AUTOMATIC SEGMENTATION OF STRUCTURES IN CT IMAGES FOR HEAD AND NECK INTENSITY-MODULATED RADIATION THERAPY

By

Antong Chen

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Electrical Engineering

August, 2012

Nashville, Tennessee

Approved:

Benoit M. Dawant

Michael J. Fitzpatrick

Bennett A. Landman

Robert L. Galloway

Kenneth J. Niermann

To my parents and my wife

ACKNOWLEDGEMENTS

The dissertation would not have been possible without the guidance and support from my committee members, and in one way or another the assistance from my colleagues in Vanderbilt University.

First and foremost, I would like to give my utmost gratitude to my advisor, Dr. Benoit Dawant. By allowing me to join the Medical Image Processing (MIP) Lab from spring 2006, he opened the door for me to the amazing world of medical imaging and image processing. Since then, he has guided me through many challenges, and supported me both financially and mentally. I cannot count how many hours he has spent on helping me prepare papers, posters, and presentation, even after work and during weekends. He has been my role model as he is always honest, positive, supportive, and thoughtful. I have been and will continue benefiting from him as his student.

I also want to express my appreciation to other members of my committee. Dr. Kenneth Niermann, since we started the project in summer 2007, has always been involving in the studies actively. He has devoted tremendous amount of time and efforts on providing manual delineations and participating in the validation studies. Dr. Michael Fitzpatrick taught me the knowledge and methodology of image registration which has served as the solid basis of my research. Dr. Bennett Landman and Dr. Robert Galloway have provided valuable suggestions which have improved the study since it was originally proposed.

This work would be much more difficult to achieve without the help from my dear colleagues. Rui Li provided assistance on the registration, segmentation, and

visualization tools. Matthew Deeley helped me select the images for the study initially. Pierre François D'Haese helped on setting up the database and the framework for data transferring, and also shared his previous experience on atlas-based methods. Jack Noble shared his experience and provided crucial opinions on the ASMs. Ryan Datteri and Andrew Asman shared valuable ideas about multiple-atlas-based methods. My other lab mates Srivatsan Pallavaram, Fitsum Reda, Yuan Liu, and Ankur Kumar inspired me from different perspectives in our discussions. Moreover, this list should also include my colleagues who have left but helped before and after their leaving: Ramya Balachandran, Yong Li, Ning Xu, Xia Li, Qingyang Shang, Luigi Moretti, Qing Xu, Siyi Ding, Zhaoying Han, Anusha Rao, and Jeremy Lecoeur.

Special thanks should be given to NIH Grant R01EB006193, which have made the research possible and supported me financially in the past five years.

Last but not the least, I want to thank my family: My parents who gave me life and have been supporting me selflessly ever since, and my wife who has been patiently sharing with me the highs and lows, sweetness and bitterness of life. To all of you I dedicate my work.

Antong Chen, June 2012

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Cancers in the head and neck region account for approximately 3 percent of all cancer cases in the United States, as it is reported by the National Cancer Institute [1]. The areas affected are generally the oral cavity, salivary glands, paranasal sinuses and nasal cavity, nasopharynx, oropharynx, hypopharynx, and larynx. Depending on the location and stage of the cancer, surgery and radiation therapy are considered to be the major treatment options. Since its introduction two decades ago, intensity-modulated radiation therapy (IMRT) has become the state of the art in head and neck radiation therapy but this technique typically requires the segmentation of head and neck structures, including the thyroid, parotids glands, level I, II, III, IV, V and VI lymph nodes, larynx, and spinal cord, from medical images, especially computed tomography (CT) images.

As manually segmenting the structures is a time-consuming task for physicians, atlas-based automatic segmentation methods have been explored as efficient alternatives to delineate head and neck structures. The approach requires selecting/constructing an atlas, precisely delineating the structures in the atlas, and a reliable registration process that can provide a correspondence between the atlas and the patient image. Once the correspondence is established, contours delineated in the atlas can be projected onto the patient images automatically. Prior work by several groups [18, 19, 21, 23] has shown the efficacy of the atlas-based approach by comparing the automatic segmentation results with the manual delineations. However, these results also show that over/under segmentations still exists due to the accuracy limitation of the registration algorithms. This may require a large amount of post-processing and human correction.

The overachieving goal of the work presented herein is to automatically and

accurately segment the structures of interest in head and neck CT images for IMRT treatment planning. Specifically, these will include the level II, III, and IV lymph node regions, the thyroid gland, and the left and right parotid glands. A series of automatic segmentation approaches are developed and the results are compared with those obtained using traditional approaches, e.g. single-atlas-based method for segmenting the lymph node regions. The potential for incorporating the automatic segmentations into the clinical treatment planning process is also accessed.

1.1. BACKGROUND ON HEAD AND NECK CANCER AND IMRT

According to the statistics provided by the National Cancer Institute, it is estimated that more than 52,000 Americans developed cancer of the head and neck in 2011 [1]. It is known that one major type of head and neck cancer is squamous cell carcinoma, which occurs in the cells on the inside of the nose, mouth, or throat. Other types that are less common are salivary gland tumors, lymphomas, and sarcomas. In addition to the ailment on and around the primary site of the cancer, spread can arise through the lymphatic channels to the lymph nodes, generally the ones located along the major blood vessels underneath the sternocleidomastoid muscles bilaterally on the neck. Also, due to the attachment of the lymph nodes and the blood vessels, cancer can spread to other parts of the body through the bloodstream.

Among the three major treatment options (radiation therapy, surgery, and chemotherapy) for head and neck cancers, radiation therapy is the method that uses ionizing radiations to kill cancer cells and make the tumors shrink. It injures or destroys cells in the target area by damaging their genetic structure, and therefore stops their

growth and division. However, as radiation can hurt both cancer and normal cells, the goal of radiation therapy is to damage as many cancer cells as possible, while limiting harm to the healthy tissue nearby. One of the state-of-the-art radiation therapy approaches meeting the precision requirement to achieve this goal is IMRT, which has become a general oncology practice since its first introduction in the 1990s [2]. Utilizing a computer-controlled linear accelerator and a precise planning technique based on 3D CT images, physicians can combine several intensity-modulated fields from different beam directions to irradiate the tumor with precise dose and geometry, while minimizing the radiation received by the surrounding normal organs. As this inverse planning technique requires radiation dose to be assigned for each structure on the CT images, an accurate delineation of the tumor as well as of the organs at risk becomes necessary. For head and neck cancers, the tumor is set as the center of the clinical target volume (CTV) and denoted as the primary target volume (CTV1) to be treated with the highest radiation dose. The peripheral area of the tumor which is likely to be affected by the cancer is set as the subclinical disease volume (CTV2 and CTV3) and is treated with relatively lower radiation dose. The organs (if not directly affected by the cancer) to be spared generally include the thyroid gland, the parotid glands, the submandibular glands, the larynx, and the spinal cord.

1.2. OVERVIEW OF MAJOR CHALLENGES FOR SEGMENTING STRUCTURES IN HEAD AND NECK CT IMAGES

For many cancers of the head and neck, there is almost always some risk of spread of cancer to the cervical (neck) lymph nodes. In many cases the lymph nodes have

microscopic disease even when they appear completely normal on CT, PET, or MRI. Instead of having a patient undergo a surgical sampling of the entire lymph node regions of the neck, it is standard practice to deliver radiation prophylactically to these regions even when there is no radiological evidence of enlargement. This requires the delineation of the lymph node regions on the same side as CTV1, as well as the lymph node regions on the opposite side. These lymph node regions are generally defined as a part of CTV2 and CTV3. This is a time consuming process. Indeed, while delineating the gross tumor volume takes on the order of 5 minutes or less, delineating the lymph node regions can typically take between 20 and 45 minutes, depending on the level of complexity of the patient image. In fact, delineating the lymph node regions is one of the most challenging and time-consuming tasks in the planning process.

Figure 1.1 shows the division of the lymph node regions on one side of the head and neck area. It can be seen that the complete lymph node regions can be divided into seven levels, from level I at the bottom of the mandible to level VII at the front of the manubrium. In clinical applications, the lymph node regions to be treated most frequently contain levels II, III, and IV, which essentially form a connected structure expanding from the jugular fossa to the clavicle bone. It can also be observed from the figure that these three levels of lymph node regions share approximately the same anterior boundary, which is defined by major blood vessels, e.g. the internal jugular vein. They also share the same posterior boundary at the end of the sternocleidomastoid muscle. Separation between these three levels is determined by locating certain anatomical landmarks at the same height on the neck, including the lower border of the hyoid separating the level II and level III, and the lower margin of cricoid cartilage dividing the level III and level IV.

For each level, sampled slices with manual delineations are shown in Figure 1.2. It can be seen that each level of lymph node regions has boundaries that are difficult to identify, e.g. the anterior boundary of level II is generally delineated by experience or based upon regularity conditions, and the posterior boundaries for all levels are defined approximately at the bending of the sternocleidomastoid muscle. To address these challenges, we are proposing to use an approach based on active shape models (ASM) which could bring a sufficient amount of prior knowledge of the shapes to assist defining these boundaries.

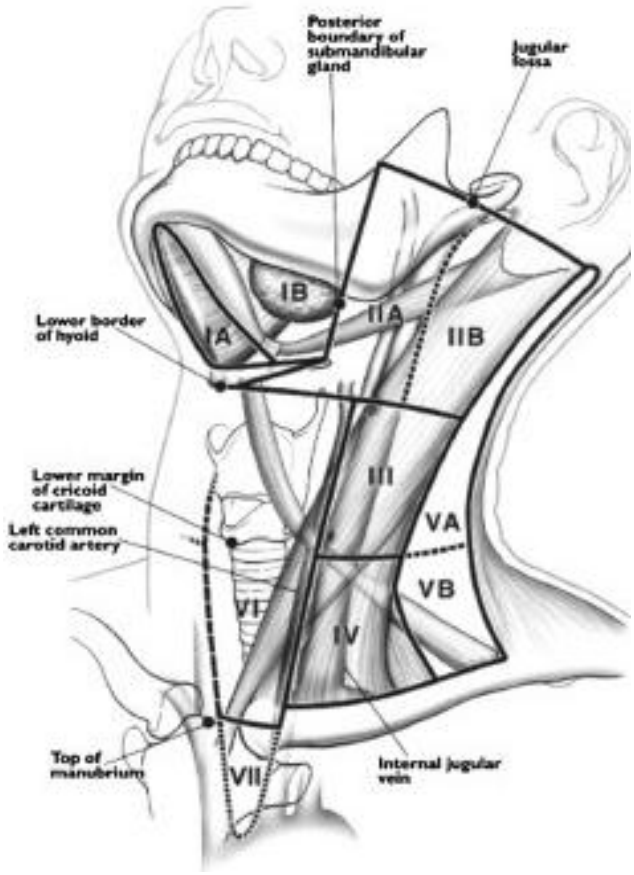

Figure 1.1. Level I to level VII lymph node regions on one side of the head and neck area (from [3]).

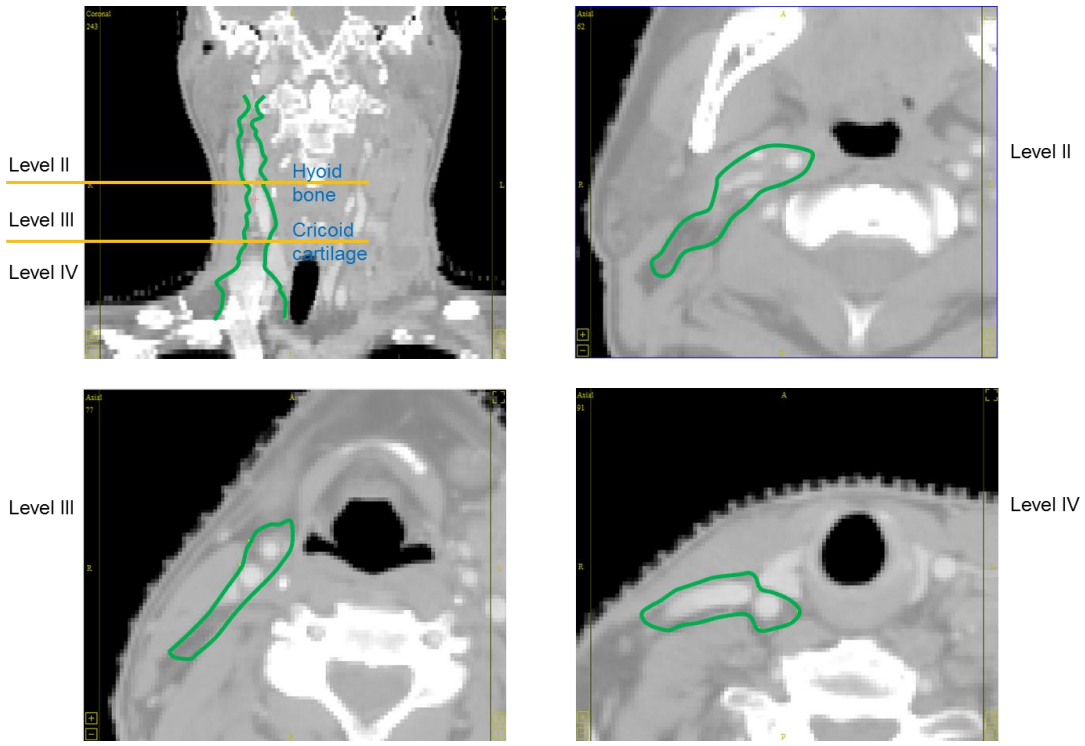

Figure 1.2. CT images and the manual delineations (green contours) overlapped on top of slices sampled for each level.

Since irradiating the thyroid gland can cause complications such as chronic hypothyroidism which is permanent and requires hormone replacement therapy on a daily basis, sparing the thyroid gland is of great importance. Figure 1.3 shows one slice of the thyroid gland in four representative patients, with the original CT images shown in the top row and the manual delineations overlapped and shown in the bottom row. It can be seen that, as opposed to the lymph node regions which consists of multiple types of soft tissues, the thyroid gland is a relatively homogeneous structure with usually clear boundaries. However, there is large shape variability between subject that exists not only because of normal anatomical differences but also due to surgical procedures, e.g. a tracheotomy which inserts a plastic tube into the larynx to help the patient breathe. Herein we propose to use a multiple-atlas-based approach to compensate for this shape variability.

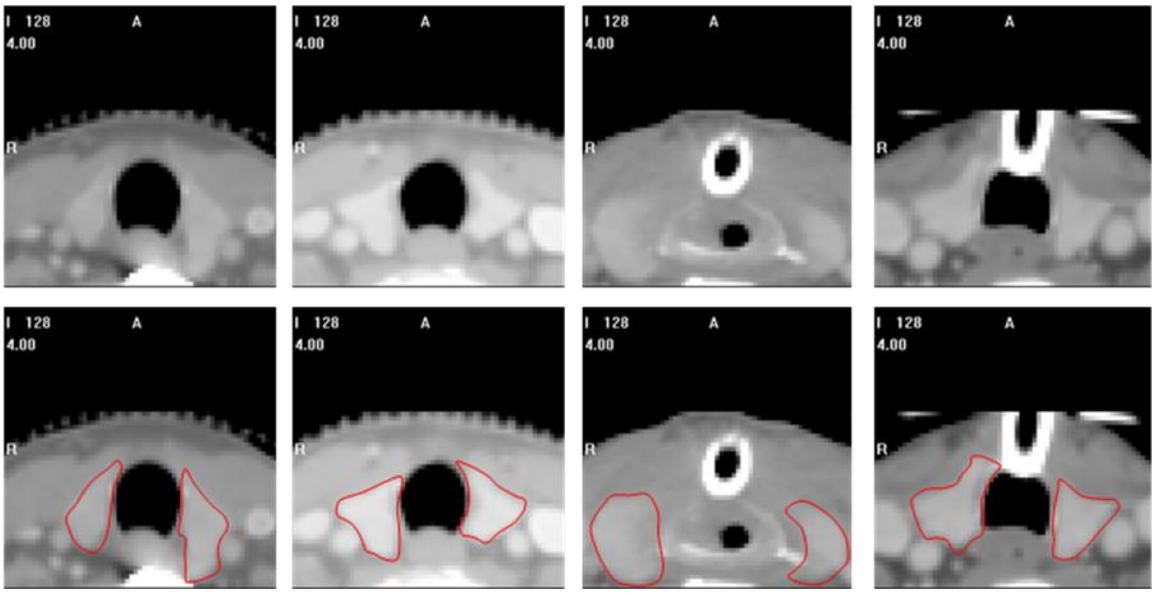

Figure 1.3. CT images and the manual delineations (red) for the thyroid gland in four cases. Top row: Original CT images. Bottom row: CT images with manual contours overlapped on top.

The left and right parotid glands are also important structures to be spared since their irradiation is the major cause of xerostomia, which is one of the most prevalent side effects of head and neck radiation therapy affecting the patients' quality of life. Figure 1.4 shows one slice of the gland for four representative examples in our dataset. In this figure, the top row shows the original CT images and the bottom row shows the overlapped contours manually delineated by the oncologist. We found the shape of the parotid glands to be more consistent across subjects than the shape of the thyroid. This makes ASM a plausible solution, but the overall intensity of the glands varies between that of the adjacent muscle groups and fat tissues, and poor contrast at some of the boundary regions, e.g. the interior boundary between the gland and the digastric muscle, complicate the segmentation process. To address these difficulties, we propose to use a constrained ASM with landmark uncertainty [61].

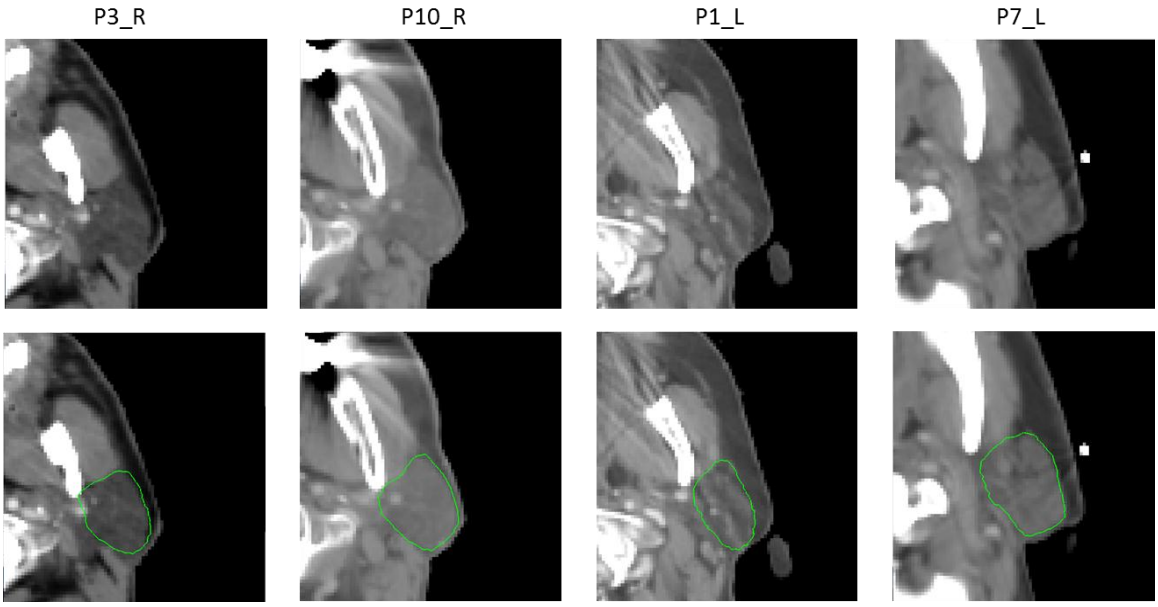

Figure 1.4. CT images and the manual delineations (green) for the parotid gland in four cases. Top row: Original CT images. Bottom row: CT images with manual contours overlapped on top.

The submandibular glands, as another set of major salivary glands, should also be spared. However, this is generally impractical because of their position (close to the middle of the neck region). Therefore segmenting the submandibular glands is not a goal in our automatic segmentation tasks. Segmentation of structures with good contrast and clear boundary against their surrounding tissues, e.g. the larynx and the spinal cord, is also unexplored in this specific study.

In the following sections, we will first review the established works on the automatic segmentation of structures in head and neck CT images, most of which use atlas-based techniques. Then we will introduce the background on the techniques we propose herein. These will include the multiple-atlas-based approach which we use for segmenting the thyroid gland, and the statistical shape models which we use for segmenting the lymph node regions and the parotid glands.

1.3. BACKGROUND ON ATLAS-BASED TECHNIQUES FOR SEGMENTATION OF HEAD AND NECK STRUCTURES

Most of the techniques used for segmenting head and neck structures involve registration of CT images. Registration is the determination of a one-to-one correspondence between the coordinates in the reference image space and those in the target image space such that the corresponding points in the two spaces can be mapped to each other [4]. Categorized by the transformation models used to relate the two spaces, the registration algorithms can be classified into rigid, affine, and nonrigid registrations. The affine registration uses affine transformations including translation, rotation, scaling and shearing to align the target and reference images, while nonrigid registration algorithm use nonrigid transformations parameterized with radial basis functions (such as B-spline, thin-plate spline (TPS), or compactly supported functions) or physical continuum models (viscous fluids, optical flows) to locally warp the target and reference images.

The standard atlas-based segmentation technique requires a template image volume, which is also known as the atlas, to be selected first. Structures of interest are then carefully segmented on the atlas by experts. The transformation that registers the atlas to the patient images is computed and used to project structure labels from the atlas to the target images. The technique has been applied to IMRT treatment planning since the last decade, generating automatic segmentations of tumors and organs at risk as an alternative to the physicians' manual delineations. Specifically, there are two major types

of applications: Automatic mapping of delineations in image-guided adaptive radiation therapy, and automatic segmentation of structures on planning CT images.

In applying atlas-based methods to adaptive radiation therapy, the planning CT is delineated and used as the atlas, while the daily images acquired during treatment for the same patient are the target images onto which the delineations are propagated. Compared with manual delineation, atlas-based methods can provide more consistent results and, most importantly, save a significant amount of time for physicians by allowing them to modify automatic contours rather than delineating the structures from scratch. Several groups have published studies that implement atlas-based segmentation for automatically generating contours on the daily treatment CT/cone-beam CT (CBCT) images from the delineations on the diagnostic/planning CT images. The major difference between their approaches is the registration. Wang *et al.* [5] developed an enhanced "Demons" registration algorithm [6, 7]. Their automatically generated contours for head and neck daily CT achieved an average 83% Dice similarity coefficient (DSC) [8] when compared with physician's manual contours, and the DSC improved to >97% when these automatically generated contours were modified by physicians and compared with the manual contours created from scratch. Lu *et al.* [9] used a deformable registration algorithm based on the calculus of variation [10] to register planning kVCT with daily MVCT in adaptive radiation therapy. Lee *et al.* [11] continued Lu's study and evaluated geometric changes in the parotid glands during the cancer treatment process. They segmented the left and right parotid glands, calculated the volume of each, and measured the distance between their center of mass (COM) in the course of 35 treatment days. The study showed that the trend in the geometric changes observed in automatically generated

parotid contours was comparable to the trend observed in manually delineated contours, thus suggesting that the method was sufficiently accurate for the task at hand. The automatic delineations are used in a study of radiation dose received by the parotids [12]. Zhang *et al.* [13] used the Lu's algorithm to register planning CT to daily CT of head-and-neck cases and compared the deformed contours with one set of manually drawn contours. Their average DSC was around 0.8 for all the structures. Also based on Lu's algorithm, Yan *et al.* [14] deformed the lymph node contours delineated on the baseline CT scan onto the follow-up images, and used a marker-controlled watershed segmentation algorithm [15] to further correct these contours. Chao *et al.* [16] used a narrow band to contain the manually delineated contour on the planning CT, and deformed the contour onto the CBCT through a B-spline based deformable registration. The registration is calculated by setting multiple control volumes on the reference image, calculating the rigid transformation for each of these regions, and using the average transformation as the rigid body transformation for the entire image. This was followed by a B-spline nonrigid registration [17].

Different from the studies discussed above, which essentially perform intra-patient image registrations, a more challenging task is to use one or a set of delineated images as the atlas(es), and provide automatic delineations of structures on the new patient image through inter-patient registrations for the initial IMRT planning. Using the same "Demons" registration algorithm as in [5-7], Chao *et al.* [18] projected contours from a template image to patient images for head-and-neck structures, and proposed them to the physicians for editing, saving delineation time up to 47%. Commowick *et al.* [19] developed a framework using an averaged symmetric atlas to delineate head-and-neck

lymph node levels. The two-step nonrigid registration included a local affine [20] and a dense deformation, which was used for atlas construction as well. These authors performed a leave-one-out experiment and reported a quality measure of error that is a combination of sensitivity and specificity of 0.253. Although it is difficult to compare the error with the DSC used in other methods directly, both average sensitivity and specificity were above 0.8, which indicated good segmentation accuracy. Commowick *et al.* [21] also proposed an improved method based on selecting locally most similar atlases via an average intensity image volume generated using the method proposed by Guimond *et al.* [22]. Using this approach, lymph node segmentation showed an improvement in specificity compared to the segmentations they obtained in [19] using an average atlas constructed with all atlas images, while the sensitivity was reduced. Gorthi *et al.* [23] also segmented all the lymph node regions with an atlas-based method. They computed the deformation with structures easily visible in the images (bones, trachea, and skin) and then applied it to the rest of the image. This led to relatively low DSC (mostly <0.6) and large average Hausdorff distance (14.52 – 21.81mm) for the for node regions of levels II, III, and IV. Han *et al.* [24] proposed a hierarchical registration framework with a linear, a poly-affine, and a dense deformable transformation and applied it on a set of 10 head and neck CT images. In their leave-one-out validation study, the method combining multiple segmentations from all atlases using an expectation maximization (EM) algorithm known as the simultaneous truth and performance level estimation (STAPLE) algorithm [25], which is commonly used for combining segmentations from multiple raters, outperformed the method selecting the most similar atlas using mutual information, and the medians of the DSC for the level II, III, and IV lymph node regions exceeded 0.65.

Han *et al.* [26] focused on the parotid glands and used the same multiple-atlas-based approach followed by a refinement step using a deformable surface model prior to fusion. The experiment using 10 training images and 8 testing images (voxel size around $0.98\times0.98\times2$ mm^3, provided by the Princess Margaret Hospital in Toronto, Canada, for the MICCAI 2010 Head and Neck Auto-Segmentation Challenge Workshop with manual segmentations of the parotid glands delineated by experts) reached an average volume DSC of 0.86. Based on the registration framework proposed in [21] involving multiple atlases, Ramus *et al.* [27] segmented the parotid glands and combined the segmentations propagated from the atlases using an intensity-weighted majority vote based on the local sum of square distances (SSD) between the transformed atlases and the patient image. The method was tested on the same dataset as the one use in [26], and an average volume DSC of 0.85 was achieved with one abnormal case eliminated. Also using this dataset for segmenting the parotid glands, Yang *et al.* [28] analyzed the intensity of the atlas images through principal component analysis (PCA), and selected a subset of most similar atlases. They combined the deformed segmentations using STAPLE and produced a volume DSC of 0.84.

From published work, one can conclude that to achieve a satisfactory level of segmentation accuracy, e.g., volume DSC > 0.8, only one planning CT image is generally required as the atlas for automatic contouring in adaptive radiotherapies. Segmentation of structures on new patient planning CT images generally requires multiple atlases. This is because the inter-patient anatomical and geometrical discrepancies are generally higher than those in the intra-patient applications even considering the same patient's weight loss in the treatment. Therefore, multiple atlases are generally needed to compensate for

differences between individual atlases and the patient images. In this dissertation we explore and evaluate the applicability of multiple-atlas-based techniques for the segmentation of the thyroid gland.

1.4. BACKGROUND ON MULTIPLE-ATLAS-BASED APPROACHES

In the standard atlas-based segmentation approach, the selected template is a single image volume. It is in general not sufficient to represent the population of images to be segmented and large registration errors can be observed when an image is extremely different from the atlas. This, in turn, causes the segmentation to be inaccurate. Methods involving a set of atlases have been proposed to solve such a problem focusing mainly on how to use information from these atlases: Wu *et al.* [29] proposed to select the single optimal atlas for each region of interest (ROI) based on local normalized mutual information (NMI) [30]. Heckermann *et al.* [31] applied the majority vote rule to fuse segmentations from up to 29 atlases, and found that using about 15 to 20 atlases was sufficient. Further increasing the number of atlases did not improve segmentation accuracy very much. The method was later improved [32] by enhancing the robustness of the nonrigid registration algorithm with an approximate tissue classification at the coarse levels of the multi-resolution implementation. Rohlfing *et al.* [33] compared three techniques: selecting the most similar atlas, using an average shape atlas, and using multiple atlases and determining the final segmentation by the majority vote rule. In their study, the last method showed the best performance. Instead of using the majority vote rule, which assigns equal weight to each atlas, Warfield *et al.* [25] weighted the segmentations through STAPLE algorithm. This algorithm has been used as a standard

technique for combining automatic or manual segmentations from multiple raters. Rohlfing *et al.* [34] developed the STAPLE algorithm for simultaneously combining labels for multiple classes. Their experiments on bee brain confocal microscopy images showed that the proposed method performed better than majority vote. Klein *et al.* [35] combined segmentations from a set of atlases using both STAPLE and an altered version of the vote rule, which weighted the contribution of each atlas with the value of the NMI between the atlas and the volume to be segmented. In their study, they showed that STAPLE did not perform better than the vote rule. They also found that using multiple atlases outperformed selecting one single optimal atlas. Aljabar *et al.* [36] studied the effect of increasing the number of atlases that were ranked by the value of the NMI between the registered atlases and the volume to be segmented. This study showed that using 20 atlases from a set of 275 was optimal. Artaechevarria *et al.* [37] compared strategies for combining segmentations by multiple atlases including STAPLE, majority vote, and weighted voting methods based on global or local similarity between patient and atlas images after affine and nonrigid registrations. The experiment on a set of 18 brain MR images showed that, among the methods that were evaluated, locally weighted voting based on measuring similarity in the neighborhood of the structure of interest performed the best. Isgum *et al.* [38] assigned weights for each atlas on a per-voxel basis using a measure of voxel-wise difference between the registered atlas and the target image. Segmentation results on heart images show that the proposed method outperformed methods including majority vote. Sabuncu *et al.* [39] proposed an iterative method to optimize the weights through EM. Different from STAPLE, which calculates the weights based only on the segmentations, the method also takes the intensity

information of the registered images into consideration. Langerak *et al.* [40] proposed a selective and iterative method for performance level estimation (SIMPLE) to combine segmentations without EM, and the experimental results on a set of 100 prostate MR volumes showed that SIMPLE outperformed STAPLE in both accuracy (statistically significant improvements on volume-wise similarity with manual segmentations) and computation time (reduction to about 1/4 to 1/3 of STAPLE). Lötjönen *et al.* [41] proposed two methods to utilize the image intensity information through graph cuts or EM in combining segmentations from 13 optimal atlases based on NMI. Comparison with results obtained without using image intensity information showed significant improvements by the proposed methods. Asman *et al*. [60] proposed a spatial STAPLE method which uses a regional confusion matrix to describe the performance level of each atlas on a per-voxel basis. The method outperformed both the original STAPLE and majority vote in the validation on whole brain segmentation with 36 ROIs.

It can be seen that the latest developments on multiple-atlas-based approaches have suggested that, instead of focusing on the segmentations themselves as is done in STAPLE and the majority vote approaches, incorporating image information from the deformed atlas and the patient image volume may improve the results. In Chapter III, we are exploring and evaluating several options for the segmentation of the thyroid gland.

1.5. BACKGROUND ON STATISTICAL SHAPE MODELS

Segmentation based on statistical shape models can be used to incorporate expected shape and appearance information about the target shape in the segmentation process. The introduction of active shape models (ASM) by Cootes *et al.* [42] set the

stage for the application of SSMs in the area of medical imaging over the past two decades. The method started with labeling a set of ordered points on the edge of 2D structures, which are called the landmarks. Then a point distribution model (PDM) is created by calculating the mean position of the corresponding landmarks over the training set, and a number of modes of variations are determined which describe the major ways the training shapes are deformed from the mean shape. It was successfully applied to modeling and segmenting 2D structures including the heart chamber from an echocardiogram image [43]. It was then extended to 3D for solving a variety of problems [44 – 46].

Achieving the proper landmark correspondence across the training images is the first step in building the SSM. Although the landmarks were defined manually in early applications [43], when the problem was expanded to 3D, establishing correspondence manually became unfeasible. Many methods were thus developed to automatically compute the correspondences, often through registration of the training shapes in the training set. Brett and Taylor [47] extracted meshes for all training shapes and found the landmark correspondences through a symmetric iterative closest point (ICP) algorithm. Then they built a binary tree of merged shapes taking the mean shape as the root and the examples from the training set as the leaves. The disadvantage was that the ICP algorithm was restricted to similarity transformations. Lorenz *et al.* [48] used TPS to nonrigidly model the deformation between corresponding landmarks, and used a mass-spring model to unfold and regularize the deformed mesh, while Paulsen *et al.* [49, 50] used a Markov Random Field-based method for the regularization after TPS deformation. As an alternative to methods based on matching landmark points directly, Fleute *et al.* [51]

deformed the landmarks from the template image surface to the grey-level training images using a multi-resolution spline-based nonrigid registration and tested the approach on reconstructing 3D vertebral surface from CT images. The transformations for finding the landmark correspondences can also be acquired from volume to volume registrations. Frangi *et al.* [52] used a quasi-affine registration to align the training images in the form of binary masks and registered them with the atlas using a nonrigid registration based on B-spline. The transformations were used to propagate the landmarks from the atlas to each mask thus establishing the correspondences. Rueckert *et al.* [53] used the same registration algorithm to register training images with the atlas directly, both in the form of grey-level MR images. With the transformations computed, they analyzed the statistics of the deformation fields rather than that of the landmarks as is usually done. After the landmark correspondence is achieved, the mean shape and the valid modes of variations are obtained, which completes the training of the shape model.

The trained model is then applied on the new patient image for finding the proper segmentation. The first step is to initialize the model to a location close to the target structure. Some applications require a large amount of human interaction, (see for instance [54]). Fripp *et al.* [55] used a rigid ICP registration to automatically initialize a knee cartilage model. Hill *et al.* [56] employed genetic algorithm search to initialize and further optimize the placement of the model. Compared with the methods based on registrations, this method is computationally costly.

Once initialized, the segmentation is refined by first updating the location of each of the landmark points to a more optimal location based on its local image characteristics, e.g. the gray level or other texture computed from the image, and then fitting a model to

the updated landmarks. The process is generally iterated until convergence, or a maximum number of iterations is reached. Cootes *et al.* [57] sampled the profiles perpendicular to the surface of the training images and used the profile as a training set to determine the movements of the landmarks. The information used to construct the profiles could be the original intensity, or the gradient along the sampled perpendicular direction, or the normalized version of either of them. In the search step the Mahalanobis distance between the profile at updated location and the corresponding profiles sampled on the training shapes of the model is used to decide the location for the best fit. van Ginneken *et al.* [58] proposed a framework to select the optimal features from the set of features including the first and second order moments and gradients using a *k*NN classifier. Toth *et al.* [59] also proposed a method in which an optimal weighted sum of features is found. By sampling voxels in the neighborhood of each landmark points, the correlation between their Euclidean distances to the landmark point and the corresponding linear combination of Mahalanobis distances based on the features is maximized. The coefficients for the optimal linear combination are used to determine the weights for each feature in the updating of the landmark points.

Training, initialization, and refinement, in turn, form a general procedure for performing ASM-based segmentation tasks. Herein we are introducing a framework automating the procedure. This framework, as it is shown in Chapter II, is used first for segmenting the lymph node regions. Then in Chapter IV it is used for training and initializing the ASM for segmenting the parotid glands. In this chapter we also present a method for updating the landmarks that builds on the work of Toth *et al.* [59] and for weighing landmarks that can be localized easily more than those that cannot.

1.6. GOALS AND CONTRIBUTIONS OF THE DISSERTATION

The goal of this dissertation is to automatically and accurately segment the structures of interest in head and neck CT images for IMRT treatment planning, which is achieved by a series of methods that are more advanced than the standard atlas-based method for the automatic segmentation of these structures. This includes an ASM-based segmentation for the lymph node regions, a multiple-atlas-based segmentation for the thyroid gland, and a constrained-ASM-based segmentation for the left and right parotid glands. The proposed methods form a framework that allows automatic and accurate segmentations of these structures which could be potentially integrated in the clinical flow used to treat patients with head and neck cancer.

Specifically the contributions of this dissertation are:

1) The construction and use of ASMs for the segmentation of the level II, III, and IV lymph node regions. The framework we developed permits automatic landmark correspondence computation, and automatic model initialization, both accomplished through nonrigid registrations with an average head and neck CT atlas. Segmentation results evaluated qualitatively and quantitatively through DSC and surface distance errors have shown that the proposed method is more accurate than an atlas-based segmentation using the average atlas.

2) A multiple-atlas-based framework for the segmentation of the thyroid gland. Instead of performing registration between each atlas and the patient image, we propose to first align all images with an average atlas volume and then

perform a soft tissue nonrigid registration between atlas and patient images in a bounding box that surrounds the thyroid gland. We show that the method combining segmentations by multiple atlases based on local correlation coefficient is more accurate than other options, including STAPLE and majority vote

3) A method to segment the parotid glands using a constrained ASM with landmark uncertainty, in which the uncertainty values are derived from optimizing a set of image features. The approach can address the issues of low contrast between the gland and the surrounding soft tissues, especially the muscle groups at the interior and anterior boundaries of the gland. The framework can be applied to other structures with partially fuzzy boundaries and can be extended to include more features when needed.

The remainder of this dissertation is organized as follows: Chapter II presents the ASM-based segmentation approach for the level II, III, and IV lymph node regions. Chapter III compares eight methods for selecting or combining segmentations from multiple atlases and shows that the method based on local correlation coefficient between registered atlas and patients images outperforms the others, including STAPLE and majority vote. Chapter IV proposes a constrained ASM with landmark uncertainty assigned using optimal features for the segmentation of the parotid glands. Segmentation results obtained by the proposed method are compared with results obtained using multiple atlases, as well as a regular ASM with optimal features. Chapter V summarizes the achievements of the research and discusses possible directions for future work.

REFERENCES

[1] http://www.cancer.gov/cancertopics/factsheet/Sites-Types/head-and-neck

[2] Chao K., Ozyigit G., Low D. *et al.* Intensity modulated radiation therapy for head and neck cancers. Lippincott Williams & Wilkins, 2002

[3] Som P, Curtin H, Mancuso A. An imaging-based classification for the cervical nodes designed as an adjunct to recent clinically based nodal classifications. Arch Otolaryngol Head Neck Surg 1999;125:388–96

[4] Maurer C. R. Jr., and Fitzpatrick J. M. A review of medical image registration. In: Interactive Image-Guided Neurosurgery American Association of Neurological Surgeons, Park Ridge, IL, 1993

[5] Wang H., Garden A. S., Zhang L., et al. Performance evaluation of automatic anatomy segmentation algorithm on repeat or four-dimensional computed tomography images using deformable image registration method. Int. J Radiat Oncol Biol Phys 2008;72:210-219

[6] Wang H., Dong L., Lii M. F., et al. Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy. Int. J Radiat Oncol Biol Phys 2005;61:725-735

[7] Wang H., Dong L., O'Daniel J., et al. Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy. Phys Med Biol 2005;50:2887-2905

[8] Dice L R 1945 Measures of the amount of ecologic association between species. Ecology 26 297–302

[9] Lu W. G., Olivera G. H., Chen Q., et al. Deformable registration of the planning image (kVCT) and the daily images (MVCT) for adaptive radiation therapy. Phys Med Biol 2006;51:4357-4374

[10] Lu W. G., Chen M. L., Olivera G. H., et al. Fast free-form deformable registration via calculus of variations. Phys Med Biol 2004;49:3067-3087

[11] Lee C., Langen K. M., Lu W. G., et al. Evaluation of geometric changes of parotid glands during head and neck cancer radiotherapy using daily MVCT and automatic deformable registration. Radiother Oncol 2008;89:81-88

[12] Lee C., Langen K. M., Lu W. G., et al. Assessment of parotid gland dose changes during head and neck cancer radiotherapy using daily megavoltage computed tomography and deformable image registration. Int. J Radiat Oncol Biol Phys 2008;71:1563-1571

[13] Zhang T. Z., Chi Y. W., Meldolesi E., et al. Automatic delineation of on-line head-and-neck computed tomography images: toward on-line adaptive radiotherapy. Int. J Radiat Oncol Biol Phys 2007;68:522-530

[14] Yan J. Y., Zhao B. S., Curran S. Automated matching and segmentation of lymphoma on serial CT examinations. Med Phys 2007;34:55-62

[15] Yan J. Y., Zhao B. S., Wang L. Maker-controlled watershed for lymphoma segmentation in sequential CT images. Med Phys 2006;33:2452-2460

[16] Chao M., Li T. F., Schreibmann E., et al. Automated contour mapping with a regional deformable model. Int. J Radiat Oncol Biol Phys 2008;70:599-608

[17] Schreibmann E., Xing L. Image registration with auto-mapped control volumes. Med Phys 2006;33:1165-1179

[18] Chao K. S., Bhide S., Chen H., et al. Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach. Int. J Radiat Oncol Biol Phys 68, 2007. 1512-1521

[19] Commowick O., Gregoire V., Malandain G. Atlas-based delineation of lymph node levels in head and neck computed tomography images. Radiother Oncol 2008;87:281-289

[20] Commowick O., Arsigny V., Costa J., et al. An efficient locally affine framework for the smooth registration of anatomical structures. Med Image Anal 2008;12: 427-441

[21] Commowick O., Warfield S. K., and Malandain G., Using Frankenstein's creature paradigm to build a patient specific atlas. In Proceedings of the 12th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'09), Part II, Vol. 5762, September, 2009. pp. 993-1000

[22] Guimond A. D., Meunier J., and Thirion J. P. Average Brain Models: A Convergence Study. Computer Vision and Image Understanding 77, 2000. 197-201

[23] Gorthi S., Duay V., and Houhou N., et al, Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration. IEEE Journal of Selected Topics in Signal Processing, Vol. 3, No. 1, February, 2009. 135-147

[24] Han X., Hoogeman M., Levendag P., et al. Atlas-based auto-segmentation of head and neck CT images. 11th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2008). LNCS, Vol. 5242. Springer, 2002. 434 - 441

[25] Warfield S, Zou K, and Wells W 2004 Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imag. 23 903–921

[26] Han X, Hibbard S, O'Connell P, and Willcut V 2010 Automatic segmentation of parotids in head and neck CT images using multi-atlas fusion. In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 297-304

[27] Ramus L and Malandain G 2010 Multi-atlas based segmentation: application to the head and neck region for radiotherapy planning. In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 281-288

[28] Yang J, Zhang Y, Zhang L, and Dong L 2010 Automatic segmentation of parotids from CT scans using multiple atlases. In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 323–330

[29] Wu M, Rosano C, Lopez-Garcia P et al. 2007 Optimum template selection for atlas-based segmentation. NeuroImage. 34 1612–1618

[30] Studholme C, Hill D L G, and Hawkes D J 1999 An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognition. 32 71 –86

[31] Heckemann R, Hajnal J, Aljabar P et al. 2006 Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage. 33 115–126

[32] Heckemann R, Keihaninejad S, Aljabar P et al. 2010 Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. Neuroimage. 51 221–227

[33] Rohlfing T, Brandt R, Menzel R et al. 2004 Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage. 21 1428–1442

[34] Rohlfing T, Russakoff D B, and Maurer C R 2004 Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans. Med. Imag. 23 983–994

[35] Klein S, van der Heide U A, Lips I M, van Vulpen M, Staring M, and Pluim J P 2008 Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. Med Phys. 35(4) 1407–1417

[36] Aljabar P, Heckermann R A, Hammers A, Hajnal J V, and Rueckert D 2009 Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage. 46 726–738

[37] Artaechevarria X, Munoz-Barrutia A, and Ortiz-de-Solorzano C 2009 Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. on Med. Imag. 28 1266–1277

[38] Isgum, I., et al. Multi-atlas-based segmentation with local decision fusion—Application to cardiac and aortic segmentation in CT scans. IEEE Trans. Med. Imag. 2009. 28(7) 1000-1010

[39] Sabuncu M R, Tomas Yeo B T, van Leemput K, Fischl B, and Golland P 2010 A generative model for image segmentation based on label fusion. IEEE Transactions on Medical Imaging. 29 1714–1729

[40] Langerak T R, van der Heide U A, Kotte A N, Viergever M A, van Vulpen M, and Pluim J P 2010 Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Transactions on Medical Imaging. 29 2000–2008

[41] Lötjönen J M P, Wolz R, and Koikkalainen J R et al. Fast and robust multi-atlas segmentation of brain magnetic resonance images. NeuroImage. 49, 2010. 2352–2365

[42] Cootes T. F., Taylor C. J., Cooper D., et al. Training models of shape from sets of examples. In D. Hogg and R. Boyle, editors, 3rd British Machine Vision Conference, 9–18. Springer-Verlag, Sept. 1992

[43] Cootes T. F., Taylor C. J., Cooper C. H., and Graham J., Active shape models-their training and application. Computer Vision and Image Understanding. 1995;61:38-59

[44] Kelemen A., Szekely G. and Gerig G., Three-Dimensional Model-Based Segmentation of Brain MRI. IEEE Trans. Med. Imag. 18(10), 1999. 828-839

[45] Lorenz C. and Krahnstover N., 3D statistical shape models for medical image segmentation. In Proc. 2nd Int. Conf. 3-D Digital Imaging and Modeling, 1999. 414-423

[46] Davies R. H., Twining C. J., and Cootes. T. F., et al, 3D statistical shape models using direct optimisation of description length. In Proc. ECCV. LNCS, vol. 2352. Springer, 2002. 3-20

[47] Brett A. D. and Taylor C. J. A method of automated landmark generation for automated 3D PDM construction. Image Vision Comput. 1999;18(9):739–748

[48] Lorenz C. and Krahnstöver N.. Generation of point-based 3D statistical shape models for anatomical objects. Comput Vis Image Underst. 2000;77(2):175–191

[49] Paulsen R., Larsen R., Nielsen C., et al. Building and testing a statistical shape model of the human ear canal. In Proc MICCAI, 2002. volume 2489 of LNCS, 373–380

[50] Paulsen R. and Hilger K. Shape modeling using Markov random field restoration of point correspondences. In Proc IPMI, 2003. volume 2732 of LNCS, 1–12

[51] Fleute M., Lavallée S., and Desbat L. Integrated approach for matching statistical shape models with intra-operative 2D and 3D data. In Proc MICCAI, 2002. volume 2489 of LNCS, 364–372

[52] Frangi A. F., Rueckert D., Schnabel J. A., and Niessen W. J. Automatic 3D ASM construction via atlas-based landmarking and volumetric elastic registration. In Proc IPMI, 2001. volume 2082 of LNCS, 78–91

[53] Rueckert D., Frangi A. F., and Schnabel J. A. Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration. IEEE Trans Med Imaging. 2003;22(8):1014–1025

[54] Kelemen A., Székely G., and Gerig G. Elastic model-based segmentation of 3-D neuroradiological data sets. IEEE Trans Med Imaging 1999;18(10):828–839

[55] Fripp J., Crozier S., Warfield S., and Ourselin S. Automatic initialization of 3D deformable models for cartilage segmentation. In Proc Digital Image Computing: Techniques and Applications, 2005. 513–518

[56] Hill A. and Taylor C. J. Model-based image interpretation using genetic algorithms. Image Vision Comput, 1992;10(5):295–300

[57] Cootes T. F., Hill A., Taylor C. J., and Haslam J. The use of active shape models for locating structures in medical images. Image and Vision Computing Vol.12, No.6 July 1994. 355-366

[58] Van Ginneken B., Frangi A. F., Staal J. J., et al. Active shape model segmentation with optimal features. IEEE Trans. on Medical Imaging, Vol. 21, August, 2002. 924-933

[59] Toth R., Doyle S., Rosen M., et al. WERITAS – Weighted ensemble of regional image textures for ASM segmentation. In Proc. SPIE Medical Imaging, 2009. 725905-725905-11

[60] Asman A. J. and Landman B. A. Characterizing spatially varying performance to improve multi-atlas multi-label segmentation. In International Conference on Information Processing in Medical Imaging. Irsee, Bavaria, 2011

[61] Baka N., de Bruijne M., Reiber J. H. C., Niessen W., and Lelieveldt B. P. F. Confidence of model based shape reconstruction from sparse data In Proc. ISBI 1077–80

CHAPTER II

COMBINING REGISTRATION AND ACTIVE SHAPE MODELS FOR THE AUTOMATIC SEGMENTATION OF THE LYMPH NODE REGIONS IN HEAD AND NECK CT IMAGES

Antong Chen[1], Matthew A. Deeley[2], Kenneth J. Niermann[2], Luigi Moretti[2], and Benoit M. Dawant[1]

[1] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235

[2] Department of Radiation Oncology, Vanderbilt-Ingram Cancer Center, 1301 22nd Avenue South, Nashville, TN 37232

[This manuscript has been published in *Med. Phys.* Vol. 37, pp. 6338-6346, Dec. 2010]

ABSTRACT

Intensity-modulated radiation therapy (IMRT) is the state of art technique for head and neck cancer treatment. It requires precise delineation of the target to be treated and structures to be spared, which is currently done manually. The process is a time-consuming task of which the delineation of lymph node regions is often the longest step. Atlas-based delineation has been proposed as an alternative but, in our experience, this approach is not accurate enough for routine clinical use. Here, we improve atlas-based segmentation results obtained for level II, III and IV lymph node regions using an active shape model (ASM) approach. An average image volume was first created from a set of head and neck patient images with minimally enlarged nodes. The average image volume was then registered using affine, global, and local nonrigid transformations to the other volumes to establish a correspondence between surface points in the atlas and surface points in each of the other volumes. Once the correspondence was established, the ASMs were created for each node level. The models were then used to first constrain the results obtained with an atlas-based approach and then to iteratively refine the solution. The method was evaluated through a leave-one-out experiment. The ASM and atlas-based segmentations were compared with manual delineations via the Dice similarity coefficient (DSC) for volume overlap and the Euclidean distance between manual and automatic 3D surfaces. The mean DSC value obtained with the ASM-based approach is 10.7% higher than with the atlas-based approach, the mean and median surface errors were decreased by 13.6% and 12.0%, respectively. The ASM approach is effective in reducing segmentation errors in areas of low CT contrast where purely atlas-based

methods are challenged. Statistical analysis shows that the improvements brought by this approach are significant.

2.1. INTRODUCTION

As one of the state of art techniques for head and neck cancer treatment, intensity-modulated radiation therapy (IMRT) requires a precise delineation of both the target volume and the structures to be spared. Manually delineating contours in CT images, which is the standard of care, is a lengthy process even for an experienced physician. One of the most time-consuming tasks is the delineation of the cervical lymph node chain and the surrounding normal anatomical structures of the head and neck. The process of bilateral lymph node definitions for the entire neck can typically take between 20 and 45 minutes, depending on the patient's level of complexity. In contrast, delineation of the gross tumor volume typically requires on the order of 5 minutes or less. Furthermore, for many cancers of the head and neck, there is almost always some risk of spread of cancer to the cervical (neck) lymph nodes. In many cases the lymph nodes have microscopic disease even when they appear completely normal on CT, PET, or MRI. Instead of having a patient undergo a surgical sampling of all the lymph nodes of the neck, it is standard practice to deliver radiation prophylactically to these regions even when there is no radiological evidence of enlargement. An automatic technique capable of segmenting normal-looking lymph nodes could thus have a significant impact on the daily clinical load.

Atlas-based segmentation has been proposed as an approach to segment the lymph nodes. In this approach, structures of interest are delineated manually in one reference volume commonly called the atlas. This reference volume is then registered to other volumes to be segmented using rigid and nonrigid registration methods. The transformation that registers the reference volume to the other volumes can then be used

to project contours from the atlas to the patient volumes. This approach is commonly used to segment brain structures in well-contrasted high resolution MR images, while only some have used it for segmenting head and neck structures in CT images. Chao *et al*. [1] used an enhanced Demons algorithm to register a template image to patient images and used the transformations to delineate the lymph nodes, the left and right parotid glands, the spinal cord and the brainstem. Instead of using these automatically generated contours directly, they presented contours to physicians for modification and then compared these edited contours with manual delineations. In another study, Commowick *et al*. [2] projected lymph node contours from an average image volume to patient CT images using global affine and local nonrigid transformations. Although the volume-based error measure showed that, overall, the atlas-based delineations were acceptable, over-segmentations of the lymph node regions were observed. In a follow up study, Commowick *et al*. [3] proposed a scheme to select the most locally similar images to the patient image from a series of reference images, thus using several atlas volumes to segment the structures of interest. The quantitative validation performed in this study showed an improvement in specificity compared to a standard atlas-based method as well as a reduction in sensitivity. Gorthi *et al*. [4] also used an atlas-based approach, but the deformation was computed with structures easily visible in the images (bones, trachea, and skin) and then applied to the rest of the image. This led to relatively large (14.52 – 21.81mm) segmentation errors for the average Hausdorff distance for node levels II, III, and IV.

Although comparison between techniques is difficult without testing them on the same image volumes, our experience indicates that lymph node segmentation is a

challenge for purely intensity-based atlas-based methods because typical CT images for head and neck IMRT generally do not have particularly high resolution and because the contrast between lymph node regions and their surrounding regions is often poor. In this study, we complement an atlas-based approach with an active shape model (ASM) approach [5] to bring a priori information about the shapes of the structures in the segmentation process and constrain the deformation. The work we present herein initializes the ASM method with the result of an atlas-based, registration-driven approach. The ASM is constructed using a technique based on the method proposed by Frangi *et al*. [6], and the search algorithm for adapting the ASM is a variation on the local gray-level model proposed by Cootes *et al*. [7]. In the remainder of this article, we describe our segmentation method and compare it to a purely atlas-based technique.

2.2. METHODS AND MATERIALS

2.2.1. Data Description

The CT images used in this study with IRB (Institutional Review Board) approval are de-identified images from patients who underwent IMRT treatment for larynx and base of tongue cancers. We selected 15 volumes with no or minimally enlarged lymph nodes. They have a voxel size of around 1 mm in the x and y directions and a slice thickness of 3 mm. The images are acquired with a Philips Brilliance Big Bore CT scanner with the patient injected with 80 mL of Optiray 320, a 68% iversol-based nonionic contrast agent (manufactured by Mallinckrodt Inc., Hazelwood, MO). Typically, the images cover the head, the neck, and the upper chest.

For all 15 volumes, the level II, III, and IV lymph node regions on the right side were delineated following published guidelines [8] by the first author and reviewed carefully by two radiation oncologists (KN and LM). These manual delineations were saved in the form of binary masks as well as contours and were used to construct the ASM and validate the results of the experiments.

2.2.2. Construction of ASM through Registration

A prerequisite for creating active shape models is to establish a correspondence between points on the training shapes. Since it is difficult to manually localize corresponding points on a set of 3D surfaces, we used a method inspired by the work of Frangi *et al*. [5] who used both affine and nonrigid registrations for model building. The transformations produced by the registration process are used to relate points representing the shapes in different training images.

2.2.2.1 Construction of an average image volume

For the construction of the reference shape onto which all training shapes are aligned, an average image volume representing the centroid of the images is first constructed using the procedure proposed by Guimond *et al*. [9]. In this procedure, one volume in the set of images is chosen as a target. All the other volumes are subsequently registered to this target by a standard intensity-based affine registration algorithm that uses the normalized mutual information (NMI) [10] as similarity measure, and then further registered by an intensity-based nonrigid registration technique. The nonrigid registration is performed by the adaptive bases algorithm (ABA) [11] we proposed in the

past. This algorithm also uses the NMI as similarity measure and models the deformation fields as a linear combination of radial basis functions with finite support:

$$\vec{v}(\vec{x}) = \sum_{i=1}^{N} \vec{c}_i \cdot \Phi(\vec{x} - \vec{x}_i) \tag{2.1}$$

where Φ is one of Wu's compactly supported positive radial basis function [12], the $\vec{c}_i$'s are the coefficients to be optimized, and *N* is the number of basis functions (more details on this algorithm can be found in [11]). The major adjustable parameters determining the performance of the algorithm include the number of basis functions to be placed, a parameter controlling the difference between the coefficients of the adjacent basis functions, which is used to adjust the stiffness of the transformation (small values for this parameter lead to transformations that are more regularized than transformations obtained with large values), and the range of intensities used to compute the intensity histograms from which the NMI between images is estimated (adjusting this range permits to compute transformations that are driven, for instance, by soft tissue regions, by bony structures, or both). The algorithm produces transformations between a source and a target volume and between the target and the source volume that are constrained to be inverses of each other. To create the average volume the forward transformations (T_1 to T_k) registering the source images (i.e., the set of available image volumes) to the current average and their inverses (T_1^{-1} to T_k^{-1}) are computed. The forward transformations are applied to the source images and the resulting images are intensity-averaged. The inverse transformations are averaged and the resulting transformation is applied to the current intensity average to produce a volume that is both an intensity and a shape average of all the volumes, and the process is repeated until convergence. It has been shown in Ref. 6

that the final image volume is not dependent on the volume chosen to initiate the process, thus reducing potential bias introduced by selecting a particular volume as the initial reference. Note that the nonrigid registrations are performed on the full scale images, with an isotropic density of basis functions at 16mm per basis function, an experimentally determined stiffness value of 0.3, and the full intensity range. This parameter setting produces adequately regularized transformations with which the bones, body boundaries, and soft tissue regions are registered.

2.2.2.2 Establishing point correspondence for the creation of the ASM

Once the average volume is created, registrations are performed to acquire the transformations needed to find point correspondence. As shown in Figure 2.1, the process starts with an affine registration (Figure 2.1(a)) to align the images with the average volume, which produces transformations T_{a1} to T_{ak}. Even though, after affine registration, the images are aligned in the same space, the head and neck areas are not aligned accurately because the neck is much more flexible than some other structures such as the head. Large discrepancies in this area also exist between patients, including neck thickness, length and bending of the cervical vertebrae, and large anatomical differences in the surrounding soft tissues. A two-step nonrigid registration process is applied to compensate for these differences. First, a registration is performed on the full scale images to align mainly the bones and the body boundaries in each image and in the average volume. The same parameter setting as the one used for constructing the average volume is used except that the value of the stiffness parameters is reduced to 0.2, such that highly regularized transformations T_{n1} to T_{nk} are obtained. Second, a bounding box surrounding the lymph node regions and extending from the inferior part of the skull to

the bottom of the clavicle is defined on the average volume, and then copied onto the other images that are registered after the first step. When registering the images in the bounding boxes, as shown in Figure 2.1(b), the density of basis functions is increased to 8mm per basis function, the stiffness value is set at 0.3, and the intensity range is set to cover soft tissues, such that transformations that are less regularized and driven mainly by soft tissue regions are obtained. This permits to register the lymph node regions and their peripheral areas more accurately. The transformations T_{nb1} to T_{nbk} are obtained as well as their inverses.

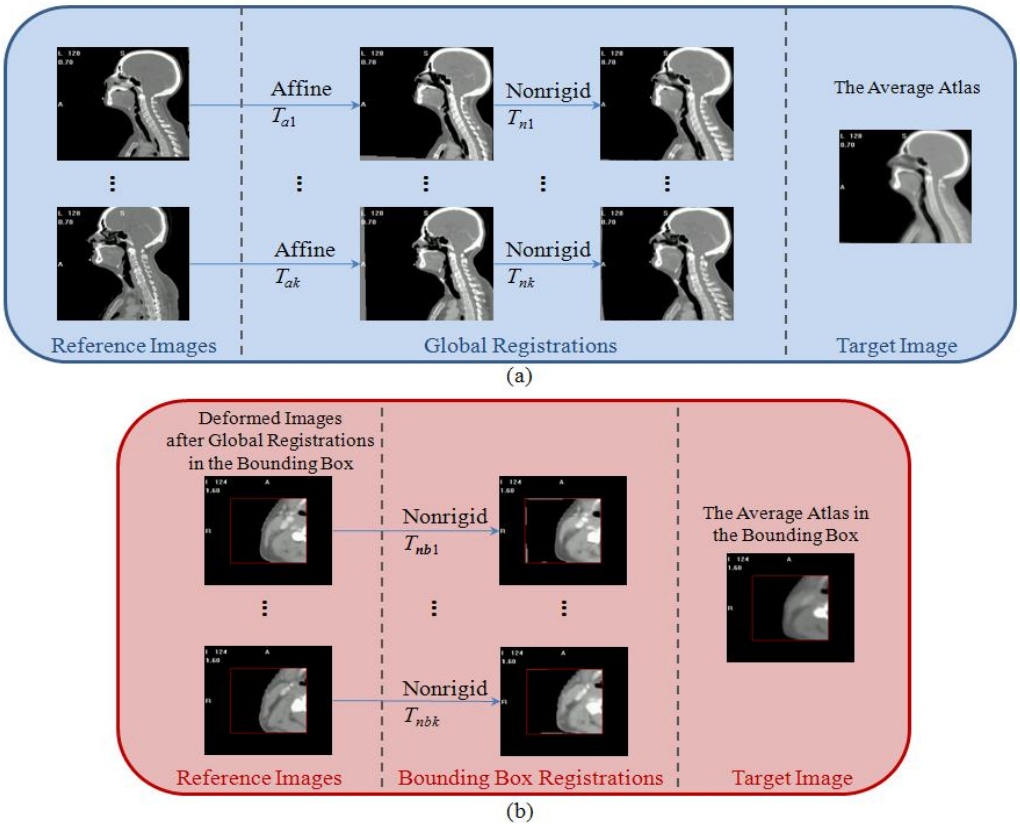

Figure 2.1. Flow charts illustrating the process used to register the training images and the average image volume. Panel (a) full scale image; panel (b) registration in the bounding box containing the nodes.

With all the transformations computed, the ASM is created following the steps shown in Figure 2.2. First, as shown in Figure 2.2(a), the manually segmented structures

from each of the volumes in the form of binary masks are projected onto the average volume by applying the affine transformations T_{a1} to T_{ak} and then the nonrigid transformations T_{n1} to T_{nk}. The projections are then averaged and thresholded to form a single binary mask representing the lymph node regions in the average volume. Several anatomical landmarks are subsequently identified manually in the average volume to separate this binary mask into three parts. These include the lower border of the hyoid, which separates the level II and level III, and the lower margin of the cricoid cartilage, which divides the level III and level IV. The landmarks define two flat surfaces separating the mask into three sections, representing the level II, III, and IV lymph node regions. As shown in Figure 2.2(b), surfaces are extracted from the binary objects in the bounding box using the ITK implementation of the marching cube algorithm [13]. This defines meshes in the space of the average volume in the bounding box for the three node levels. Correspondence between these points and points in each of the other images is established in two steps. First, the inverses of the transformations obtained from the nonrigid registrations in the bounding boxes, which are denoted as T_{nb1}^{-1} to T_{nbk}^{-1}, are applied to these points. Next, the inverses of the transformations computed in the global nonrigid registration step, which are denoted as T_{n1}^{-1} to T_{nk}^{-1}, are applied. For each vertex X, this produces a point X', which is mapped back into the space where all the images are affinely registered. Second, the correspondence is refined for the points X' that do not belong to the separating surfaces by finding the closest point X'' to X' on the manually segmented surfaces projected into the affinely aligned image space. This compensates for the inaccuracy in the registration process. The X'' points are then used to compute the

modes of variations for the structures of interest following the method described by Cootes *et al*. [5].

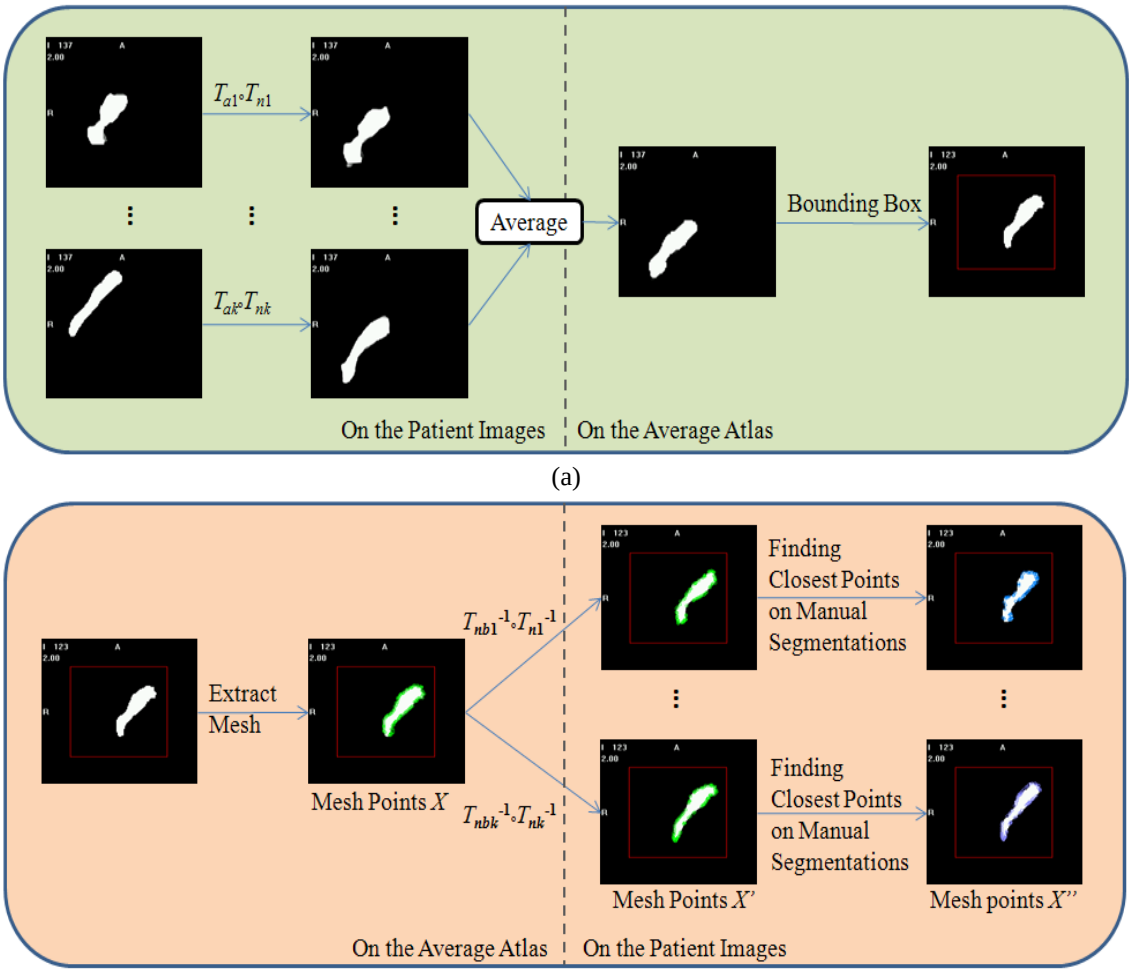

Figure 2.2. Flow charts illustrating the process used for the construction of the ASM using transformations obtained from registrations. Panel (a) creation of the mask in the average volume; panel (b) establishing point correspondence to create the ASM.

The *x*, *y* and *z* coordinates of the landmarks are concatenated into *k* vectors $\vec{x}_i$'s, for which a principal component analysis (PCA) is performed to obtain a linear model of shapes for each node level in the form of

$$\vec{x} = \bar{\vec{x}} + \Phi\vec{b} \tag{2.2}$$

where $\bar{\vec{x}}$ is the mean shape, $\Phi = [\vec{\phi}_1, \vec{\phi}_2, \ldots, \vec{\phi}_t]$ is the matrix of the first t eigenvectors associated with the highest eigenvalues of the covariance matrix, and $\vec{b} = [b_1, b_2, \ldots, b_t]^T$ is

the vector of model parameters. Notice that the models are built in the space where all images are aligned using affine registrations, this is because the ASM should focus on describing shape variations while excluding discrepancies due to large differences in patient orientation.

2.2.3. Segmentation of New Images

A new image is segmented as follows. First, the image is registered to the average volume using an affine and then a nonrigid global transformation computed as described earlier. The bounding boxes are extracted and nonrigid registration is performed again locally on these regions. The two nonrigid registrations produce displacement vectors for all the vertices on the structure of interest and map them from the average volume onto the affinely aligned image, forming a new shape. The active shape model is used to constrain the new shapes to conform to a shape compatible with the training set. This is accomplished by computing the linear combination of the modes of variation that best captures the new shape. Suppose the new shape is denoted as $\vec{x}_{new}$, the goal is to find the transformation *T* and model parameters $\vec{b}$ such that the new shape can be estimated as

$$\tilde{\vec{x}}_{new} = T(\bar{\vec{x}} + \Phi\vec{b}) \tag{2.3}$$

which is solved as a least squares estimation problem.

The resulting first segmentation is subsequently refined. For each vertex, a search vector is computed in the direction of the structure's surface normal, and possible boundary points are localized along this surface normal. A local gray-level model is then applied to determine which one of these candidate points can be selected as the best boundary point.

The local gray-level model is built based on that proposed by Cootes *et al.* [7]. For each point sitting on the manually delineated surface of a training image, the surface normal is calculated, and the intensity profile along the normal direction is sampled. For the corresponding points over the *k* training images, a total of *k* profiles are computed. The profiles can be built using a number of image properties such as the intensity or the intensity gradient. We explored several options, including the image intensity, gradient, and normalized gradient which is calculated as

$$\vec{g}_{iN} = \frac{\vec{g}_i}{\sqrt{g_{xi}^2 + g_{yi}^2 + g_{zi}^2}} \tag{2.4}$$

where $\vec{g}_i$ is the original gradient vector for point *i* with values g_{xi}, g_{yi}, and g_{zi} on each direction, and $\vec{g}_{iN}$ is the normalized gradient vector. We opted to use the normalized gradient of the training images because of its relative insensitivity to intensity variations caused by contrast agent washout in different patient images.

Figure 2.3 illustrates how the best boundary point is chosen from a set of *M* candidate points along the surface normal search vector. At each point along the search vector we extract a profile of length *N* for the candidate point and points on either side. We then compare this profile to the *k* profiles in the gray-level model. Cost is computed for each profile and the lowest cost is stored as the cost of the candidate point in question. The cost C_j for the *j*th candidate point is computed as the Euclidean distance between the profile of the candidate point, $\vec{p}_j$, and the profiles in the model, $\vec{p}_l$,

$$C_j = \arg\min_{1 \le l \le k} d(\vec{p}_j, \vec{p}_l) \tag{2.5}$$

where $d(\vec{p}_j, \vec{p}_l)$ is the Euclidean distance between two vectors. The candidate point along the search vector with the lowest cost is chosen as the new boundary. After all vertices are updated through this process, the displacements are constrained by the ASM, generating the shape for the next iteration. The process is repeated until successive iterations converge on a shape or a maximum number of iterations is reached. After the meshes for the three levels of lymph node regions are obtained, they are converted into binary masks and combined into one mask as the union of the three.

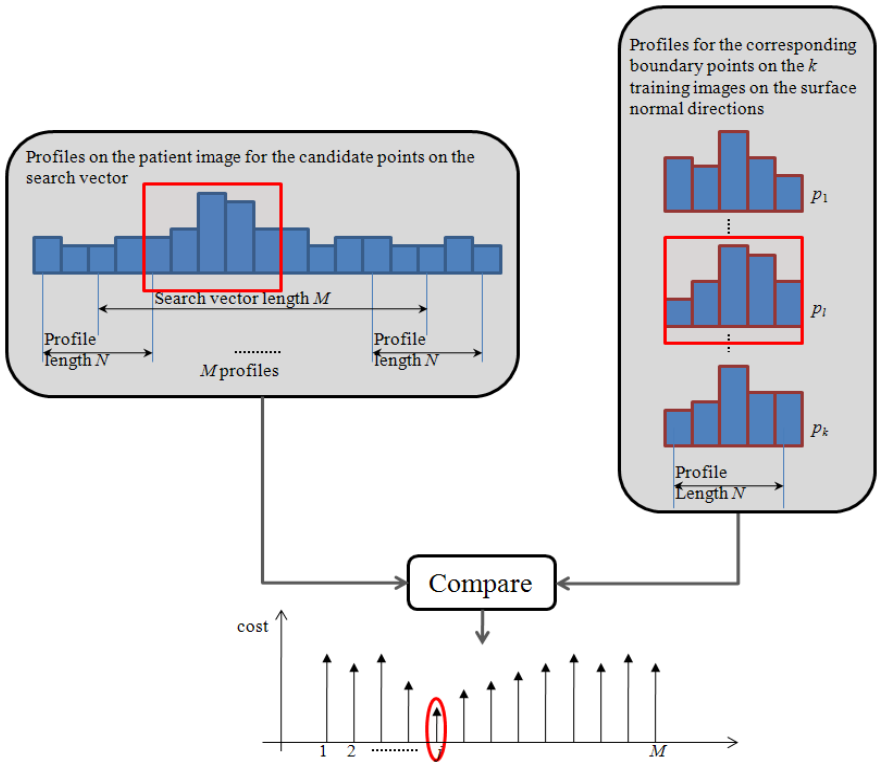

Figure 2.3. Search for the point on the search vector with the best fit to the gray-level model. The entire search vector consists of *M* candidates points. At each candidate point on the search vector, a profile of length *N* is calculated. These profiles are compared to analogous profiles in the training images.

2.2.4. Running Time

The affine and nonrigid registration algorithms used in this study are implemented in C and C++. Typical running time on a computer with a 2.93 GHz Intel Xeon quad-core

PC with the 64-bit Windows OS and 16 GB of memory is 2 min for the global affine component, 10 min for the global nonrigid component, and 3 min for the local nonrigid component. The model-fitting component is still implemented in MATLAB and takes on the order of 6 min.

2.3. RESULTS

All 15 volumes were used to create the average volume because the final volume is not sensitive to the volume chosen to initialize the process and to generate the average node mask in the average volume. Then, a leave-one-out strategy was used to create the ASMs and the intensity models and to evaluate these. For each run, one volume was eliminated from the image set, and the model was created using the remaining 14 volumes. This model was used to segment the 15*th* volume and the process was repeated 15 times. Validation was performed by comparing the automatic segmentation with a manual delineation used as the reference standard for comparison. The Dice similarity coefficient (DSC) [14] was used to evaluate the volumetric overlap between the manual and automatic segmentations. DSC is defined in Equation (2.6) as the overlap of two volumes normalized to their mean volume, where A and B represent the two binary segmentations and notation $N(A)$ represents the number of voxels contained in segmentation A

$$DSC = \frac{N(A \cap B)}{\frac{1}{2}(N(A) + N(B))} \tag{2.6}$$

The DSC is defined on [0, 1], where 0 indicates no overlap and 1 indicates identical segmentations with exact overlap. Volumetric measures such as the DSC can be

insensitive to boundary displacements that are small compared to the structure's size. To provide additional information, we calculated the Euclidean distance between the surfaces of the ASM-based and manual segmentations. To gauge the effect of the model on the segmentation, we also compared results obtained with the method we propose and results obtained solely with an atlas-based approach.

Table 2.1. Dice Similarity Coefficient (DSC) comparing atlas-based and manual segmentations, and ASM-based and manual segmentations.

Patient	DSC_atlas	DSC_ASM	Improvement in %
1	0.563	0.678	20.43
2	0.528	0.599	13.45
3	0.689	0.723	4.93
4	0.696	0.731	5.03
5	0.663	0.717	8.14
6	0.603	0.689	14.26
7	0.642	0.705	9.81
8	0.632	0.711	12.50
9	0.657	0.684	4.11
10	0.667	0.764	14.54
11	0.524	0.546	4.20
12	0.689	0.748	8.56
13	0.651	0.760	16.74
14	0.612	0.728	18.95
15	0.646	0.680	5.26
Mean	0.631	0.698	10.73

Table 2.1 shows that in all 15 cases the ASM-based segmentations have a higher DSC than those obtained with a purely atlas-based method; the improvement brought by the method we propose ranges from 4.1% to 20.4% with an overall improvement of 10.7%.

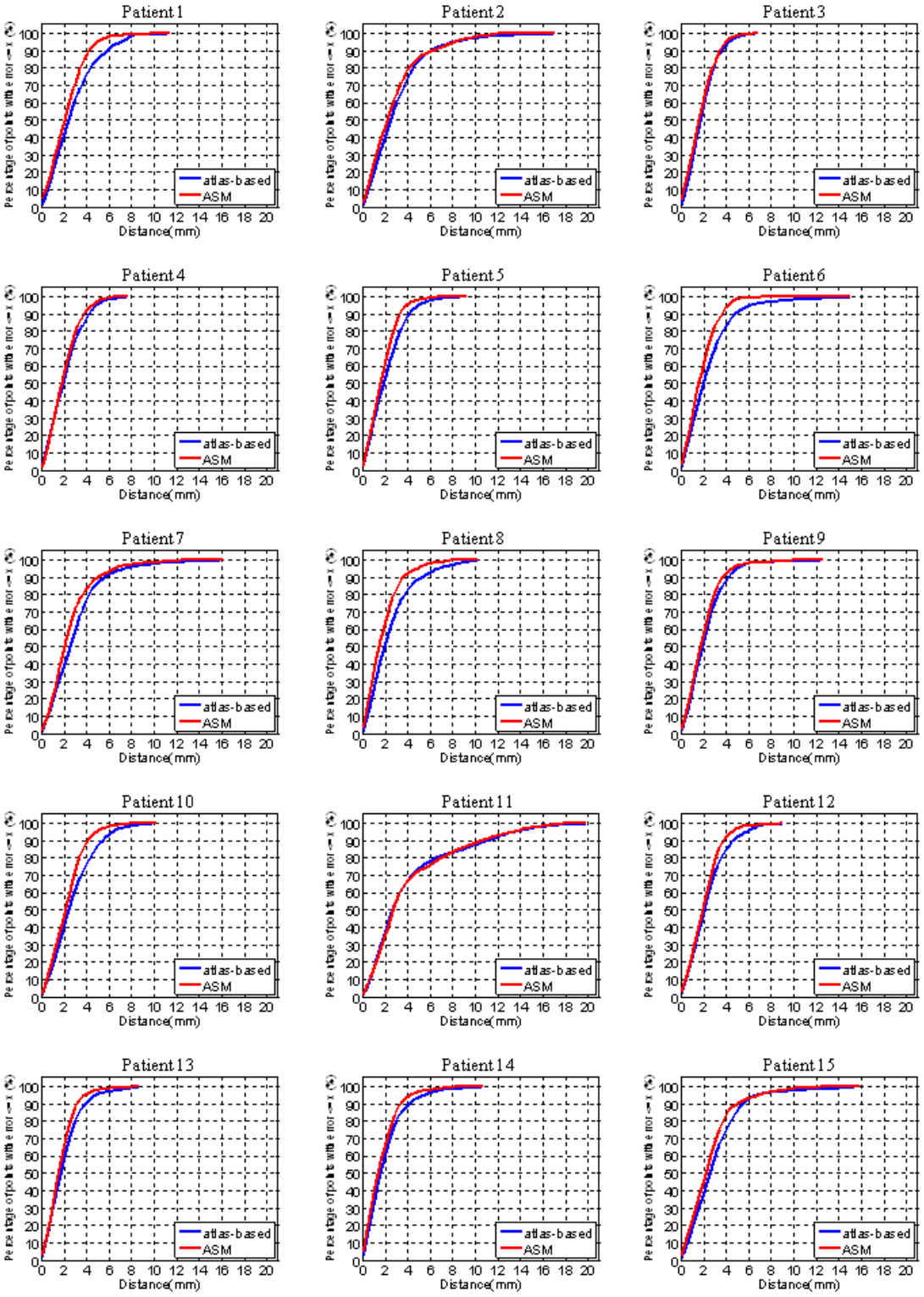

Figure 2.4. Cumulative distributions for the surface errors for each volume, with the *x*-axis showing a distance error and the *y*-axis showing the percentage of points for which the error is smaller than this distance.

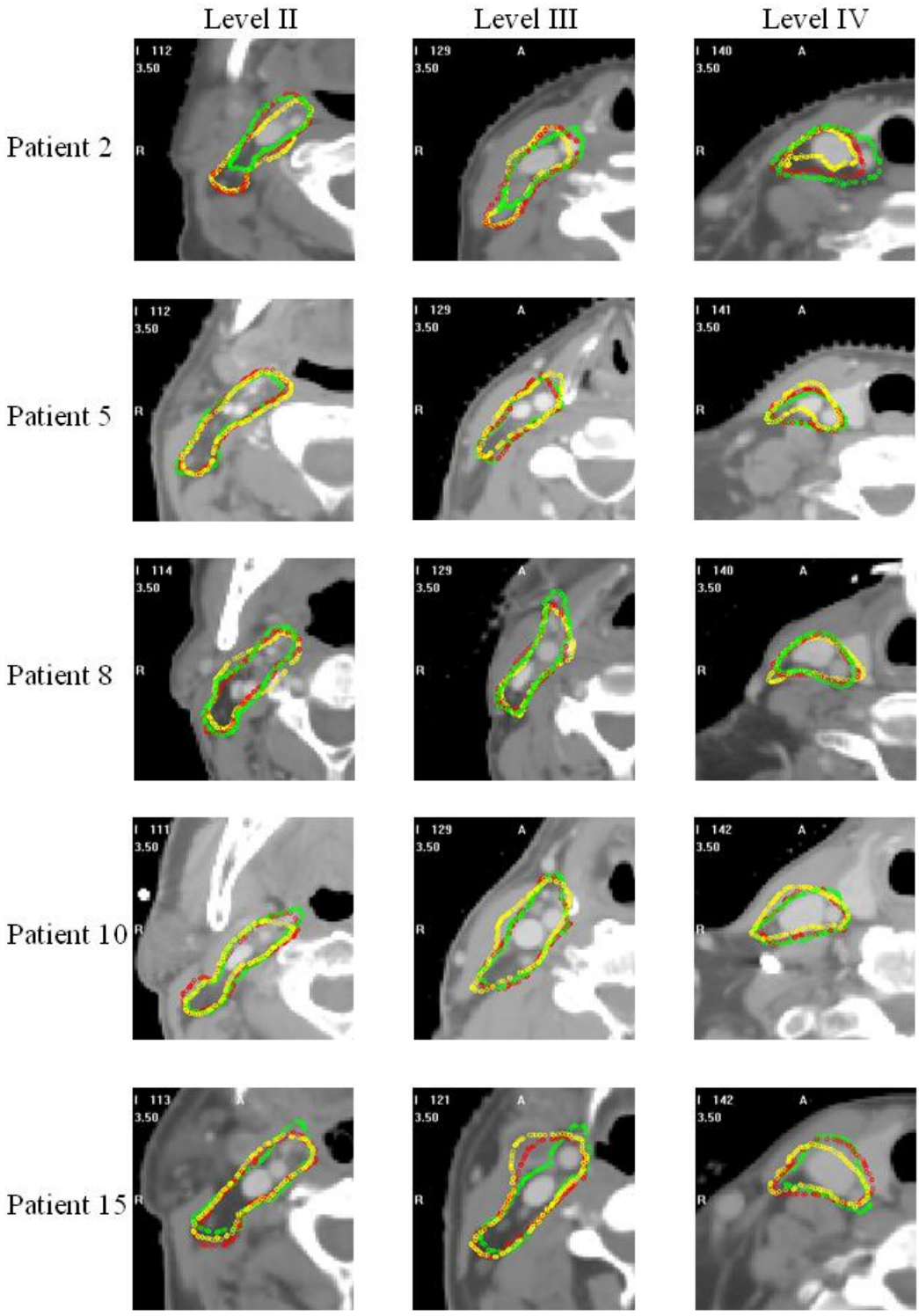

Figure 2.5. 2D Contours for the manual, atlas-based, and ASM-based segmentations for patient 2, 5, 8, 10 and 15. Shown from the left to the right are contours in the level II, III, and IV regions; the manual contour is in green, the atlas-based in yellow, and the ASM-based in red.

Table 2.2. Mean, Median, and Max errors for atlas-based (Mean_atlas, Median_atlas, and Max_atlas) and ASM-based methods (Mean_ASM, Median_ASM, and Max_ASM) in mm.

Patient	Mean_ atlas	Mean_ ASM	Improvement in %	Median_ atlas	Median_ ASM	Improvement in %	Max_ atlas	Max_ ASM	Improvement in %
1	2.87	2.29	20.20	2.48	2.13	13.96	11.53	10.02	13.10
2	3.10	2.95	4.89	2.56	2.25	12.09	17.10	11.95	30.09
3	1.96	1.82	7.27	1.78	1.65	7.07	6.94	6.49	6.46
4	2.18	2.03	6.73	1.99	1.86	6.78	7.72	6.37	17.51
5	2.17	1.80	16.94	1.93	1.67	13.70	9.24	6.59	28.67
6	2.51	1.85	26.29	2.02	1.63	19.28	15.10	7.01	53.56
7	3.00	2.56	14.55	2.61	2.10	19.72	16.10	12.92	19.75
8	2.50	1.84	26.52	2.02	1.54	24.00	10.32	7.92	23.22
9	2.22	2.02	9.18	1.97	1.79	9.05	12.64	9.91	21.59
10	2.82	2.33	17.34	2.48	2.21	11.07	10.29	7.84	23.77
11	4.22	4.22	-0.09	2.70	2.80	-3.67	19.88	18.40	7.46
12	2.40	2.12	11.60	2.13	1.96	7.86	9.07	8.86	2.37
13	2.05	1.77	13.64	1.78	1.57	11.92	8.65	8.38	3.09
14	2.04	1.72	15.58	1.66	1.45	12.78	10.79	8.35	22.66
15	3.02	2.64	12.60	2.64	2.28	13.69	15.93	12.79	19.70
Mean	2.60	2.26	13.55	2.18	1.92	11.95	12.09	9.59	19.53

The mean, median and maximum distances to the manual surface are shown in Table 2.2. In this table, Mean_atlas, Median_atlas, and Max_atlas, refer to the results obtained with atlas-based segmentation alone. Mean_ASM, Median_ASM, and Max_ASM are results obtained with the method we propose. For these three measures, Table 2.2 also presents the percent improvements brought by the model-based approach. These range from 0% to 26.3% for the means, -3.7% to 24.0% for the medians, and 2.3% to 53.6% for the max errors. The only case for which no improvement has been observed is case 11. This is a special case because a tumor located in the trachea pushed the thyroid to where vessels are normally located in the level IV lymph node region. Because of this, the registration step was inaccurate. The model was thus also initialized incorrectly and converged to the wrong solution. One-sided t-tests were performed to test the statistical significance of the differences observed for the DSC, mean, median, and max values. In

all cases these differences were significant ($p < 0.01$). Figure 2.4, which shows cumulative distributions for the surface error for each volume (i.e., the *x*-axis is a distance error and the *y*-axis is the percentage of points for which the error is smaller than this distance), illustrates the effect of the model on the segmentation error. In all cases except for patient 11, the cumulative distribution curve for the model-based approach is above the curve for the atlas-based approach.

The slice thickness in the volumes used in this study is 3mm. One observes that the model-based approach leads to results with more than 90% of the surface points having a distance error less than 4mm, which is on the order of one voxel in the axial direction, for all cases except cases 2, 7, 11, and 15. The model-based approach leads to substantial improvements for cases 1, 5, 6, 8, 10, 12, 13, and 14. For cases 3, 4, and 9 the 90% threshold was reached with the atlas-based approach alone. A close look at patient 2 (see Figure 2.5) shows errors at levels II, III and IV. At level II the manual contour was drawn smaller than the contour produced by our algorithm. Retrospective discussions with the radiation oncologists determined that the observed difference was within the inter-rater variability. At level III a reactive but not pathological enlarged node occupies the place normally occupied by interstitial fat, which has a lower intensity than other tissues in CT images. The automatic contour includes the node when it is excluded in the manual contour. Again, retrospective discussion with the radiation oncologists determined that both were acceptable and a function of the physician's preferences. The error at the level IV is caused by the size of the thyroid, which is smaller than usual. The main segmentation error for patient 7 occurred at level IV. This subject has a thyroid that is much larger than usual and the contrast between the thyroid and surrounding tissues is

low. As a consequence, the registration was inaccurate, the model was initialized incorrectly, and the ASM component of the system became trapped in a local minimum.

In patient 15 (see Figure 2.5), the largest error was at the end of level II towards level III. At this place, a large node infiltrated by metastatic cancer was visible. This also produced registration and initialization errors that could not be recovered. Retrospective discussion with the radiation oncologists established that this area should have been treated with a higher dose and not part of a prophylactic regimen. A substantial error was also visible at level IV because of an enlarged thyroid.

Figure 2.5 shows contours superimposed on the images for 5 subjects. In each case, one representative slice per level has been chosen. The green, yellow, and red contours are the manual, atlas-based, and model-based contours, respectively. This figure confirms what is shown in Figure 2.5. Subjects 5, 8, and 10 are cases for which the cumulative distributions show a clear improvement. For these three subjects, the red contours are indeed closer to the green contours than the yellow ones. For subjects 2 and 15, the cumulative distributions do not show a substantial difference between the two approaches. In these two cases, the model could not compensate for registration errors caused by abnormal anatomy.

2.4. DISCUSSION AND CONCLUSIONS

We have developed a method for the segmentation of normal-looking lymph nodes in clinically acquired head and neck CT scans that improves upon atlas-based approaches, which have been proposed to solve this problem. As discussed in the background section, prophylactic treatment of normal-looking lymph nodes is within

standard practice for many head and neck cancers, and their delineation is a time-consuming process. Reliable methods designed for their automatic segmentation may thus have a substantial impact on the daily clinical load. Previously, we had used a single model for all three node levels (see [15]). This approach led to mixed results, i.e., in some cases the model-based approach led to better results while in others it did not. Separating the model into three models, one for each level, improved the results. As reported in this study, the model-based approach now leads to better results than a purely atlas-based method in all cases with normal-looking anatomy. In all these cases, more than 90% of the surface points have a distance error of less than 4mm, which is on the order of one voxel in the axial direction.

Comparison to other studies is difficult not only because the data sets are different but also the evaluation methods vary amongst studies. In the work of Gorthi *et al.*[4] CT images are of similar size (0.9375mm×0.9375mm×3mm) and the sensitivity, DSC and Hausdorff distance are reported. In their leave-one-out experiment, the mean DSC reported for level IIA, IIB, III, and IV are 0.53, 0.46, 0.43 and 0.36 respectively, while the mean DSC for our ASM-based segmentation is 0.698 for level II, III and IV segmented as a single structure. The mean Hausdorff distance that are reported are 14.52mm, 15.06mm, 18.68mm, and 21.81mm for level IIA, IIB, III, and IV. The comparable average maximum distance error is 9.59 mm for our ASM-based approach. In similar work Commowick *et al.* [3] reported mean sensitivity of 0.692, specificity of 0.813, and combined error of 0.360. But sensitivity and specificity numbers are difficult to interpret for segmentation tasks. Indeed, sensitivity is defined as $TP/(TP+FN)$ and specificity as $TN/(TN+FP)$ with *TP* (true positive) the number of voxels included in both

the manual and automatic contours, *TN* (true negative) the number of voxels excluded by both methods, *FP* (false positive) the number of voxels in the automatic segmentation but not in the manual segmentation, and *FN* (false negative) the number of voxels in the manual segmentation but not in the automatic segmentation. Sensitivity and specificity need to be reported together because the former does not measure over-segmentation and the latter does not measure under-segmentation. In addition, the definition of the specificity involves *TN,* which for segmentation tasks is the intersection of the manual and automatic background regions. These can be made arbitrarily large thus leading to large specificity values and requiring heuristic criteria to define *TN,* as discussed by Isambert *et al.*[16].

The results we have presented also show shortcomings of the current approach. Abnormal anatomy and/or pathology (cases 2, 7, 11, and 15) lead to poorer results, mainly because inaccuracy in the registration results in poor initialization of the model. While one may expect increased robustness to anatomical variations with larger training sets, segmenting volumes with large pathologies is more challenging and will, in all likelihood, require different strategies. One possibility is to use a mixed approach in which the pathology is delineated by hand first and then used as constraint to guide the segmentation process. Based on our observation, the thyroid region remains challenging because of large variations in the shape and size of this organ. Reducing the sensitivity of our approach to initialization errors may be possible by modifying the algorithm that is used to update the boundary points. For instance, Van Ginneken *et al*. [17] have proposed a technique in which optimal features are selected for each landmark using a *k*NN-classifier. Toth *et al*. [18] also propose a method in which an optimal weighted

average of texture features is used to establish correspondence. Using neighborhood information, as proposed by these authors, instead of line information may permit the algorithm to escape from local minima. Whether or not these improvements are able to compensate for inaccuracy in the registration process caused by anatomical variations will need to be determined.

REFERENCES

[1] K. S. Chao et al., "Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach," Int. J. Radiat. Oncol., Biol., Phys. 68,1512–1521 2007.

[2] O. Commowick, V. Grégoire, and G. Malandain, "Atlas-based delineation of lymph node levels in head and neck computed tomography images," Radiother. Oncol. **87**, 281–289 2008.

[3] O. Commowick, S. K. Warfield, and G. Malandain, "Using Frankenstein's creature paradigm to build a patient specific atlas," Proceedings of the 12th International Conference on Medical Image Computing and Computer Assisted Intervention _MICCAI '09_, Vol. 5762, Pt. II, pp. 993–1000, September 2009 .

[4] S. Gorthi, V. Duay, N. Houhou, M. B. Cuadra, U. Schick, M. Becker, A. S. Allal, and J. P. Thiran, "Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration," IEEE J. Sel. Top. Signal Process. **3**, 135–147 2009.

[5] T. F. Cootes, C. J. Taylor, C. H. Cooper, and J. Graham, "Active shape models-their training and application," Comput. Vis. Image Underst. **61**, 38–59 1995.

[6] A. F. Frangi, D. Rueckert, J. A. Schnabel, and W. J. Niessen, "Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling," IEEE Trans. Med. Imaging **21**, 1151–1166 2002.

[7] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam, "The use of active shape models for locating structures in medical images," Image Vis. Comput. **12**, 276–285 1994.

[8] G. Ausili Cefaro, C. A. Perez, D. Genovesi, and A. Vinciguerra, *A Guide for Delineation of Lymph Nodal Clinical Target Volumes in Radiation Therapy* Springer, New York, 2008.

[9] A. D. Guimond, J. Meunier, and J. P. Thirion, "Average brain models: A convergence study," Comput. Vis. Image Underst. **77**, 192–210 2000.

[10] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," Pattern Recogn. **32**, 71–86 1999.

[11] G. K. Rohde, A. Aldroubi, and B. M. Dawant, "The adaptive bases algorithm for intensity-based nonrigid image registration," IEEE Trans. Med. Imaging **22**, 1470–1479 2003.

[12] Z. Wu, “Compactly supported positive definite radial functions,” Adv. Comput. Math. **4**, 283–292 1995.

[13] Insight Segmentation and Registration Toolkit, http://www.itk.org/.

[14] L. R. Dice, “Measures of the amount of ecologic association between species,” Ecology **26**, 297–302 1945.

[15] A. Chen, M. A. Deeley, K. J. Niermann, L. D. Moretti, and B. M. Dawant, “Segmentation of lymph node regions in head-and-neck CT images using a combination of registration and active shape model,” Proceedings of SPIE Medical Imaging, Vol. 7623, p. 76231Q, 2010.

[16] A. Isambert *et al.*, “Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context,” Radiother. Oncol. **87**, 93–99 2008.

[17] B. van Ginneken *et al.*, “Active shape model segmentation with optimal features,” IEEE Trans. Med. Imaging **21**, 924–933 2002.

[18] R. Toth *et al.*, “WERITAS–Weighted ensemble of regional image textures for ASM segmentation,” Proceedings of SPIE Medical Imaging, pp. 725905, 2009.

CHAPTER III

EVALUATION OF MULTIPLE-ATLAS-BASED STRATEGIES FOR THE SEGMENTATION OF THE THYROID GLAND IN HEAD AND NECK CT IMAGES FOR IMRT

Antong Chen[1], Kenneth J. Niermann[2], Matthew A. Deeley[3], and Benoit M. Dawant[1]

[1] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235

[2] Department of Radiation Oncology, Vanderbilt-Ingram Cancer Center, 1301 22nd Avenue South, Nashville, TN 37232

[3] Medical Physics Division, Fletcher Allen Health Care and Department of Radiology, University of Vermont, Burlington, VT 05401

[This manuscript has been published in *Phys. Med. Biol.* Vol. 57, pp. 93-111, Jan. 2012]

ABSTRACT

Segmenting the thyroid gland in head and neck CT images is of vital clinical significance in designing intensity-modulated radiation therapy (IMRT) treatment plans. In this work, we evaluate and compare several multiple-atlas-based methods to segment this structure. Using the most robust method, we generate automatic segmentations for the thyroid gland and study their clinical applicability. The various methods we evaluate range from selecting one single atlas based on one of three similarity measures, to combining the segmentation results obtained with several atlases and weighting their contribution using techniques including a simple majority vote rule, a technique called STAPLE that is widely used in the medical imaging literature, and the similarity between the atlas and the volume to be segmented. We show that the best results are obtained when several atlases are combined and their contributions are weighted with a measure of similarity between each atlas and the volume to be segmented. We also show that with our data set, STAPLE does not always lead to the best results. Automatic segmentations generated by the combination method using the correlation coefficient (CC) between the deformed atlas and the patient volume, which is the most accurate and robust method we evaluated, are presented to a physician as 2D contours and modified to meet clinical requirements. It is shown that about 40% of the contours of the left thyroid and about 42% of the right thyroid can be used directly. An additional 21% on the left and 24% on the right require only minimal modification. The amount and the location of the modifications are qualitatively and quantitatively assessed. We demonstrate that, although challenged by large inter-subject anatomical discrepancy, atlas-based segmentation of the thyroid gland in IMRT CT images is feasible by involving multiple

atlases. The results show that a weighted combination of segmentations by atlases using the CC as the similarity measure slightly outperforms standard combination methods, e.g., the majority vote rule and STAPLE, as well as methods selecting one single most similar atlas. Results we have obtained suggest that using our contours as initial contours to be edited has clinical value.

3.1. INTRODUCTION

Intensity modulated radiation therapy (IMRT) requires a precise delineation of structures to be treated and of organs to be spared on the pre-treatment planning CT images. For head and neck cancer IMRT, the thyroid gland is one of the most important organs to be spared. Irradiating the thyroid gland may result in thyroid dysfunction, which is an important clinical complication of radiation treatment that may be manifest as either chronic hypothyroidism requiring long-term daily hormone replacement therapy or, less commonly, the potentially fatal acute thyroid storm clinical syndrome. The reported incidence of thyroid dysfunction in patients undergoing conventional methods of radiation treatment ranges from 6 to 48% in retrospective studies [1-10].

Although most external beam radiation treatment plans require entry or exit beams to pass directly through the thyroid gland to reach the primary tumor region and/or the region of surrounding cervical lymph node regions which are at risk for cancer involvement, damage to the gland still needs to be controlled to avoid severe complications. Since the risk for "collateral damage" to the thyroid gland appears to be related to the volume exposed to radiation [4,9,11] as well as the overall intensity of radiation passing through the gland [4, 12-14], precise delineation serves as the basis for constraining the radiation fluence resulting from the inverse treatment planning process.

Developing automatic or semi-automatic methods for the segmentation of the thyroid is important because precise manual delineation of this structure is time-consuming, even for experienced radiation oncologists. Because the thyroid is inhomogeneous, and because it is surrounded by structures that have similar intensity, its

segmentation using standard intensity-based methods such as level-set [15] or graph-cut techniques [16] is challenging. In the recent past we have used model-based methods to segment the lymph node regions [17] but the anatomical variability we have observed for the thyroid makes the creation of reliable statistical shape models difficult for this structure. Atlas-based methods have been proposed as an alternative for the segmentation of head and neck structures but not for the thyroid. These methods require the selection or construction of one or multiple atlas images, with the structures of interest delineated precisely on the atlas(es). Image registration is then used to compute the transformations required to propagate the segmentations onto the patient image. Chao *et al.* [18] delineated the clinical target volume (CTV), left and right parotid glands, spinal cord, brainstem and optical track using one atlas. The automatically generated contours were modified and then compared with manual delineations. Commowick *et al.* [19] used an average atlas to segment the mandible, parotid glands, submandibular glands, spinal cord, brainstem, and lymph node regions with good precision (sensitivity and specificity higher than 0.8). However, over-segmentations of lymph node regions were observed, and a limitation of the method was revealed when the patient had large anatomical discrepancy compared with the atlas. To overcome these problems, Commowick *et al.* [20] proposed a scheme to select the locally most similar images from the set of atlases to construct a piecewise most similar atlas. This reduced the over-segmentation problem, which resulted in an improvement in specificity, but the sensitivity was reduced. Gorthi *et al.* [21] used a single-atlas-based method to segment the lymph node regions, but the accuracy of the segmentations was limited (average Dice similarity coefficient (DSC) [22] <0.5 and average Hausdorff distance (HD) [23] >17mm). Han *et al.* [24] segmented

several muscle groups in addition to the same set of structures as those segmented by Commowick *et al.* in [19]; these did not include the thyroid. The median DSC for five out of seven structures to be spared were 0.8 or above using a method combining segmentations from multiple atlases. It was also shown that the method outperformed the method using the most similar atlas selected based on mutual information (MI) [25, 26]. In the later approach only three out of seven structures reached the 0.8 mark. Both methods had a median DSC above 0.55 for all structures.

Compared with the conventional technique that uses a single atlas, the major advantage of using a multiple-atlas-based approach as we do in this study is its ability to reduce the possibly large discrepancy between a single atlas and the patient image. In general, atlases are registered to the patient image, and the segmentations are propagated from each atlas to the patient image using the transformations obtained. The final segmentation on the patient image is then established through either combining or selecting among these propagated segmentations. Wu *et al.* [27] proposed to select the single optimal atlas for each region of interest (ROI) based on the measure of local normalized mutual information (NMI) [28]. Heckermann *et al.* [29] applied the majority vote rule to fuse segmentations from up to 29 atlases, and found that using about 15 to 20 atlases was sufficient. Further increasing the number of atlases did not improve segmentation accuracy very much. The method was later improved [30] by enhancing the robustness of the nonrigid registration algorithm with an approximate tissue classification at the coarse levels of the multi-resolution implementation. Rohlfing *et al.* [31] compared three techniques: selecting the most similar atlas, using an average shape atlas, and using multiple atlases and determining the final segmentation by the majority vote rule. In their

study, the last method showed the best performance. Instead of using the majority vote rule, which assigns equal weight to each atlas, Warfield *et al.* [32] weighted the segmentations through an expectation maximization (EM) algorithm known as the simultaneous truth and performance level estimation (STAPLE) algorithm. This algorithm has been used as a standard technique for combining automatic or manual segmentations from multiple raters. Rohlfing *et al.* [33] expanded the original STAPLE algorithm such that it could be used to simultaneously combine labels for multiple classes. Experiments they performed on bee brain confocal microscopy images showed that the proposed method performed better than majority vote. Klein *et al.* [34] combined segmentations from a set of atlases using both STAPLE and an altered version of the vote rule, which weighted the contribution of each atlas with the value of the NMI between the atlas and the volume to be segmented. In their study, they showed that STAPLE did not perform better than the vote rule. They also found that using multiple atlases outperformed selecting one single optimal atlas. Aljabar *et al.* [35] studied the effect of increasing the number of atlases that were ranked by the value of the NMI between the registered atlases and the volume to be segmented. This study showed that using 20 atlases from a set of 275 was optimal. Artaechevarria *et al.* [36] compared strategies for combining segmentations by multiple atlases including STAPLE, majority vote, and weighted voting methods based on global or local similarity between patient and atlas images after affine and nonrigid registrations. The experiment on a set of 18 brain MR images showed that, among the methods that were evaluated, local weighted voting based on measuring similarity in the neighborhood of the structure of interest performed the best. In addition to the typical local weighted voting which assigns the weights only once,

Sabuncu *et al.* [37] proposed an iterative method to optimize the weights through EM. Different from STAPLE, which calculates the weights based only on the segmentations, the method also takes the intensity information of the registered images into consideration. Langerak *et al.* [38] proposed a selective and iterative method for performance level estimation (SIMPLE) to combine segmentations without EM, and the experimental results on a set of 100 prostate MR volumes showed that SIMPLE outperformed STAPLE in both accuracy (statistically significant improvements on volume-wise similarity with manual segmentations) and computation time (reduction to about 1/4 to 1/3 of STAPLE).

In our experience, inter-subject variations are large in head and neck CT images, especially for the areas near the thyroid gland, where differences can be caused by the existence of tumors, surgical procedures, or simply normal tissue variation. These problems challenge conventional single-atlas-based approaches and may be a very good application for multiple-atlas-based approaches. Because there is no consensus on the best atlas combination method to use, we investigate the use of several approaches involving multiple atlases to segment the thyroid gland in a set of clinical scans. In the study discussed herein, different methods for selecting or combining segmentations were compared. Results obtained with these various methods were evaluated quantitatively and qualitatively. The method that generated the most accurate results was then employed, and its resulting contours were presented to a practicing radiation oncologist. The subsequent amount of contour-editing was then assessed to measure the clinical usefulness of the automatic contouring method.

3.2. METHODS AND MATERIALS

3.2.1. Description of data

The 20 CT images used in this study with institution review board (IRB) approval are de-identified images from patients undergoing IMRT treatment for larynx and base of tongue cancers. Sixteen of them show normal anatomy in the area of the thyroid gland. The four remaining ones had a tracheotomy, which involves inserting a plastic tube into the larynx to help breathing. One of these four (patient 11) also had the trachea filled by a large tumor that substantially altered the anatomy around the thyroid gland. These images have a voxel size of approximately 1 mm in the *x* and *y* directions and a slice thickness of 3 mm. For all 20 volumes, the thyroid glands were manually delineated by the first author and reviewed by a radiation oncologist (KN). Changes were made carefully to meet the anatomical and clinical requirements (this was done with greater care than what is done for routine clinical cases), and the final manual delineations were saved in the form of contours and binary masks.

3.2.2. Registration programs

The affine registration program used in this study has been developed in-house. It is a standard algorithm that is intensity-based and uses NMI as the similarity measure. The nonrigid registrations are performed using the adaptive bases algorithm (ABA) [39] also developed in-house. This algorithm also uses NMI defined below as the similarity measure,

$$NMI(A, B') = \frac{H(A) + H(B')}{H(A, B')} \tag{3.1}$$

where H(A) is the marginal entropy of the source image, H(B') is the marginal entropy of the transformed target image, and H(A, B') is their joint entropy. This similarity measure is used to optimize a deformation field modeled as a linear combination of radial basis functions with local support,

$$\vec{v}(\vec{x}) = \sum_{i=1}^{N} \vec{c}_i \cdot \Phi(\vec{x} - \vec{x}_i) \tag{3.2}$$

where Φ is one of Wu's compactly supported positive radial basis functions [40], $\vec{c}_i$ is a vector of coefficients to be optimized, and N is the number of basis functions. Three major parameters are used in this algorithm: (1) The density of basis functions, which determines the scale of the transformation. Few basis functions with large support lead to transformations that are more global than transformations obtained with many basis functions with small support. (2) One parameter that constrains the difference between the coefficients of the adjacent basis functions, which we call the elasticity parameter. This parameter is used to control the regularity of the transformation. A small value for this parameter leads to transformations that are more regularized than transformations obtained with large values. (3) The range of intensities, which is used to compute the intensity histograms from which the NMI between images is estimated. It is used, for instance, to specify whether the deformations are driven by soft tissue regions or bony structures, or both. The algorithm produces forward (from the source image, i.e., the atlas, to the target image, i.e., the patient image volume) and backward (from the target image to the source image) transformations that are inverses of each other.

3.2.3. Registration and segmentation procedure

Instead of directly performing registrations between each atlas and the patient image at full resolution, which can be a very time-consuming process given the size of the CT image volumes, we define a common space in which all the volumes are registered with affine and heavily regularized nonrigid transformations. Nonrigid registration can then be obtained in smaller bounding boxes, thus speeding up the calculations. To create this common space we first compute an average volume to which all the other volumes are registered.

3.2.3.1 Construction of an average image volume

We follow the procedure proposed by Guimond *et al.* [41] for the construction of the average image volume. In this method, one image in the set of images is selected randomly as the initial target. All the other volumes are then registered affinely to this first target. This compensates for large differences in pose or shape between the volumes. Nonrigid registrations are then computed between each of the affinely registered volume and the initial target, producing forward and backward deformation fields. The images are deformed using the forward fields and the resulting images intensity-averaged. This produces an intensity-average volume. The backward deformation fields are averaged and this average field is used to deform the intensity-average volume. At the end of this step a shape- and intensity-average volume representing the population as a whole has been produced. The process is repeated, taking the current shape- and intensity-average volume as the target, until convergence is reached. In our experience, convergence is typically reached in 3 to 4 iterations. The nonrigid registrations are computed with a parameter setting that is suitable for aligning all classes of tissues (an isotropic density of

basis functions at 16 mm per basis function, a moderate value for the elasticity parameter of 0.3, and the entire intensity range was used to compute 32-bin histograms that were used to estimate the entropy of the images). We have observed that because anatomical variations in the region of the thyroid are large, the accurate registration of CT images in this area is difficult. As a consequence, the average volume we obtain is very blurry in this area, which undermines its potential to be used as the atlas. The average volume is thus only used as a common space where all atlas images are globally aligned.

3.2.3.2 Segmentation of patient images

To segment a new patient image volume, this volume is first registered to the average image volume using affine and nonrigid transformations following the procedure shown in the top row of Figure 3.1. The parameters for the nonrigid registration (an isotropic density of basis functions at 16 mm per basis function, a smaller value for the elasticity parameter of 0.2, and the full intensity range for histogram computation) are set to produce a transformation T_n that is highly regularized. These two registrations result in the alignment of the bony structures and the outside body contours between the patient image and the atlas.

A bounding box surrounding the thyroid gland is defined on the average image volume and copied onto the patient and atlases. As shown in the bottom row of Figure 3.1, nonrigid registrations are then performed between each atlas and the patient image inside the bounding box using parameters (an isotropic density of basis functions at 6.25 mm per basis function, a moderate value for the elasticity parameter of 0.3, and intensity range to compute the intensity histograms limited to the soft tissue range) permitting more flexible transformations T_{nbk}'s for aligning the soft tissue regions. With these

transformations, segmentations of the thyroid gland are propagated from each atlas to the patient image. These segmentations are then combined using various methods described below.

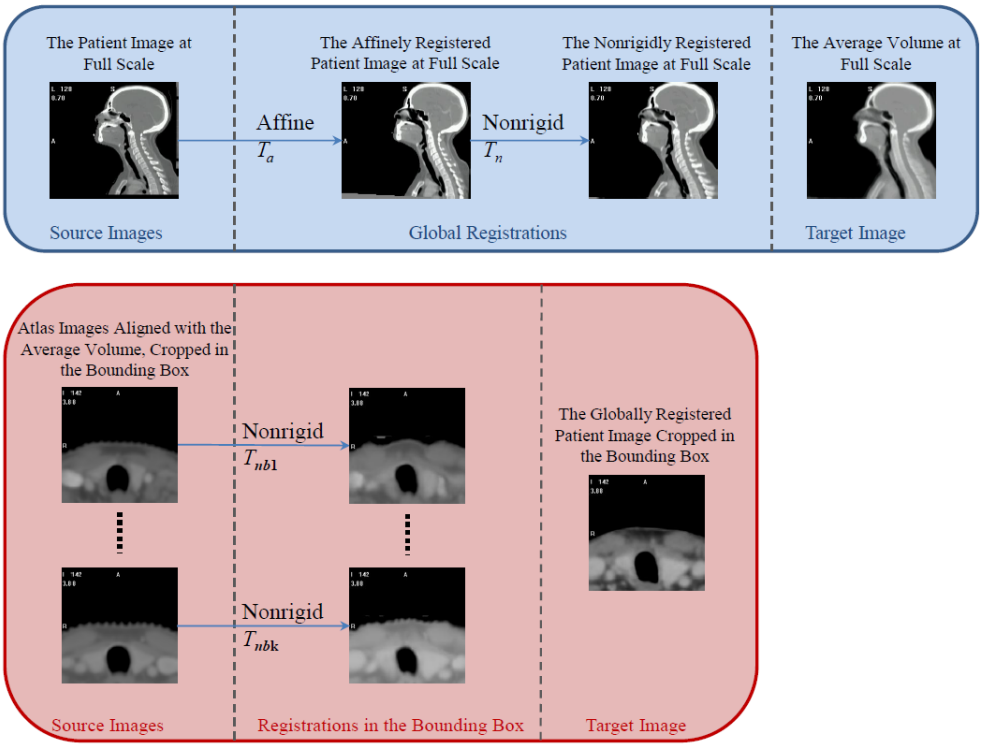

Figure 3.1. Flow charts illustrating the registration process. Top panel: Registration of patient image with the average image volume at full scale. Bottom panel: Registration in the bounding box containing the thyroid between the patient image and the atlases.

The first category of approaches proposed to take advantage of multiple segmentations selects the atlas volume that is most similar to the patient image according to some similarity criterion. The criteria we use are similar to those proposed by Rohlfing *et al.* (2004a), i.e., the correlation coefficient (CC) between the volumes after nonrigid registrations, the average magnitude of the deformation field (AVG_df), and the maximum magnitude of the deformation field (MAX_df). The first criterion is a measure of the similarity between the volume to be segmented and the deformed atlas, i.e., it can be viewed as a measure of the registration quality. The other two criteria measure how

much "work" the registration algorithm has to do to register the image volumes. It is thus a measure of similarity between volumes before registration. We note that Rohlfing *et al.* [31] and Klein *et al.* [34] used NMI as a measure of similarity after registration. Here we have preferred to use the CC because NMI is the quantity being optimized by our registration algorithm. The CC is defined as follows [42, 43]:

$$CC = \frac{\sum_i (A(i) - \bar{A})(B'(i) - \bar{B}')}{(\sum_i (A(i) - \bar{A})^2 \sum_i (B'(i) - \bar{B}')^2)^{1/2}} \quad \forall i \in A \cap B' \tag{3.3}$$

where $\bar{A}$ is the mean intensity value of the voxels in the source image and $\bar{B}'$ is the mean intensity value of the voxels in the transformed target image.

To avoid measurements of CC and deformation fields in irrelevant areas, a region that contains the thyroid gland is created by dilating the union of segmentations propagated from all the atlases using a 3×3×3 structuring element such that only the CC and deformation fields that are close to the thyroid gland are considered in the computation. The atlas with the highest CC is selected as the most similar atlas. In the other two methods the atlas with the lowest average or maximum deformation field magnitude is selected. Segmented structures propagated from this volume are used to segment the patient image.

The second category of approaches uses the entire set of atlases, but their contributions are weighted using the same three similarity measures introduced above. To be used as weights in the combination, the similarity measures are normalized. For CC, the weights are calculated as

$$w_{ncc_i} = \frac{cc_i}{\sum_{i=1}^{k} cc_i} \tag{3.4}$$

where cc_i is the average CC over the region of interest for the ith atlas, and k is the number of atlases. For AVG_df and MAX_df, the weights are determined as

$$w_{df_i} = \frac{1/d_i}{\sum_{i=1}^{k} 1/d_i} \tag{3.5}$$

in which for AVG_df, d_i is the average deformation field magnitude in the region defined above, while for MAX_df it is the maximum field magnitude in this box. The combined segmentation L is calculated as

$$L = \sum_{i=1}^{k} w_i L_i \tag{3.6}$$

where L_i is the segmentation produced by the ith atlas and w_i is its weight calculated using either Eq. (4) or (5). The segmentation L is then rescaled into the intensity range of [0, 255], and the final segmentation is obtained by thresholding the rescaled image at intensity $I > 127$ and saving it as a binary image in which 0 is the background and 255 is the structure.

3.2.4. Running time

The affine and nonrigid registration algorithms used in this study are implemented in C and C++. When segmenting a new patient image, the typical running time on a computer with a 2.93 GHz Intel Xeon quad-core PC with the 64-bit Windows OS and 16

GB of memory is 2 min for the global affine component and 10 min for the global nonrigid component used for registering the image to the average image volume. After the bounding box is defined, each of the k (k=20 for segmentation of new patient with all 20 atlases, and k=19 for leave-one-out experiments) local nonrigid registrations between one atlas and the patient image takes around 1 min, but the process can be parallelized.

3.3. RESULTS

A leave-one-out strategy is used to compare the various atlas-based segmentation methods introduced above. For each run, one image is eliminated from the set of atlas images, and the segmentations of the thyroid gland are obtained using the remaining 19 volumes. The results are evaluated on the 20th image by comparing the atlas-based and manual segmentations. For each patient the thyroid gland on the left and right sides are segmented separately, because for patients undergoing tracheotomy the thyroid gland is fully transected in the midline to facilitate installation of tracheotomy tube, resulting in two totally separated sections (one left and one right) of the thyroid gland.

For patients with tracheotomy, combining or selecting segmentations propagated from patients with normal anatomy may result in over-segmentation of the thyroid into the area of the air tube. Segmentations for these patients were post-processed by excluding areas with Hounsfield units (HU) > 270 on the corresponding CT images, since the plastic tracheotomy tube in general have intensity of 300 HU and above, while the soft tissues are lower than 250 HU.

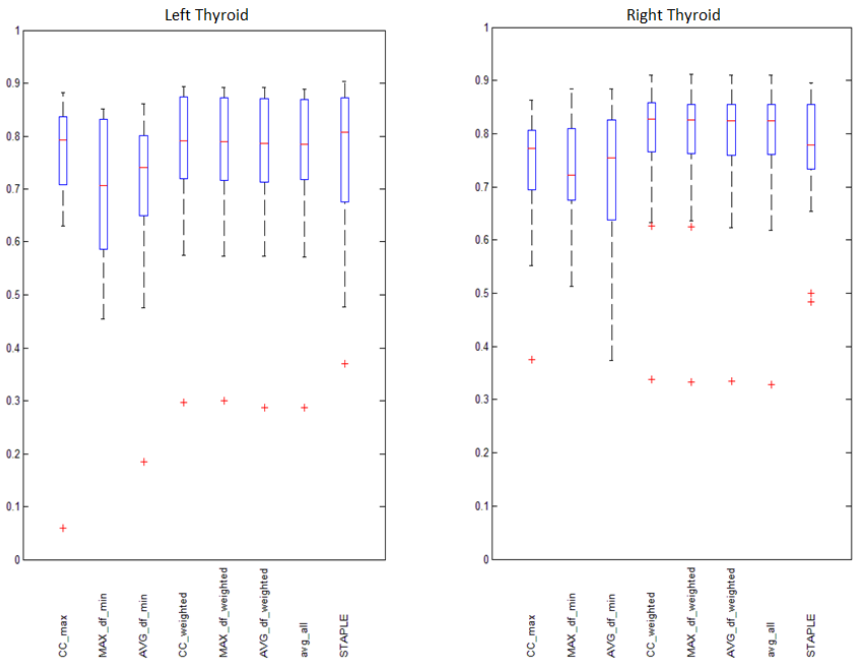

Figure 3.2. Boxplots showing the sample minimum, Q1, Q2, Q3, and the sample maximum of volume DSC's obtained using the most similar atlas selected by maximum CC (CC_max), minimum MAX_df (MAX_df_min) , and minimum AVG_df (AVG_df_min), the combination of all segmentation weighted by CC (CC_weighted), MAX_df (MAX_df_weighted), AVG_df (AVG_df_weighted), the average of all segmentations (avg_all), and the combination of all segmentations by STAPLE. Left panel: Left thyroids. Right panel: Right thyroids.

The DSC, which is defined as the overlap of two segmentation volumes normalized to their mean volume, is the primary measure used to assess the accuracy of the segmentations. It ranges from zero to one with zero indicating no overlap and one complete agreement. Statistics of DSCs obtained between the automatic and manual segmentation volumes using the various atlas-based methods are summarized in Figure 3.2. In this figure, the range between the minimum and the maximum whiskers show the data range, the bottom and top of the box shows the 25^{th} and 75^{th} percentile, the line in the middle shows the median, and the "+" signs show the outliers. High DSC values indicate high similarity between manual and automatic contours. The left panel shows the results for the left thyroid and the right panel for the right thyroid. For each panel, results

obtained when using a single atlas selected with one of the three metrics discussed above (these are labeled CC_max, MAX_df_min, and AVG_df_min) are shown on the three first left columns. The next three columns show results obtained with methods combining segmentations weighted using the three metrics (these are called CC_weighted, MAX_df_weighted, and AVG_df_weighted). The last two columns show the average of all segmentations (avg_all) which is equivalent to the majority vote rule and the results obtained using STAPLE. When calculating the combined segmentation using STAPLE, a bounding box of minimum size containing the dilated union of all propagated segmentations is defined for each thyroid. The method is thus evaluated in regions that are similar to those used for evaluating the other methods. The STAPLE implementation provided by the Computational Radiology Laboratory of Warfield *et al.* (http://crl.med.harvard.edu/software/STAPLE/index.php) was used, with the stationary priors of the background and thyroid set to 0.9 and 0.1 according to the approximate ratio of each class in the bounding box. The average DSC values for the various approaches are presented in numerical form in Table 3.1. Since the anatomy of patient 11 is substantially different from the anatomy of the other patients due to pathology, the DSCs obtained for this patient are generally much lower than those for other patients. Therefore the average DSCs were also calculated without the results for patient 11 and shown in Table 3.1. Since CC_weighted showed the highest volume DSC among all the eight methods, we statistically compared the DSCs of CC_weighted with all the other seven methods by performing a one sided paired t-test. The *p*-values that were obtained are shown in Table 3.2.

Table 3.1. The average DSCs for volumes calculated with various methods with and without patient 11.

		CC_max	MAX_df_min	AVG_df_min	CC_weighted	MAX_df_weighted	AVG_df_weighted	avg_all	STAPLE
Left Thyroid	Avg.	0.747	0.698	0.703	0.768	0.767	0.765	0.761	0.756
	Avg. w/o 11	0.783	0.705	0.730	0.793	0.792	0.790	0.786	0.771
Right Thyroid	Avg.	0.737	0.726	0.719	0.784	0.783	0.781	0.779	0.765
	Avg. w/o 11	0.743	0.737	0.737	0.808	0.806	0.805	0.803	0.779

Table 3.2. The *p*-values for t-tests on DSCs of CC_weighted compared with the other seven methods, with and without DSCs for patient 11. *p*-values greater than 0.05 are italic, indicating statistical insignificance.

		CC_max	MAX_df_min	AVG_df_min	MAX_df_weighted	AVG_df_weighted	avg_all	STAPLE
Left Thyroid	Avg.	*0.180*	0.021	0.001	*0.087*	0.002	0.001	*0.312*
	Avg. w/o 11	*0.317*	0.003	0.001	*0.059*	0.005	0.002	*0.178*
Right Thyroid	Avg.	*0.061*	0.002	0.002	0.026	0.004	0.005	*0.215*
	Avg. w/o 11	0.007	0.000	0.001	0.043	0.008	0.008	*0.114*

As shown in Table 3.1, among the three methods we evaluated for selecting the most similar atlas, CC_max has the highest mean DSC. Among the three combination methods we evaluated, CC_weighted has the highest mean DSCs albeit the boxplot shows that the three combination methods perform quite similarly. Although the difference is small, the *p*-values in Table 3.2 show that the results for CC_weighted are significantly better than those for AVG_df_weighted. The difference between the results obtained with CC_weighted and MAX_df_weighted are statistically insignificant. Comparing DSCs of CC_weighted with CC_max, avg_all, and STAPLE, CC_weighted outperforms avg_all with significant difference, while the difference between CC_weighted and CC_max, as well as the difference between CC_weighted and STAPLE, are statistically insignificant. Note that although STAPLE has the highest median DSC for the left thyroid, as shown in Figure 3.2, its mean DSC is lower than

CC_weighted, AVG_df_weighted, MAX_df_weighted, and avg_all because of lower DSCs for cases in the lower range. Based on these results, we selected CC_max as the representative method for the three methods relying on the most similar atlas and CC_weighted as the representative method for the three methods relying on a weighted combination of segmentations along with avg_all and STAPLE that are the two standard methods for combining segmentations. We then analyzed further the performance of these methods. Since in clinical applications structures are delineated as contours on axial slices, and the accuracy of automatic segmentations on axial slices may not be directly reflected by volume-wise comparisons, we refined our analysis of these methods by comparing automatic and manual segmentations on a slice-by-slice basis.

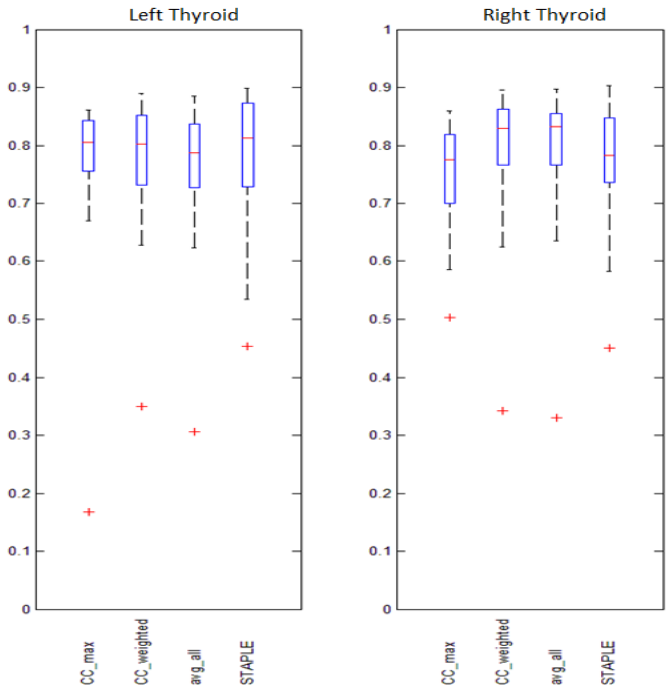

Figure 3.3. Boxplots showing the sample minimum, Q1, Q2, Q3, and the sample maximum of the averages of slice DSC's obtained using CC_max, CC_weighted, avg_all, and STAPLE. Left panel: Left thyroids. Right panel: Right thyroids.

In this analysis, we first calculate the DSC on each image slice. Then we compute the average among all slices for each method and for each patient, thus generating 20 averages for each method. Figure 3.3 shows the range, the 25th percentile, the 75th percentile, the median, and the outliers for each technique. It can be seen that when compared on a slice-by-slice basis all methods perform comparably on the left thyroid. Although STAPLE has the highest median value, it also has a lower minimum. On the right thyroid the leading methods are CC_weighted and avg_all.

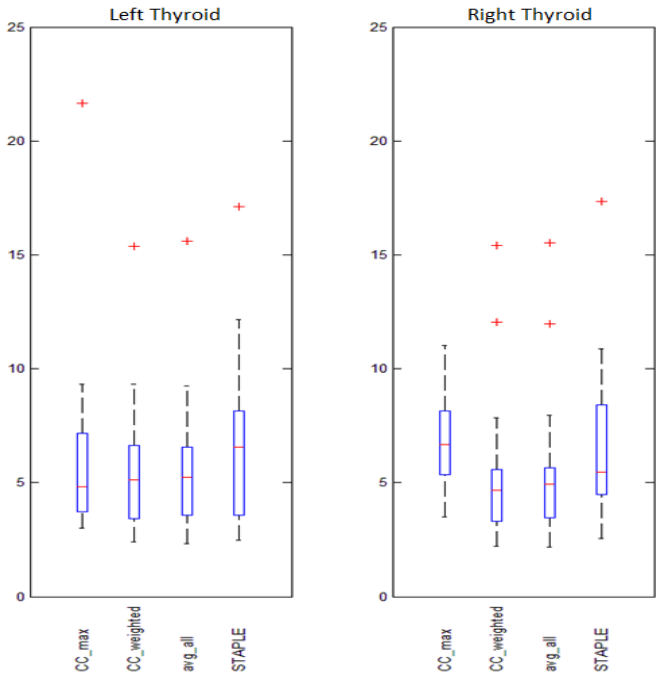

Figure 3.4. Boxplots showing the sample minimum, Q1, Q2, Q3, and the sample maximum of the averages of Hausdorff distance in mm on 2D slices obtained using CC_max, CC_weighted, avg_all, and STAPLE. Left panel: Left thyroids. Right panel: Right thyroids.

The Hausdorff distance (HD) which is defined as

$$HD(L_A, L_M) = \max\{\max_{v_A \in L_A}(\min_{v_M \in L_M}(\|v_A - v_M\|)), \max_{v_M \in L_M}(\min_{v_A \in L_A}(\|v_M - v_A\|))\} \quad (3.7)$$

where v_A represents voxels belonging to the automatic segmentation, v_M represents voxels belonging to the manual segmentation, and $\|\cdot\|$ is the Euclidian distance, is also calculated on the 2D slices for these four representative methods. The range, 25th percentile, 75th percentile, median, and the outliers of the averages for each method are shown in Figure 3.4. It can be seen that CC_max, CC_weighted, and avg_all show similar Hausdorff distances which are generally lower than STAPLE on the left side, while on the right side CC_weighted and avg_all are comparable and show more cases toward the lower end of the Hausdorff distance measure than the other two methods.

Because a Hausdorff distance of 3 mm on a slice generally indicates that the automatic contour is clinically acceptable, we also counted, for each method, the number of slices on which the contours are at a Hausdorff distance of 3 mm or less from the manual contours. CC_weighted had 108 out of 267 slices on the left side and 122 out of 292 slices on the right side falling in this range. These numbers were 103, 82, and 98 out of 267 on the left side and 119, 51 and 74 out of 292 on the right side for the avg_all, CC_max, and STAPLE methods, respectively. Thus, 41.1% of the contours produced by CC_weighted were in this range when only 39.7%, 23.8%, and 30.8% were in the same range with the avg_all, CC_max, and STAPLE methods, respectively.

The slice-wise comparisons, together with the volume-wise comparison show that CC_weighted and avg_all are more consistent than CC_max and STAPLE which perform more poorly on the right side. CC_weighted is slightly better than avg_all in all comparisons.

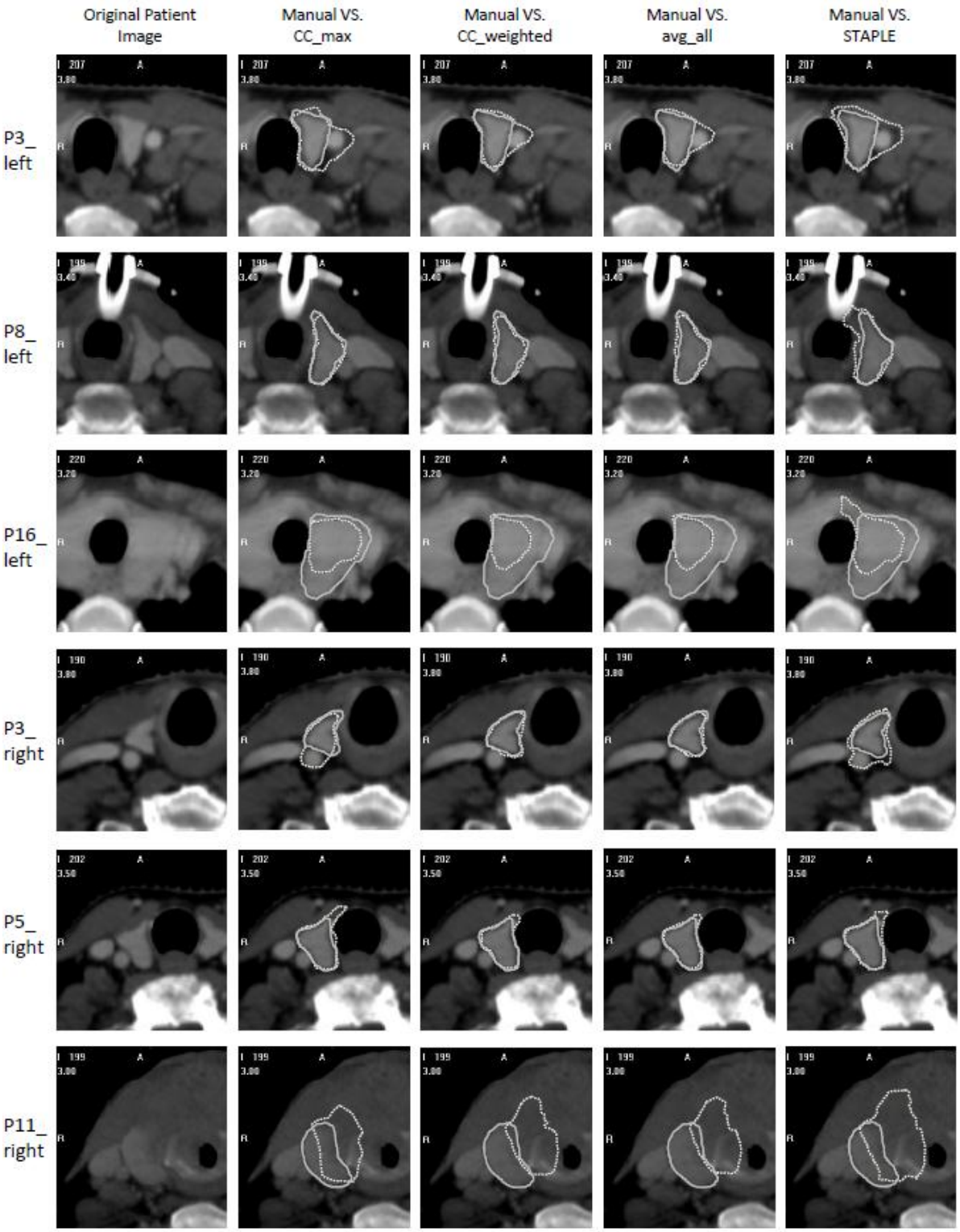

Figure 3.5. Segmentations obtained using the four representative methods shown in contours with dotted lines compared with the manual segmentation shown in solid lines. For each row, from the left to the right: The original patient image, images with contours obtained using CC_max, CC_weighted, avg_all, and STAPLE compared with the manual contours. Rows from top to bottom: Left thyroids for patients 3, 8, and 16, right thyroids for patients 3, 5, and 11.

Figure 3.5 shows qualitative results for six representative cases. From top to bottom, segmentation results are shown for the left thyroid obtained on one slice in the data set of patients 3, 8, and 16 using the four representative methods discussed above, as well as results for the right thyroid of patients 3, 5, and 11. It can be observed that for patients with normal anatomy, e.g., the left thyroid of patient 8 and the right thyroid of patient 3 and 5, CC_weighted and avg_all showed more consistent agreement with the manual delineations than CC_max and STAPLE. Using all methods, over-segmentations into the blood vessel are observed for the left thyroid of patient 3, while under-segmentations are seen for the left thyroid of patient 16. A substantial segmentation error is observed for patient 11, in which the anatomy is altered by a large tumor.

To further study the clinical usefulness of the automatically generated contours, we presented the segmentation results of CC_weighted to a physician (KN) and instructed him to modify the contours to make them clinically useable. In this experiment, the contours are shown superimposed to the patient images presented along the axial direction and modified using a tool which removes/adds a part from/to the contour by brushing over the target area. The modified contours, which are denoted CC_weighted_mod, are saved as binary masks and compared with the original automatic segmentations to measure the amount of modifications made by the physician.

Table 3.3 shows the volume-wise DSC between the original automatic segmentations (CC_weighted) and the modified segmentations (CC_weighted_mod) for all 20 patients. It can be observed that for each side, 13 cases out of the 20 reached a volume-wise DSC of 0.9 or higher. We also compared the 2D axial contours of CC_weighted and the CC_weighted_mod. 112 out of 281 slices (40%) on the left side

and 125 out of 297 (42%) on the right side were accepted without modification. For 173 slices on the left side and 196 slices on the right side, the DSC between the automatic and the modified contours was 0.9 or above, indicating that in about 61% of cases on the left side and 66% of cases on the right side, none or minimal changes were made.

Table 3.3. The DSCs computed between volumes of CC_weighted and volumes of the modified segmentations CC_weighted_mod for all patients.

Patient	1	2	3	4	5	6	7	8	9	10
Left Thyroid	0.930	0.669	0.782	0.955	0.982	0.908	0.923	0.987	0.728	0.992
Right Thyroid	0.950	0.702	0.956	0.885	0.980	0.948	0.893	0.981	0.972	0.985
Patient	11	12	13	14	15	16	17	18	19	20
Left Thyroid	0.202	0.926	0.764	0.946	0.982	0.787	0.876	0.983	0.936	0.980
Right Thyroid	0.304	0.860	0.949	0.953	0.879	0.695	0.900	0.909	0.953	0.981

To illustrate qualitatively the amount and location of modifications made by the physician, we calculated the distance from the surfaces of the modified segmentations to the surfaces of the original automatic segmentations. Colored 3D surfaces of the modified segmentations are shown in Figure 3.6, where the blue color represents zero or small distance, indicating none or minimum modification, while the red color represents large distances, indicating substantial modification. The patients shown are the same as those shown in Figure 3.5. It can be seen that for the left thyroid of patient 8 and the right thyroid of patients 3 and 5, the automatic segmentations received little or none modification for most of the areas, except for several slices on the top and bottom. This is because there is some variability on the extent of the gland along the *z*-direction delineated on the atlases, which in turn causes disagreements in the combination. Also, large portions of the surface of the left thyroid for patients 3 and 16 and the right thyroid for patient 11 are red, indicating extensive modifications. These are also the cases for

which CC_weighted showed inaccurate results when compared with the manual segmentations, as shown in Figure 3.5.

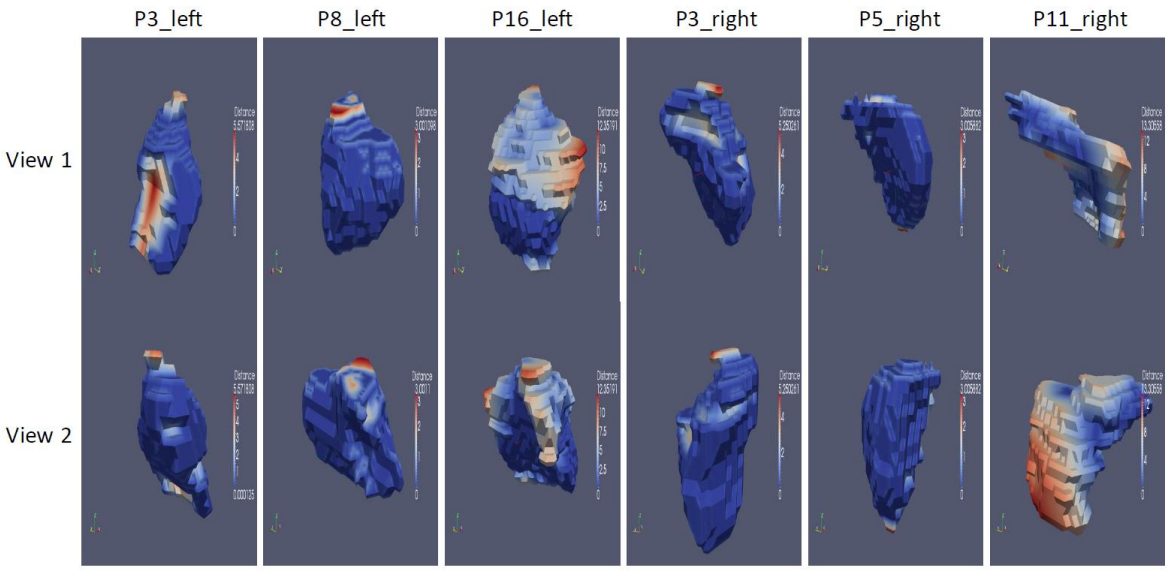

Figure 3.6. 3D surfaces of the modified segmentations, with blue color representing zero or little distance to the surface of the original automatic segmentation obtained using CC_weighted, and red color representing large distance. Columns from left to right: Left thyroids for patients 3, 8, and 16, and right thyroids for patients 3, 5, and 11. For each column, the top and bottom rows show the same surface viewed from two different angles.

3.4. DISCUSSION AND CONCLUSIONS

Although the conventional atlas-based segmentation of the thyroid gland in head and neck CT images is challenged by large anatomical differences, we demonstrated that the automatic segmentation of the gland may be achievable by using a multiple-atlas-based approach. We show that combining segmentations obtained from multiple atlases tends to perform better than methods selecting a single most similar atlas, especially for segmenting images that do not show drastic anatomical differences with the majority of the atlases. This is in agreement with the conclusions of the study conducted by Klein *et al.* [34], in which combining segmentations from atlases with large normalized similarity

performed better than selecting the most similar atlas. Among the methods that combine segmentations, the method based on CC showed its stability by staying in the top-ranked methods when evaluated with all three criteria (volume DSC, slice DSC, slice Hausdorff distance). Its overall performance is substantially better than CC_max, which is the method on the lower end of the accuracy scale. It shows comparable performance with STAPLE when evaluated by volume and slice-wise DSCs, and a significantly better performance when evaluated with the Hausdorff distance. When compared with avg_all, the volume DSC shows that CC_weighted performs better than avg_all, and the difference is statistically significant. The slice-wise differences between the two methods with the DSC and the Hausdorff distance are statistically insignificant, but as it is shown in Figure 3.3 and 3.4, CC_weighted is never outperformed by avg_all. These results suggest that when both methods are available, CC_weighted should be preferred to avg_all. The clinical acceptance of the method was further assessed in the modification study, in which the segmentations generated by CC_weighted were presented to a physician and modified to meet clinical requirements. A comparison between the original and modified automatic segmentations shows that a large portion of segmentations (about 61% of 2D contours on axial slices for the left thyroid, and 66% for the right thyroid) required zero or very little modification, which is an indication of the clinical usefulness of the approach.

Although CC_weighted is shown to be the best method overall in this study, its superiority over other methods is not always significant, which indicates that determining the optimal multiple-atlas-based strategy remains an open problem. This is in line with contradictory results that have been reported in the literature. Indeed, Rohlfing *et al.* [33]

report that majority vote was outperformed by STAPLE on their data. However, in the study of Artaechevarria *et al.* [36], STAPLE was shown to lead to results that are substantially worse than those obtained with a series of alternative methods including majority vote. These authors report that voting weighted by local similarity was found to be the most accurate, while the difference between the top two (weighted by local NMI or mean square distance (MSD)) was subtle. Sabuncu *et al.* [37] indicated that both STAPLE and majority vote were inferior when compared with the three weighted voting methods they used, with the majority vote being the clear worst. The best local weighted voting method using EM improved DSC by only 0.006 or less for most structures compared with a method similar to CC_weighted at the cost of multiplying the CPU time by about 17. On the other hand, Klein *et al.* [34] did not find STAPLE to be significantly different from the weighted voting method based on NMI, as well as majority vote. While there is converging evidence that multiple-atlas-based strategies lead to better results, there is clearly no agreement on the best way to achieve it. Comparison between methods also remains difficult due to many factors, e.g. image modalities, image quality, size of dataset, registration accuracy, size of the structure of interest, and parameter setting and implementation of standard algorithm (especially STAPLE). It would thus be useful to perform more comprehensive comparative study of all available methods on a series of openly accessible data sets for which a ground truth is known.

Even though we have shown the potential clinical usefulness of the approach, i.e., CC_weighted is accurate for most of the 20 patients with normal anatomy included in this study and exemplified by the left side of patient 8 and the right side of patients 3 and 5 shown in Figure 3.5, shortcomings still need to be addressed. We have observed three

major categories of problems: First, as shown in Figure 3.5, for the left thyroid of patient 3, over-segmentation into the blood vessel was observed by all automatic methods including CC_weighted. This is mainly because the thyroid gland on the left side of this patient is smaller than those in the atlases. In this patient, the vessel and thyroid together match the size and shape of a left thyroid in a regular volume. The large anatomical discrepancy between this particular patient and the other volumes in our atlases resulted in a systematic false segmentation. A similar problem is also observed on both the left and right sides of patient 2, the left side of patient 13, and the right side of patient 15. Second, anatomical discrepancies can also be caused by structures that are larger than usual, which is the case for patient 16 on the left side. This patient had a thyroid gland that was considerably larger than those in normal patients. Also, the gland extended into the chest cavity, while for a normal patient the gland does not extend lower than the clavicle level. Since none of the atlases could match the gland with similar size and extent, the results showed obvious under-segmentation. Both the left and right sides of patient 7 and patient 17 fall into the same category. Third, pathology is another major cause for anatomical differences, as shown in patient 11. The tumor filling the trachea pushed its surrounding tissues into the area that should normally be occupied by the thyroid, and subsequently caused completely false registrations.

Over-segmentations may be corrected by applying anatomical constraints, i.e. segmenting the falsely included structures individually and removing them from the original automatic segmentation. For the left thyroid of patient 3, an accurate segmentation of the blood vessel may be achievable [44], and removing the vessel could drastically improve the accuracy of the segmentation of the thyroid. A more general

solution to false segmentations caused by anatomical discrepancies in normal patients is to expand the set of atlases. Cases like patient 3 on the left side and patient 16 may benefit from using a subset of atlases with similar anatomy, which could be selected automatically via certain similarity measures. However, false segmentations caused by large anatomical discrepancies in patients with pathology, e.g., large tumors and tissue resections, may not be corrected by expanding the atlas set, since the structural alteration in each patient may be unique. In these cases, models of the tumors or resections may be needed to simulate the deformation, and practically manual delineations may be more suitable than automatic approaches.

Increasing the number of atlases may not only provide the anatomical variability required for segmenting patients with rare anatomy, but also optimize the number of atlases involved in the combination. Aljabar *et al.* [35] opted to use a fixed number of atlases (20 out of 275) ranked by their similarity to the patient image. This method may be confounded when there are too few atlases that are similar to the volume to segment. In this case, a number of very dissimilar atlases could be selected to reach the preset number of atlases and thus negatively affect the results. Klein *et al.* [34] also studied the impact of using a subset of atlases which was selected by thresholding their normalized similarity φ (the NMI of each atlas divided by the maximum NMI in all atlases) and found an optimum threshold value that corresponded to 23 out of 49 atlases in the experiments they performed. The advantage of the method can be limited when most atlases are similar. The SIMPLE approach by Langerak *et al.* [38] essentially reduced the number of atlases in an iterative process by eliminating the worst performers, i.e., those leading to segmentations that are different from the consensus at the current iteration.

Ultimately around 35 atlases out of a set of 99 were used to compute the combined segmentation. The major obstacle to conducting studies with a very large set of atlases is to obtain a ground truth, i.e., volumes in which the thyroid has been segmented with an accuracy that exceeds the accuracy of delineation performed in the clinical setting under time constraints.

Even though automatic segmentations obtained with the CC_weighted combination approach may not be directly applicable for clinical application (i.e. they need to be modified by physicians), this study has shown its potential to reduce delineation efforts compared to a fully manual delineation of the structure. For most cases with normal anatomy the clinician who evaluated the results did not need to make changes for the majority of the contours generated automatically. In this study, modifications have been done by one physician using in-house developed software instead of the clinical radiation oncology planning station. What this study does not yet address is the accuracy of the results compared to intra- and inter-rater variability. In a recent study [45] performed with eight raters on twenty volumes, we have shown that a single-atlas-based method performs as well as a human rater for the segmentation of the eyes, optic nerves, optic chiasm, and brainstem. The anatomical variability we have observed in the thyroid led us to explore a multi-atlas procedure. Based on the encouraging results we have obtained with this approach for the thyroid and with a model-based approach for the lymph node regions [17] we are planning a multi-rater validation study for the thyroid, the lymph node regions, and the parotid. Early results we have obtained with the parotid indicate that a model-based approach as the one we have

used for the lymph node regions may be better than a multi-atlas-based approaches as proposed by Ramus *et al.* [46], Yang *et al.* [47], or Han *et al.* [48].

Finally, we note that this work is focused on the segmentation of structures of interest in the planning CT images. Adaptation of these contours to the on-board CT, e.g. Cone Beam CT (CBCT), when these are acquired during the course of therapy is required to take into account change in tumor and normal anatomy (e.g. shrinkage of tumor and body size) that may occur between acquisitions. Techniques have been proposed for this purpose (see for instance studies by Wang *et al.* [49], Lu *et al.* [50], Chao *et al.* [51], Xie *et al.* [52], and Lee *et al.* [53]). The comparison of these techniques with the registration of the planning CT to the CBCT using an intensity-based nonrigid registration method as we have used herein will need to be done.

REFERENCES

[1] Bethge W, Guggenberger D, Bamberg M et al. 2000 Thyroid toxicity of treatment for Hodgkin's disease. Ann Hematol. 79 114–8

[2] Colevas A D, Read R, Thornhill J *et al.* 2001 Hypothyroidism incidence after multimodality treatment for stage III and IV squamous cell carcinomas of the head and neck. Int J Radiat Oncol Biol Phys. 51 599–604

[3] Garcia-Serra A, Amdur R J, Morris C G *et al.* 2005 Thyroid function should be monitored following radiotherapy to the low neck. Am J Clin Oncol. 28 255–8

[4] Grande C 1992 Hypothyroidism following radiotherapy for head and neck cancer: Multivariate analysis of risk factors. Radiother Oncol. 25 31–6

[5] Lo Galbo A M, de Bree R, Kuik D J *et al.* 2007 The prevalence of hypothyroidism after treatment for laryngeal and hypopharyngeal carcinomas: Are autoantibodies of influence? Acta Otolaryngol.127 312–7

[6] Liening D A, Duncan N O, Blakeslee D B *et al.* 1990 Hypothyroidism following radiotherapy for head and neck cancer. Otolaryngol Head Neck Surg. 103 10–3

[7] Mercado G, Adelstein D J, Saxton J P *et al.* 2001 Hypothyroidism: A frequent event after radiotherapy and after radiotherapy with chemotherapy for patients with head and neck carcinoma. Cancer. 92 2892–7

[8] Nishiyama K, Tanaka E, Tarui Y *et al.* 1996 A prospective analysis of subacute thyroid dysfunction after neck irradiation. Int J Radiat Oncol Biol Phys. 34 439–44

[9] Tell R, Sjodin H, Lundell G *et al.* 1997 Hypothyroidism after external radiotherapy for head and neck cancer. Int J Radiat Oncol Biol Phys. 39 303–8

[10] Tell R, Lundell G, Nilsson B *et al.* 2004 Long-term incidence of hypothyroidism after radiotherapy in patients with head-andneck cancer. Int J Radiat Oncol Biol Phys. 60 395–400

[11] Alterio D, Jereczek-Fossa B A, Franchi B *et al.* 2007 Thyroid disorder in patients treated with radiotherapy for head-and-neck cancer: A retrospective analysis of seventy-three patients. Int J Radiat Oncol Biol Phys. 67 144–50

[12] Bhandare N, Kennedy L, Malyapa R S *et al.* 2007 Primary and central hypothyroidism after radiotherapy for head-and-neck tumors. Int J Radiat Oncol Biol Phys. 68 1131–9

[13] Constine L S, Donaldson S S, McDougall I R *et al.* 1984 Thyroid dysfunction after radiotherapy in children with Hodgkin's disease. Cancer. 53 878–83

[14] Norris A A, Amdur R J, Morris C G, *et al.* 2006 Hypothyroidism when the thyroid is included only in the low neck field during head and neck radiotherapy. Am J Clin Oncol. 29 442–5

[15] Sethian J A 1999 Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science. Cambridge University Press

[16] Boykov Y, Veksler O, and Zabih R 2001 Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell.. 29 1222–39

[17] Chen A, Deeley M A, Niermann K J, Moretti L, and Dawant B M 2010 Combining registration and active shape models for the automatic segmentation of the lymph node regions in head and neck CT images. Med Phys 37 6338–46

[18] Chao K S, Bhide S, Chen H *et al.* 2007 Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach. Int. J Radiat Oncol Biol Phys. 68 1512–21

[19] Commowick O, Grégoire V, and Malandain G 2008 Atlas-based delineation of lymph node levels in head and neck computed tomography images. Radiother Oncol. 87 281–9

[20] Commowick O, Warfield S K, and Malandain G 2009 Using Frankenstein's creature paradigm to build a patient specific atlas. In Proceedings of the 12th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'09), Part II, Lecture Notes in Computer Science. 5762 993–1000

[21] Gorthi S, Duay V, Houhou N, Bach Cuadra M, Schick U, Becker M, Allal A S, and Thiran J P 2009 Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration. IEEE Journal of Selected Topics in Signal Processing. 3(1) 135–47

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

[23] Huttenlocher D, Klanderman D, and Rucklige A 1993 Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 23(7) 850–63

[24] Han X, Hoogeman M, Levendag P *et al.* 2008 Atlas-based auto-segmentation of head and neck CT images. 11th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2008), Lecture Notes in Computer Science. 5242 434–41

[25] Wells W M, Viola P, Atsumi H, Nakajima S and Kikinis R 1996 Multi-modal volume registration by maximization of mutual information. Med. Image Anal. 1 35–51

[26] Maes F, Collignon A, and Suetens P 1997 Multimodality image registration by maximization of mutual information, IEEE Transaction on Medical Imaging.16(2) 187–98

[27] Wu M, Rosano C, Lopez-Garcia P *et al.* 2007 Optimum template selection for atlas-based segmentation. NeuroImage. 34 1612–8

[28] Studholme C, Hill D L G, and Hawkes D J 1999 An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognition. 32 71 –86

[29] Heckemann R, Hajnal J, Aljabar P *et al.* 2006 Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage. 33 115–26

[30] Heckemann R, Keihaninejad S, Aljabar P *et al.* 2010 Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. Neuroimage. 51 221–7

[31] Rohlfing T, Brandt R, Menzel R *et al.* 2004 Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage. 21 1428–42

[32] Warfield S, Zou K, and Wells W 2004 Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging. 23 903–21

[33] Rohlfing T, Russakoff D B, and Maurer C R 2004 Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Transactions on Medical Imaging. 23 983–94

[34] Klein S, van der Heide U A, Lips I M, van Vulpen M, Staring M, and Pluim J P 2008 Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. Med Phys. 35(4) 1407–17

[35] Aljabar P, Heckermann R A, Hammers A, Hajnal J V, and Rueckert D 2009 Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage. 46 726–38

[36] Artaechevarria X, Munoz-Barrutia A, and Ortiz-de-Solorzano C 2009 Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. on Med. Imag. 28 1266–77

[37] Sabuncu M R, Tomas Yeo B T, van Leemput K, Fischl B, and Golland P 2010 A generative model for image segmentation based on label fusion. IEEE Transactions on Medical Imaging. 29 1714–29

[38] Langerak T R, van der Heide U A, Kotte A N, Viergever M A, van Vulpen M, and Pluim J P 2010 Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE Transactions on Medical Imaging. 29 2000–8

[39] Rohde G K, Aldroubi A, and Dawant B M 2003 The adaptive bases algorithm for intensity-based nonrigid image registration. IEEE Trans. on Medical Imaging. 22(11) 1470-9

[40] Wu Z 1995 Compactly supported positive definite radial functions. Adv. Comput. Math. 4 283–92

[41] Guimond A D, Meunier J, and Thirion J P 2000 Average Brain Models: A Convergence Study. Computer Vision and Image Understanding. 77 197–201

[42] Lewis J P 1995 Fast Normalized Cross-Correlation. Vision Interface

[43] Fitzpatrick J M, Hill D L G, and Maurer Jr. C R 2000 Chapter 8 Image Registration of the Handbook of Medical Imaging, 2. Medical Image Processing and Analysis (SPIE Press Monograph Vol. PM80/SC)

[44] Noble J H, Warren F M, Labadie R F, and Dawant B M 2008 Automatic segmentation of the facial nerve and chorda tympani in CT images using spatially dependent feature values. Med Phys. 35(12) 5375–84

[45] Deeley M A, Chen A, and Datteri R *et al.* 2011 Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. Phys. Med. Biol. 56 4557–77

[46] Ramus L and Malandain G 2010 Multi-atlas based segmentation: application to the head and neck region for radiotherapy planning. In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 281-8

[47] Yang J, Zhang Y, Zhang L, and Dong L 2010 Automatic segmentation of parotids from CT scans using multiple atlases. In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 323–30

[48] Han X, Hibbard S, O'Connell P, and Willcut V 2010 Automatic segmentation of parotids in head and neck CT images using multi-atlas fusion. In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 297-304

[49] Wang H, Dong L, and O'Daniel J *et al.* 2005 Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy. Phys. Med. Biol. 50 2887–905

[50] Lu W, Olivera G H, and Chen Q *et al.* 2006 Deformable registration of the planning image (kVCT) and the daily images (MVCT) for adaptive radiation therapy. Phys. Med. Biol. 51 4357–74

[51] Chao M, Li T, Schreibmann E, Koong A, and Xing L 2007 Automatic contour mapping with a regional deformable model. Int. J. Radiation Oncology Biol. Phys., 70 599–608

[52] Xie Y, Chao M, Lee P and Xing L 2008 Feature-based rectal contour propagation from planning CT to cone beam CT. Med. Phys. 35 4450–9

[53] Lee C, Langen K M, and Lu W *et al.* 2008 Evaluation of geometric changes of parotid glands during head and neck cancer radiotherapy using daily MVCT and automatic deformable registration. Radiother Oncol. 89 81–8

CHAPTER IV

SEGMENTATION OF PAROTID GLANDS USING A CONSTRAINED ACTIVE SHAPE MODEL WITH LANDMARK UNCERTAINTY AND OPTIMAL FEATURES IN HEAD AND NECK CT IMAGES FOR IMRT

Antong Chen[1], Jack H. Noble[1], Kenneth J. Niermann[2], Matthew A. Deeley[3], and Benoit M. Dawant[1]

[1] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235

[2] Department of Radiation Oncology, Vanderbilt-Ingram Cancer Center, 1301 22nd Avenue South, Nashville, TN 37232

[3] Medical Physics Division, Fletcher Allen Health Care and Department of Radiology, University of Vermont, Burlington, VT 05401

[This manuscript has been submitted to *Phys. Med. Biol.*]

ABSTRACT

In head and neck intensity-modulated radiation therapy (IMRT), irradiating the left and right parotid glands causes xerostomia which impacts the patients' quality of life (QOL) permanently. Since it is clinically feasible to spare at least the gland on one side of the patient, segmenting the parotid glands is of great significance in the treatment planning process. In this article, we propose a constrained active shape model (ASM) with landmark uncertainty and optimal features to segment the glands automatically. This approach permits to weigh boundary points that are easy to localize more than those that cannot be reliably identified. Via a leave-one-out experiment, we compare results obtained using the proposed model with results obtained using a regular ASM with the same combination of image features, as well as results obtained using a multiple-atlas-based approach. Volumetric and slice-by-slice differences between methods are quantitatively evaluated. Results show that the constrained ASM consistently performs the best among the three approaches. This is also confirmed by qualitative comparisons using 2D contours and 3D surfaces whose color encodes the distance to the manual segmentations. Segmentations by the proposed approach are also presented to a radiation oncologist and modified to meet clinical requirements. This component of the study shows that about 87.8% of the slices are accepted without any modification. This indicates the proposed method's potential usefulness in the clinical workflow.

4.1. INTRODUCTION

Irradiation to the salivary glands, especially the parotid glands, is the major cause of xerostomia, which is one of the most prevalent side effects of head and neck radiation therapy (RT) [1–11]. As a consequence, patients could experience a reduction in salivary flow, which can lead to a lower quality of life (QOL) in the forms of difficulties in mastication, deglutition, and speech, and be predisposed to mucosal fissures and ulcerations, dental caries, and osteoradionecrosis [1, 3, 7, 8]. With irradiation of more than 50 Gy to the glands, the damages become irreversible and the xerostomia is permanent [4, 5]. Since intensity modulated radiation therapy (IMRT) provides the technique to precisely deliver the radiation dose to the structures to be treated, while reducing radiation to the normal tissues, sparing at least partially the parotid glands [1, 2, 12], or more practically the contralateral gland with respect to the position of the tumor site in the treatment becomes feasible [4–6, 8, 10, 11, 13–15]. The sparing requires maintaining the mean radiation dose of the gland at $\leq$ 26 Gy, which is a threshold for preserving its function after the treatment [4, 11]. Studies have shown that because of the sparing of the contralateral parotid gland in patients with lateralized tumors, the salivary rate for the spared gland can be recovered almost fully within one year after the treatment [4, 5, 15], and improvements on the QOL have been reported via questionnaires evaluating the categorized life quality status, e.g. eating, communication, pain, and emotion, of the patients after the treatment [9, 14, 15]. On the other hand, patients undergoing conventional RT (non-IMRT) tend to experience more severe xerostomia-related symptoms and have lower QOL [8, 11, 12, 14].

Since sparing the parotid glands in IMRT plans requires precise delineation of both left and right parotid glands on the planning CT images, which is a laborious task generally done manually, automatic segmentation methods have been proposed in recent years, with the multiple-atlas-based approaches being the most accurate. The multiple-atlas-based approaches generally involve first selecting a set of template images known as the atlases. These are then segmented by experts, and registered with the patient image to be segmented through affine and nonrigid transformations. Segmentations are propagated onto the patient image by the transformations and fused to form the segmentation of the structure. Ramus *et al.* [16] performed registrations between the atlases and the patient image via an average intensity image volume generated using the method proposed by Guimond *et al.* [17], and combined the segmentations using an intensity-weighted majority vote based on the local sum of square distances (SSD) between the transformed atlases and the patient image. Yang *et al.* [18] analyzed the intensity of the atlas images through principal component analysis (PCA) and selected a subset of most similar atlases. They combined the deformed segmentations using the STAPLE algorithm [19] that is commonly used for combining segmentations from multiple raters based on an expectation maximization (EM) algorithm. Han *et al.* [20] also used STAPLE but combined the segmentations propagated from all the available atlases. Each of the projected segmentations was refined by a deformable surface model before they were fused. All three multiple-atlas-based approaches were evaluated using a common set of CT images (10 training images, 8 testing images, voxel size around $0.98\times0.98\times2$ mm^3, provided by the Princess Margaret Hospital in Toronto, Canada, for the MICCAI 2010 Head and Neck Auto-Segmentation Challenge Workshop with manual

segmentations of the parotid glands delineated by experts) on both volumetric and slice-by-slice basis, since the manual segmentations are generated on axial slices in general clinical practice. Comparable segmentation accuracy was reached for these methods, with the average Dice similarity coefficient (DSC) [21] for entire volumes around 0.85 and the average slice-wise DSC around 0.82. Assessed using the same data set, other proposed methods led to considerably lower accuracy. Gorthi *et al.* [22] also compared several multiple-atlas-based approaches involving three methods: Selecting the most similar atlas, combining using STAPLE, and combining using majority voting. The best results obtained by majority voting had an average volume DSC around 0.76 and an average slice-wise DSC around 0.71. Hollensen *et al.* [23] proposed to use the common volume of the propagated segmentations as the initialization for an algorithm based on level sets driven by gradient, and an average volume DSC around 0.60 and an average slice-wise DSC around 0.50 were achieved. Gering *et al.* [24] proposed a situated Bayesian classification approach reaching an average volume DSC around 0.71 and an average slice-wise DSC around 0.66, while under-segmentation was observed in most cases. An expanded image set with 15 training images was used by Qazi *et al.* [25] in their model-based approach followed by refinement using a *k*NN classifier based on local texture features, and an average volume DSC of around 0.77 was reached for the 10 testing images. By improving the initialization of the model, introducing additional texture features, and combining multiple features, Qazi *et al.* [26] obtained segmentations with an average volume DSC of 0.83, which was comparable to the best results obtained by multiple-atlas-based approaches.

Although the multiple-atlas-based approaches and the model-based approach followed by refinement have achieved high segmentation accuracy (volumetric DSC > 0.8), evaluation on a slice-by-slice basis has shown that the automatic segmentations are not yet applicable directly in the clinical practice. Using the slice-wise Hausdorff distance (HD) [27] as the measure of accuracy, the model-based approach by Qazi *et al.* [26] led to average median HD of 5.82 mm for the left parotid and 5.70mm for the right parotid, and the most accurate multiple-atlas-based approach by Han *et al.* [20] led to an average HD of 5.93 mm for the left parotid and 5.69mm for the right parotid, all exceeding the 3mm threshold under which segmentations are considered to be clinically acceptable (Pekar *et al.* [28]). These results indicate that the automatic segmentations may need extensive manual corrections before they can be used in clinical treatment plans. Qualitative observation of the segmentation results also shows that, because of the streak dental filling artifacts and the poor contrast between the glands and their adjacent tissues, precise auto-segmentations are difficult to achieve in certain regions, e.g. the interior boundary against the sternomastoid and digastric muscle groups. In this study, we begin to address these issues and introduce a constrained active shape model (ASM) with landmark uncertainty [29] for the automatic segmentation of the parotid glands. We extend the ASM segmentation framework we proposed previously [30]. To do so, we estimate the reliability of each landmark using a combination of texture features in their surrounding neighborhoods following the method proposed by Toth *et al.* [31]. We use landmarks of low uncertainty to drive the ASM model and infer the location of landmarks with high uncertainty from the fitted model. Segmentations obtained using this constrained ASM are evaluated via a leave-one-out experiment by comparing them with

manual segmentations. We compare results obtained using this constrained ASM approach with the results obtained using a multiple-atlas-based approach in which we weigh the segmentations using local correlation coefficient (CC) [32]. We also compare the constrained ASM approach with a regular ASM approach that treats all landmarks equally. Quantitative validation includes a volumetric comparison using DSC and surface distance error, as well as a slice-by-slice comparison using both DSC and HD as measures of accuracy. Qualitative results will also be shown in the form of 2D contours superimposed in the images and 3D surfaces.

4.2. METHODS AND MATERIALS

4.2.1. Data Description

A total of 15 de-identified CT images for IMRT planning are used. Images are selected such that generally normal anatomy is observed in the areas of the left and right parotid glands. All images have voxel size of around 1mm×1mm within slices and a slice thickness of 3mm. The parotid glands on both left and right sides are delineated manually by the first author and reviewed carefully by a radiation oncologist. These manual delineations are saved as binary masks for both ASM construction and validation purposes.

The quality of the images used in this study varies. Patients 5 and 10 are affected by moderate dental filling artifacts, while the artifacts in patient 1 are severe and visible streaks pass directly through the interior boundary of the glands. Although the intensity of the parotid glands is ideally lower than that of muscle groups and higher than that of fat tissues, the actual tissue density of the gland varies between patients. Therefore the

contrast between the glands and their surrounding tissues can be weak, especially at the interior boundary against the muscle groups. Four examples of parotid glands in the head and neck CT images and their manual delineations are shown in Figure 4.1. It shows that the left parotid of patient 1 is affected by streak artifacts, while all four cases show a lack of contrast against the surrounding digastric muscle at the interior boundary and the masseter muscle at the anterior boundary to different levels.

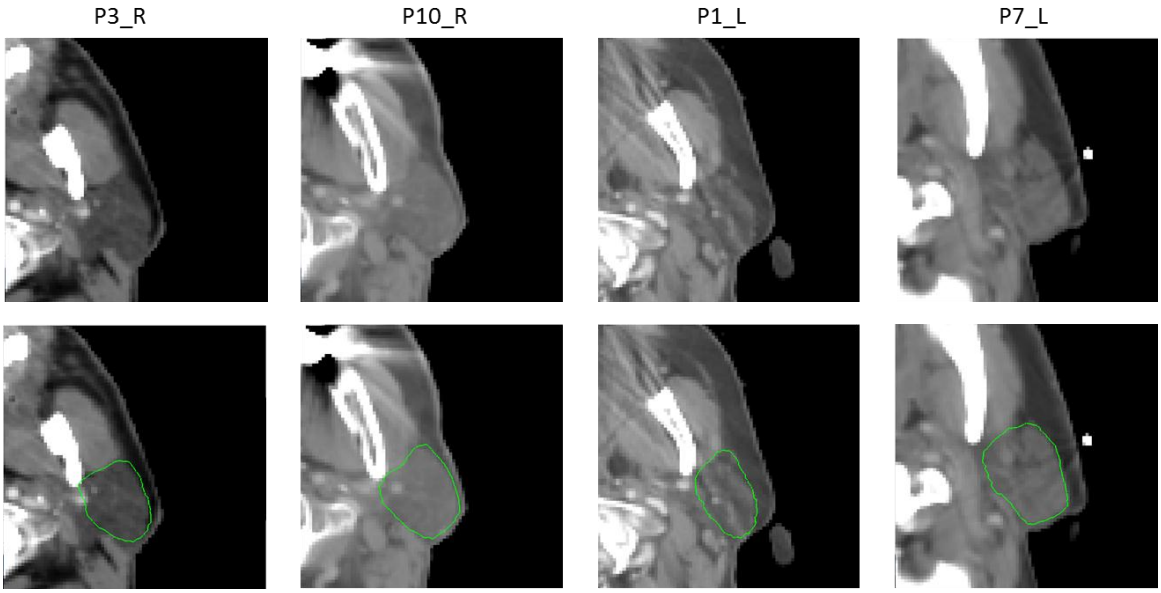

Figure 4.1. Examples of parotid glands in CT images of four patients. Top row: The original images. Bottom row: Images with manual delineations in green contours overlapped on top. Left to right: Right parotid of patient 3, right parotid of patient 10, left parotid of patient 1, and left parotid of patient 7.

4.2.2. Segmentation of the parotid glands using a constrained ASM with landmark uncertainty

An automatic approach based on a constrained ASM [29] is proposed for the segmentation of the parotid glands. Using the framework we proposed previously for the segmentation of lymph node regions [30], we construct the ASM using an automated nonrigid registration method to establish landmark correspondence. As proposed by Toth

et al. [31], local landmark adjustment is done using a locally optimal combination of image features. The set of features is also used to derive the uncertainty of each landmark, which is a property describing the reliability of the landmark and used for deriving the model parameters for the constrained ASM.

4.2.2.1 ASM construction – Establishing landmark correspondence

Following the procedure we proposed previously [30], the correspondence for the landmark points on the training shapes of the parotid glands is achieved through registration. Firstly, each training image I_i (i=1, …, K where K is the size of the training set) is affinely registered to one training image that is selected arbitrarily. Based on the set of affinely registered images, an average atlas volume representing the intensity and shape average of the training images is generated via iterative nonrigid registrations following the procedure proposed by Guimond *et al.* [17]. Applying the affine and nonrigid transformations from the training images to the atlas, the manual segmentations are projected into the space of the average atlas. These projections are averaged to generate an average segmentation. A mesh that describes the average surface is extracted using the marching cube algorithm implemented in ITK (www.itk.org) and the vertices of the mesh are defined as the landmarks. Projecting the landmarks back into the space where the images were affinely registered using the inverses of the nonrigid transformations, and finding the closest points on the surfaces of the corresponding manual segmentations, a set of training shapes representing the parotid glands are generated. In this work, instead of building ASM for the left and right parotid glands individually, we applied a mirror transformation to map the images and shapes of the left parotid onto the right side such that the number of training images and shapes are both

doubled from 15 to 30. Performing principal component analysis on the shapes, the mean shape $\bar{\vec{x}}$ is obtained, as well as the matrix $\Phi = [\vec{\phi}_1, \ldots, \vec{\phi}_t]$ consisting of the first *t* eigenvectors associated with the *t* largest eigenvalues of the covariance matrix. Therefore, a new shape $\vec{x}$ can be represented as

$$\tilde{\vec{x}} = \bar{\vec{x}} + \Phi\vec{b} \tag{4.1}$$

where $\tilde{\vec{x}}$ is the estimate to $\vec{x}$ by the model and $\vec{b}$ is the vector of the model parameters.

4.2.2.2 Segmentation using ASM – Model initialization and refinement

When segmenting a new patient image, the image is registered with the average atlas by applying in sequence an affine transformation T_a, a nonrigid transformation T_n aligning the body boundary and bones, and a nonrigid transformation T_{nb} aligning the soft tissues surrounding the parotid glands. Applying in turn the inverses of the two nonrigid transformations T_{nb}^{-1} and T_n^{-1}, the landmarks can be projected from the average atlas volume onto the patient image in the affinely registered image space, which produces $\vec{x}_p$ that is a rough registration-based segmentation of the patient's parotid gland. The ASM is initialized by finding the optimal (in a least squares sense) transformation *T* and parameter $\vec{b}$ such that

$$\tilde{\vec{x}}_p \approx T \circ (\bar{\vec{x}} + \Phi\vec{b}). \tag{4.2}$$

After the model is initialized, the segmentation is refined by first displacing each landmark to a more desirable location. For each landmark, a search vector of length *L* is extracted along the surface normal direction. This produces a set of *L* candidate points. A feature profile $\vec{g}_l$ is then extracted for the *l*th point of the *L* points also along the surface

normal direction, and then compared with the G profiles $[\vec{g}_1, \ldots, \vec{g}_G]$ extracted for the corresponding landmark in the training set using the Mahalanobis distance defined as

$$d_l = \sqrt{\left(\vec{g}_l - \overline{\vec{g}}\right)^T S_G^{-1} \left(\vec{g}_l - \overline{\vec{g}}\right)} \tag{4.3}$$

in which $\overline{\vec{g}}$ is the mean of the G profiles, and S_G^{-1} is the inverse of their covariance matrix. The candidate point yielding the lowest d_l is selected as the updated landmark. After all landmarks are updated a new shape model is fitted to these points. The updating-fitting process is repeated until convergence, or the maximum number of iterations is reached.

Note that various features can be used for updating the position of the landmarks, e.g. the original intensity, the intensity gradient, or their smoothed versions. Combining the features to find the best landmark position could also be more reliable than using only one feature. With a set of K features, the optimal point in the L candidate points could be determined as

$$l_{opt} = \arg\min_l \sum_{k=1}^{K} w_k \, d_{l,k} \tag{4.4}$$

where $d_{l,k}$ is the Mahalanobis distance associated with feature k for the lth candidate point, and w_k's are a set of weights for the features, which need to be assigned properly.

4.2.2.3 Feature extraction and optimization

In this subsection, we introduce the set of features we have used to update the position of the landmarks, as well as the method used to determine their weights in the combination. Four image features are computed on a per-voxel basis, such that for every

training image Ii each of its features is represented as a feature image that has the same dimensions as Ii. Specifically, the features include:

a. Intensity: The original intensity of the training images is used directly as a feature.
b. Local intensity mean (in a box): For each voxel, except for those on the image boundary, the intensity average in a 7×7×7 box centered at this voxel is calculated as a feature. The feature image is a smoothed version of the original intensity image.
c. Local intensity standard deviation (in a box): For each voxel, except for those on the image boundary, the standard deviation of intensity in a 7×7×7 box centered at this voxel is calculated. Regions near the boundary of a structure tend to have higher standard deviation.
d. Local entropy (in a box): For each voxel, except for those on the image boundary, a 7×7×7 box centered at this voxel is extracted and the intensity distribution is analyzed using a local intensity histogram. The local entropy is calculated based on the histogram, which describes the local intensity variability.

To ensure that the features are compared on the same scale, each feature is normalized such that it has a mean of zero and a standard deviation of one.

Following the method proposed by Toth *et al.* [31] the features' weights are determined by correlating Mahalanobis distance and the Euclidean distance, i.e, a feature for which Mahalanobis and Euclidean distances are correlated is weighed more than a features for which these distances are not. To compute the weigts for the *i*th landmark $\vec{p}_{i,j}$ on the *j*th training image, a set of *M* voxels are randomly sampled in a *D*×*D*×*D*

neighborhood surrounding it. For each of the *M* sampled voxels $\vec{p}_{i,j,m}$, the Euclidean distance to $\vec{p}_{i,j}$ is computed as

$$e_{i,j,m} = \left\| \vec{p}_{i,j} - \vec{p}_{i,j,m} \right\| \tag{4.5}$$

The Euclidean distances for all *M* voxels form a vector $\vec{e}_{i,j}$. Correspondingly, for the *k*th ($k \in \{1, \ldots, K\}$, and *K*=4 here) feature image, at each of the *M* sampled voxels, the feature vector $\vec{t}_{i,j,k,m}$ is obtained, and its Mahalanobis distance to the features at the *i*th corresponding landmark points in the training set is computed as

$$d_{i,j,k,m} = \sqrt{\left(\vec{t}_{i,j,k,m} - \bar{\vec{t}}_{i,k}\right)^T S_{i,k}^{-1} \left(\vec{t}_{i,j,k,m} - \bar{\vec{t}}_{i,k}\right)} \tag{4.6}$$

where $\bar{\vec{t}}_{i,k}$ is the mean feature vector at all landmarks corresponding to the *i*th landmark points on the *k*th feature, and $S_{i,k}^{-1}$ is their covariance matrix, both associated with the *k*th feature. The Mahalanobis distances for the *M* sampled landmark points form a vector $\vec{d}_{i,j,k}$. The calculation is performed for the *N* training images. Concatenating all $\vec{e}_{i,j}$'s for all training images, a vector of Euclidean distances $\vec{E}_i$ is formed. For each feature, concatenating all $\vec{d}_{i,j,k}$'s, a vector $\vec{D}_{i,k}$ of Mahalanobis distances is formed. Therefore, for the *K* features, there is a set of vectors $\{\vec{D}_{i,1}, \ldots, \vec{D}_{i,K}\}$ for the Mahalanobis distances. With the correlation coefficient of two vector $\vec{x}$ and $\vec{y}$ with same length defined as

$$cc(\vec{x}, \vec{y}) = \frac{\sum(\vec{x} - \bar{\vec{x}})(\vec{y} - \bar{\vec{y}})}{\sqrt{(\sum(\vec{x} - \bar{\vec{x}})^2 \sum(\vec{y} - \bar{\vec{y}})^2)}} \tag{4.7}$$

The linear combination for the $\vec{D}_{i,k}$'s that maximizes the correlation with the Euclidean distance vector $\vec{E}_i$ is obtained by finding the coefficients $\alpha_{i,k}$'s forming a vector $\vec{\alpha}_i$

$$\vec{\alpha}_i = \arg\max_{\alpha_{i,k}}[cc_i(\vec{E}_i,\ \alpha_{i,1}\vec{D}_{i,1} + \cdots + \alpha_{i,K}\vec{D}_{i,K})] \tag{4.8}$$

This vector $\vec{\alpha}_i$ provides the weights for combining the features to determine the updated location of the landmark on the search vector. Following the optimization procedure introduced by Toth *et al.* [31], for each landmark point, a specific $\vec{\alpha}_i$ is calculated and stored, as well as the maximized correlation cc_i.

4.2.2.4 Constrained ASM and assignment of landmark uncertainty

Depending on the quality of the image, the actual displacement of the landmark points may not be reliable in certain areas, even when the optimal set of features is used. Including these landmarks in fitting the ASM could affect the fitted shapes, and subsequently lead the updating to suboptimal directions. To address the problem, a constrained ASM with landmark uncertainty is used. The model was originally proposed by Baka *et al.* [29] for the fitting of noisy sparse landmarks. For each landmark, a new term named uncertainty is used as a property describing the reliability of its position. For the purpose of image segmentation, for instance, landmarks on strong edges tend to have low uncertainty, while landmarks in the regions without reliable boundaries tend to have high uncertainty. Instead of fitting the shape using the entire set of landmarks, the constrained ASM essentially fits landmarks with low uncertainty and then use their locations to derive the locations for landmarks with high uncertainty.

According to Cootes *et al.* [33], fitting the ASM is equivalent to maximizing the probability of

$$P(ASM|Data) \propto P(Data|ASM)P(ASM) \tag{4.9}$$

in which the $P(Data|ASM)$ term can be described by a covariance matrix Σ_U, after the notion of uncertainty is introduced for each landmark. Σ_U is a diagonal matrix, whose corresponding diagonal element for each landmark is its uncertainty. The term $P(ASM)$ can be described by the covariance of the ASM. The maximization of the probability $P(ASM \mid Data)$ can be expresses as the minimization of a negative log likelihood function that is as a function of the model parameters

$$E(\vec{b}) = \left(\vec{x} - \tilde{\vec{x}}(\vec{b})\right)^T \Sigma_U^{-1} \left(\vec{x} - \tilde{\vec{x}}(\vec{b})\right) + \vec{b}^T C_m^{-1} \vec{b} \tag{4.10}$$

where $\vec{x}$ is the target shape to be fitted, $\tilde{\vec{x}}(\vec{b})$ is the fit by the model using the model parameter $\vec{b}$, and C_m is the covariance matrix in the reduced parameter space that can be derived from the original covariance matrix Σ_m as

$$C_m = \Phi^T \Sigma_m \Phi \tag{4.11}$$

Taking the first order derivative of $E(\vec{b})$ with respect to $\vec{b}$ and setting the result to zero, the optimal model parameter can be found as

$$\vec{b}_{opt} = (\Phi^T \Sigma_U^{-1} \Phi + C_m^{-1})^{-1} \Phi^T \Sigma_U^{-1} \left(\vec{x} - \overline{\vec{x}}\right) \tag{4.12}$$

Substituting the parameter $\vec{b}_{opt}$ into Equation 4.1, the optimal fitting for the constrained ASM can be obtained. Note that by inverting the matrix Σ_U, landmarks with high

uncertainty will have lower weights, while landmarks with low uncertainty will have higher weights in determining $\vec{b}_{opt}$.

Previously (Chen *et al.* 2012) we have introduced a semi-automatic method for assigning uncertainty. This scheme assigns high uncertainty to a fixed set of landmarks corresponding to the interior boundary defined manually by a binary mask. Here we propose to use an automatic approach based on the optimal features. We calculate the cumulative distribution of the optimized correlation coefficients for all landmarks, and obtain a cumulative distribution function $F(cc) \subseteq [0,1]$. Let cc_{th_l} be a lower limit for the correlation coefficient such that $F(cc_{th\_l}) = p_l$ where p_l is a percentile of landmarks with low correlation, and let cc_{th_h} be a higher limit for the correlation coefficient such that $F(cc_{th\_h}) = p_h$ where p_h is a percentile of landmarks with high correlation. In this study, p_l and $p_h$were experimentally set at 20% and 80%, respectively. With the uncertainty values 0.1 (low uncertainty points) and 10 (high uncertainty points) used in our previous experiments [34] assigned as the lower and higher bounds, we compute the uncertainty u_i for the *i*th landmark as

$$u_i = \begin{cases} 0.1, & cc_i \geq cc_{th\_h} \\ 10^{1-2\left(\frac{cc_i - cc_{th_l}}{cc_{th_h} - cc_{th_l}}\right)^{\gamma}}, & cc_{th\_l} < cc_i < cc_{th\_h} \\ 10, & cc_i \leq cc_{th\_l} \end{cases} \tag{4.13}$$

where $\gamma \geq 0$ is a parameter controlling the rate at which the uncertainty value decreases when the correlation coefficient increases. The higher γ is, the slower the rate. Based upon this assignment, landmarks with low cc will receive high uncertainty values, while landmarks with high cc will receive low uncertainty values. After a few tests, we set γ at

0.012 in our following experiments. The uncertainty u_i's obtained using Equation 4.13 are used to form the matrix Σ_U for estimating the model parameters.

4.2.2.5 Implementing the constrained ASM with landmark uncertainty and optimal features to update landmarks and fit shapes

After the uncertainty for each landmark is assigned, the points $\vec{x}_p$ representing the rough segmentation obtained from registrations with the average atlas are fitted to the model by the constrained ASM algorithm using the optimal model parameters computed by Equation 4.12. Each landmark of the fitted shape is then updated. This is achieved firstly by using $\vec{\alpha}_i$ as the weights in combining the Mahalanobis distances provided by the features using Equation 4.4, and selecting the candidate point on the search vector which yields the lowest combined Mahalanobis distance. A model is fitted again to the updated landmarks by the constrained model and the process is repeated until convergence or the maximum number of iterations is reached. Note that although the location of landmarks with high uncertainty is updated at each iteration, their impact on the fitted shape is always reduced due to the use of the constrained model. Their location is mainly derived from landmarks with low uncertainty.

4.3. RESULTS

We carry out a leave-one-out experiment to evaluate the effectiveness of the proposed model. In each run, one image volume is eliminated from the training set. We also eliminate the volume corresponding to the opposite side of the same patient. This is done because, although we have not observed shape symmetry that is sufficient to derive the shape of the patient's gland on the opposite side, the image features may be similar on

both sides. This may introduce a bias in the updating of the landmarks. Therefore, the model is constructed using the 28 volumes that are obtained from the remaining 28 training images. Results obtained using the proposed model are compared with results obtained using a regular ASM approach. The regular ASM approach used here relies on the same combination of features as the constrained model, but all landmarks are treated equally when fitting the model to the updated landmarks. We also compare results obtained using a multiple-atlas-based approach. The approach is adapted from the method we developed previously for segmenting the thyroid gland [32], which weighs each atlas by its local correlation coefficient with the patient image after nonrigid registration. Since it is not always desirable to use all the atlases [35, 36], we are using half of the available atlases, i.e., 14 atlases with the highest local correlation coefficients out of the 28 atlases in the training set. We expect the performance of the multiple-atlas-based approach to be comparable to the multiple-atlas-based methods proposed by Han *et al.* [20] and Yang *et al.* [18]. Based on quantitative and qualitative comparisons, the automatic segmentation method with the best overall performance is identified, and its segmentations are presented to a radiation oncologist as contours on 2D axial slices. The amount of modifications made by the oncologist to make the segmentations clinically acceptable is measured and evaluated.

4.3.1. Quantitative Results

Quantitative validations are performed volumetrically, i.e., we compare automatic segmentation volumes obtained by the three approaches with the manual segmentation volumes, using DSC and surface distance error, i.e. the distance from the surface of the automatic segmentations to the surface of the manual segmentation. Since clinically the

structures are delineated on a slice-by-slice basis, slice-wise comparisons between the automatic and manual segmentations are also performed using DSC and the Hausdorff distance (HD). The DSC is defined as

$$\mathrm{DSC}(L_A, L_M) = \frac{2|L_A \cap L_M|}{|L_A| + |L_M|} \tag{4.14}$$

where L_A represents the automatic segmentation volume/slice and L_M represents the manual segmentation volume/slice. HD is defined as

$$HD(L_A, L_M) = \max\{\max_{v_A \in L_A}(\min_{v_M \in L_M}(\|v_A - v_M\|)), \max_{v_M \in L_M}(\min_{v_A \in L_A}(\|v_M - v_A\|))\} \tag{4.15}$$

where v_A represents voxels inside the automatic segmentation, v_M represents voxels inside the manual segmentation, and $\|\cdot\|$ is the operator for the Euclidian distance.

Due to severe streak artifacts in the region of the parotid glands, the results for patient 1, which are denoted case 1 for the right side and case 16 for the left side, are poor. These two cases are outliers and have thus been eliminated from the results shown here. Results presented in this section thus include 28 cases.

4.3.1.1. Volumetric comparison

The left panel of Figure 4.2 compares the DSCs for the automatic segmentations obtained using the three methods, which are the multiple-atlas-based (MultiAtlas) approach, the regular ASM with optimal features (RegASM), and the constrained ASM with optimal features (ConsASM) from left to right, respectively. In this figure, the data range for each method is shown as the range between the minimum and the maximum whiskers in the boxplot. The bottom and top of the box shows the 25th and 75th percentile,

the line in the middle represents the median, and the "+" signs show the outliers. It can be seen that all three methods perform comparably when they are compared with volumetric DSC. The median DSC of ConsASM is 0.804, which is slightly higher than the median DSC of MultiAtlas (0.802) and the median DSC of RegASM (0.791). Also, the average DSC for ConsASM is 0.798, which is higher than the average DSC for MultiAtlas (0.796) and the average DSC for RegASM (0.792), but the difference is insignificant. One-sided paired *t*-tests with α=0.05 indicate that none of the methods shows a difference that is statistically significant when compared with any of the other two.

In addition to DSC, for each method, the distances from the surface of the automatic segmentation to the surface of the manual segmentation are calculated. The average and maximum distance errors for each case are calculated, and the results in the form of range, 25th percentile, 75th percentile, median, and outliers for the 28 cases are shown in the middle and right panels of Figure 4.2. The middle panel shows the plots for the mean values; the right panel shows the plot for the maximum values. It can be seen that the ConsASM has the lowest average and maximum distance errors among the three methods. RegASM performs better than MultiAtlas in general, while the median of the average distance errors for RegASM is higher than that for MultiAtlas. *T*-tests show that both average and maximum distance errors of ConsASM are significantly lower than those of RegASM and MultiAtlas with α=0.05.

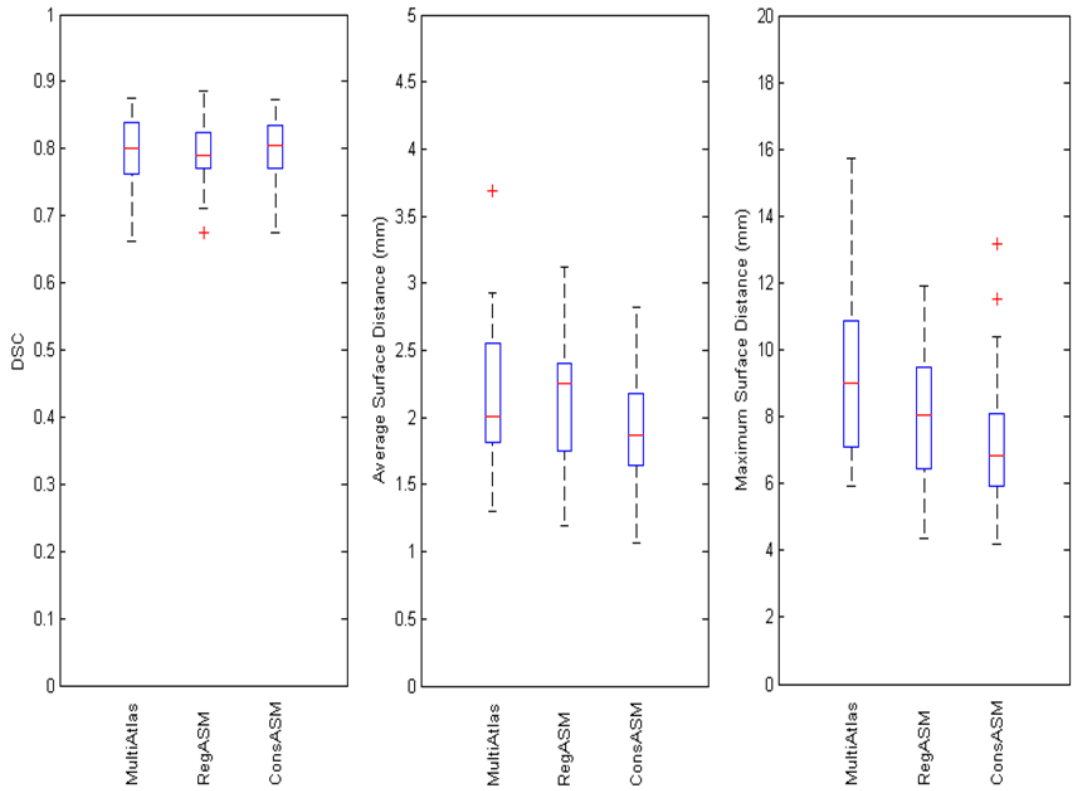

Figure 4.2. Volumetric comparisons between automatic and manual segmentations. Boxplots show the sample minimum, Q1, Q2, Q3, sample maximum, and outliers. Left panel (left to right): DSC of MulitAtlas, RegASM, and ConsASM. Middle panel (left to right): Average surface distance errors of MulitAtlas, RegASM, and ConsASM. Right panel (left to right): Maximum surface distance errors of MulitAtlas, RegASM, and ConsASM.

4.3.1.2. Slice-by-slice comparison

For each case, the DSC between each of the automatic segmentations and the manual segmentation is calculated on a slice-by-slice basis. Since it is difficult for the top and bottom slices in the automatic segmentation to be as flat as those in the manual segmentations, results for all slices except the top slice and bottom slice are averaged. Repeating this for all 28 cases, the 28 averages for each automatic segmentation method are obtained. The range, 25th percentile, 75th percentile, median, and outliers for the mean distributions are shown in the left panel of Figure 4.3. Similarly, the slice-by-slice HDs are calculated and their statistics are shown in the right panel of Figure 4.3. All three

methods show similar performances and their differences are not statistically significant when compared with DSC. However, when measured by HD, ConsASM has the lowest median and more cases are in the lower range than with RegASM and MultiAtlas. The average for the 28 cases for ConsASM is 5.01mm, which is lower than the 5.18mm of RegASM and the 5.56mm of MultiAtlas. The difference between the HD of ConsASM and MultiAtlas is statistically significant (p=0.017).

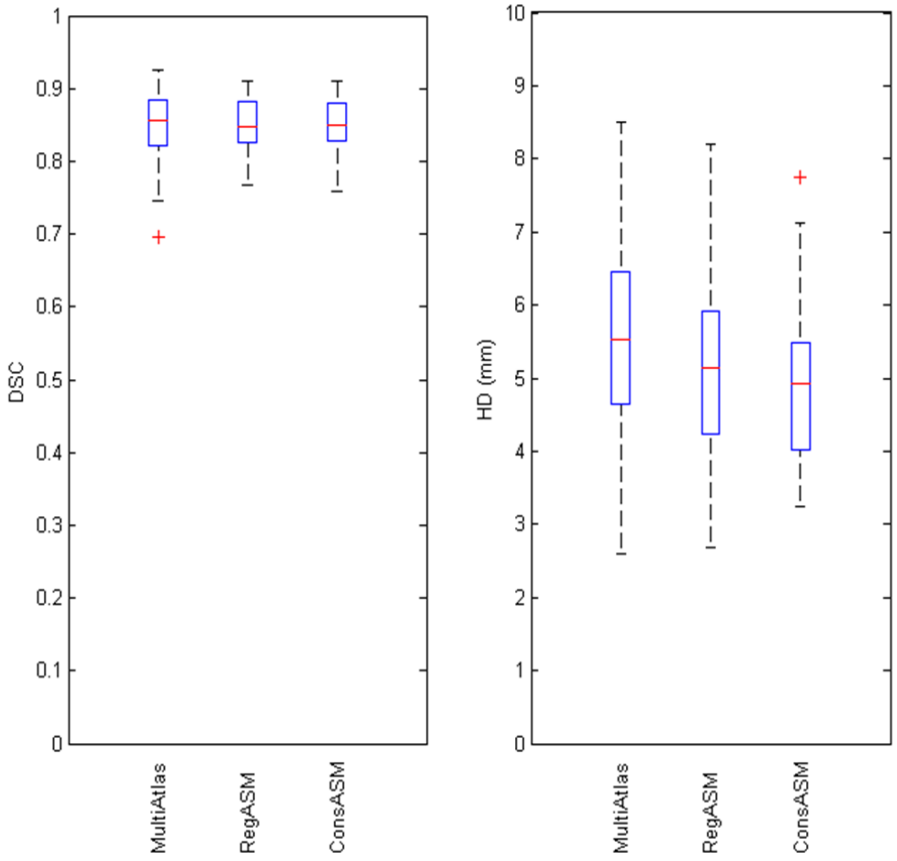

Figure 4.3. Slice-by-slice comparisons between automatic and manual segmentations. Boxplots show the sample minimum, Q1, Q2, Q3, sample maximum, and outliers. Left panel (left to right): Average DSC of MulitAtlas, RegASM, and ConsASM. Right panel (left to right): Average HD of MulitAtlas, RegASM, and ConsASM.

4.3.2. Qualitative Results

Qualitative results are shown for six representative cases in the forms of 2D contours on axial slices, as well as 3D surfaces. The color in each image encodes the distance from the surface of the automatic segmentation to the surface of the corresponding manual segmentation. The six cases are: the right parotid of patient 3 (P3_R), the right parotid of patient 6 (P6_R), the right parotid of patient 10 (P10_R), the left parotid of patient 4 (P4_L), the left parotid of patient 7 (P7_L), and the left parotid of patient 12 (P12_L).

4.3.2.1. 2D contours

In Figure 4.4, contours obtained with automatic methods are compared to those obtained manually. The original CT images are shown in the leftmost column. Contours obtained with the MultiAtlas technique are shown as yellow contours in the second column, contours obtained with RegASM are shown as blue contours in the third column, and contours obtained with ConsASM are shown as red contours in the last column. In general, it can be seen that, among the three sets of contours generated automatically, the red contours are the closest to the green. Compared with the yellow and blue contours, most of the improvements shown by the red contours appear at the interior and anterior boundaries between the parotid glands and the surrounding soft tissues. For almost all cases shown here, ConsASM increases the segmentation accuracy at the interior boundary, except for P4_L in which ConsASM was initialized at a wrong location and was subsequently trapped in a local minimum (none of the three methods reached the correct interior boundary). Also, the identification of the anterior boundary against the masseter muscle has been improved in some of the cases, e.g. P6_R, P10_R, P4_L, and P12_L.

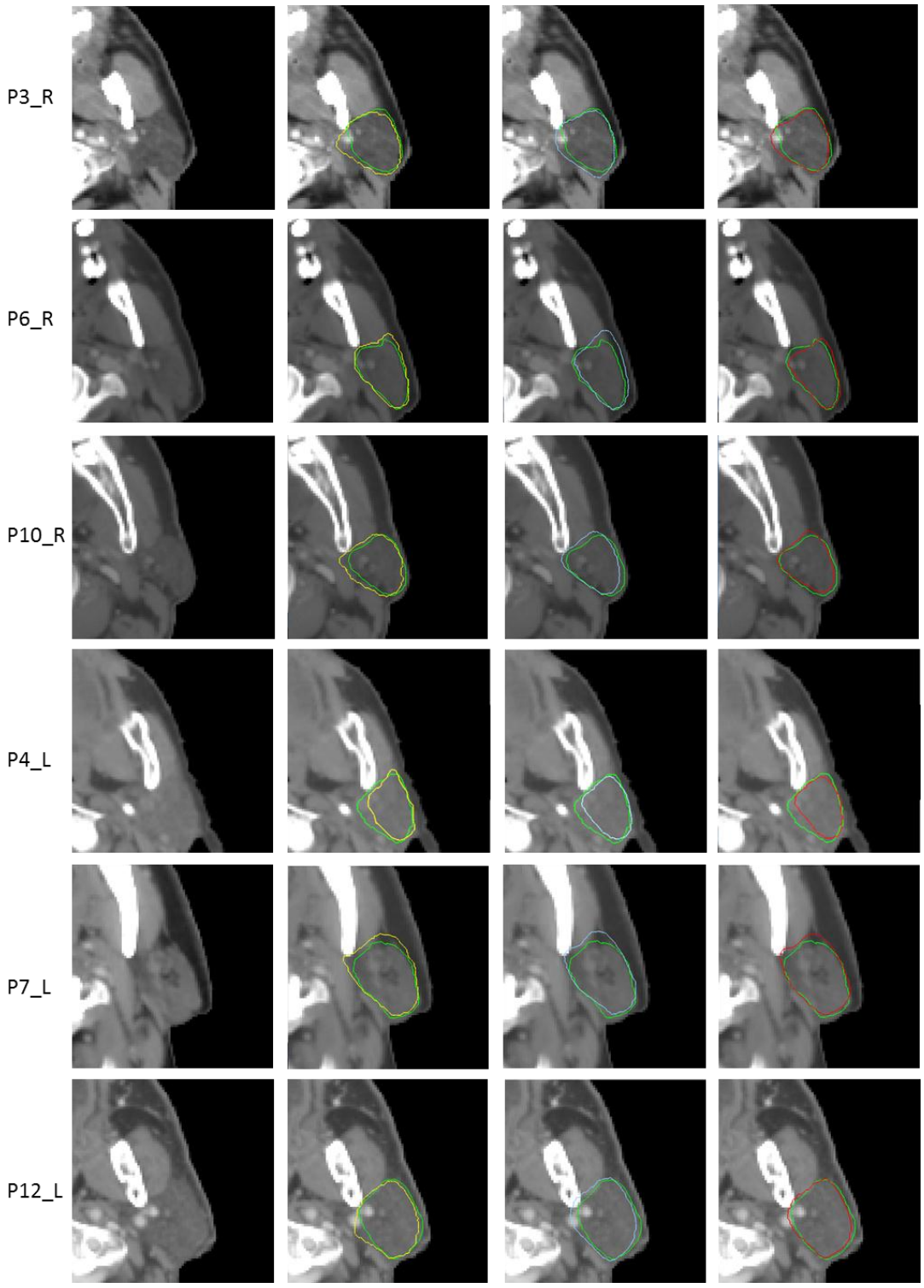

Figure 4.4. Segmentations shown as contours on 2D axial slices. Manual segmentations are in green. Each column from left to right: The original image, MulitAtlas in yellow, RegASM in blue, and ConsASM in red. From top to bottom rows: right parotid of patient 3, right parotid of patient 6, right parotid of patient 10, left parotid of patient 4, left parotid of patient 7, and left parotid of patient 12.

4.3.2.2. 3D surfaces

For the same six cases, the automatic segmentations are shown in Figure 4.5 as 3D surfaces, on which red means a large distance from the manual segmentation, and blue means a small error. The warmer the color is, the higher the distance. To focus on the boundaries that are more difficult to segment, the surfaces are rotated such that the interior boundary, which is the left facet of the structure, and the anterior boundary, which is the right facet of the structure, are facing front. Comparing the three methods, which are MultiAtlas, RegASM, and ConsASM from the top to the bottom rows, it can be observed that ConsASM performs the best in general. Indeed, fewer regions are red and the area of each red region has been substantially reduced. Also, the color of these regions tends to be less warm when compared with RegASM and MultiAtlas.

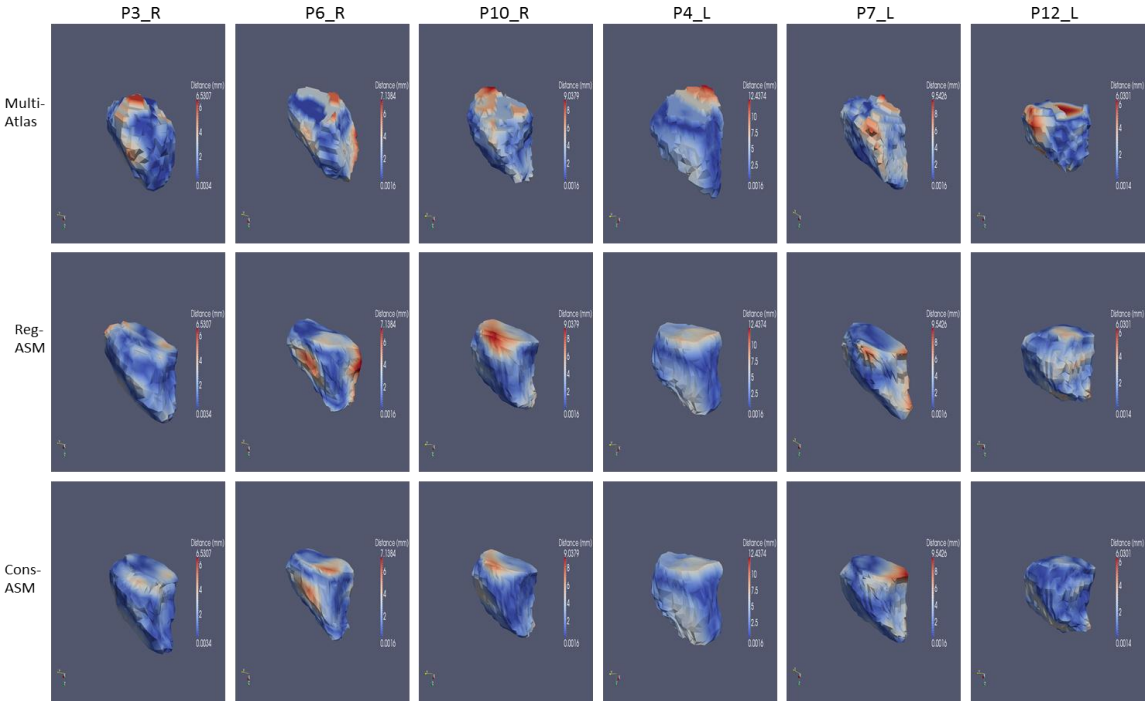

Figure 4.5. Segmentations shown as 3D surfaces for MultiAtlas (top row), RegASM (middle row), and ConsASM (bottom row), where red represents large distance error and blue means small distance error. For each column, from left to right: right parotid of patient 3, right parotid of patient 6, right parotid of patient 10, left parotid of patient 4, left parotid of patient 7, and left parotid of patient 12.

4.3.3. Modification of Automatic Segmentations

To study the clinical usefulness of the segmentations generated by ConsASM, we present the segmentations of ConsASM to a radiation oncologist such that he could modify the contours shown on axial slices in 2D to let them meet clinical requirements. Using a tool developed in house, the oncologist can add a part to the contour by brushing over the targeted area while pressing the left button of the mouse, or remove a part by brushing over and pressing the right button. The modified segmentation, denoted as ConsASM_mod, are saved as binary masks and then compared with the segmentations obtained by ConsASM.

The average volumetric DSC for the 28 cases is 0.984, which is an indication of good segmentation accuracy. Further comparison on a slice-by-slice basis using DSC has shown that among the 353 slices of ConsASM_mod, 310 slices have a DSC of 1. This means that these contours were accepted without modification. This represents about 87.8% of all slices. No modification of any kind was made on two out of the 28 cases.

4.4. DISCUSSION AND CONCLUSIONS

We have introduced a constrained ASM with landmark uncertainty and optimal features for the automatic segmentation of the parotid glands. Correspondence between the training shapes is obtained automatically via nonrigid registration. Updating the landmarks is facilitated by combining four image features which are intensity, local intensity mean, local intensity standard deviation, and local entropy. The coefficients for combining the features at each landmark are found by first sampling a set of points in the landmark's neighborhood, and then optimizing the correlations between their Euclidean

distances and their Mahalanobis distances. Moreover, the optimized correlation coefficients are used to assign an uncertainty to each of the landmarks. This property is then used to drive a constrained ASM model. Landmarks with low uncertainty have a higher impact in the fitting process than landmarks with high uncertainty. This is a general method that can be implemented in other ASM-based segmentation problems when parts of the object boundaries cannot be localized reliably. The automated framework is highly flexible and permits changing the size of the training set, the features used, or the mapping between the correlation coefficients and the uncertainties, depending on the application.

A direct comparison between the proposed approach and the established approaches for segmenting the parotid glands is difficult because a common dataset is unavailable. As an alternative, we have compared result obtained with ConsASM with results obtained with MultiAtlas, which is a multiple-atlas-based method whose accuracy has been shown to be better than majority vote and STAPLE [32]. Notice that results obtained using MultiAtlas show an average volumetric DSC of 0.796, which is lower than the average volumetric DSC of 0.851 obtained by the multiple-atlas-based approach of Han *et al.* [20]. However, the average slice-wise HD obtained by MultiAtlas is 5.56 mm, which is lower than the 5.93 mm by Han's method. This indicates that the performances of MultiAtlas and Han's method may actually be comparable. The worse volumetric DSC of MultiAtlas might be due to the larger slice thickness in our image set (3mm) than that in the images used in the MICCAI Head and Neck Auto-Segmentation Challenge Workshop (2mm). Thinner slices lead to more well-segmented slices in the middle part of the gland and thus improve the statistics. The significantly lower average

HD of 5.01 mm obtained with ConsASM provides evidence that the proposed method can effectively reduce the slice-by-slice distance error observed with multiple-atlas-based approaches. This is also supported by the reduction in surface distance errors of about 22% (ConsASM's average of maximum distance errors in all cases is 7.22mm, while MultiAtlas's average of maximum distance errors in all cases is 9.21mm. $p<0.05$) and the qualitative results shown as 2D contours in Figure 4.4 and 3D surfaces in Figure 4.5.

Compared with RegASM whose landmarks are updated with equal weights, ConsASM also effectively reduces the large segmentation errors at the fuzzy boundaries, e.g. the interior boundaries of P3_R and P12_L. It is seen from Figure 4.4 that RegASM for P12_L caused over-segmentation and the fitted shape entered the area of the mandible. This is corrected when using ConsASM whose landmarks at the interior boundary have low weights. The average of the maximum volumetric surface distance errors is 7.92 mm for RegASM, and the reduction by ConsASM is about 9%. A 0.17 mm reduction of average slice-by-slice HD is also observed, although the reduction is statistically insignificant ($p=0.17$). When comparing ConsASM with the feature-driven model-based method proposed by Qazi *et al.* [26], one notices that Qazi's method results in a higher average volumetric DSC of 0.83, but the data also has a slice thickness at 2mm. However, the HD of 5.82mm for the left parotid and 5.70mm for the right parotid are both higher than 5.01mm obtained with ConsASM. Because only one case was shown qualitatively by Qazi *et al*. in [26], and because the quality of the image on that slice is good with a high contrast between the parotid and the surrounding tissues, it is difficult to identify regions or cases that are challenging for their approach.

As is the case for RegASM, the capability of ConsASM to accurately segment structures in patient images is limited by the availability of shapes and image features in the training set. For example, P7_L shows a case in which the parotid is not closely connected to the masseter muscle. This has not been observed in other volumes and the dark gap between the two structures was not identified as the proper boundary. This resulted in over-segmentation by the ASM. A possible solution could be to construct an ASM specifically for this type of patients but grouping training images in coherent groups is not easy. This may be facilitated by exploring the relation between the patients' metadata and the texture features in the images, e.g. patients' age information since parotid density loss caused by aging could result in darkness and inhomogeneity of the parotid glands. This kind of information could help in dividing a large set of training images into groups and construct group-specific ASMs.

Since the interval between the time the oncologist was first asked to manually delineate the parotid to train the models and the time at which the modification study was done is over 1 year, he could not remember specifically how contours were initially drawn. As reported, about 87.8% of the ConsASM contours have been accepted directly by the radiation oncologist despite the fact that the average HD for the slice-by-slice evaluation is larger than 5mm. This indicates the amount of variability that can occur when delineating contours manually, the complexity of the task and the difficulty to come up with good gold standards that can be used to evaluate automatic segmentation algorithms. It also suggests that the method we propose would be of clinical value if integrated into the normal clinical flow. It has been observed that most of the modifications are performed on slices near the top and bottom of the glands. This could

be addressed by more carefully delineating these slices in the atlases and identifying better the top and bottom of the glands. To better define the upper boundary of the glands in 3D, one option is to detect the mastoid process at the bottom of the skull. In most of the cases, once this structure is seen in the axial view of the CT image, the adjacent parotid gland is considered to have reached its upper boundary. Because the mastoid process is a porous structure filled with air, a rough segmentation just for positioning purpose could probably be achieved automatically by a local threshold in CT images. The identification of the lower boundary of the glands in 3D is more challenging due to the lack of anatomical landmarks. But since the size of a gland becomes smaller when approaching the lower boundary, it could be defined as the slice with a sharp drop in the number of pixels within the 2D contour. To do this, the identification of both boundaries will initially require a little over-segmentation of the glands, followed by post-processing discarding the redundant volumes. This will require the manual delineations on the training images to exceed their general upper and lower boundaries in the first place.

In conclusion, we have demonstrated that the proposed constrained ASM with optimal features and landmark uncertainty can segment the parotid glands at higher accuracy than both a regular ASM approach with the same optimal features and a multiple-atlas-based approach. The most significant improvement is on the volumetric maximum surface distance errors and slice-by-slice HD. Although an average slice-by-slice HD of 5.01 mm by ConsASM is still considered to be substantial, a study conducted with a radiation oncologist has shown that more than 87% of the contours were clinically acceptable.

REFERENCES

[1] Chao K S, Deasy J O, Markman J, Haynie J, Perez C A, Purdy J A, and Low D A 2001 A prospective study of salivary function sparing in patients with head-and-neck cancers receiving intensity-modulated or three-dimensional radiation therapy: initial results Int. J. Radiat. Oncol. Biol. Phys. 49 907–16

[2] Chao K S 2002 Protection of salivary function by intensity-modulated radiation therapy in patients with head and neck cancer Seminars in Radiation Oncology 12 20–5

[3] Eisbruch A, Ship J A, and Martel M K *et al.* 1996 Parotid gland sparing in patients undergoing bilateral head and neck irradiation: techniques and early results Int. J. Radiat. Oncol. Biol. Phys. 36 469–80

[4] Eisbruch A, Ten Haken R K, Kim H M, Marsh L H, and Ship J A 1999 Dose, volume and function relationships in parotid salivary glands following conformal and intensity-modulated irradiation of head and neck cancer Int. J. Radiat. Oncol. Biol. Phys. 45 577–87

[5] Eisbruch A, Kim H M, Terrell J E, Marsh L H, Dawson L A, and Ship J A 2001 Xerostomia and its predictors following parotid-sparing irradiation of head-and-neck cancer Int. J. Radiat. Oncol. Biol. Phys. 50 695–704

[6] Eisbruch A, Rhodus N, Rosenthal D, Murphy B, Rasch C, Sonis S, Scarantino C and Brizel D 2003 The prevention and treatment of radiotherapy-induced xerostomia Seminars in Radiation Oncology 13 302–308

[7] Eisbruch A 2007 Reducing xerostomia by IMRT: what may, and may not, be achieved Journal of Clinical Oncology 25 4863–4

[8] Nutting C M, Morden J P, and Harrington K J *et al.* 2011 Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial Lancet Oncol 12 127–36

[9] Parliament M B, Scrimger R A, Anderson S G, Kurien E C, Thompson H K, Colin Field G, and Hanson J 2004 Preservation of oral health-related quality of life and salivary flow rates after inverse-planned intensity-modulated radiotherapy (IMRT) for head-and-neck cancer Int. J. Radiat. Oncol. Biol. Phys. 58 663–73

[10] Saarilahtia K, Kouria M, and Collana J *et al.* 2005 Intensity modulated radiotherapy for head and neck cancer: evidence for preserved salivary gland function Radiother. Oncol. 74 251-8

[11] van Rij C M, Oughlane-Heemsbergen W D, Ackerstaff A H, Lamers E A, Balm A J, and Rasch C R 2008 Parotid gland sparing IMRT for head and neck cancer improves xerostomia related quality of life Radiat Oncol 3:41

[12] Hsiung C Y, Ting H M, Huang H Y, Lee C H, Huang E Y, and Hsu H C 2006 Parotid-sparing intensity-modulated radiotherapy (IMRT) for nasopharyngeal carcinoma: preserved parotid function after IMRT on quantitative salivary scintigraphy, and comparison with historical data after conventional radiotherapy Int. J. Radiat. Oncol. Biol. Phys. 66 454-61

[13] Eisbruch A, Marsh L H, Martel M K, *et al.* 1998 Comprehensive irradiation of head and neck cancer using conformal multisegmental fields: Assessment of target coverage and noninvolved tissue sparing. Int. J. Radiat. Oncol. Biol. Phys. 41 559–568

[14] Jabbari S, Kim H M, and Feng M *et al.* 2005 Matched case-control study of quality of life and xerostomia after intensity-modulated radiotherapy or standard radiotherapy for head-and-neck cancer: initial report Int. J. Radiat. Oncol. Biol. Phys. 63 725–31

[15] Lin A, Kim H M, Terrell J E, Dawson L A, Ship J A, and Eisbruch A 2003 Quality of life after parotid-sparing IMRT for head-and-neck cancer: a prospective longitudinal study Int. J. Radiat. Oncol. Biol. Phys. 57 61–70

[16] Ramus L and Malandain G 2010 Multi-atlas based segmentation: application to the head and neck region for radiotherapy planning. In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 281–8

[17] Guimond A D, Meunier J, and Thirion J P 2000 Average Brain Models: A Convergence Study. Computer Vision and Image Understanding. 77 197–201

[18] Yang J, Zhang Y, Zhang L, and Dong L 2010 Automatic segmentation of parotids from CT scans using multiple atlases. In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 323–30

[19] Warfield S, Zou K, and Wells W 2004 Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging. 23 903–21

[20] Han X, Hibbard S, O'Connell P, and Willcut V 2010 Automatic segmentation of parotids in head and neck CT images using multi-atlas fusion. In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 297-304

[21] Dice L R 1945 Measures of the amount of ecologic association between species. Ecology 26 297–302

[22] Gorthi S, Cuadra M B, and Schlick U *et al.* 2010 Multi-atlas based segmentation of head and neck CT Images using active contour framework In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 313–21

[23] Hollensen C, Hansen M F, Hojgaard L, Specht L, and Larsen R 2010 Segmenting the parotid gland using registration and level set methods. In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 305–12

[24] Gering D and Kalinosky B 2010 Automatic segmentation of the parotid glands by situated Bayesian classification. In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 289–296

[25] Qazi A A, Kim J J, Jaffray D A, and Pekar V 2010 Probabilistic refinement of model-based segmentation: Application to radiation therapy planning of the head and neck. In Porc. MIAR, LNCS 6326. 403–10

[26] Qazi A A, Pekar V, Kim J J, Xie J, Breen S, and Jaffray D A 2011 Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach. Med. Phys. 38 6160–70

[27] Huttenlocher D, Klanderman D, and Rucklige A 1993 Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 23(7) 850–63

[28] Pekar V, Allaire S, Qazi A A, Kim J J, and Jaffray D A 2010 Head and neck auto-segmentation challenge: Segmentation of the parotid glands. In proc. MICCAI 2010 Workshop Head & Neck Autosegmentation Challenge. 273–80

[29] Baka N, de Bruijne M, Reiber J H C, Niessen W, and Lelieveldt B P F 2010 Confidence of model based shape reconstruction from sparse data In Proc. ISBI 1077–80

[30] Chen A, Deeley M A, Niermann K J, Moretti L, and Dawant B M 2010 Combining registration and active shape models for the automatic segmentation of the lymph node regions in head and neck CT images. Med Phys 37 6338–46

[31] Toth R, Doyle S, and Rosen M *et al.* 2009 WERITAS – Weighted ensemble of regional image textures for ASM segmentation. In Proc. SPIE Medical Imaging 725905-725905-11

[32] Chen A, Niermann K J, Deeley M A, and Dawant B M 2011 Evaluation of multiple-atlas-based strategies for segmentation of the thyroid gland in head and neck CT images for IMRT. Phys. Med. Biol. 57 93–111

[33] Cootes T F and Taylor C J 2001 Constrained active appearance models. In Proc. ICCV, 1 748–54

[34] Chen A, Noble K J, Niermann K J, Deeley M A, and Dawant B M 2012 Segmentation of parotid glands in head and neck CT images using a constrained

active shape model with landmark uncertainty. In Proc. SPIE Medical Imaging 83140P

[35] Aljabar P, Heckermann R A, Hammers A, Hajnal J V, and Rueckert D 2009 Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage. 46 726–38

[36] Heckemann R, Hajnal J, Aljabar P *et al.* 2006 Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage. 33 115–26

CHAPTER V

SUMMARY AND FUTURE WORK

This dissertation introduces a set of innovative approaches for the automatic segmentation in CT images of structures at risk and structures to be spared during head and neck IMRT treatment. Specifically, method have been developed and evaluated for the level II, III, and IV lymph node regions that are typically regions that need to be irradiated, and for the thyroid gland and the parotid glands that typically need to be spared. In Chapter II, we propose a method for the segmentation of the lymph node regions that relies on a combination of nonrigid registration and ASM. The landmark correspondence for the training shapes is achieved automatically via nonrigid registration, and the initialization of the model on the patient image is established by applying the nonrigid deformation fields. In Chapter III, we describe a framework for multiple-atlas-based approaches that is applied to the segmentation of the thyroid. In this chapter, we compare eight methods for combining the structures projected from the atlases, as well as a method that relies on selecting the segmentation obtained with the most similar atlas. These eight methods include STAPLE and majority vote methods which are the standard techniques. Through volumetric and slice-by-slice comparisons, we identified the best performer as the one combining the segmentations based on local correlation coefficients between registered atlases and the patient image. In a modification study performed with a trained radiation oncologist on contours obtained with thus method, we show that about 40% of the contours on 2D slices for the left thyroid and about 42% for the right thyroid are clinically acceptable and do not require manual modification. An additional 21% for the left and 24% for the right require only minimal modifications. In Chapter IV, we propose a constrained ASM approach for the segmentation of the left and right parotid glands. Based on the set of image features used

to update the landmarks, we associate a reliability term to each landmark. The constrained model allows the locations of the landmarks with high uncertainty to be derived from the locations of landmarks with low uncertainty. The approach leads to better identification of structure boundaries that are partially fuzzy due to the lack of contrast against surrounding tissues. We note that the framework can be extended easily to other ASM-based segmentation problems. Although a 5.01 mm average slice-by-slice HD from the manual delineations used to validate the algorithm is substantial, a modification study also performed by a trained radiation oncologist has shown that about 87.8% of the contours shown on the 2D slices are clinically acceptable without any modification. These results suggest that the proposed method would be of clinical value if integrated into the normal processing flow of IMRT treatment planning.

Options for further improving the segmentation accuracy will involve expanding the training image set. For both ASM-based approaches and multiple-atlas-based approaches, this will allow us to better address the problems caused by inter-patient anatomical discrepancies. For instance, for patients whose thyroids are larger than normal, using a specific set of atlases with large thyroid could improve the registration accuracy and hence lead to more accurate segmentations. Moreover, group-specific ASMs could improve the identification of lymph node and parotid boundaries in patients with different texture features. However, based on our experience, large anatomical discrepancies caused by pathologies, e.g. existence of large tumors and hence large distortion of their surrounding soft tissues, or surgical procedures such as resections and tracheotomy, may not be addressable by simply expanding the training set. In general, pre- or post-processing will be needed in these cases, such as delineating the tumors

manually and use them as hard-constraints for updating landmarks in ASM-based approaches, or eliminating the plastic tubes from patients with tracheotomy by thresholding for more accurate segmentation of the thyroid glands. Not all of these methods are fully automatic at the current stage but semi-automatic methods that would be clinically acceptable can be implemented. .

Automatic segmentation of certainty structures could also be used for constraining the automatic segmentations of nearby structures. For example, the automatic segmentation of the thyroid gland could be used as a constraint when updating the landmarks of the adjacent level IV lymph node region. Further improvements will also require the development of algorithms for automatically segmenting other structures that are not generally delineated for IMRT treatment planning purposes. For instance, segmenting the left and right common carotid arteries and the internal jugular veins could improve the segmentation accuracy for both the thyroid gland and the level IV lymph node regions, and this could be achieved automatically by applying a vessel segmentation algorithm, e.g. an algorithm proposed by Noble *et al*. [1]. These various algorithms could be integrated into a processing pipeline that could be integrated into the clinical flow and facilitate the task of the physicians and medical physicists. Figure 5.1 shows a flow chart for a pipeline that would automate the segmentation process. The main automatic segmentation algorithms are shown in orange. Three major steps would need to be taken care of:

1. Segmentation of constraining structures. Several structures that are not generally delineated for head and neck IMRT treatment planning wouldneed to be segmented and used as constraints for other segmentation tasks. These include the

major vessels, which will help segment the thyroid glands and lymph node regions, the mastoid process, which will help define the upper boundary of the parotid glands, and the hyoid bone and the cricoid cartilage, which are used as landmarks for dividing the lymph node regions into three levels. Other constraining structures could include the tumor site, which could be segmented manually, and the structures introduced by surgical procedures, e.g., the plastic tube implanted when performing a tracheotomy.

2. Segmentation of structures to be spared. In addition to the thyroid glands and the parotid glands, structures such as the larynx, the spinal cord, and the mandible also need to be spared during IMRT treatment. The spinal cord is relatively easy to segment using a conventional single-atlas-based approach, since it is inside the cervical vertebrae, which are significantly brighter than the spinal cord in CT images. It is anticipated that the same approach can be used to segment the larynx since in CT the throat is filled with air and appears much darker than the surrounding soft tissues. We believe a single-atlas-based approach with global affine and nonrigid registrations will be sufficient to segment the two structures. The mandible is a bony structure and could be segmented by either a single-atlas-based approach or local thresholding.
3. ASM segmentation of lymph node regions. The segmentation of the lymph node regions is performed later than those for other structures, since the lymph node regions form the largest and the most challenging structure in our segmentation tasks. Structures such as the tumor, the blood vessels and the thyroid gland could be segmented and used as anatomical constraints for the refinement of the ASM.

Also, for the identification of fuzzy boundaries, e.g. the boundary at the higher portion of the level II lymph node regions, a constrained ASM with landmark uncertainty that has been used for segmenting the parotid glands could be applied.

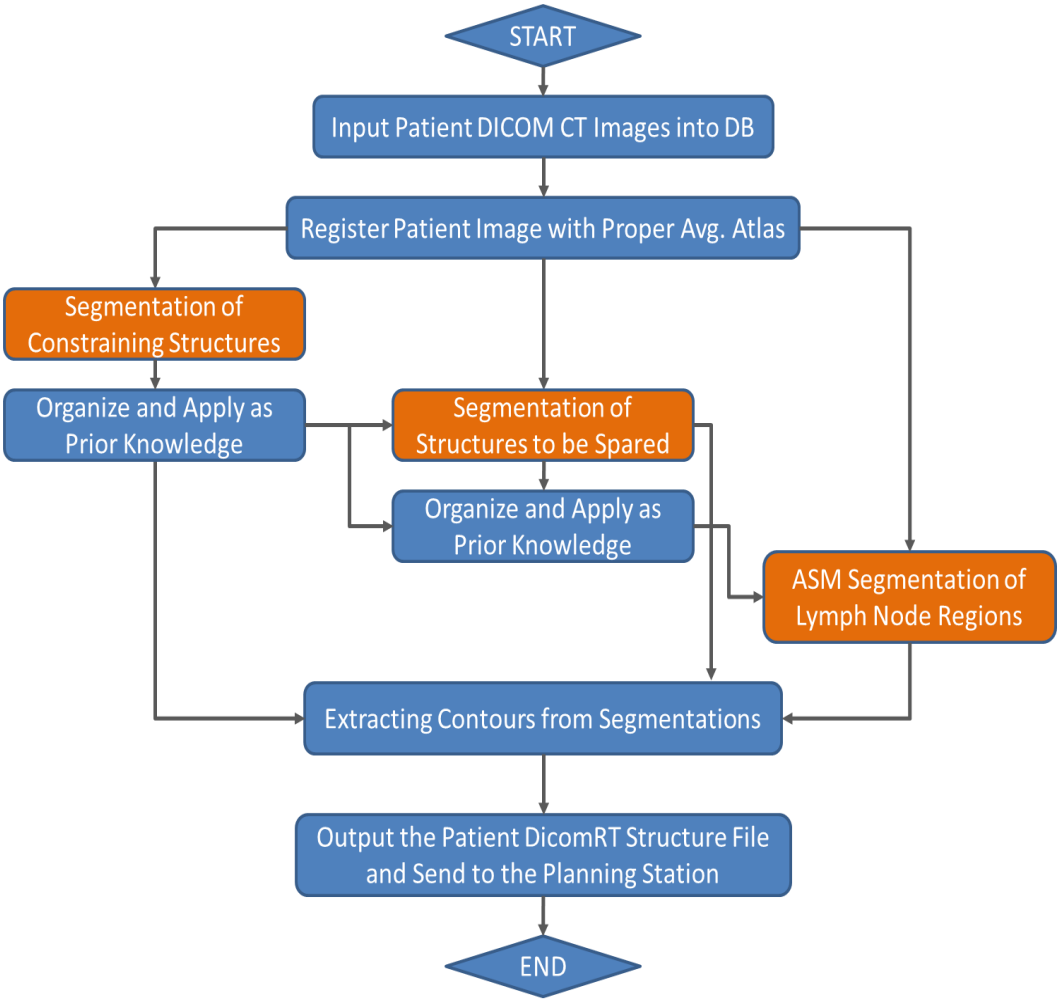

Figure 5.1. Pipeline flow chart of an automatic delineation system for head and neck IMRT treatment planning

The complete segmentation flow will also be supported by several crucial technical components. The system starts with reading the patient's CT image in DICOM format and converting the DICOM slices into a volume. The patient UID's and other information from the metadata will be entered into a database. A proper atlas will be selected and registrations will be executed between the patient and atlas images following

the procedures we proposed. Segmentations will then be achieved using the algorithms we have developed, and the results will be saved. The pipeline will link them with the UID's of the patient and generate the corresponding DICOMRT structure file, and send the files to the IMRT planning stations such as the Eclipse planning systems (Varian Medical Systems, Inc., Palo Alto, CA) used in the Radiation Oncology Department of Vanderbilt University. Our goal is for the physicians to be able to receive a complete set of delineations of head and neck structures in 1~1.5 hours after sending the patient image to the pipeline.

Although the approaches we have provided herein may not be the final solutions yet, we hope that the proposed algorithms have made valuable contribution toward solving the problem of automatic delineation of structures in head and neck CT images, and could be integrated into the clinical flow of head and neck IMRT treatment planning on a regular basis.

REFERENCES

[1] Noble J H, Warren F M, Labadie R F, and Dawant B M 2008 Automatic segmentation of the facial nerve and chorda tympani in CT images using spatially dependent feature values. Med Phys. 35(12) 5375–84