in the deformation field.

For the diffeomorphic demons registration, we use the Insight Toolkit implementation of the symmetric log-domain diffeomorphic demons algorithm [6].

3.4 Label Fusion

Registration produces N transformed structure labels $\tilde{L}_n = L_n \circ \phi_n$, one from each atlas subject and corresponding deformation. Figure 3.5 shows how these N labels do not agree on the same label for each voxel. Our goal is to find a single label L for the structure in the target image. We must then decide how to fuse these N suggested segmentations into a single label.

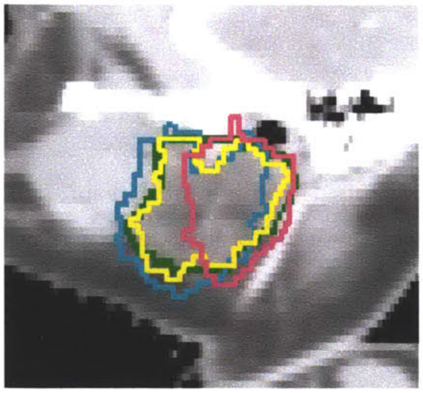


Figure 3.5: Transformed atlas labels failing to align.

We determine for each voxel whether or not it is in the target structure. The problem then reduces to a voxel-wise decision on how to integrate the N binary indicators from each of the transformed atlas structures into a single label $L(x)$.

3.4.1 Weighted Voting

In deciding on the label L, we weight the votes from each atlas image, to give higher weight to images that we believe are more likely to indicate the true target label. While we can not know directly which structures are better aligned, we can use clues from the intensities of the images. The idea is that when the images are better aligned their intensities will be more similar, and when they are poorly aligned the intensities likely will not match.

In addition to this weighting based on differences in intensity, we also consider the distance from the boundary of $\tilde{L}_n$. The intuition here is that we are less certain about labels on voxels near the edge of a structure. This is because boundaries are where human error can occur in drawing manual labels on the axial slices.

We select the label for each voxel by choosing the label that maximizes the joint likelihood of the label and the intensity of the target image, given the label and intensities of all the transformed atlas images.

$$\text{Vote}(x) = \max\{p_1(x), p_{-1}(x)\} \tag{3.2}$$

$$p_l(x) = \sum_{n=1}^{N} p(I(x)|I_n, \phi_n)p(L(x) = l|L_n, \phi_n) \tag{3.3}$$

The first term of p_l represents the difference in intensities. It is equivalent to the probability that one intensity was generated from the other. That is, $I(x)$ is sampled from a Gaussian distribution centered at $(I_n \circ \phi_n)(x)$. This probability is lower if the intensities are very different.

$$p(I(x)|I_n, \phi_n) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(I(x) - (I_n \circ \phi_n)(x))^2}{2\sigma^2}) \tag{3.4}$$

The second term encapsulates the distances from the contour of the atlas structure. It gives an exponentially lower weight to voxels that are very near the boundary of $\tilde{L}_n$, the transformed structure from atlas image n. ρ is the rate parameter of the exponential distribution.

$$p(L(x) = l|L_i, \phi_n) = \frac{\exp(\rho D_n^l(x))}{\exp(\rho D_n^1(x)) + \exp(\rho D_n^{-1}(x)))} \tag{3.5}$$

$D_n^l(x)$ is the signed distance transform, defined as the minimum distance from x to a point on the boundary or contour of $\tilde{L}_n$, denoted $C(\tilde{L}_n)$. We let $D_n^l(x)$ be positive if $\tilde{L}_n(x) = l$, meaning that x is within the structure, and negative if $\tilde{L}_n(x) \neq l$, x is outside the structure. The superscript l indicates which is the structure of interest.

$$D_n^l(x) = \tilde{L}_n(x) * l * \min_{y \in C(\tilde{L}_n)} \text{dist}(x, y) \tag{3.6}$$

where dist is the Euclidean distance between voxels x and y. This term is illustrated in Figure 3.6. For a voxel x, we see the distance $d = D_n^l(x)$ is the distance from x to the closest point y on the boundary of the structure $\tilde{L}_n$.

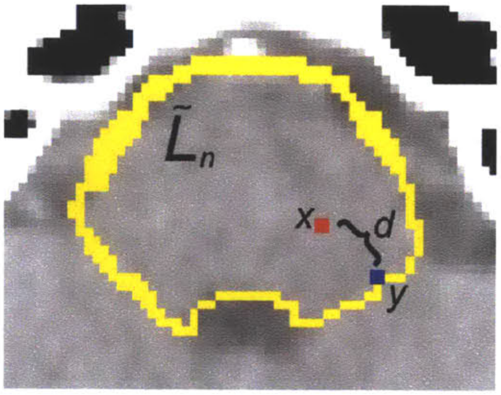


Figure 3.6: Boundary distance illustration.

■ 3.4.2 Thresholding

Initial experiments revealed that strictly maximizing Vote(x) resulted in consistent undersegmentation of the target structure. To compensate for this, we introduce an additional parameter t as a threshold for the likelihood. When we maximize Vote(x) we calculate values p_1 and p_{-1} for $L(x) = 1$ and for $L(x) = -1$, respectively. If we normalize $p_1(x)$ and $p_{-1}(x)$ such that we have $\tilde{p}_1 = \frac{p_1}{p_1+p_{-1}}$, we are effectively thresholding the $\tilde{p}_1(x)$ at $\frac{1}{2}$. By varying the threshold t we can make it more likely to select $L(x) = 1$ to overcome the undersegmentation.

For each structure that we are attempting to segment in the target image, we use this weighted voting method at each voxel x to determine whether or not that voxel is part of the structure. This leaves us with a single label, indicating a set of voxels, for each anatomical structure in the target image.

Chapter 4

Experiments

In this section we describe the dataset used in evaluating the methods, and explain the setup of the experiments. We then present the results.

■ 4.1 Dataset

We evaluate the method on a set of sixteen CT scans of the head and neck, each one depicting a different patient. Each image was labeled by a trained anatomist for treatment planning. There were over 60 unique structures labeled across the patients, but most patients have only a subset of all 60 labels, depending on which structures were most relevant for that patient's treatment.

■ 4.1.1 CT Scans

Computed tomography (CT) scans are comprised of multiple axial slices of the subject. These slices are generated by shooting X-rays at the patient in the plane of the slice. The image for each slice can then be reconstructed algorithmically based on how much of the X-ray is blocked at each point along the perimeter of the slice. These slices are then combined to form a three-dimensional image.

Each voxel in the resulting image contains a single value representing the radiodensity of tissue at that position. The units of this value are Hounsfield units (HUs), where -1000 is the density of air and 0 is water. The scale is cut off at 3000 which is about the HU of dense bone. Soft tissue, which comprises most of the structures we are interested in ranges from -300 to 100 HUs.

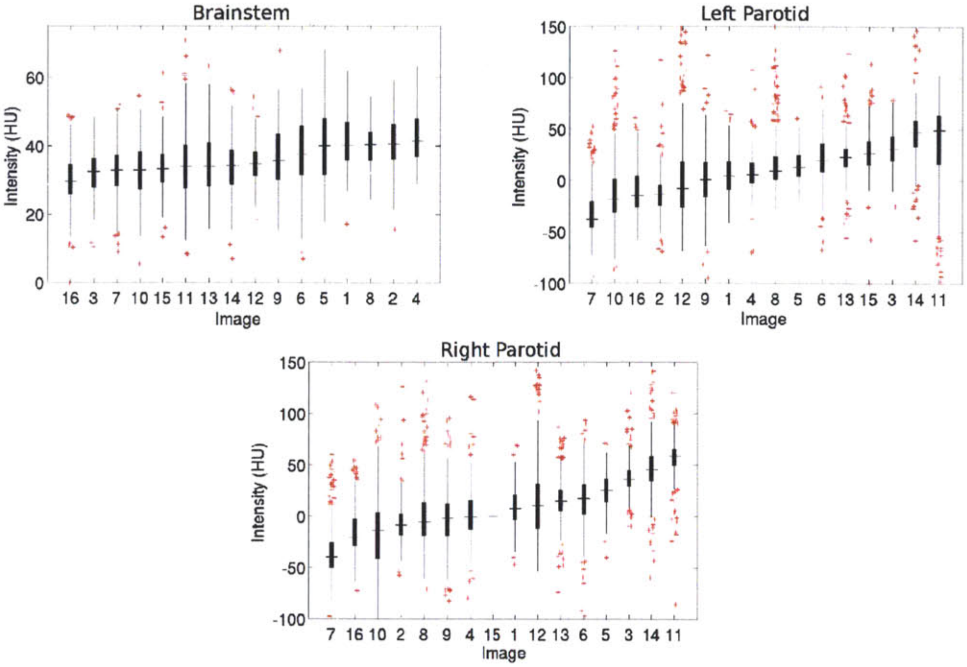


Figure 4.1: Histograms of structure intensities.

Figure 4.1 shows the intensity distribution within each structure, brainstem and parotid glands, for each of the sixteen images. For each structure, these boxplots show the intensity distribution of a 250 voxel sample of voxels within a 3mm margin inside the boundary of the structure. The plots are sorted by median intensity. These intensities coincide approximately with reported soft tissue regions, but more notably, the intensities within the structures are very different. That is, they differ more from patient to patient than the intensities differ within any given patient's region of interest, and likewise for the surrounding tissue.

Each image consists of somewhere between 80 and 200 axial slices, each containing 512x512 pixels. The number of slices varies from patient to patient because not all images include exactly the same field of view. For example, some images are truncated at the top of the skull while others include the entire head. Also, some scans include the patients' shoulders while others stop at the neck.

The resolution of the images is slightly different for each patient, but is usually

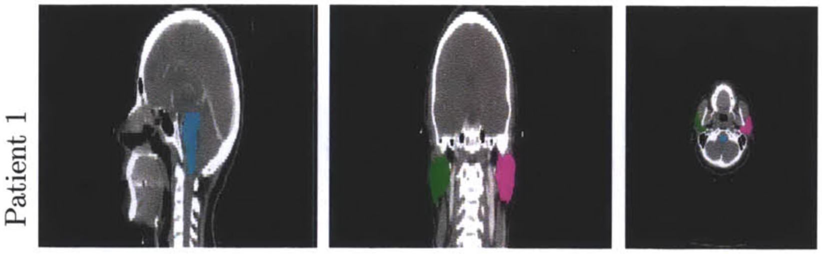


Figure 4.2: A labeled CT scan.

around .9mm per voxel in the axial plane, except for one patient whose resolution is only .48mm/voxel. Each slice is 2.5mm thick. Each of the labels is a separate image with the same resolution and dimensions as its corresponding patient's CT scan. The labels contain only binary values, simply indicating whether or not each voxel is contained within the relevant structure. All sixteen images have all three structures labeled, except Patient 15, whose right parotid has been consumed by a tumor. Figure 4.2 shows cross sections of a CT scan with brainstem (cyan), left parotid (green), and right parotid (magenta) labels overlaid.

4.2 Experimental Setup

We perform sixteen experiments in which we remove the labels from one of the images and use the other fifteen images as the atlas. We then evaluate the results of voting using the Dice score [5], Hausdorff distance [2], and median distance between boundaries, as explained below.

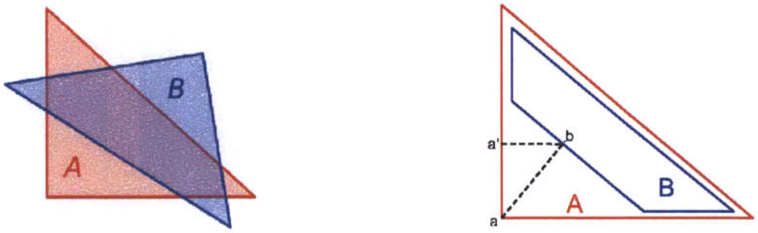
4.2.1 Evaluation

Because we have manual labels for all patients, in each experiment we compare the estimated label with the manual one. We employ several metrics to provide a quantitative evaluation of the segmentation accuracy.

Dice

The Dice score is a general metric for representing the similarity of sets.

$$\text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{4.1}$$



(a) Dice Score (b) Hausdorff Distance

Figure 4.3: Illustrations for Dice and Hausdorff metrics.

In our case A and B are the two labels we are attempting to compare, the manually labeled structure and the automatically estimated label for the same structure. We consider them as sets of unique voxels contained within the structure. The Dice score can be thought of as a measure of volume overlap between the two labels. Figure 4.3(a) illustrates the Dice score for two 2D triangles. The metric indicates the ratio of the area of the overlapping region of the triangles and the sum of both their areas.

Distance Metrics

While not a true distance, because it is not symmetric, the Hausdorff distance indicates the maximum distance between the contours of the two structures:

$$\text{Hausdorff}(A, B) = \max_{x \in C(A)} |D^B(x)|, \tag{4.2}$$

where $C(A)$ is the contour of label A. $D^B(x)$ is the minimum distance from voxel x to the nearest point in the contour of B. The Hausdorff distance is then the maximum of these distances over the points in the counter of A, $C(A)$.

In addition to this maximum distance, it can also be useful to look at the median distance between the two boundaries. This gives us a better idea of how close the boundaries are in general, while Hausdorff just gives us the worst case.

Figure 4.3(b) illustrates the Hausdorff distance. We can see how the asymmetry arises, where $\text{Hausdorff}(A, B) \neq \text{Hausdorff}(B, A)$. This is because the point on the contour of B, b, farthest from any point, a, on the contour of A, may be closer to another point on $C(A)$, a'. Thus, the Hausdorff distances can be different.

Because the Hausdorff and median distance metrics are not symmetric, we summarize the scores by taking the maximum of Hausdorff(A, B) and Hausdorff(B, A) and the average of the two median distances.

4.2.2 Parameter Selection

Voting Parameters

The two parameters for the voting method, the intensity difference standard deviation σ and the contour distance scaling ρ, needed to be set. In addition, we had to set the threshold t for Vote(x).

For each experiment, we ran a grid search of the parameter space, calculating average Dice and Hausdorff metrics for voting on each of the remaining fifteen images, using the last fourteen as the atlases.

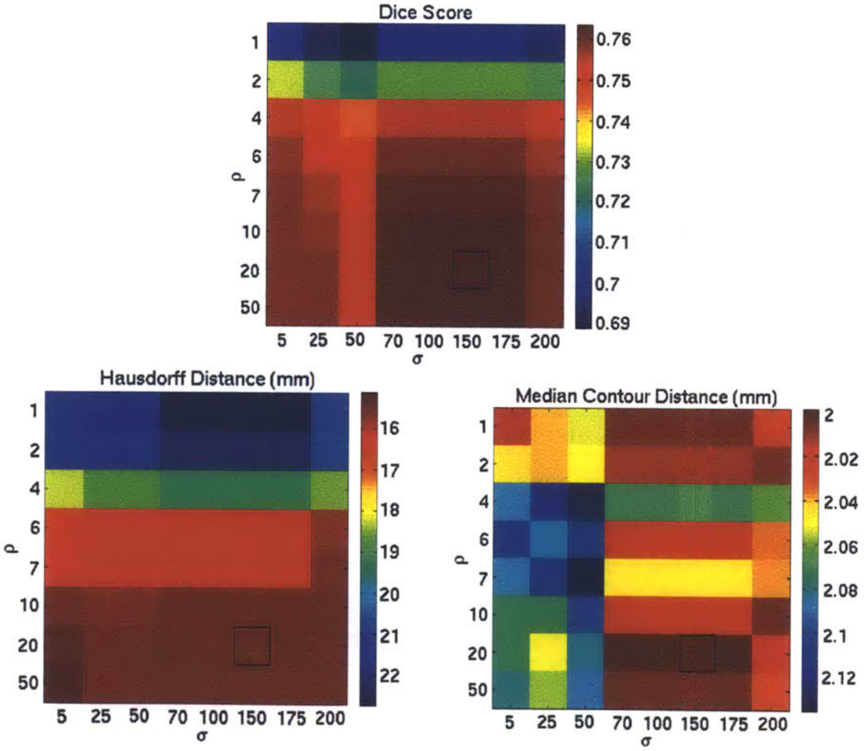


Figure 4.4: Voting parameter selection colormaps.

For the threshold t, a value of 0.2 was consistently optimal across all selections of the other two parameters. Results of voting for different values of σ and ρ with $t = 0.2$ are shown in Figure 4.4. The colormaps show, for an example target image, the mean Dice scores, Hausdorff distances, and median contour distances for the left parotid, using different values of σ and ρ, with $t = 0.2$. Each mean was calculated from the results of 15 voting experiments, using an atlas of 14 images.

Because maximizing the three metrics yields different values for σ and ρ, we selected values where the hot spots of the colormaps coincided, with high values for Dice scores and low values for distances, resulting in values around $\sigma = 150$ and $\rho = 20$ for the parotids, indicated by the black square in the colormaps, and $\sigma = 50$ and $\rho = 3$ for the brainstem.

Registration

The registration method did not lend itself as easily to simple parameter grid search because there were too many parameters, when considering which type of registration to do at each step, at what resolution, and for how many iterations. Thus, we trained by hand on a subset of five images, computing 20 pairwise registrations for each parameter setting. We began by finding resolutions and iteration levels for translational alignment alone, and then moved on to affine. We evaluated the optimality of a given registration by calculating the average Dice and distance metrics between the transformed atlas structures and the manually labeled target structure.

Number of Iterations at Each Resolution Level

Registration Type	40x40x10	10x10x4	4x4x2	2x2x1	1x1x1
Translational	30	30	30	None	None
Affine	None	10	30	None	30
Demons	None	None	20	20	50

Table 4.1: Registration pipeline.

Once we found a good pipeline for non-deformable registration, we optimized the demons registration. For the additional parameter s for the smoothness regularization kernel, we found a smoothness of $s = 3$ to produce the best label overlap measured by the Dice and distance metrics, on the same subest of five images.

The registration pipeline we used is summarized in Table 4.1. It shows, for each type of registration, the number of iterations run at each resolution. Resolution is shown as XxYxZ for subsampling rate in millimeters, in the X, Y, and Z dimensions. The steps were run in the order listed in the table.

■ 4.3 Results

This section presents the results of the sixteen voting experiments, using the registration method described above.

■ 4.3.1 Voting Results

The results of the experiments are shown below in Figure 4.5 compared to the results of using a simple majority voting scheme. The blue bars indicate the average results of weighted voting, the red indicate the same for majority voting. The black error bars indicate the standard deviation of the results. Both methods are fairly comparable in their performance across all metrics in all three structures.

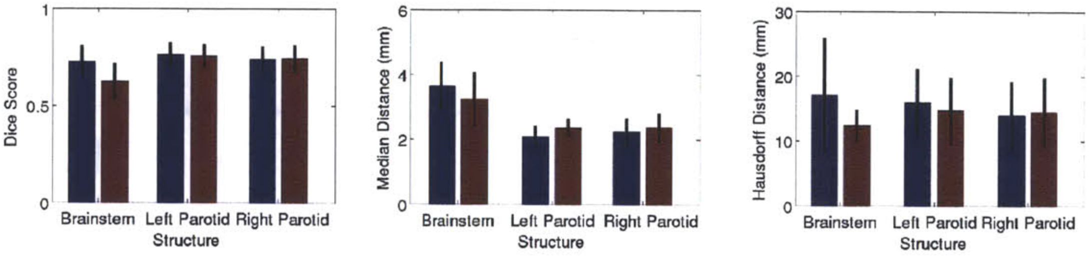
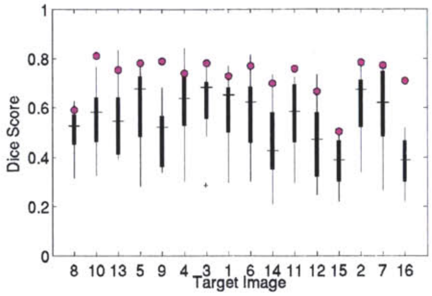


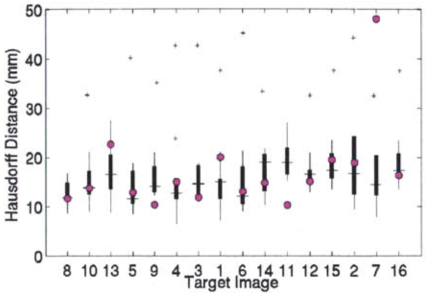
Figure 4.5: Weighted versus majority voting results.

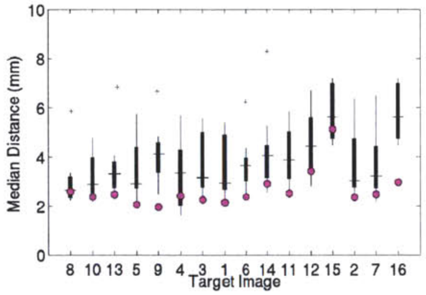
■ 4.3.2 Single Atlas Results

In Figure 4.6 we compare the results of weighted voting to Dice and distance metrics for using a single atlas. That is, for each target image, we evaluate the fifteen warped atlas labels, and compare those distributions, shown as boxplots, to the results of voting, the pink circles. The label produced by the weighted voting method does consistently better than any single atlas could have done in terms of Dice score in almost all cases. It also does consistently better than the median single atlas for the distance metrics.

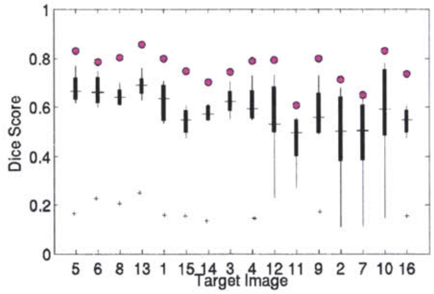
Brainstem

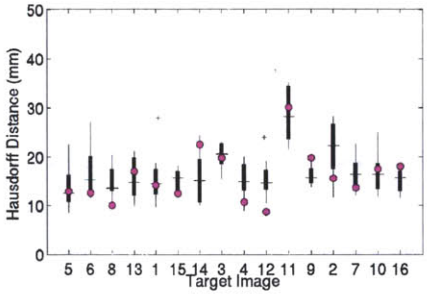


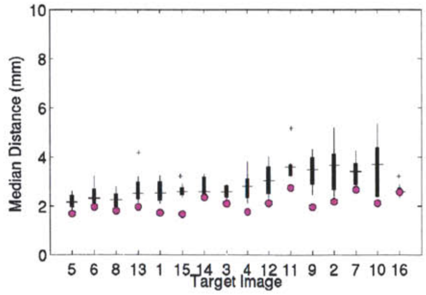




Left Parotid







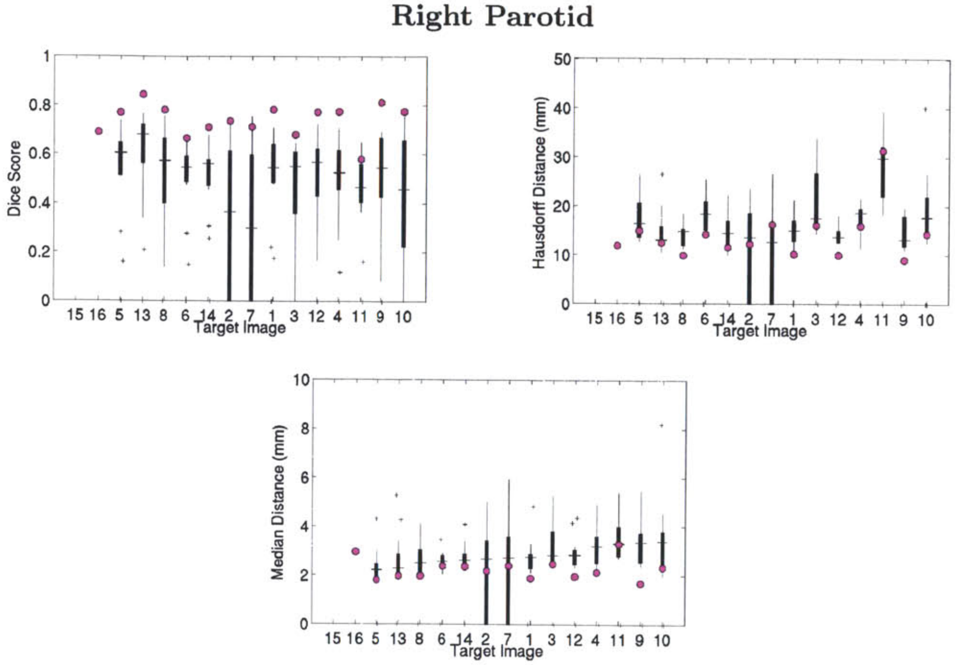


Figure 4.6: Single atlas boxplots vs. voting results.

This result suggests that adding more atlas images might be helpful, because the voting method is capable of doing consistently better than the median of all single atlases. With more training subjects in our atlas, it is likely that a single atlas image will correspond very well to the target image, which will in turn improve the performance of the voting result.

4.4 Segmentation Examples

Here we present several examples of the voting algorithm in action and discuss the results.

Figure 4.7 illustrates the undersegmentation of the weighted voting algorithm compared to a simple majority voting scheme, where each atlas image has the same weight at each voxel. The upper images show the values of p_1 for voxels in sagittal, coronal, and axial slices of the left parotid from Patient 5. Brighter colors of magenta indicate higher probability of being included in the structure. Similarly for the lower images, the colors indicate the proportion of atlas images that voted for each voxel to be included

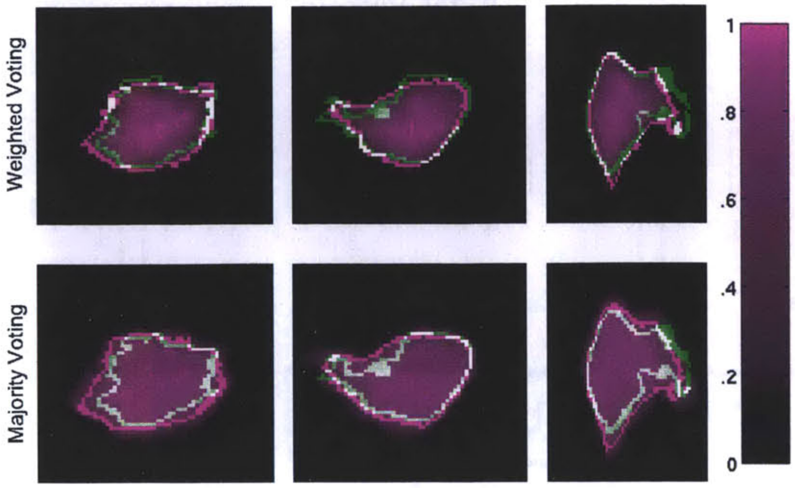


Figure 4.7: Comparison of majority and weighted voting probability values for the left parotid.

in the structure. The manual contour is shown in green, and a bright pink contour indicates the contour resulting from thresholding the weighted and majority votes at $t = 0.2$ and $t = 0.5$ respectively. While the majority voting probabilities fade smoothly from zero to a wide plateau of values greater than 0.5, the weighted voting probabilities increase sharply just inside the contour of the manual label and then maintain a higher plateau within, closer to 0.8. This is due to the contribution of the distance term to the weight of an atlas image's vote at a given voxel. Because the voxels closer to the edge of target structure are likely to be near the edge of the transformed atlas structures, they have lower weight, and consequently have a lower probability of being assigned as part of the target structure.

In Figure 4.8 the contours of segmentations resulting from three different values of the threshold t. The intensity map beneath the contours indicates the values of $\tilde{p}_1$ for the corresponding voxels. The outermost contour in green is the contour resulting from $t = 0$. In addition the manual contour is shown in blue. This segmentation includes all voxels that had a value of $\tilde{p}_1 > 0$, meaning they were classified as part of the structure by at least one warped atlas label. The innermost contour in yellow is the contour of the segmentation resulting from a threshold of $t = 0.5$. This segmentation is clearly much smaller than the segmentation illustrated by the blue contour. The pink contour

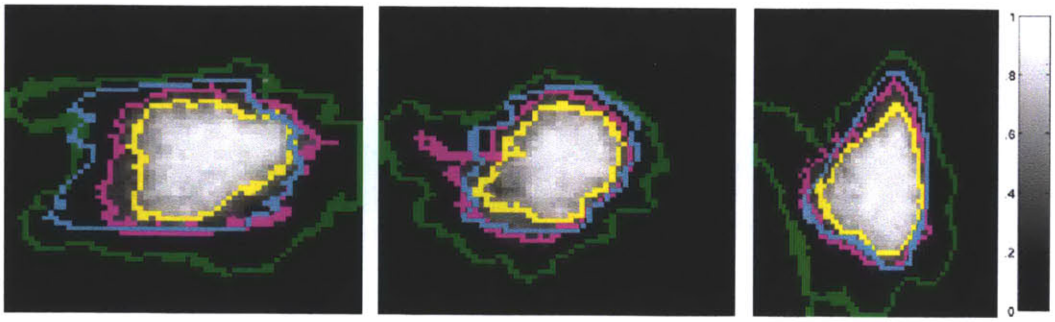


Figure 4.8: Contours for different values of t in weighted voting for the left parotid.

outlines the boundary of the region thresholded at $t = 0.2$, which we found to be most similar to the manual segmentation.

Voxels in the target structure that were not included in at least one transformed atlas structure are particularly problematic. This is because it is impossible for the voting algorithm to classify them as part of the target structure because none of the atlas images are voting to include them. The registration step would need to be changed to potentially include these voxels. Fortunately, only a few images contain such large regions of voxels in the target structure not in any transformed atlas labels. A worst case example is shown in Figure 4.9, where the manual label is shown in green with the $t = 0$ segmentation overlaid in magenta and the overlap shown in white. The green region of the manual label is guaranteed to be excluded from the segmentation since no atlas label includes it when registered to the target image.

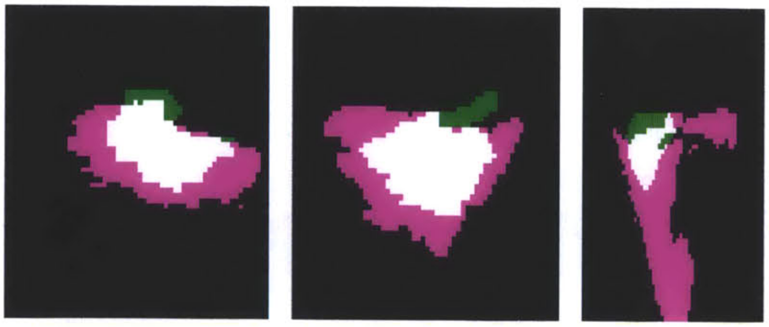


Figure 4.9: Worst case undersegmentation.

Chapter 5

Discussion and Conclusions

In this section we discuss the potential impact of this work and directions for future research.

■ 5.1 Contributions

Through this experiment in automatic segmentation, we have found that the method is very accurate at finding the location of anatomical structures, though not as accurate at delineating the boundaries of the structures. For radiation treatment planning, it is important to know where the structures are, but it is also necessary to know the contour with an error of less than a few millimeters. While our method often produces results within this margin of error, we offer no way to tell, without manual verification, when this is not the case. However, the manual verification and correction of existing contours can take significantly less time than creating a contour manually.

The deficiencies of the method lie in its failure to identify the boundaries of structures. The root of this problem is in the registration step, in that many atlas images do not align perfectly with the boundaries of the target structure. The demons algorithm is powerful enough to recognize and accurately match these boundaries in many cases, but it is difficult to find parameters that work uniformly across patients. This may be due to the variation in intensity, because the velocity field updates are affected by the difference between the two structures' intensities in the two images.

Ideally the label fusion method would be able to distinguish between patients whose structure boundaries aligned correctly and those that did not, but this is not always the case with our weighted voting scheme. While there is often a difference in intensity right on the boundary of the structure, the surrounding tissue can be of a very similar density and image intensity. The voting method has no way to tell which side of the boundary it is on, and while it may assign lower weight to voxels on the boundary itself, which likely have a larger intensity difference, it will assign just as much weight to a voxel on the wrong side of the boundary as to a voxel inside the structure. One potential way to compensate for this would be to include a term in the voxel weight proportional to its distance from a large intensity gradient in the target image, which likely indicate boundaries. This would allow us to give higher weight to atlas images whose boundaries aligned correctly by penalizing atlas labels whose boundaries do not align with gradients in the target image.

■ 5.2 Future Work

There are many extensions that could be made to this work, both to improve results and to further analyze the challenges this application presents for segmentation methods. In this section we present some of the directions that promise to be most useful for improving segmentation accuracy.

■ 5.2.1 Intensity Normalization

One option that we explored was modifying the intensities of the atlas images prior to alignment. Because the structures of interest consist of soft tissue and are primarily surrounded by soft tissue, it would be ideal to enhance intensity differences in this region. This can be helpful in registration because intensity differences in the soft tissue region will be penalized more heavily, forcing the registration algorithm to focus more effort on aligning those regions correctly. In the weighted voting stage as well, these enhanced intensity differences can help better indicate which atlas images registered well and should be given higher weight.

To create this exaggerated soft tissue intensity contrast, we tried applying an intensity mapping function, show in Figure 5.1, to all of the images before registration. The parameters a, b, c, d and a', b', c', d' were set by examining the intensity distributions of the regions surrounding and including the target structures.

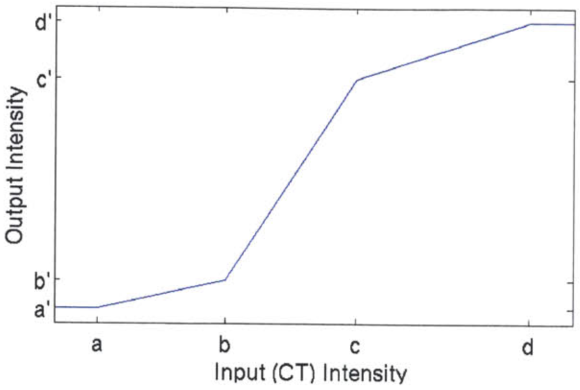


Figure 5.1: A graph of the intensity transform.

Unfortunately, registration performed worse on the images after this transform was applied. This is likely because the intensities across images are not similar enough for the same types of tissue, and consequently, the intensity modification function exacerbates those discrepancies. Even though CT scans ideally produce absolute intensities for the same densities across images, this is not always a safe assumption. Not all machines are calibrated correctly and identically, so the intensities may be linearly shifted.

To effectively apply this intensity modification, one would first have to normalize the intensity distributions between the images. If the intensities of the images are scaled linearly, it may be possible to rescale the images back into the same intensity space. While it is not immediately obvious how to do this from the intensities of the structures themselves nor from the intensity distribution of the whole image, it is likely that there are regions of the image more consistent in density that would be better candidates for determining intensity shifts.

In addition, we used ad hoc values for the parameters of the transform function, selected from the known ranges of intensity for soft tissue in CT scans, which did not correspond exactly to the intensity distributions (Figure 4.1) of the structures in our dataset. If these parameters were found more methodically from the intensity distributions of the images and their structures, the transform function may have a more positive impact, because it is actually focusing on the intensities of the relevant structures.

■ 5.2.2 Adaptive Standard Deviation Demons

The smoothness of the demons algorithm is based on the global standard deviation of the image. To make the deformation flexible enough in the soft tissue region of the image, we are also allowing for even more deformation in other regions of the image where the difference in intensities is more pronounced. If we could set this smoothness parameter based on a local standard deviation instead of the global standard deviation, it may lead to a better registration result.

■ 5.2.3 Weighted Voting Improvements

In the weighted voting algorithm, we only take into account the difference in intensities at the exact voxel that we are weighting. We might get a more robust signal by looking at a neighborhood of intensities around the voxel in question. For example, we might be able to tell if we were at a voxel near the edge of a structure in the target image, which could be helpful in better segmenting the structure's boundary.

■ 5.3 Summary

This work presents a system for labeling sensitive structures to plan for radiation treatment therapy. When an unlabeled target CT scan is received, each labeled training image is registered to the target subject. The atlas labels are then transformed by the resulting deformation into the space of the target image. Finally, the weighted voting algorithm is applied to these warped labels to determine the final label for the target structure. The experimental results presented in this thesis show that the resulting segmentations are often within only one or two millimeters of the manual segmentations. As such, the contours produced by this method can greatly aid doctors and save considerable time when constructing models of patients for radiation treatment therapy.

Bibliography

[1] V. Bevilacqua, A. Piazzolla, and P. Stofella. Atlas-based segmentation of organs at risk in radiotherapy in head mris by means of a novel active contour framework. *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*, pages 350–359, 2010.

[2] P. Cignoni, C. Rocchini, and R. Scopigno. Metro: measuring error on simplified surfaces. In *Computer Graphics Forum*, volume 17, pages 167–174. Wiley Online Library, 1998.

[3] O. Commowick, V. Grégoire, and G. Malandain. Atlas-based delineation of lymph node levels in head and neck computed tomography images. *Radiotherapy and Oncology*, 87(2):281–289, 2008.

[4] M. Depa, M. Sabuncu, G. Holmvang, R. Nezafat, E. Schmidt, and P. Golland. Robust atlas-based segmentation of highly variable anatomy: left atrium segmentation. *Statistical Atlases and Computational Models of the Heart*, pages 85–94, 2010.

[5] L.R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[6] F. Dru, P. Fillard, and T. Vercautereen. An itk implementation of the symmetric log-domain diffeomorphic demons algorithm. `http://hdl.handle.net/10380/3060`. Accessed: 2012.

[7] X. Han, L.S. Hibbard, N.P. Oconnell, and V. Willcut. Automatic segmentation of parotids in head and neck CT images using multi-atlas fusion. In *Proc. MICCAI 2010 Workshop Head and Neck Autosegmentation Challenge*, pages 297–304, 2010.

[8] E.J. Schmidt P. Golland M. Depa, G. Holmvang and M.R. Sabuncu. Towards efficient label fusion by pre-alignment of training data. In *Proc. MICCAI Workshop on Multi-atlas Labeling and Statistical Fusion*, pages 38–46, 2011.

[9] D. Mattes, D.R. Haynor, H. Vesselle, T.K. Lewellen, and W. Eubank. PET-CT image registration in the chest using free-form deformations. *Medical Imaging, IEEE Transactions on*, 22(1):120–128, 2003.

[10] A. Neumann and C. Lorenz. Statistical shape model based segmentation of medical images. *Computerized Medical Imaging and Graphics*, 22(2):133–143, 1998.

[11] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11:285–296, 1975.

[12] Plastimatch. `http://plastimatch.org`. Accessed: 2012.

[13] L. Ramus, G. Malandain, et al. Multi-atlas based segmentation: Application to the head and neck region for radiotherapy planning. In *MICCAI Workshop Medical Image Analysis for the Clinic-A Grand Challenge*, pages 281–288, 2010.

[14] R.K. Sharma and M. Pavel. Multisensor image registration. In *SID International Symposium Digest of Technical Papers*, volume 28, pages 951–954. Society for Information Display, 1997.

[15] J.P. Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical image analysis*, 2(3):243–260, 1998.

[16] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.

[17] 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. *Medical Imaging, IEEE Transactions on*, 23(7):903–921, 2004.

[18] W.M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical image analysis*, 1(1):35–51, 1996.