MACHINE LEARNING-BASED TECHNIQUES FOR AUTOMATING IMAGE-GUIDED COCHLEAR IMPLANT PROGRAMMING

By

Dongqing Zhang

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

March 31st, 2019

Nashville, Tennessee

Committee members:

Professor Benoit M. Dawant, Ph.D

Professor Richard Alan Peters, Ph.D

Professor Robert F. Labadie, M.D, Ph.D

Professor Jack H. Noble, Ph.D

Professor Yuankai Huo, Ph.D

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Benoit M. Dawant, for accepting me as a member of Medical Image Processing (MIP) laboratory at the beginning of my graduate program. I have been learning from him in every aspect of doing research since then. These include developing important ideas, elaborating experimental approaches, conducting rigorous validation and presenting the work with clarity and fluency. Also, his friendliness, professionalism, passion and diligence in work are what I wish to emulate, in not only my Ph.D period but also my future career.

I would like to thank my co-advisor Dr. Jack H. Noble of Biomedical Image Analysis for Image Guided Interventions Laboratory (BAGL). Dr. Noble has given me direct guidance during my early years in MIP lab, from which I have benefit tremendously. When I have a technical problem, he is the first one I will seek advice from, thanks to his both big pictures in mind and mastery of the very details of the techniques in this field. His constructive suggestions in both experimenting and writing contribute a lot to the solidity of the studies in this dissertation. His patience to learners and enthusiasm in research set him a great role model for me as well.

I would like to thank Dr. Robert F. Labadie, our expert in cochlear implantation on the medical side for his committee duties. I have also benefit a lot from the works by him and other medical experts on the clinical background of cochlear implantation. They guide me to what we need to do on the engineering side. I would like to thank Dr. Richard Alan Peters and Dr. Yuankai Huo for their committee duties. Dr. Peters provided valuable suggestions in my Ph.D proposal. Dr. Huo is very inspiring and I really benefit from his advice on my deep learning projects.

I appreciate the help from my former and current colleagues. Priyanka Prasad, Dr. Raul Wirz, Rui Li and Bill Rodriguez are always ready to help when I need any technical support. I

would like to thank other former and current Ph.D students in MIP lab and BAGL, including Dr. Yuan Liu, Dr. Yiyuan Zhao, Jianing Wang, Ahmet Cakir, Srijata Chakravorti, Rueben Banalagay and Xiaochen Yang. I value the useful discussions with them. I would like to thank my mentor, Dr. Antong Chen, during my internship at Merck in 2017, for a rewarding internship experience and inspiring communications.

Special thanks to NIH grants R01DC014037, R01DC008408 and R01DC014462 for the financial assistantships for my research. I would also like to thank Advanced Computing Center for Research and Education (ACCRE), Vanderbilt University for providing computing resources and NVIDIA Corporation for its generous donation of a GPU to support our studies.

Finally, I would like to express my gratitude to my parents and my sister for their love and unconditional support to whatever I decide to do.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Table Page

Chapter I INTRODUCTION

1.1 The cochlea and cochlear implants

The cochlea is a snail-shaped cavity in the bony labyrinth and is the auditory portion of the inner ear. In natural hearing, the cochlea transforms mechanical sound waves into electrical stimulation in the auditory nerves, thus creating a sense of hearing. The cochlea is encircled in **Fig 1.1**(a). The spiral ganglion (SG) nerves cells are strung from the bony core of the cochlea and are housed by a structure called the modiolus. The SG nerves are shown in **Fig. 1.1**(b). A characteristic frequency (CF) is associated with each SG nerve. The SG nerves are tonotopically ordered along the length of the cochlea by decreasing CF. An SG nerve is stimulated when the incoming sound includes components that have a frequency that correspond to its CF. When one's intra-cochlear anatomy is damaged, the natural electro-mechanical transduction mechanism cannot function normally and induces hearing loss.

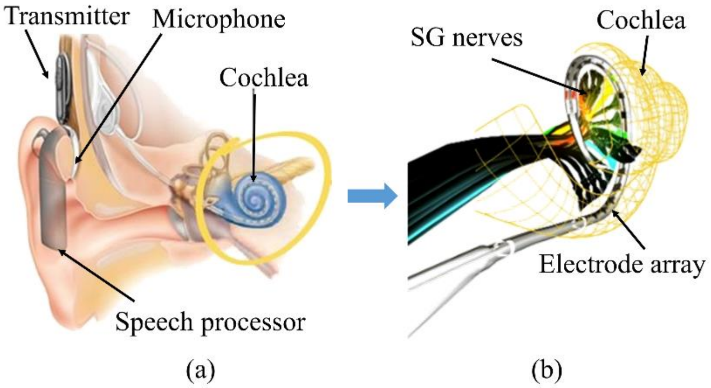

Fig. 1. 1 (a) The cochlea and a cochlear implant. (b) The geometrical relationship between CI electrodes and the intra-cochlear anatomy

Cochlear Implants (CIs) are neural prosthetic devices that are used to restore hearing in patients with severe-to-profound hearing loss. In a CI surgery, an electrode array is threaded into the cochlea. A microphone, a signal processor, and a transmitter are placed externally to receive a sound signal, to process it, and to transmit it to an electrode array which stimulate the SG nerves, thus creating a sense of hearing. Post-operatively, the CI processor needs to be programmed. The programming procedure includes the definition of a set of instructions which we call "MAP". "MAP" specifies the electrodes that are activated when a certain frequency component is detected in the sound and determines these electrodes' stimulation levels. In general, CIs lead to substantial hearing improvements and are often life-changing but a non-negligible number of recipients only get marginal hearing benefit. One important reason is the hearing artifacts caused by electrode interaction. Specifically, since the surgeons do not have full control of the electrode array insertion process, the electrodes may not have a spatial distribution along the neural band as is expected by the audiologists. Some electrodes could be too close to their neighboring electrodes so they could stimulate the same neural region. We call this effect "electrode/channel interaction".

1.2 Image-guided cochlear implant programming (IGCIP)

The traditional CI programming workflow assumes that all electrodes are placed within the cochlea at predefined positions and the audiologists use a default frequency allocation table to program the CIs. Recent studies have shown that hearing outcomes with CIs are correlated with the spatial relationship between the electrode array and the intra-cochlear anatomy (**Fig. 1.2** (a)) [1-6]. One phenomenon that could negatively impact the hearing outcomes, as we have introduced, is electrode interaction. Electrode interaction happens when multiple electrodes are simultaneously stimulating the same area of the neural region. It can be reduced by deactivating a subset of the electrodes which cause it. To select the electrodes to deactivate, the relative positions of each electrode to the SG nerves and their neighboring electrodes need to be determined (See **Fig. 1.2** (b)).

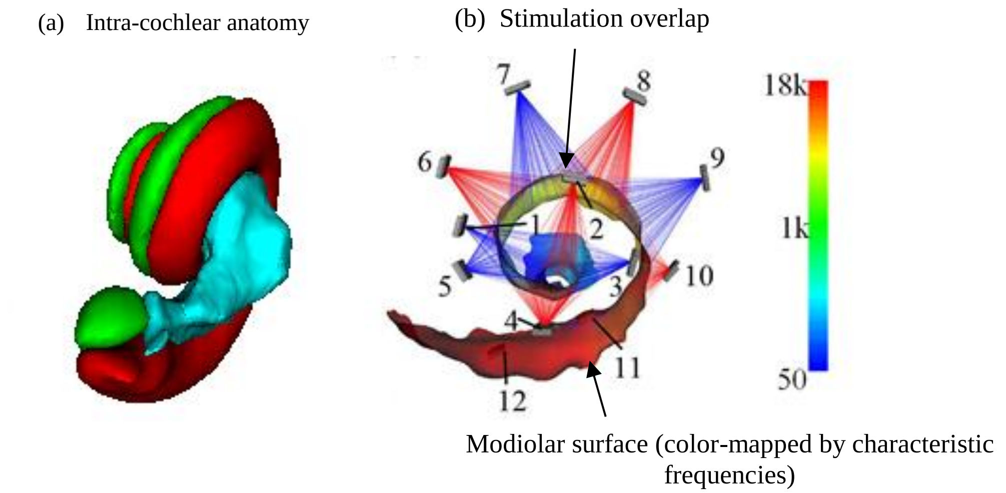
(a) Intra-cochlear anatomy
(b) Stimulation overlap
7
8
6
9
1
2
5
10
3
4
11
12
18k
1k
50
Modiolar surface (color-mapped by characteristic frequencies)

Fig. 1. 2 (a) The intra-cochlear anatomy: red: scala tympani, green: scala vestibuli, cyan: spiral ganglion nerves (b) the modiolus and CI electrodes.

In recent years, our group has developed a system which we call IGCIP for image-guided cochlear implant programming. In IGCIP, we are able to obtain the spatial relationship between the cochlea and the electrode array and recommend patient-specific CI configurations for audiologists, via accurate intra-cochlear anatomy segmentation and electrode localization techniques using head computed tomography (CT) images. Though involving many meticulous technical designs, this system is used at three main stages: (1) for a CI recipient, prior to the surgery, a pre-implantation CT scan, denoted as preCT, is obtained. We segment the intra-cochlear anatomy in the preCT. (2) After the CI surgery, a post-implantation CT scan, denoted as postCT, is obtained. We localize individual contacts in the implanted electrode array in the postCT. (3) An image registration between the preCT and the postCT is done. By bringing the segmented anatomy and the localized contacts into the same coordinate system, we are able to show the spatial relationship between the neural regions and each electrode, to analyze the risk that channel interaction happens, and to customize CI settings to minimize the negative hearing artifacts it causes. In the following paragraph, more technical details on these steps are provided. Issues that hinders the large-scale evaluation and deployment of IGCIP are also identified and discussed.

In step (1), a head CT typically has a field of view (FOV) of ~200mm in each of the three spatial dimensions while the cochlea typically has sizes in the order of a few *mm*. So applying any algorithm directly to the whole image to attempt to accurately segment the intra-cochlear anatomy is difficult. To address this in the IGCIP system, (I) an image registration step is done first to localize the region of interest and (II) a deformable model is then initialized and evolves according to local intensity information to obtain accurate segmentation. Specifically in (I), we use an atlas image where the intra-cochlear anatomy is segmented. To register the atlas with the target preCT, we find the transformation which maximizes their mutual information (MI) [7], i.e., a measure that quantifies the similarity between two images. Specifically, we perform the registration sequence in a hierarchical way, i.e., from global registrations to local ones, and from affine registration to elastic registration. The parameters we estimate for the transformation field thus varies from an affine matrix, parameterized by 12 numbers, including translation (3 numbers), rotation (3 numbers), scaling (3 numbers) and skew (3 numbers), to a non-rigid deformation field parameterized by a set of Wu's compactly supported positive radial basis functions [8]. After such a sequence of registration steps, the intra-cochlear anatomy that is manually defined on the atlas is projected to the target CT via the transformation fields. In (II), we first build a deformable cochlear model that can approximate the shapes of a wide range of cochlea variations, controlled by a set of parameters, each representing the weight of one independent mode of variation. We then use the model to search an intra-cochlear anatomical shape that can best fit the target image, using local image intensity clues. There are several variants of the above ASM-based algorithm. These variants are used when no preCT is available. In these cases, the segmentation of the cochlea has to be done in the postCT and the image quality degradation caused by the electrode imaging artifacts needs to be considered. Specifically, if a patient is not scanned pre-operatively, the anatomy has to be segmented in the postCT and it is difficult due to streaking artifacts caused by the electrodes. In [9], to deal with intra-cochlear anatomy segmentation of an implanted ear of which the contralateral ear is not implanted with a CI, Reda et al. leveraged the information from the contralateral ear because it was not contaminated by the metal artifacts. A shape library-based method [10] is also proposed that leverages information from other patients' data to do postCT segmentation. Later on, Wang et al. [11] used a conditional Generative Adversarial Networks (cGANs) model to learn to generate an artifact-free image from the postCT and applied the ASM to the artificial image.

In step (2), to localize individual contacts, a registration step that is similar to the one used in the intra-cochlear anatomy segmentation step is used first to find the region of interest. In [12], Zhao et al. first extracted the centerline of the electrode array using the approach in [13] and refined it, according to the intensity and 'blobness' feature [14] of the image volume. The next step is to find the endpoints of the centerline and to use a geometric model of the electrode array which are known from the manufacturers to determine the positions of individual electrodes. However, the distance between neighboring electrodes varies depending on the manufacturers of the CIs. This method only works for CI brands of which the electrodes are closely-spaced. For those of which the electrodes are distantly spaced, the centerline cannot be reliably extracted. To solve this problem, in [15], Noble et al. proposed a method in which a set of candidate positions for electrodes are first found according to intensity information. They are then screened for the true positions according to the *a-priori* knowledge on the electrodes' spatial distribution using a graph-based method. In [16], an improved version of this method is presented.

In step (3), since the intra-cochlear anatomy and the electrode positions are both determined in the preCT space and the postCT space, the two images are registered into the same space. Characteristic frequencies are assigned to each point on the neural regions. For each electrode, the distance from it to each point on the neural regions can be computed. A distance-vs-frequency curve (DVF) can thus be plotted. By combining the DVFs of all electrodes, an expert is able to assess the risk of electrode interaction happening. By properly deactivating certain electrodes that cause interaction, such effect can be minimized to improve the patient's hearing outcomes.

1.3 Issues in IGCIP and contributions of this thesis to resolving them

Though the IGCIP technology has proven its effectiveness in improving hearing outcomes for both adult and pediatric CI recipients, several steps in the system still need human intervention. This can be a long and tedious process that requires expertise and training. Also, although the techniques that are introduced above are currently in production, we want to continue improving their accuracy and robustness by employing state-of-the-art approaches.

For each stage discussed in section 1.2, image registration plays an indispensable role in automatically locating the region of interest. Specifically, in step (1), a registration from the atlas to the target preCT is required for initializing the intra-cochlear anatomy segmentation algorithm. In step (2), a registration from the atlas to postCT is required for initializing the search for electrodes. In step (3), the preCT and postCT need to be registered to create DVFs. There are two issues in this process. First, the images that are used in the IGCIP processes can come from multiple institutions and are obtained using different acquisition protocols. Because of that they vary in quality, FOV and head orientation. For instance, the image we obtain could contain both ears, only one (left/right) inner ear, or no inner ear at all. Prior to applying the IGCIP processing pipeline, an image documentation step used to determine whether the image includes the left or right inner ear is needed and is currently done manually. After image documentation, since the registration process is an optimization problem, the parameters that need to be estimated have to be initialized. It means that, at the beginning of the registration process, the two images have to be roughly registered. Currently in the IGCIP system, this is done by manual translation and rotation in our software system. Without this step, the iterative parameter searching algorithm for images registration could be trapped in a local minimum or simply diverge to inaccurate registrations.

After initializing and running a sequence of image registrations steps, we segment the intra-cochlear anatomy using the ASM-based method. So far, this method has stable performance and in most cases, is satisfactory. However, sub-optimal results can still be observed. We are interested in exploring the state-of-the-art deep learning algorithms to improve segmentation accuracy in preCTs.

In step (3), when the spatial information of the neural regions and that of the electrodes are visualized in the DVF curves, electrode interaction effects are detected by visual checking and the CI settings are recommended. However, this process could only be done empirically by a specialist who is trained for this task. For large-scale evaluation and deployment, it needs to be automated. In our group, Zhao et al. [17] have recently developed a method for this purpose. In this method, the empirical knowledge of the specialists is quantified using a set of rules. The weight of each rule is learned from a training set. However, there are more than 10 parameters in this model. Also, the training process which takes long, has to be repeated when new examples are added to the training set. It is thus interesting to explore an alternative that is easier to train and requires fewer parameters.

Focused on resolving the issues in IGCIP that we raise above, this dissertation makes the following contributions. First, we propose a series of works including: (1) a slice-wise convolutional neural network (CNN) model that can automatically classify 3D head CT images according to the content it includes [18], and (2) an algorithm based on random forests [19,20] that can localize a series of pre-defined landmarks surrounding the ear region to initialize a registration matrix for the subsequent MI-based algorithms, and (3) a system that can simultaneously classify head CT and localize a landmark set using a multi-tasking deep neural network model (part of this work is published in [21]). Second, we propose a 2-level training scheme for a 3D convolutional neural network to segment the intra-cochlear anatomy in head CTs when only a limited set of ground truth training data is available [22]. A quantitative evaluation on CTs of 6 ear specimens shows that it is more accurate than the current ASM-based method. A large-scale qualitative evaluation also shows that it produces segmentations of better quality in the vast majority of the test images. Third, we propose a template matching-based method [23] to do automatic electrode configuration selection for CI recipients by using existing patients' electrode configuration data. Compared with the current method in use which requires the estimation of more than 10 weights, it only involves two parameters. It provides an alternative to the current method.

References

[1] A. Aschendorff, R. Kubalek, B. Turowski, F. Zanella, A. Hochmuth, M. Schumacher and R. Laszig, "Quality Control after Cochlear Implant Surgery by means of Rotational Tomography, " Otol. Neurotol., vol. 26, no. 1, pp. 34–37, 2005.

[2] C. C. Finley and M.W. Skinner, "Role of electrode placement as a contributor to variability in cochlear implant outcomes," Otol Neurotol, vol. 29, pp. 920–928, 2008.

[3] L. K. Holden, C. C. Finley, J. B. Firszt, T. A. Holden, C. Brenner, L. G. Potts and M. W. Skinner, "Factors affecting open-set word recognition in adults with cochlear implants," Ear Hear., vol. 34, no. 3, p. 342, 2013.

[4] G. B. Wanna, J. H. Noble, R. H. Gifford, M. S. Dietrich, A. D. Sweeney, D. Zhang and R. F. Labadie, "Impact of Intrascalar Electrode Location, Electrode Type, and Angular Insertion Depth on Residual Hearing in Cochlear Implant Patients: Preliminary Results," Otol Neurotol, vol. 36, pp. 1343–8, 2015.

[5] K. F. Nordfalk, K. Rasmussen, E. Hopp, R. Greisiger, and G. E. Jablonski, "Scalar position in cochlear implant surgery and outcome in residual hearing and the vestibular system," Int. J. Audiol., vol. 53, no. 2, pp. 121–127, 2014.

[6] G. B. Wanna, J. H. Noble, M. L. Carlson, R. H. Gifford, M. S. Dietrich, D. S. Haynes, B. M. Dawant and R. F. Labadie, "Impact of electrode design and surgical approach on scalar location and cochlear implant outcomes," Laryngoscope, vol. 124, no. S6, pp. S1–S7, 2014.

[7] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information." IEEE Trans. Med. Imaging, vol. 16, no. 2, pp. 187–98, 1997.

[8] Z. Wu, "Compactly supported positive definite radial functions," Adv. Comput. Math., vol. 4, no. 1, pp. 283–292, 1995.

[9] F. A. Reda, T. R. McRackan, R. F. Labadie, B. M. Dawant, and J. H. Noble, "Automatic segmentation of intra-cochlear anatomy in post-implantation CT of unilateral cochlear implant recipients," Med. Image Anal., vol. 18, no. 3, pp. 605–615, 2014.

[10] F. A. Reda, J. H. Noble, R. F. Labadie and B. M. Dawant, "An artifact-robust, shape library-based algorithm for automatic segmentation of inner ear anatomy in post-cochlear-implantation CT." In Medical Imaging 2014: Image Processing (Vol. 9034, p. 90342V). SPIE.

[11] J. Wang, J. H. Noble, and B. M. Dawant, "Conditional Generative Adversarial Networks for Metal Artifact Reduction in CT Images of the Head," in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 3-11, 2018.

[12] Y. Zhao, B. M. Dawant, R. F. Labadie, and J. H. Noble, "Automatic Localization of Cochlear Implant Electrodes in CT," in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2014, pp. 331–338.

[13] S. Bouix, K. Siddiqi, and A. Tannenbaum, "Flux driven automatic centerline extraction," Med. Image Anal., vol. 9, no. 3, pp. 209–221, 2005.

[14] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale Vessel Enhancement Filtering," in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 130–137, 1998.

[15] J. H. Noble and B. M. Dawant, "Automatic graph-based localization of cochlear implant electrodes in CT," in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 152–159, 2015.

[16] Y. Zhao, S. Chakravorti, R. F. Labadie, B. M. Dawant, and J. H. Noble, "Automatic graph-based method for localization of cochlear implant electrode arrays in clinical CT with sub-voxel accuracy," Med. Image Anal., vol. 52, pp. 1–12, 2019.

[17] Y. Zhao, B. M. Dawant, and J. H. Noble, "Automatic selection of the active electrode set for image-guided cochlear implant programming," J. Med. Imaging, vol. 3, no. 3, pp. 035001–035001, 2016.

[18] D. Zhang, J. H. Noble, and B. M. Dawant, "Automatic detection of the inner ears in head CT images using deep convolutional neural networks," in Proceedings of SPIE conference on Medical Imaging Conference, 2018, p. 10574.

[19] D. Zhang, Y. Liu, J. H. Noble, and B. M. Dawant, "Automatic localization of landmark sets in head CT images with regression forests for image registration initialization," in Proceedings of SPIE Medical Imaging Conference, 2016.

[20] D. Zhang, Y. Liu, J. H. Noble, and B. M. Dawant, "Localizing landmark sets in head CTs using random forests and a heuristic search algorithm for registration initialization," J. Med. Imaging, vol. 4, no. 4, p. 44007, 2017.

[21] D. Zhang, J. Wang, J. H. Noble, and B. M. Dawant, "Accurate Detection of Inner Ears in Head CTs Using a Deep Volume-to-Volume Regression Network with False Positive Suppression and a Shape-Based Constraint," in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 703–711, 2018.

[22] D. Zhang, R. Banalagay, J. Wang, Y. Zhao, J. H. Noble and B. M. Dawant, "Two-level Training of a 3d U-Net for Accurate Segmentation of the Intra-cochlear Anatomy in Head CT with Limited Ground Truth Training Data", In Proceedings of SPIE Medical Imaging Conference 2019.

[23] D. Zhang, Y. Zhao, J. H. Noble and B. M. Dawant. "Selecting electrode configurations for image-guided cochlear implant programming using template matching. Journal of Medical Imaging", *5*(2), 021202, 2017.

Chapter II

LOCALIZING LANDMARK SETS IN HEAD CTS USING RANDOM FORESTS AND A HEURISTIC SEARCH ALGORITHM FOR REGISTRATION INITIALIZATION

Dongqing Zhang, Yuan Liu, Jack H. Noble, Benoit M. Dawant

Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA, 37235

This chapter has been published in the Journal of Medical Imaging, 4.4 (2017): 044007

Abstract

Cochlear Implants (CIs) are electrode arrays that are surgically inserted into the cochlea to stimulate frequency-mapped nerve endings. They are used to treat patients with profound hearing loss. CIs are programmed post-operatively by audiologists using behavioral tests without information on electrode-cochlea spatial relationship. We have recently developed techniques to segment the intra-cochlear anatomy and to localize individual contacts in clinically acquired CT images. Using this information, we have proposed a new programming strategy that we call IGCIP for Image-Guided Cochlear Implant Programming and we have shown that it significantly improves hearing outcomes for both adult and pediatric recipients. One obstacle to large-scale deployment of this technique is the need for manual intervention in some processing steps. One of these is the rough registration of images prior to the use of automated intensity-based algorithms. Although seemingly simple, the heterogeneity of our image set makes this task challenging. In this article, we propose a solution that relies on the automated random forest-based localization of multiple landmarks used to estimate an initial transformation with a point-based registration method. Results shows that it produces results that are equivalent to a manual initialization. This work is an important step toward the full automation of IGCIP.

2.1 Introduction

Cochlear implants are electrode arrays that are surgically inserted into the cochlea. Individual contacts in the array stimulate frequency-mapped nerve endings, thus replacing the natural electro-mechanical transduction mechanism. Post-operatively, the CI is programmed. Key programming parameters are the assignment of a frequency range to each contact, i.e., what contact to activate when a range of frequency is present in the input signal, and the level of activation. In current clinical practice this is done by an audiologist who, blind to the position of the electrodes, relies on patients' subjective response to stimuli, e.g., whether or not they can hear a signal or rank pitches. This is a trial-and-error process that has remained essentially unchanged since the mid-80s and can be frustratingly long (dozens of programming sessions is not unusual). In the recent past, we have introduced a technique that we call IGCIP for Image-Guided Cochlear Implant Programming [1].

This technique relies on a series of image processing steps that permit the segmentation of the inner ear structures in pre-operative CT images and the localization of the implant in post-operative CT images. Using this information, we have designed techniques that assist the audiologists in programming the implants. We have shown in retrospective studies performed on long-term recipients in both adults and children that image-guided programming leads to settings that are preferred over clinical settings that were considered to be optimum [2,3]. IGCIP thus has the potential to profoundly change the way CIs are programmed. However, one barrier to the large-scale clinical deployment and evaluation of our methods is the lack of full automation. Indeed, several steps in our sequence of algorithms still require manual intervention. Specifically, the segmentation of the inner ear anatomy in pre-operative CTs relies on the registration between an atlas and pre-operative CT images. The initial segmentation is obtained by projecting manual segmentations in the atlas to pre-operative CTs using the computed transformation and is refined using an active shape model based method [4]. Pre- and post-operative scans also need to be registered for segmenting inner ear anatomy in post-operative CTs. The electrode arrays are localized using methods described in [5-9] in post-operative CTs. In addition, if a patient is not scanned pre-operatively, the atlas and the post-operative CT need to be registered and a shape

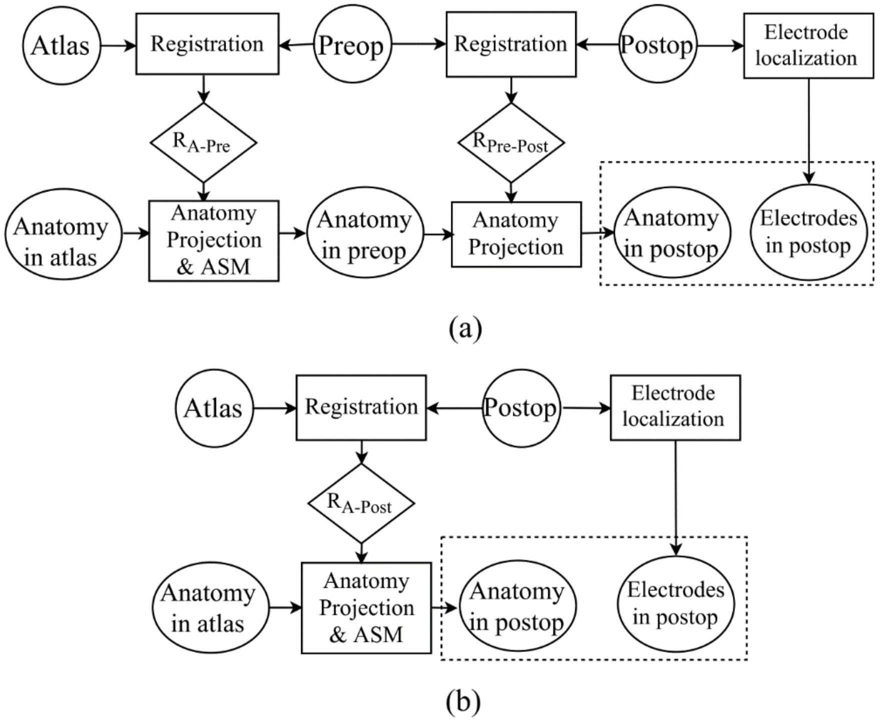

Fig. 2. 1 Pipelines used to segment the inner ear anatomy and to localize electrode arrays in postCTs when (a) the preCT scan is available and (b) when the pre-operative CT scan is not available. Circles represent CT images (Preop: preCT, Postop: postCT). Rectangles represent operations and ellipsoids represent structures of interest. $R_{A\text{-}Pre}$, $R_{A\text{-}Post}$ and $R_{Pre\text{-}Post}$ represent the transformations from atlas to pre-operative CT, from atlas to postCT and from preCT to postCT, respectively. ASM represents active shape model-based methods for segmenting inner ear anatomy given an initial segmentation projected from the atlas. The rectangles drawn with dashed lines highlight the pipeline's final outcomes

library-based algorithm is used for the segmentation of the anatomy [10]. The processing pipelines that are used to segment the inner ear anatomy and localize electrodes in post-operative CTs are shown in **Fig. 2.1**. For all atlas-to-subject registration tasks we use a sequence of algorithms ranging from rigid-body to non-rigid registrations. In all cases we rely on intensity-based techniques that use mutual information (MI) [11,12] as similarity measure. Intensity-based algorithms need to be initialized to converge and properties of our data set make this process challenging. First, the image sets often have very different fields of view. Instead of covering the whole head, they can be limited to a slab covering the left and right temporal bones, or one half of the head, or even just one temporal bone. This is because for CI surgery, the main structure of interest is the ear and coverage is limited to avoid exposing patients to unnecessary radiation. Second, the orientation of the scans can vary substantially, exceeding the capture range of traditional registration algorithms. **Fig. 2.2** shows a few illustrative examples.

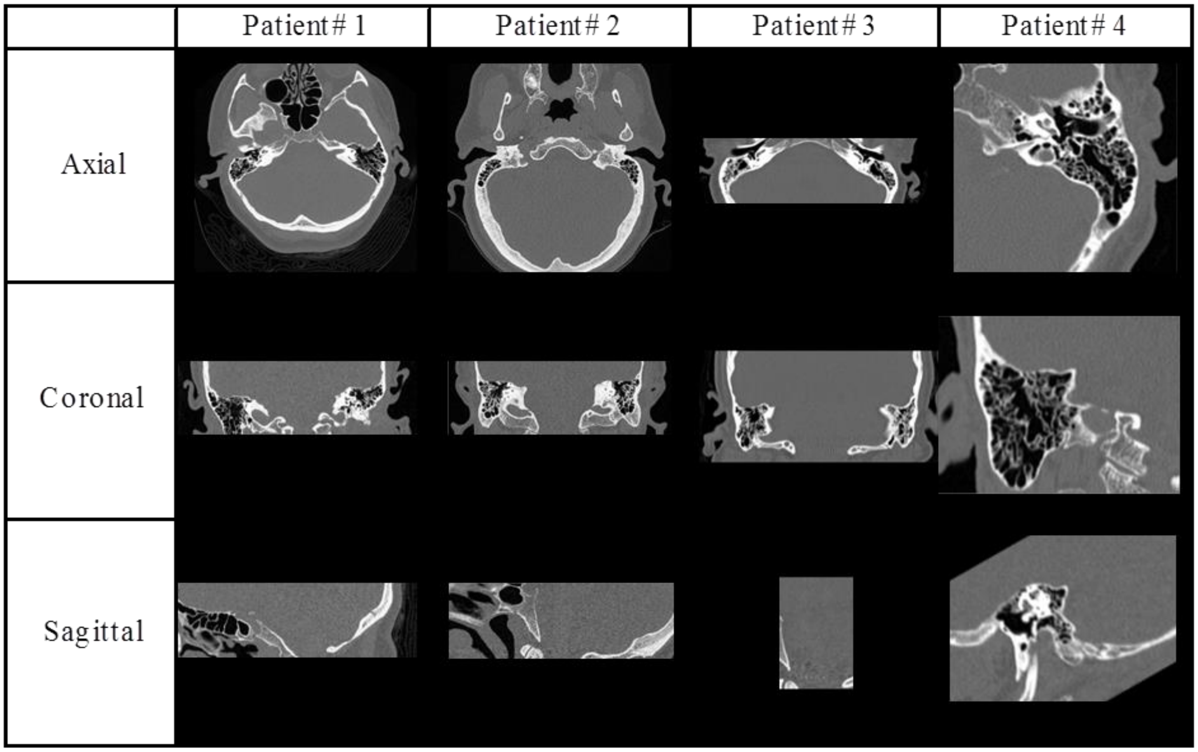

Fig. 2. 2 Transverse (top panels), coronal (middle panels), and sagittal (lower panels) view of four representative image volumes in our data set. This shows the range of coverage and orientation that can be encountered. Images of patient #3 only cover the left and right temporal bones; Images of patient #4 only covers one temporal bone.

As can be seen, patient #4's scan is unusually tilted because a large gantry angle has been used to avoid radiation to sensitive organs like the eyes. Third, in our study CT images can be acquired with a standard clinical scanner or with a Xoran xCAT® scanner, which is a low-dose flat-panel scanner. Images produced with this scanner typically have lower signal-to-noise ratio than images produced by conventional CT scanners. **Fig. 2.3** shows side-by-side one image acquired with a Xoran xCAT® scanner and one acquired with a standard scanner. Fourth, in post-operative CT volumes, the metallic implants have very high intensity and cause severe imaging artifacts. **Fig. 2.4** shows a post-operative CT image in which this is visible. All these factors make our dataset very heterogeneous and, as mentioned above, make the automatic initialization of intensity-based registration algorithms challenging. Our current approach s to initialize these algorithms

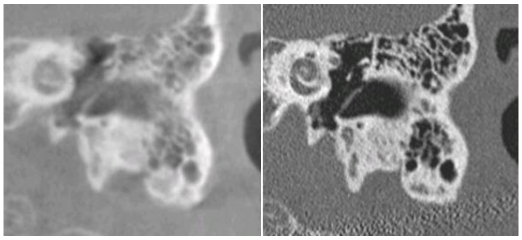

Fig. 2. 3 A comparison of images acquired with a Xoran xCAT® scanner (left) and with a conventional scanner (right). Both of them are transverse views of the ear region

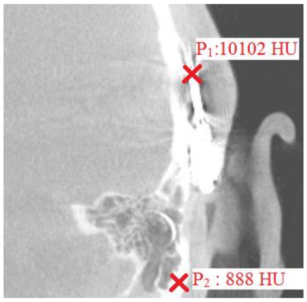

Fig. 2. 4 An example of post-operative CT image showing the imaging artifact caused by the electrode array. $P_{1,}$ which is a voxel near electrode array, has much higher intensity than $P_{2,}$ which is inside the bone

interactively either by manually rotating and translating the images or by selecting homologous points that are used to compute a rigid-body transformation. While effective, these methods necessitate loading images in a viewer and interacting with the images. This may not appear to be a major hindrance but it requires skills and experience with this type of images and with 3D image manipulation software. As a result, it cannot easily be done by an end user such as an audiologist. More importantly, it prevents the development of fully automatic pipelines that are necessary for clinical translation, which is our ultimate goal. In this article we present a method to replace this interactive steps to bring us closer to full automation.

Our method relies on the detection of landmarks and a large body of work has been dedicated to the localization of landmarks in medical images (see [13] for instance). A complete review of the literature on the topic would be beyond the scope of this article but some work that is particularly germane to what is presented herein includes the work by Potesil et al. [14] and by Donner et al. [15] in which spatial compatibility between landmarks was taken into consideration using a parts-based graphical model. To localize eight aortic valve landmarks in C-Arm CT images, Zheng et al. [16] imposed spatial constraints between these by fitting to the images a model that comprises the complete set of landmarks. It leads to an approximate position for each landmark that is then refined for each of them individually using probabilistic boosting tree classifiers. In [17], for the purpose of localizing a target point in Deep Brain Stimulation procedures using MR images, Liu et al. first automatically detected the Anterior Commissure, the Posterior Commissure and the Mid-Sagittal Plane using the method described in [18], and perform a rigid registration of all images based on the three detected structures. They then used the coordinates of the voxels as a feature along with intensity contextual features for the training and testing of a random forest-based [19] detector. Here, as others have proposed [18,20-24], we use random forest regressors to automatically localize a number of landmarks. These landmarks that have also been localized in an atlas are utilized to compute a rigid-body transformation that is used to initialize our sequence of intensity-based algorithms. Because our landmarks are used for initialization purposes, we do

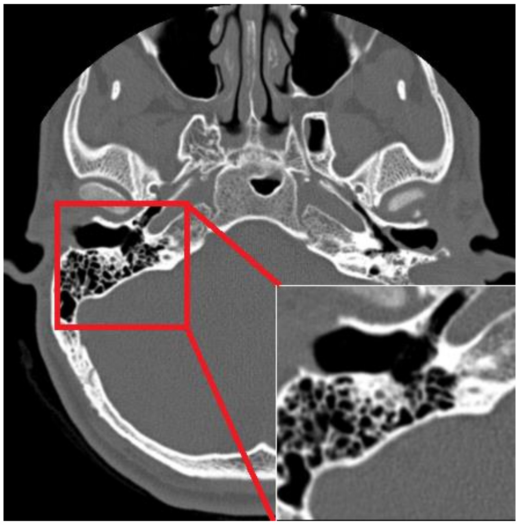

Fig. 2. 5 The red bounding box shows the region of interest in a transverse cross-section.

not need the accuracy others may require and we show that our approach is sufficient for the task at hand. But, as will be shown, our method also lacks specificity, i.e., it is not capable of unequivocally localizing each landmark. Rather, it produces several likely candidates from which the correct landmark set needs to be identified. To do so, we find among the possible candidates the set of landmarks whose spatial configuration matches the known spatial configuration of the landmarks in the atlas. Our technique is validated on the three aforementioned registration tasks (see **Fig. 2.1**), i.e, atlas to pre-operative images, atlas to post-operative images, and pre-operative to post-operative images. We use a data set that includes both conventional and xCAT® images. We show that our initialization technique leads to final results, i.e, results obtained after the intensity-based steps, that are equivalent to those obtained when initialization is performed manually, which is the procedure we currently use.

Image Type	Pre-operative CT		Post-operative CT	
	Conventional	Xoran	Conventional	Xoran
Number	228	59	16	80

Table 2. 1 Dataset used in the study

	kVp (V)	Exposure time (ms)	X-ray tube current (mA)	Slice size	In-plane resolution (mm)	Slice thickness (mm)
Conventional CT	100-140	500-7450	30-450	410×410 -768×768	0.13-0.49	0.22-2.50
Xoran CT	120	9600 / 11200	6	640×640 -812×812	0.30 / 0.40	0.30 / 0.40

Table 2. 2 Acquisition parameters for conventional and xCAT® CT images.

2.2 Methods

2.2.1 Data

Table 2.1 shows the data that have been used in this study. They consist of 383 image volumes (one ear/volume is used) that are available at the time the experiments are made. 287 of those are pre-operative images that have been acquired with a conventional scanner (228 volumes) and with an xCAT® scanner (59 image volumes). The remaining 96 volumes are post-operative images also acquired with a conventional scanner (16 image volumes) and with an xCAT® scanner (80 image volumes). The acquisition parameters of the images in our dataset are shown in **Table 2.2**. As discussed below, two sets of experiments have been performed. In the first set of experiments the training set consists of 83 pre-operative image volumes acquired with a conventional scanner; the testing set consists of 145 pre-operative volumes acquired with a conventional scanner and of 30 pre-operative volumes acquired with an xCAT® scanner. In the second set of experiments, the training set also consists of 83 pre-operative image volumes but 54 of those are acquired with a conventional scanner and 29 with an xCAT® scanner. In this set of experiments, the testing set consists of 175 pre-operative CT volumes and of 96 post-operative CT volumes. 145 of the pre-operative CT volumes are acquired with a conventional scanner and 30 with an xCAT® scanner (these are the same volumes that are used in experiment one). 16 of the post-operative images are acquired with a conventional scanner and 80 with an xCAT® scanner. All images are downsampled to $2.25 \times 2.25 \times 2.25$ mm^3 voxels both for the training of the regressors and for testing. This may appear coarse but, as the results will show, is sufficient for our main goal, i.e., the initialization of intensity-based registration algorithms.

The models are trained on the left ears. In each test image, localization of landmarks and registration are done for one ear. This is done because we have shown a high degree of symmetry between the right and the left ear [25]. When testing a volume that contains a right ear, the volumes

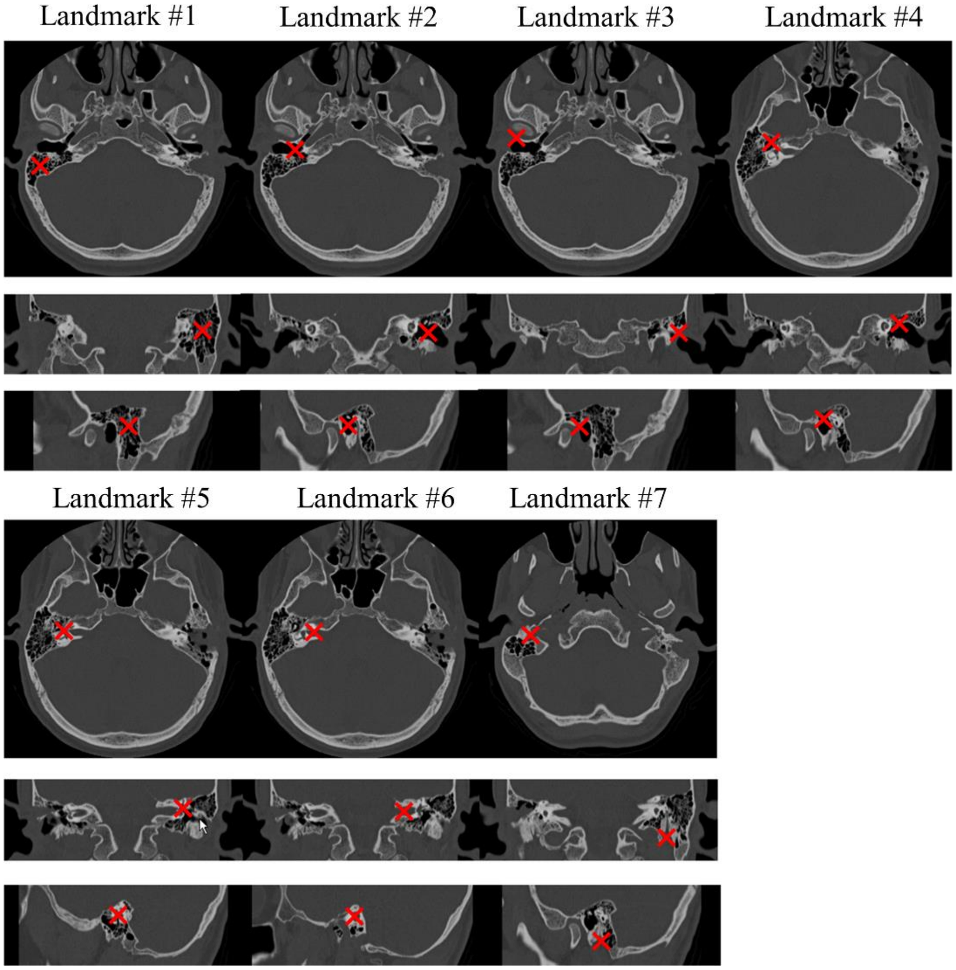

Fig. 2. 6 All seven landmarks, marked by crosses shown on transverse (top), coronal (middle) and sagittal (bottom) views

are mirrored with respect to the mid-image plane, which is typically the mid-sagittal plane of the head.

As discussed, our technique relies on the localization of landmarks. To produce good initialization results around the ear, these should surround it. They should also be visually distinguishable to permit their manual localization. The seven landmarks that meet these criteria and have been selected are shown in **Fig. 2.6**. They represent the positions of the mastoid, the external auditory canal, the spine of henle, the ossicles, the cochlear labyrinth, the internal auditory canal, and the stylomastoid foramen. Each of these has been localized manually in all volumes.

2.2.2 Random Forest-based localization of candidate landmarks

As proposed by Pauly *et al.* [26], we compute a vector of textural features for each voxel. Specifically, we apply a displacement to a voxel x, calculate the mean intensities of a 3D cuboidal region R_x^s centered on x and of a similar region $Q_x^{s,m}$ of the same size but centered on the displaced voxel, and subtract these two:

$$f^m = \frac{1}{|R_x^s|}\left(\sum_{x' \in Q_x^{s,m}} I(x') - \sum_{x' \in R_x^s} I(x')\right) \tag{2.1}$$

where I is the intensity, and s is the current scale, i.e., a particular size of the cuboidal region. Four scales are used with corresponding window sizes of 2, 4, 8, and 16. This process is repeated $M = 2000$ times (500 times at each scale) to obtain the feature set $\{f^m\}_{m=1}^{M}$. At each scale, the displacements are obtained by uniformly sampling a cube centered at voxel x. With a truncated Gaussian function, each point is assigned a probability p to be the landmark as follows,

$$p(x) = e^{-\frac{d^2}{2\sigma^2}} \tag{2.2}$$

where σ is the standard deviation of the Gaussian function and d is the Euclidian distance between the voxel and the true landmark. We truncate this function at $p = 0.1$ to limit the range of values and speed up the training process. Given a number of training pairs $\{\vec{f_n}, p_n\}_{n=1}^{N}$, random forest regressors are trained to learn a nonlinear mapping from the feature space $\{\vec{f}\}$ to the probability space $\{p\}$.

In the training phase, all the voxels in a $21 \times 21 \times 21$ cuboidal region centered on the manually selected landmark are used as training samples to create 7 models (one per landmark). We use 20 regression trees to construct the forest. For each tree, all the training samples are fed to the root node. Given the training samples $\{\overrightarrow{f_n}, p_n\}_{n=1}^{N'}$, a feature f^m and a threshold *t* are selected at each node to best split the data. This is achieved by minimizing the Mean Squared Error (MSE). i.e.,

$$t, m = \arg_{t,m} \min MSE(\{p_n : f_n^m < t\}) + MSE(\{p_n : f_n^m \geq t\}) \tag{2.3}$$

By design, when the number of samples arriving at leaf nodes is smaller than 5, or the best split threshold cannot be found or a maximum tree depth of 10 is reached, the tree stops growing. By preventing the regression trees from growing when they are deep enough or when the number of samples in one node is small enough, the risk of overfitting is reduced. Each leaf stores the mean probability of all training samples associated with it to be the landmark. When a test sample is fed to the forest, the probabilities of the leaf node to which it is associated in each tree are averaged and used to determine the probability that it is the landmark for which the forest is trained.

In the testing phase the regression forests are applied to the entire volumes to produce response maps, one for each of the seven landmarks. In these maps, the value of a voxel is its probability to be the landmark of interest. Ideally, each map would have one single maximum that corresponds to the landmark for which the forest is trained. Results will show that this is not the case. Each map typically contains multiple local maxima and the global maximum does not always correspond to the actual landmark position. To select the correct landmark set, i.e, the best landmark in each of the seven response maps, we use the known spatial relation between landmarks as discussed in the next section.

2.2.3 Selection of the landmark set using spatial constraints

To select the correct maximum in each map we use the relative position of the maxima in all the maps. To do so, we rely on the fact that (1) the landmarks we use surround the inner ear and (2) the shape of the inner ear does not change very substantially across subjects. If two sets of points in two image volumes correspond to the actual landmarks, these two sets should thus be approximately related by a rigid-body transformation. To select the correct set of landmarks for a particular volume, we first threshold each of the probability maps to eliminate spurious local maxima. The threshold is the maximum probability of the response map divided by a positive number d ($d \geq 1$) that will be discussed in section 2.3.3. All local maxima in each of the thresholded maps are then localized, which produces a set of possible positions for each landmark. Here, a local maximum is identified as a voxel with higher probability than any other voxel inside a $17 \times 17 \times 17$ cuboidal region. Next, a rigid-body transformation is computed between each landmark set configuration, i.e., one particular set of landmark candidates, and the landmark positions in a reference volume (the atlas) using a standard least squares method [27]. The landmark set that produces the configuration that leads to the smallest Fiducial Registration Error (FRE) [28], i.e., the distance after registration between homologous landmarks that have been used to estimate the transformation is selected as the solution.

To find the best configuration, the simplest and most reliable way would be to perform an exhaustive search, i.e., to generate all possible landmark configurations and test each of them. This is, however, computationally prohibitive. Indeed, suppose that for the i^{th} landmark, the number of possible positions, i.e., the number of local maxima in the response map that have been kept, is N_i (i = 1, 2,…, L). In this case, the total number of configurations and thus of point-based registrations that need to be computed is

$$N = \prod_{i=1}^{L} N_i \tag{2.4}$$

Rather than considering all possible configurations we rely on a heuristic search technique: We begin with identifying the first two landmarks by computing the Euclidean distance between all pairs of landmarks in their respective thresholded response maps. These distances are compared to the distance between the first two landmarks in the atlas. The k landmark configurations with distances closest to the reference distance are kept to form the current solution set. Next, we take the third landmark into consideration. The k configurations in the current solution set are augmented with each candidate for the third landmark. This produces a set of candidate triplet configurations. A rigid-body transformation is computed between each of these and the landmark points in the atlas. The transformation is computed using the algorithm in [25], which is a commonly used technique to estimate a rotation matrix and a translation vector to minimize the root mean squares error between homologous points. The FRE is then computed for each configuration. Because the smaller the FRE the better the landmarks can be aligned with a rigid body transformation, we use the FRE to rank the solutions and we again keep the k best ones. This procedure is repeated for all subsequent landmarks. When all seven landmarks have been included, the configuration that leads to the smallest FRE is kept as the final solution. Using this approach, the maximum number of configurations we need to evaluate is

$$N_h = N_1 N_2 + k \sum_{i=3}^{L} N_i \tag{2.5}$$

The parameter k permits to balance the robustness and the computation cost of the algorithm. When k is small, the algorithm is fast but the correct solution may be discarded in the process. When k is large, the correct solution is more likely to be found but the computational cost will increase. Even for relatively large values of k this approach is much faster than an exhaustive search.

After the landmark set is localized in the target image, it is used to compute a rigid-body transformation between the atlas and the target image.

2.2.4 Timing

The random forest regressors are trained in Python using scikit-learn [29]. The testing has been first prototyped in Matlab and Python and then implemented in C++. While it takes roughly 30 min to test each image in the prototype code (most of the time is spent on feature extraction), our C++ multithreaded implementation takes an average of 13.08 seconds per case on a standard PC (Intel (R), Xeon (R) CPU X5570, 2.93GHz, 48GB RAM, 16 cores). In the 13.08*s*, the computation of the feature vectors for all the voxels in one image requires 4.43*s*. Creation of probability maps requires 6.03*s*. Identifying the true landmarks among the possible candidates with our heuristic search algorithm requires 20*ms*. Loading image and random forest models, resampling images, and writing results to the output file require 2.6s. We also calculate the average time t_p to compute one rigid body transformation and found it to be $2.31\times10^{-6}s$. With an average number of candidates for each landmark equal to 43, the number of possible combinations is $43^7 = 2.72 \times 10^{11}$. The time to perform an exhaustive search rather than using our heuristic approach to find the best set of landmarks would thus be $6.28 \times 10^5 s = 174.4\ h$, which is prohibitive.

2.2.5 Evaluation

To test the proposed method we both measure landmark localization error by comparing our results with manual localization and assess the effect differences between manual and automatic initializations of the intensity-based algorithms have on the final results. To do the latter, we register images using three different schemes illustrated in **Fig. 2.7** In scheme one, the process is initialized with the identity transformation, i.e., we use the volumes as they are acquired. This is followed by an intensity-based step applied to the entire volume to estimate a global affine transformation. In the last step, we crop a region of interest (ROI) that contains the inner ear as shown in **Fig. 2.5** and we apply the same intensity-based algorithm we used in step two to this ROI to compute a local affine transformation. Scheme two is similar to scheme one but the process is initialized manually (this is the procedure we currently use when scheme one fails). In scheme three we use our automatic method to initialize the process. Note that because the automatic method produces a transformation that registers the ears, there is no need to estimate a global

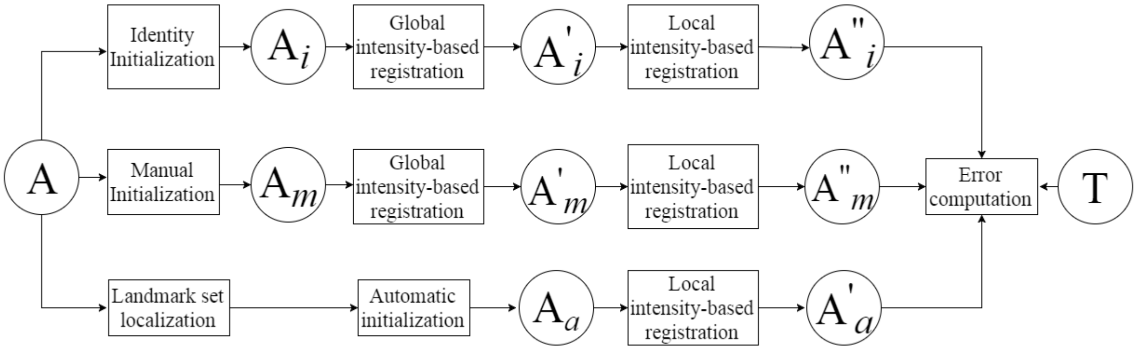
Identity Initialization
A_i
Global intensity-based registration
A'_i
Local intensity-based registration
A''_i
A
Manual Initialization
A_m
Global intensity-based registration
A'_m
Local intensity-based registration
A''_m
Error computation
T
Landmark set localization
Automatic initialization
A_a
Local intensity-based registration
A'_a

Fig. 2. 7 A flowchart of the schemes used to evaluate the proposed automatic initialization method transformation. Scheme three thus only requires a local intensity-based registration step.

With the transformations computed with each of the schemes, landmark points are projected from the atlas onto the test volumes and the position of the projected landmarks is compared to the manually selected positions for each of the landmarks in each of our test volumes. The mean Euclidean distance between projected and manually selected landmark positions is used as the quality measurement.

2.3 Results

2.3.1 Landmark set localization

Fig. 2.8 shows the response maps for the seven landmarks for one representative volume and it illustrates the multiple maxima problem. **Table 2.3** reports the landmark localization errors we have observed. In this table and in the remainder of this article a failure is a case with localization error larger than 10mm and failed cases are not included when computing error statistics. In our first set of experiments with random forests trained on images acquired with a conventional scanner, we do not observe failures when the testing set contains only pre-operative CT images of this type. It however rises to 10% when tested on pre-operative images acquired with an xCAT® scanner. A likely reason for this discrepancy is the difference in intensity characteristics between

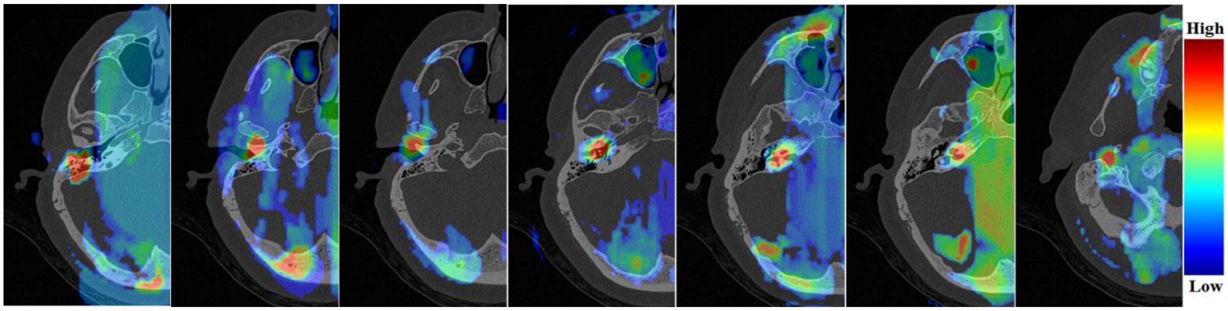

Fig. 2. 8 The axial view of the probability maps of one test CT, illustrating the multiple-maxima problem

two types of scanner. As discussed in the method section, this is addressed by modifying the training set to include images acquired both with conventional and xCAT$^{\circledR}$ scanners. With this the training set and the same testing set the failure rate drops to zero for both the conventional and the xCAT$^{\circledR}$ images. The same regressors are used to localize landmarks in post-implantation volumes acquired both with a conventional scanner and with an xCAT$^{\circledR}$ scanner. We do not observe any failure for the former. Our approach failed for only two xCAT$^{\circledR}$ volumes (2.5%), despite the presence of artifacts caused by the electrode array.

Regression forest	Model trained using conventional CTs		Robust model trained using conventional + Xoran CTs			
Dataset (# of volumes)	Conventional Pre-op (145)	Xoran Pre-op (30)	Conventional Pre-op (145)	Xoran Pre-op (30)	Conventional Post-op (16)	Xoran Post-op (80)
Mean (in mm)	3.11	3.55	3.01	4.17	4.52	4.58
SD (in mm)	0.91	0.5	0.8	1.01	1.74	1.24
Failure rate	0	10%	0	0	0	2.50%

Table 2. 3 Statistics of the mean landmark set localization error when different regressors are used for both pre-op and post-op CT images.

2.3.2 *Effect of landmark localization error on final registration*

Registration from atlas to pre-op/post-op CTs						
Dataset (# of volumes)	Conventional pre-op CT (145)			Xoran pre-op CT (30)		
Schemes	Mean (in mm)	SD (in mm)	Failure rate	Mean (in mm)	SD (in mm)	Failure rate
I	1.6	0.85	15.2%	2.75	0.74	16.70%
II	1.47	0.45	0.0	2.76	0.83	0.0
III	1.49	0.49	0.0	2.78	0.84	0.0
Dataset (# of volumes)	Conventional post-op CT (16)			Xoran post-op CT (80)		
Schemes	Mean (in mm)	SD (in mm)	Failure rate	Mean (in mm)	SD (in mm)	Failure rate
I	3.1	0.59	31.3%	2.17	0.51	61.30%
II	3.58	0.98	0.0	2.24	0.74	0.0
III	3.04	0.69	0.0	2.46	1.09	0.0
Registration from pre-op CT to post-op CT						
Dataset (# of volumes)	Conventional CT (7)			Xoran CT (76)		
Schemes	Mean (in mm)	SD (in mm)	Failure rate	Mean (in mm)	SD (in mm)	Failure rate
I	1.34	0.63	28.6%	1.36	1.25	49.70%
II	1.65	1.02	0	1.11	0.37	0
III	1.71	1.01	0	1.12	0.36	2.60%

Table 2. 4 Statistics of the registration error when no initialization (scheme I), manual initialization (scheme II), and automatic initialization (scheme III) is used for the intensity-based registration.

1) Registration between atlas and pre-operative/post-operative images

In our IGCIP technique, a registration is typically done between the atlas and a pre-operative CT. When a pre-operative CT is not available, registration between the atlas and the post-operative CT is necessary. In both scenarios, the same atlas is used and the manually labeled landmark set on the atlas and the automatically localized landmark set on the target image are utilized to compute the initial rigid-body transformation. **Table 2.4** reports differences between manually localized landmarks in each of the test volumes and the position of the atlas landmarks projected onto each of the test volume using the three strategies described in the method section. Without any initialization, we observe a high failure rate indicating that differences in orientation between the atlas and the acquired images exceed the capture range of our global intensity-based algorithm. There is no statistically significant difference between manual initialization (scheme two) and our automatic technique (scheme three) except for the conventional post-operative CT test set for which the automatic error is lower ($p < 0.05$ for a two-sided Wilcoxon signed rank test).

2) Registration between pre-operative images and post-operative images

As discussed earlier, in the normal IGCIP process pre-operative and post-operative images are available. The intra-cochlear anatomy is segmented in the pre-operative image and the electrode array is localized in the post-operative image. And both images are registered to reveal the position of the array with respect to the anatomy. Registering pre- and post-operative images involves a rigid-body transformation the computation of which may appear to be trivial. But **Table 2.4** shows that because of differences in coverage and head orientation between pre- and post-operative scans, relying only on intensity-based algorithms leads to a high failure rate. Intensity-based algorithms thus need to be initialized either manually or using the set of landmarks localized in the pre-operative and post-operative images. Here we treat the pre- and post-operative images differently to reflect what is currently done in our processing pipeline. In this pipeline the affine intensity-based registration step described above to register the atlas and the pre-operative images is followed by a non-rigid intensity-based registration step [25] that accurately registers the atlas and the pre-operative images. When evaluating our method on the pre- to post-operative registration task, the position of the landmarks is obtained by projecting them from the atlas to the pre-operative volume using the composite (rigid, affine and non-rigid) transformation that registers these volumes. We assume that through this process, landmarks are localized accurately in the pre-operative images and this is confirmed by doing a visual check. In the post-operative images, we localize the landmarks using the proposed method. We then compare the same three approaches to compute the preop-to-postop intensity-based rigid-body transformation that are used to compute the atlas-to-volume affine transformations, i.e., no initialization, manual initialization, and landmark-based initialization. The bottom section of **Table 2.4** reports the results that are obtained. In this experiment, all pre-operative images are conventional CTs. The proposed approach failed for two post-operative xCAT® volumes.

2.3.3 Sensitivity to parameter values

The performance of the heuristic search algorithm we use to localize the set of landmarks is influenced by two main parameters: (1) the value that is used to threshold the probability map to identify potential candidates, which we call T and (2) the number of solutions we keep in the solution set at each iteration, which we call *k*. The value of the threshold T is defined as follows:

$$T = \frac{M}{d}, (d \geq 1) \tag{2.6}$$

In this expression, *M* is the maximum value in the probability map and *d* is a parameter. When *d* = 1, only one landmark candidate is kept, i.e., the point at which the probability map is maximum. As *d* increases, so does the number of candidates. The value of *d* and *k* were set heuristically to 8 and 32, respectively for all the results presented in this study.

To study the sensitivity of the results to these parameter values we retrospectively repeat the localization experiments with a range of values. **Fig. 2.9** and **Fig. 2.10** show the ratio of localization failures among all cases we have observed as a function of the parameter values. These figures show that the performance of our approach reaches an asymptote for values of *d* larger than 4 and values of *k* larger than 16.

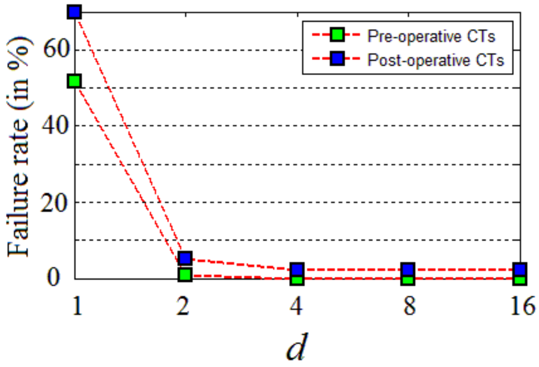

Fig. 2. 9 Localization failure rate as a function of *d*

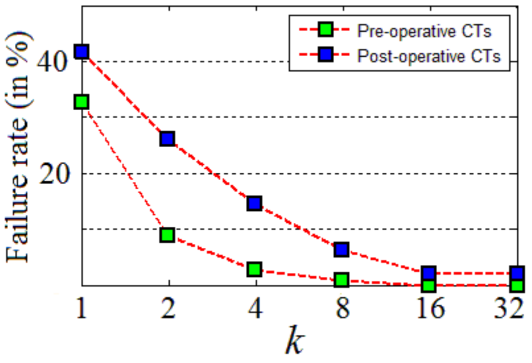

Fig. 2. 10 Localization failure rate as a function of *k*

2.4 Conclusions and discussions

In this paper, we propose a method to automatically localize a set of landmarks in head CT images to assist image registration initialization. Random forest regression is used to produce a number of candidates, which are then pruned using *a priori* information about the spatial relationship between landmarks. For the pruning process, we develop a fast and reliable heuristic technique to find the right configuration. Our technique is validated on a large heterogeneous dataset, made of both conventional and xCAT® pre-operative CTs, as well as conventional and xCAT® post-operative CTs. After the localization is done, we have shown on this dataset that the automatically localized landmark set can be used to estimate an initial transformation for registrations that could replace the manual method we currently use. We evaluate this technique by doing registrations from atlas to pre-operative CTs, from atlas to post-operative CTs, and from pre-operative CTs to post-operative CTs. This is an important step toward the full automation of our processing pipeline, which is required for the large-scale evaluation of our IGCIP technique.

References

[1] J. H. Noble, R. F. Labadie, R. H. Gifford, and B. M. Dawant, "Image-Guidance enables new methods for customizing cochlear implant stimulation strategies," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 21, no. 5, pp. 820–829, 2013.

[2] J. H. Noble, R. H. Gifford, A. J. Hedley-Williams, B. M. Dawant, and R. F. Labadie, "Clinical evaluation of an image-guided cochlear implant programming strategy," Audiol. Neurotol." vol. 19, no. 6, pp. 400–411, 2014.

[3] J. H. Noble, A. J. Hedley-Williams, L. Sunderhaus, B. M. Dawant, R. F. Labadie, S. M. Camarata and R. H. Gifford, "Initial Results With Image-guided Cochlear Implant Programming in Children, " Otol. Neurotol., vol. 37, no. 2, pp. e63–e69, 2016.

[4] J. H. Noble, R. F. Labadie, O. Majdani, and B. M. Dawant, "Automatic segmentation of intracochlear anatomy in conventional CT," IEEE Trans. Biomed. Eng., vol. 58, no. 9, pp. 2625–2632, 2011.

[5] J. H. Noble and B. M. Dawant, "Automatic graph-based localization of cochlear implant electrodes in CT," Int. Conf. Med. Image Comput. Comput. Interv., pp. 152–159, 2015.

[6] Y. Zhao, B. M. Dawant, R. F. Labadie, and J. H. Noble, "Automatic Localization of Cochlear Implant Electrodes in CT," Int. Conf. Med. Image Comput. Comput. Interv., 2014, pp. 331–338.

[7] Y. Zhao, B. M. Dawant, and J. H. Noble, "Automatic localization of cochlear implant electrodes in CTs with a limited intensity range," SPIE Med. Imaging. Int. Soc. Opt. Photonics, vol. 10133, 2017.

[8] Y. Zhao, B. M. Dawant, R. F. Labadie, and J. H. Noble, "Automatic localization of closely-spaced cochlear implant electrode arrays in clinical CTs," Med. Phys., vol. 45, no. 11, pp. 5030–5040, 2018.

[9] Y. Zhao, S. Chakravorti, R. F. Labadie, B. M. Dawant, and J. H. Noble, "Automatic graph-based method for localization of cochlear implant electrode arrays in clinical CT with sub-voxel accuracy," Med. Image Anal., vol. 52, pp. 1–12, 2019.

[10] F. A. Reda, J. H. Noble, R. F. Labadie, and B. M. Dawant, "An artifact-robust, shape library-based algorithm for automatic segmentation of inner ear anatomy in post-cochlear-

implantation CT," in Proceedings of SPIE Medical Imaging Conference, 2014, no. March 2014, p. 90342V.

[11] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information.," IEEE Trans. Med. Imaging, vol. 16, no. 2, pp. 187–98, 1997.

[12] G. K. Rohde, A. Aldroubi and B. M. Dawant, "The adaptive bases algorithm for intensity-based nonrigid image registration," IEEE Trans. Med. Imaging, vol. 22, no. 11, pp. 1470–1479, 2003.

[13] K. Rohr, "Landmark-based image analysis: using geometric and intensity models," vol. 21. 2001.

[14] V. Potesil, T. Kadir, G. Platsch, and M. Brady, "Improved Anatomical Landmark Localization in Medical Images Using Dense Matching of Graphical Models," in British Machine Vision Conference (BMVC), 2010, no. January 2010.

[15] R. Donner, B. H. Menze, H. Bischof, and G. Langs, "Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization," Med. Image Anal., vol. 17, no. 8, pp. 1304–1314, 2013.

[16] Y. Zheng, M. John, R. Liao, A. Nottling, J. Boese, J. Kempfert, G. Brockmann, and D. Comaniciu, "Automatic aorta segmentation and valve landmark detection in C-Arm CT for transcatheter aortic valve implantation," IEEE Trans. Med. Imaging, vol. 31, no. 12, pp. 2307–2321, 2012.

[17] Y. Liu and B. Dawant, "Multi-modal Learning-based Pre-operative Targeting in Deep Brain Stimulation Procedures," in IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2016, pp. 17–20.

[18] Y. Liu and B. M. Dawant, "Automatic Localization of the Anterior Commissure, Posterior Commissure, and Midsagittal Plane in MRI Scans using Regression Forests," IEEE J. Biomed. Heal. Informatics, vol. 19, no. 4, pp. 1362–1374, 2015.

[19] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[20] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui, "Regression forests for efficient anatomy detection and localization in computed tomography scans," Med. Image Anal., vol. 17, no. 8, pp. 1293–1303, 2013.

[21] M. A. Dabbah, S. Murphy, H. Pello, R. Courbon, E. Beveridge, S. Wiseman, D. Wyeth, and I. Poole, "Detection and location of 127 anatomical landmarks in diverse CT datasets," in Proceedings of SPIE Medical Imaging Conference, 2014, no. March 2014.

[22] D. Han, Y. Gao, G. Wu, P. T. Yap, and D. Shen, "Robust anatomical landmark detection with application to MR brain image registration," Comput. Med. Imaging Graph., vol. 46, pp. 277–290, 2015.

[23] C. Lindner, C. W. Wang, C. T. Huang, C. H. Li, S. W. Chang, and T. F. Cootes, "Fully Automatic System for Accurate Localisation and Analysis of Cephalometric Landmarks in Lateral Cephalograms," Sci. Rep., vol. 6, no. Sept, 2016.

[24] T. Ebner, D. Stern, R. Donner, H. Bischof, and M. Urschler, "Towards automatic bone age estimation from MRI: Localization of 3D anatomical landmarks," Int. Conf. Med. Image Comput. Comput. Interv, 2014.

[25] F. A. Reda, T. R. McRackan, R. F. Labadie, B. M. Dawant, and J. H. Noble, "Automatic segmentation of intra-cochlear anatomy in post-implantation CT of unilateral cochlear implant recipients," Med. Image Anal., vol. 18, no. 3, pp. 605–615, 2014.

[26] O. Pauly, B. Glocker, A. Criminisi, D. Mateus, A. M. Moller, S. Nekolla, and N. Navab, "Fast multiple organ detection and localization in whole-body MR dixon sequences," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6893 LNCS, no. PART 3, pp. 239–247, 2011.

[27] J. Farrell, "Problem 65-1: A least squares estimate of satelite attitute," SIAM Rev., vol. 8, pp. 384–386, 1966.

[28] J. M. Fitzpatrick, J. B. West, and C. R. Maurer, "Predicting error in rigid-body point-based registration." IEEE Trans. Med. Imaging, vol. 17, no. 5, pp. 694–702, 1998.

[29] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2012.

Chapter III

AUTOMATIC DETECTION OF INNER EARS IN HEAD CTS USING CONVOLUTIONAL NEURAL NETWORKS

Dongqing Zhang, Jack H. Noble, Benoit M. Dawant

Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA, 37235

This chapter has been published in the Proceedings of the 2016 SPIE Medical Imaging Conference, vol 1057427

Abstract

Cochlear implants (CIs) use electrode arrays that are surgically inserted into the cochlea to stimulate nerve endings to replace the natural electro-mechanical transduction mechanism and to restore hearing for patients with profound hearing loss. Post-operatively, the CI needs to be programmed. Traditionally, this is done by an audiologist who is blind to the positions of the electrodes relative to the cochlea and relies on the patient's subjective response to stimuli. This is a trial-and-error process that can be frustratingly long (dozens of programming sessions are not unusual). To assist audiologists, we have proposed what we call IGCIP for image-guided cochlear implant programming. In IGCIP, we use image processing algorithms to segment the intra-cochlear anatomy in pre-operative CT images and to localize the electrode arrays in post-operative CTs. We have shown that programming strategies informed by image-derived information significantly improve hearing outcomes for both adults and pediatric populations. We are now aiming at deploying these techniques clinically, which requires full automation. One challenge we face is the lack of standard image acquisition protocols. The content of the image volumes we need to process thus varies greatly and visual inspection and labelling is currently required to initialize processing pipelines. In this work we propose a deep learning-based approach to automatically detect if a head CT volume contains two ears, one ear, or no ear. Our approach has been tested on a data set that contains over 2,000 CT volumes from 153 patients and we achieve an overall 95.97% classification accuracy.

3.1 Introduction

Cochlear implants (CIs) have been one of the most successful prosthetics in the past decades [1]. With a CI, an array of electrodes that is surgically inserted into the cochlea is used to stimulate auditory nerve endings, thus replacing the natural electro-mechanical transduction mechanism and restoring hearing for patients with profound hearing loss. Post-operatively, CIs need to be programmed to tune the implant for each recipient, e.g., to assign a frequency range to individual contacts such that they are activated when a frequency within that range is detected in the input signal and to adjust activation levels. In clinical practice, this is done by an audiologist who is blind to the positions of the electrodes relative to the cochlea and relies on the subjective patients' response to stimuli, e.g., whether they can hear a signal or rank pitches. This is a trial-and-error process that has remained essentially unchanged since the mid-80s and can be frustratingly long (dozens of programming sessions are not unusual). In recent years, we have introduced what we call IGCIP for image-guided cochlear implant programming [2]. In IGCIP, we use image processing algorithms to segment the intra-cochlear anatomy in pre-operative CT images and to localize the electrode array in post-operative CTs [3-8]. Using this information, we have designed techniques to recommend CI processor settings to assist audiologists in programming the implants. We have shown that this leads to improvement in patient outcomes [9-11].

Our long-term objective is to automate the series of image processing steps that support IGCIP to permit its clinical deployment. One barrier to full automation is the lack of standardized image acquisition protocols. In Chapter II, we have developed a random forest-based method to detect a set of landmarks around the inner ear. However, that study is based on a screened dataset, in which each image contains the inner ear region. In reality, our dataset is more heterogeneous. Images could include very different portions of the head. They can include both inner ears, one inner ear (left or right), or sometimes neither. **Fig. 3.1** shows 4 examples to illustrate the range of images we need to be able to process. CT #1 includes the whole head so both ears are visible. In CT #2, although both the full right half and a fairly large portion of the left half of the head are included, only the right inner ear is visible. CT #3 includes only a very narrow portion of the head, but the whole right inner ear is visible. CT #4 includes only an anterior portion of the head and neither inner ear is visible.

In our current IGCIP process, when a new volume is received, prior to all registration steps, it is visually inspected and assigned a label to document its content for proper processing in subsequent steps. In this work, we aim to replace this visual inspection step.

3.2 Data and methods

In recent years, convolutional neural networks (CNN or ConvNets) have been proposed as a solution for a wide range of problems such as image classification, object detection, semantic segmentation and other high-level computer vision tasks. The impressive performance they have achieved makes them the preferred solution for an increasing number of applications (see [12-15] for representative examples). CNN designed specifically for detection tasks include R-CNN [16], fast R-CNN [17], SPP-net [18], faster R-CNN [19] and YOLO [20]. These networks permit detecting the presence of a set of pre-defined objects and localizing each of them with a bounding box in 2D images. Extending these algorithms to 3D data sets requires substantially more resources in terms of hardware, model complexity, training data and training time but solutions to this problem have been proposed. Work that is particularly relevant to or own work is presented in [21, 22] in which authors use a 2D CNN to detect whether the anatomical structure of interest is present using slices extracted from axial, coronal and sagittal views of the 3D CT volumes. The algorithm is validated in three different CT datasets [22]. Similarly, Mamani et al. [23] used 2D multi-label convolutional neural networks for each orthogonal view of the thorax-abdomen CT scans for the detection of four human organs. In the work described herein, we propose to use axial slices of the head CTs to train a CNN to determine whether a new head CT volume includes one or both inner ears, to facilitate automating our IGCIP pipeline.

The data set that we used in this study consists of 1,593 CT volumes obtained from 322 patients. We have more volumes than patients because it is common that multiple acquisitions are performed and that multiple reconstructions, and thus volumes, are produced from the same acquisition. For each patient, both the pre-operative and the post-operative CTs are included. The CTs are acquired with several scanners, including conventional and Xoran scanners. The volumes in our data set also cover regions of very different sizes, ranging from 10mm to 256mm in the left-right and anterior-posterior dimensions, and from 52mm to 195mm in the superior-inferior dimension. The voxel dimensions range from 0.14mm to 2.00mm in left-right and anterior-posterior directions and from 0.14 mm to 2.50 mm in superior-inferior direction. In **Fig. 3.1**, we show four examples to illustrate the range of images we need to be able to process. CT #1 includes the whole head so both ears are visible. In CT #2, although both the full right half and a fairly large portion of the left half of the head are included, only the right inner ear is visible. CT #3 includes only a very narrow portion of the head, but the whole right inner ear is visible. CT #4 includes only an anterior portion of the head so neither inner ear is visible. We use CTs of half of the patients to train the CNN and to select parameters for 3-D volume classification and CTs of the other half

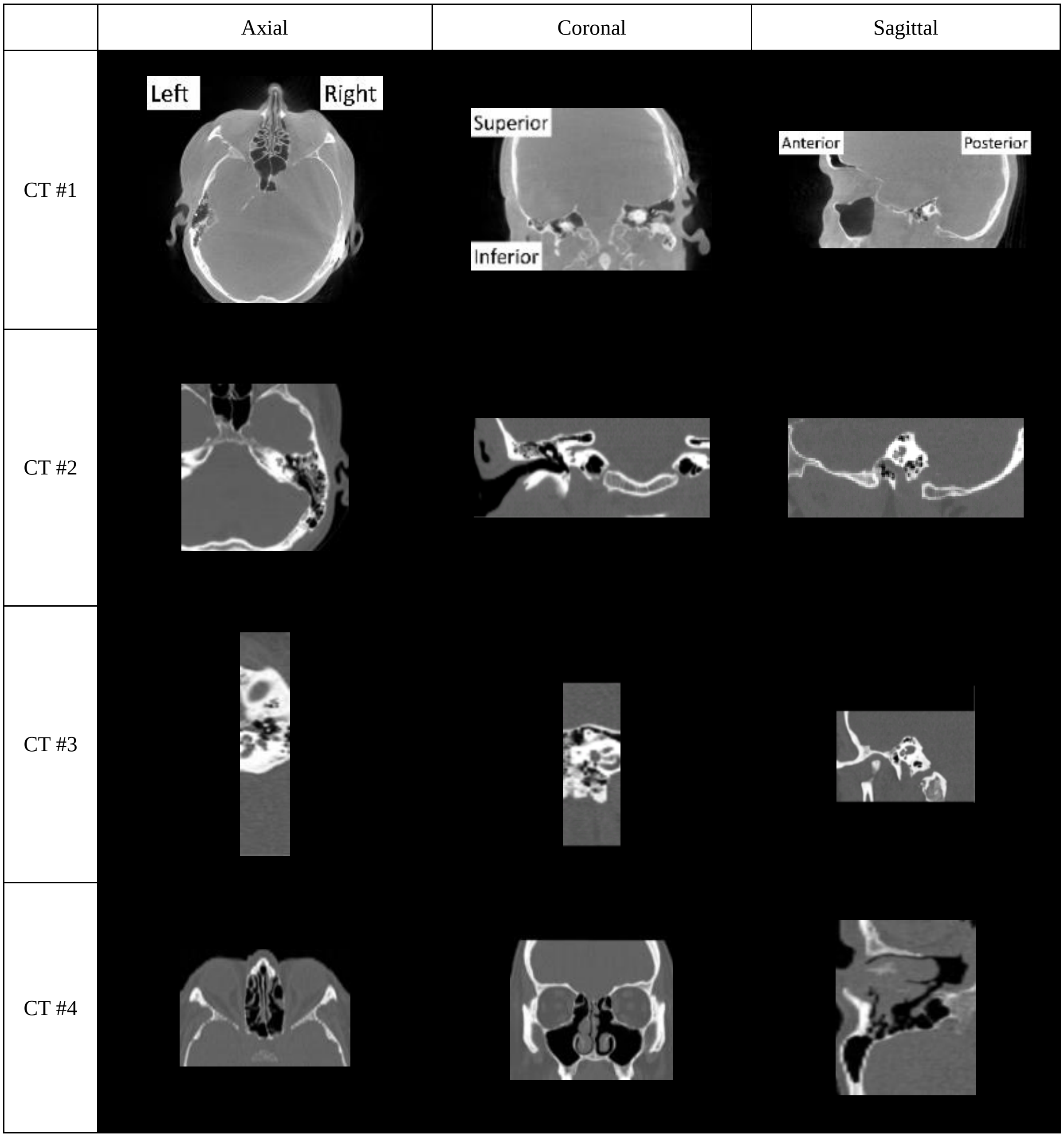

Fig. 3. 1 Four examples from our dataset. Orientations of the slices are labeled in CT #1 and apply to other examples as well.

Group	Training	Validation	Testing
Number of patients	120	49	153
Number of original CTs	563	235	795
Number of CTs after adding artificial CTs	563	732	2484

Table 3. 1 Numbers of patients and CT volumes in each group

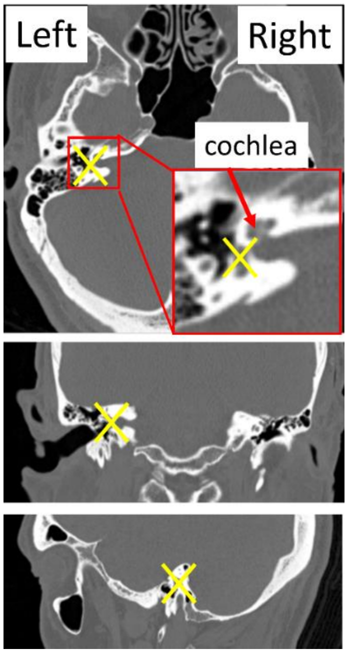

Fig. 3. 2 The yellow marker is the landmark we use as the position of the left inner ear. From top to bottom, they are the axial, coronal and sagittal views, respectively

of the patients to do testing. Specifically the first half is split into a training set and a validation set, for training the CNN model and for optimizing parameters of volume-wise classification, respectively.

We first resample all CT volumes to 0.8×0.8×0.8 mm^3 per voxel using trilinear interpolation. All CT volumes are visually checked and are assigned to one of four categories: category 1, no ear; category 2, both ears; category 3, only the right ear; and category 4, only the left ear. As we have mentioned, we split the image volumes into (1) a training set, (2) a validation set and (3) a test set. The number of patients and number of CT volumes in each set are shown in the second and third rows of **Table 3.1**. Unfortunately, the data set we currently have at our disposal is very unbalanced in terms of the content. About 80% of the image volumes include both ears and about 20% include a single left or right ear. Image volumes which do not include any ear do exist but are very rare. If we build a machine learning system and search for the best parameters to maximize the overall accuracy using unbalanced training and validation sets, the optimal setting will tend to classify all images into the majority class. To tackle this problem, we need to balance the number of samples in the four categories. To do so, we cropped the original CT volumes in the validation set to make more image volumes that include a single left ear, or right ear, or no ear and add them back to make the validation set have roughly equal numbers of volumes from each category. The same balancing operation is done for the test set. Since we use 2-D slices to train the network, in the training set, we only need to make sure the number of slices that we sampled, instead of the number of CT volumes is the same for each of the four categories. No artificial data thus needs to be added to balance the training set. After adding the artificial CT volumes, the total numbers of CT volumes in each set are shown in the fourth row of **Table 3.1**. For each image volume in the training set, we manually localize the inner ears. This is done by selecting one point around the cochlea, as shown in **Fig. 3.2**. As we have mentioned, the images are obtained with different protocols. This results in different intensity ranges. We normalize each image's intensity to a uniform range, i.e., (0, 1).

At the current stage of the work, we assume that we know the orientation of the volume and we base our approach on axial images. To train the network, we use slices in the training volumes and we assign each slice to one of the four previously mentioned categories, i.e., category 1, no ear; category 2, both ears; category 3, only the right ear; and category 4, only the left ear. A slice is assigned to category 1 if either the CT volume it belongs to is in category 1, or if the CT volume belongs to categories 2-4 but the distances between the ears and the slice are larger than a threshold d_t. Here, we empirically choose d_t = 10mm. A slice is assigned to category 2 if it comes from a CT volume in category 2 and the distances between its ears and the slice are less than d_t. A slice is assigned to category 3 if it comes from a CT volume in category 3 and the distance between its ear's distance and the slice is less than d_t. Finally, a slice is assigned to category 4 if it comes from a volume in category 4 and the distance between its ear and the slice is less than d_t. We augment the training volumes by applying reasonable translations, scaling and rotations to existing CT volumes and extract additional slices from them. By doing data augmentation, we have generated 100,000 slices from the training CT volumes to train the network. Because the size of the regions covered by the images varies from volume to volume, resampling to isotropic pixels leads to slices with different number of pixels, which cannot be accommodated by the network we use. To address this issue we symmetrically crop or pad the slices to make them have 224 × 224 pixels which is the size of the network's input layer.

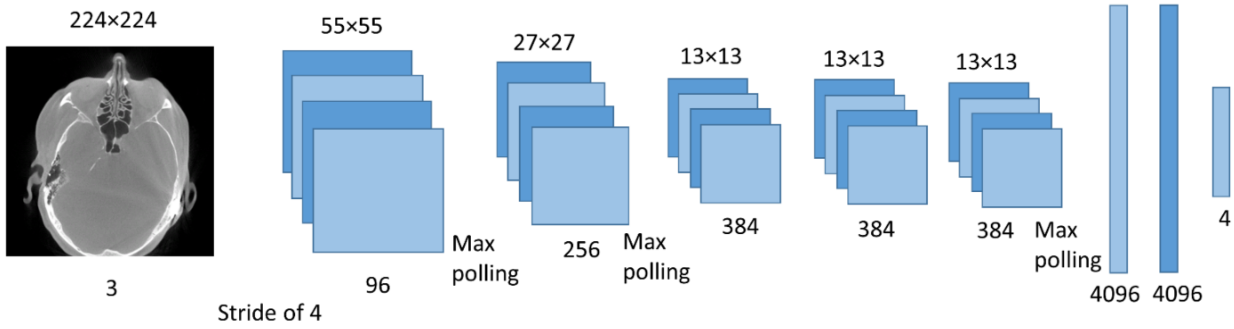

Fig. 3. 3 Architecture of the AlexNet

Here, we use the AlexNet [15] that is pre-trained on the ImageNet data set. **Fig. 3.3** shows the architecture of this network (more details can be found in [15]). It has five convolutional layers and three fully-connected layers. At each convolutional layer, multiple filters are used for convolution with the input. The output feature maps are shown in the figure as stacked squares. The number of feature maps obtained after each layer is shown at the bottom of the feature maps. The size of the feature maps is shown at the top. Following convolution, max pooling is applied to reduce the dimensionality of the feature space. Finally, a non-linear activation function, here a rectified linear unit (ReLU) is applied to the feature maps. The following fully-connected layers are the same as layers used in traditional artificial neural networks. A Softmax function is applied to the output of the third fully-connected layer to probabilities which sum to 1. In the AlexNet architecture, the size of the output layer is 1000.

To adapt the architecture to our needs, we change the size of the output layer from 1000 to 4 and we re-initialize the weights of the last fully-connected layer. Since the first layers of the pre-trained CNN are generic feature extractors, they do not need substantial update. We thus fine-tune the CNN by keeping the learning rate of the first 7 layers 1/10 of that of the last layer. We use the categorical cross-entropy between ground truth labels and the output as the loss function and minimize it. The network is trained using stochastic gradient descent using a batch size of 256. We adopt the simplified learning rate adjustment strategy in the original AlexNet paper. The initial learning rate of the last layer is 0.01 and gets 10 times smaller after each 10,000 iterations. AlexNet is designed for RGB images. Here we test two strategies for generating 3-channel inputs: (1) for each position, we simply use three copies of the same axial slice at this position, one per channel. (2) For each position, besides the slice at this position, we also extract the slice that is above it and the slice below it to constitute the 3-channel input. By using neighboring slices, we are able to capture more spatial information.

When using the trained network to label a new volume, we preprocess it the way we do for image volumes in the training set, i.e., we resample it, and crop or zero-pad it as required. Slices at each position are then input to the CNN to obtain the probabilities that it belongs to each of the four categories. Suppose the number of slices in the test volume is q, the output we produce is a q×4 matrix $[\boldsymbol{p}^n, \boldsymbol{p}^b, \boldsymbol{p}^r, \boldsymbol{p}^l]$. Here, the four column vectors of dimensionality q, i.e., $\boldsymbol{p}^n, \boldsymbol{p}^b, \boldsymbol{p}^r$ and $\boldsymbol{p}^l$ represent the probabilities that the slices belong to the “no ear”, “both ears”, "right ear" and "left ear" categories, respectively. Each row in the matrix represents the probabilities of the corresponding slice in the volume. **Fig. 3.4** shows four representative examples. The images show coronal views of four CT volumes. The two four-curve groups on the right show $\boldsymbol{p}^n, \boldsymbol{p}^b, \boldsymbol{p}^r$ and $\boldsymbol{p}^l$, respectively. The group on the left is produced by the model using input generating strategy (1) and the group on the right is produced by the model using input generating strategy (2). The *x*-axis is the probability. The *y*-axis is the slice number. The images and the plots have been aligned to help relating the content of the image and the curves. In the example shown in (a), the image volume covers the right ear. The probability curves show that for those slices close to the inner ear, the "right ear" probability is nearly 1. For other slices, the "no ear" probability is nearly 1. The CT volume in (b) covers the whole head. The probability curves show that for those slices close to the inner ear, the "both ears" probability is nearly 1. For other slices, the "no ear" probability is nearly 1. (c) & (d) show two examples in which the probability curves are not as neat as those in (a) & (b). The CT volume in (c) contains only the left ear. The probabilities of the slices being "left ear" are higher but the values are not close to 1 and the number of consecutive slices having high "left ear" probabilities are fewer compared to that in (a) & (b). We can see a similar phenomenon in (d), in which the CT contains only the right ear. This could be due to the visually noticeable image noise in (c) and imaging artifact in (d). However, in both (c) and (d), the overall responses at the ground-truth channel produced by the model using input generating strategy (2) are stronger than those produced by the model using input generating strategy (1). This could be attributed to the incorporation of the extra inter-slice information.

The last step is to assign each volume to a category based on the probability curves. A straightforward criterion, which we currently use, is to find the class *c* (*c* = *l*, *r* or *b*) such that there exist a threshold probability p_t and k consecutive indices i, i+1,…, i+k-1, such that, $min\{p_i^c, p_{i+1}^c, \ldots, p_{i+k-1}^c\} \geq p_t$. If there is no such c, we predict that the volume does not include

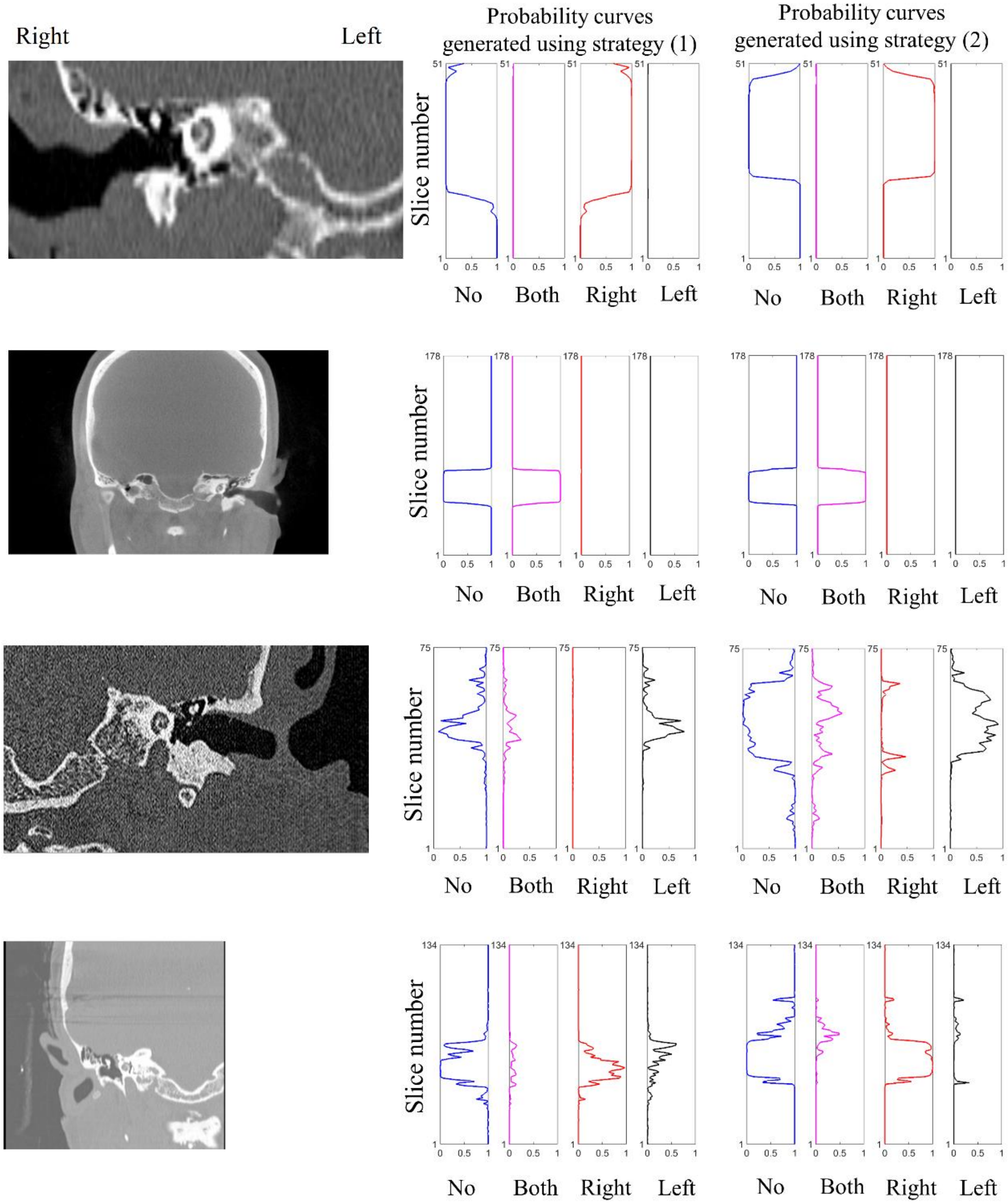

Fig. 3. 4 4 examples, (a)-(d). For each example, the image on the left is a coronal slice of the CT. The two plots represent the probabilities of the slice series containing no ear, both ears, right ear and left ear, generated by models using strategy (1) and (2).

any ear. If there are multiple cs, we choose the category c_{opt} for which the probability curve has the maximal average value. The performance of our algorithm depends on the value of k and $p_t$and, to find the optimal values of them, we do a grid search in the validation set. The optimal values for k and p_t are: k = 3 and p_t= 0.56 for the model trained using input generating strategy (1) and k = 4 and p_t= 0.63 for the model trained using input generating strategy (2).

	Predicted: no ear	Predicted: both ears	Predicted: right ear	Predicted: left ear
Actual: no ear	171	1	4	7
Actual: both ears	0	183	0	0
Actual: right ear	6	2	167	8
Actual: left ear	2	2	2	177

Table 3. 2 Detection results in the validation set produced by the model using input generating strategy (1)

	Predicted: no ear	Predicted: both ears	Predicted: right ear	Predicted: left ear
Actual: no ear	177	1	2	3
Actual: both ears	0	183	0	0
Actual: right ear	6	1	176	0
Actual: left ear	3	11	1	168

Table 3. 3 Detection results in the validation set produced by the model using input generating strategy (2)

	Predicted: no ear	Predicted: both ears	Predicted: right ear	Predicted: left ear
Actual: no ear	579	10	14	18
Actual: both ears	1	619	0	1
Actual: right ear	27	14	554	26
Actual: left ear	17	11	3	590

Table 3. 4 Detection results in the test set produced by the model using input generating strategy (1)

	Predicted: no ear	Predicted: both ears	Predicted: right ear	Predicted: left ear
Actual: no ear	596	9	5	11
Actual: both ears	2	617	0	2
Actual: right ear	12	13	595	1
Actual: left ear	14	28	3	576

Table 3. 5 Detection results in the test set produced by the model using input generating strategy (2)

3.3 Results

In the validation set, the classification error rates for models trained using strategy (1) and (2) are 4.64% and 3.83%, respectively. **Table 3.2 & 3.3** are the results we have obtained with our validation set under the optimal k & p_t settings, when the input generating strategy (1) and (2) are used, respectively. Similarly, **Table 3.4 & 3.5** show the results we obtain with the test set when input generating strategy (1) and (2) are used, respectively. In the test set, using input generating strategy (1), we achieve an overall labelling accuracy of 94.28%. Using input generating strategy (2), we achieve an overall labelling accuracy of 95.97%. The detection accuracy when using strategy (2) is thus slightly higher than that of using strategy (1).

3.4 Conclusions

Automatic labelling of head CT images with CNNs appears achievable. So far we have tested our approach on 2,484 image volumes and we reach a very encouraging success rate. We achieve a higher accuracy when using three consecutive slices as three input channels to the CNN, compared to that when we replicate a single slice twice. This improvement could be attributed to the consideration of inter-slice information.

References

[1]. NIDCD Fact Sheet: Cochlear Implants, "National institute on deafness and other communication disorders," NIH Publication No. 11-4798, https://www.nidcd.nih.gov/sites/default/files/Documents/health/hearing/FactSheetCochlearImplant.pdf (2011).

[2]. J. H. Noble, R. F. Labadie, R. H. Gifford, B. M. Dawant. "Image-guidance enables new methods for customizing cochlear implant stimulation strategies." IEEE Trans Neural. Syst. Rehabil. Eng. 21.5 (2013): 820-829.

[3]. J. H. Noble and B. M. Dawant, "Automatic graph-based localization of cochlear implant electrodes in CT," Med Image Comput Comput Assist Interv, vol. 9350, pp. 152-159, Oct 2015.

[4]. F. A. Reda, T. R. McRackan, R. F. Labadie, B. M. Dawant, and J. H. Noble, "Automatic segmentation of intra-cochlear anatomy in post-implantation CT of unilateral cochlear implant recipients," Med Image Anal, vol. 18, pp. 605-615, Apr 2014.

[5]. F. A. Reda, J. H. Noble, R. F. Labadie, and B. M. Dawant. "An artifact-robust, shape library-based algorithm for automatic segmentation of inner ear anatomy in post-cochlear-implantation CT." In Proceedings of SPIE Medical Imaging Conference, 2014.

[6]. J. H. Noble, R. F. Labadie, O. Majdani, and B. M. Dawant, "Automatic segmentation of intracochlear anatomy in conventional CT," IEEE Trans Biomed Eng, 58(9), pp. 2625-2632, Sep 2011.

[7]. Y. Zhao, B. M. Dawant, R. F. Labadie, and J. H. Noble, "Automatic localization of cochlear implant electrodes in CT," Med Image Comput Comput Assist Interv, vol. 17, pp. 331-338, 2014.

[8]. Y. Zhao, B. M. Dawant, and J. H. Noble, "Automatic localization of cochlear implant electrodes in CTs with a limited intensity range", Proc. SPIE 10133, Medical Imaging 2017: Image Processing, 101330T (February 24, 2017)

[9]. J. H. Noble, R. H. Gifford, A. J. Hedley-Williams, B. M. Dawant. "Clinical evaluation of an image-guided cochlear implant programming strategy." Audiology and Neurotology 19.6 (2014): 400-411

[10]. J. H. Noble, A. J. Hedley-Williams, L. Sunderhaus, B. M. Dawant, R. F. Labadie, S. M. Camarata, and R. H. Gifford. "Initial results with image-guided cochlear implant programming in children." Otology & neurotology: 37.2 (2016): e63.

[11]. R. F. Labadie, J. H. Noble, A. J. Hedley-Williams, L. W. Sunderhaus, B. M. Dawant, and R. H. Gifford, "Results of Postoperative, CT-based, Electrode Deactivation on Hearing in Prelingually Deafened Adult Cochlear Implant Recipients," Otol Neurotol, vol. 37, pp. 137-45, Feb 2016.

[12]. A. Krizhevsky, I. Sutskever, and G.E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012. (pp. 1097-1105).

[13]. K. Simonyan, and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[14]. K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. (pp. 770-778).

[15]. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. "Densely connected convolutional networks." arXiv preprint arXiv:1608.06993 (2016).

[16]. R. Girshick, J. Donahue, T. Darrell and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. (pp. 580-587)

[17]. K. He, X. Zhang, S. Ren, and J. Sun. "Spatial pyramid pooling in deep convolutional networks for visual recognition." European Conference on Computer Vision. Springer, Cham, 2014. (pp. 346-361)

[18]. R. Girshick. "Fast r-cnn." arXiv preprint arXiv:1504.08083 (2015).

Chapter IV

HEADLOCNET: DEEP 3D CONVOLUTIONAL NEURAL NETWORKS FOR ACCURATE CLASSIFICATION AND MULTI-LANDMARK LOCALIZATION OF HEAD CTS

Dongqing Zhang, Jianing Wang, Jack H. Noble, Benoit M. Dawant

Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA, 37235

Part of this chapter has been published in the Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention 2018, Lecture Notes in Computer Science, vol 11073, pp 703-711.

Abstract

Cochlear implants (CIs) are used to treat patients with hearing loss. In a CI surgery, an electrode array is inserted into the cochlea to stimulate auditory nerves. After surgery, CIs need to be programmed. Studies have shown that the cochlea-electrode spatial relationship derived from medical images can guide CI programming and lead to significant improvement in hearing outcomes. We have developed a series of algorithms that permit the segmentation of the inner ear anatomy and the localization of the electrodes. But, because clinical head CT images are acquired with different protocols, the field of view and orientation of the image volumes vary greatly. As a consequence, visual inspection and manual image registration to an atlas image are needed to document their content and to initialize intensity-based registration algorithms used in our processing pipeline. To facilitate large-scale evaluation and deployment of our methods these steps need to be automated. In this article we propose to achieve this with a deep convolutional neural network (CNN) that can be trained end-to-end to classify a head CT image in terms of its content and to generate probability maps that show landmarks' positions. Positions of the landmarks can then be used to estimate a point-based registration with the atlas image in which the same landmark set's positions are known. We achieve a classification accuracy of 99.5% and a localization error of 3.5mm for all 7 landmarks around each inner ear. These results are better than what we obtained with an earlier method that was designed for the same tasks.

4.1 Introduction

Cochlear implants (CIs) have been among the most successful neural prosthetics developed in the past few decades [1]. They are used to treat patients with severe-to-profound hearing loss. During a cochlear implantation surgery, an array of electrodes is threaded into the cochlea to replace the natural sound transduction mechanism of the human hearing system. After surgery, the CI needs to be programmed for hearing outcome optimization. This process includes the assignment of a

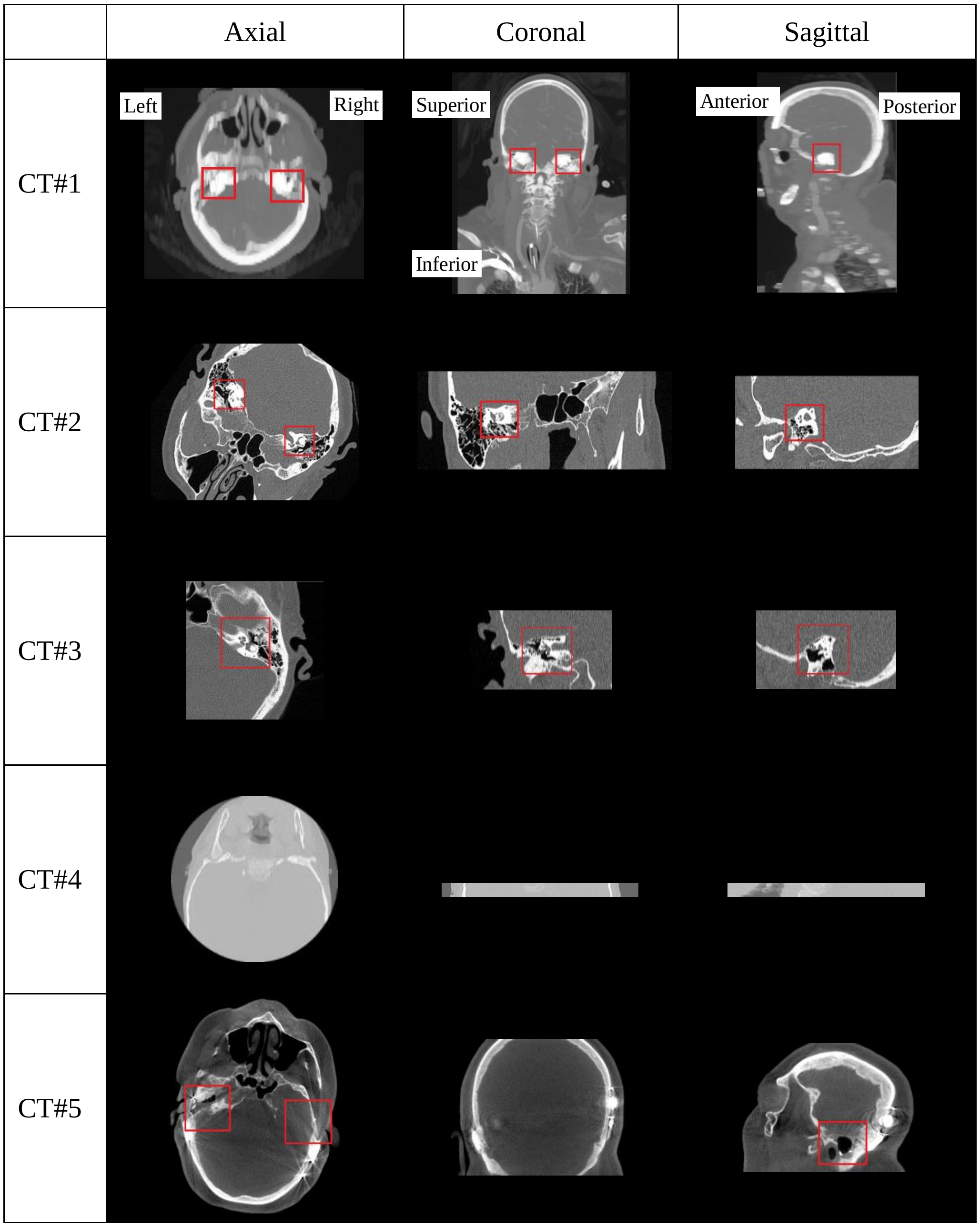

Fig. 4. 1 Four exemplar CTs from our dataset. The inner ears are shown in red boxes if they are present in the volume and visible in the slice

frequency range to each individual contact in the array so that it is activated when the incoming sound includes frequency components in such a range. Traditionally, the programming is done by an audiologist who can only rely on the recipients' subjective response to certain stimuli, e.g., whether they can hear a signal or rank pitches, without other clues. Accurate localization of electrodes in the CI relative to the intra-cochlear anatomy can provide useful guidance to audiologists to adjust the CI programming. Recently, our group has developed an image-guided cochlear implant programming (IGCIP) system [2]. It includes algorithms that permit the accurate segmentation of the intra-cochlear anatomy [3] and the localization of CI electrodes in clinical head CTs [4,5]. Studies in [6,7] have shown that the use of image guidance to program the CI leads to a significant improvement in hearing outcomes for both adults and children.

At the time of writing our IGCIP system is not yet fully automated, which hampers its large scale clinical deployment. One hurdle is the heterogeneity of the clinical head CTs that can be acquired on a variety of scanners with a range of acquisition protocols from multiple sites. Because of this the field of view (FOV) and orientation of the CT volumes vary greatly. **Fig. 4.1** shows several representative examples. Here, CT#1 covers a very large FOV, including both the whole head and the upper part of the torso. The FOV of CT#2 is representative of most (~80%) CTs in our image repository, but the head orientation deviates a lot from the most common pose shown in CT#1. CT#3 has a smaller FOV which only includes the right inner ear. In CT#4, only a narrow horizontal portion of the head is imaged and neither inner ear is included. The lack of a standard acquisition protocol also affects image contrast and quality as shown in **Fig. 4.1** where CT#1 is visibly of lower quality than the three others. CT #5 is a post-operative CT and serious beam hardening artifacts caused by the electrode arrays are visible. This poses an additional challenge.

Because of the heterogeneity we describe above manual intervention is often needed in our system prior to applying automatic algorithms to the images. First, when a new CT volume is received, a human operator needs to document manually its content, i.e., what ear(s) is/are shown in the volume; this is needed by model-based algorithms we use to segment the inner ear structures. Second, because our processing pipeline includes several rigid and non-rigid Mutual Information (MI)-based registration algorithms that require a reasonable initial alignment, manual intervention, i.e., manual translation and rotation of the images is required. Our ultimate goal is to develop a series of algorithms that are robust and fully automatic. In this work, we focus on developing methods to document the image content and to localize a set of landmarks that can be used for the estimation of an initial transformation that registers an atlas with a new volume using point-based registration techniques. We use a deep multi-task learning algorithm in an end-to-end fashion. The algorithm we propose can map a CT volume to a four-way classifier which accurately predicts whether the volume includes both inner ears, only the right inner ear, only the left inner ear or neither, and simultaneously generates probability maps that indicate the positions of a set of landmarks around each inner ear.

This chapter is an extension of our MICCAI conference paper [8]. The novel aspects we present in this chapter are: (1) the number of landmarks surrounding each inner ear is augmented from one to seven. By doing this, instead of only being able to find the positions of the inner ears, we can estimate a local rigid-body transformation between the image volume and an atlas to initialize our MI-based registration. (2) Instead of only using the maximum response of the final network output to document the volume content, we add a classification branch that is specially designed to perform this task. We use feature maps at different levels as input to the classification branch and train the classification branch. We show that such hierarchical features improve the image classification accuracy. (3) In our conference paper, we reported observing false positives that were eliminated using a post-processing step. Herein, we address the issue by using an adaptively weighted loss function in the training of the regression task.

	Training data					Test data				
	convCT w/ CI	lCT w/ CI	convCT w/o CI	lCT w/o CI	Total	convCT w/ CI	lCT w/ CI	convCT w/o CI	lCT w/o CI	Total
Before	125	146	527	0	798	102	140	553	0	795
After	155	253	1871	323	2602	150	239	1776	321	2486

Table 4. 1 Distributions of our CT data w.r.t. presence of CI and scanner type before and after augmentation

4.2 Materials

The data we use in this study include head CTs from 322 patients. Since images were acquired both pre-operatively and post-operatively, and multiple reconstructions can be performed for one acquisition, one patient can have several CT volumes. In total, we have 1,593 CT volumes. The scanners include both conventional and Xoran xCAT® scanners. Xoran xCAT® scanners are flat-panel, low-dose scanners. Compared to CTs acquired with conventional scanners, images acquired with such scanners typically have lower quality and suffer from intensity inhomogeneity. We refer to CTs they produce as lCTs and CTs acquired with conventional CTs as convCTs. One typical lCT and one typical convCT obtained from the same patient and rigidly registered are shown in **Fig. 4.2** to illustrate differences between images. The volumes in our data set also cover regions of different sizes and have different resolutions. The size ranges from 10 mm to 256 mm in the left-right and anterior-posterior directions and from 52 mm to 195 mm in the inferior-superior direction. The resolution varies from 0.14 mm to 2.00 mm in the left-right and anterior-posterior

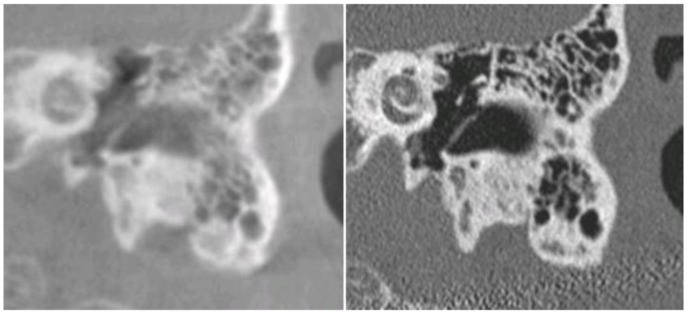

Fig. 4. 2 A comparison of lCT (left) and convCT (right) from the same patient, in the left ear region.

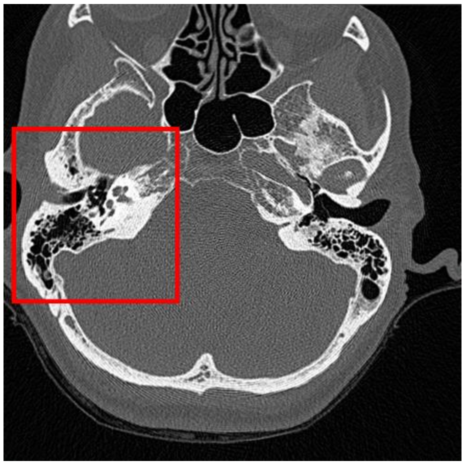

Fig. 4. 3 The region in which we select our landmarks

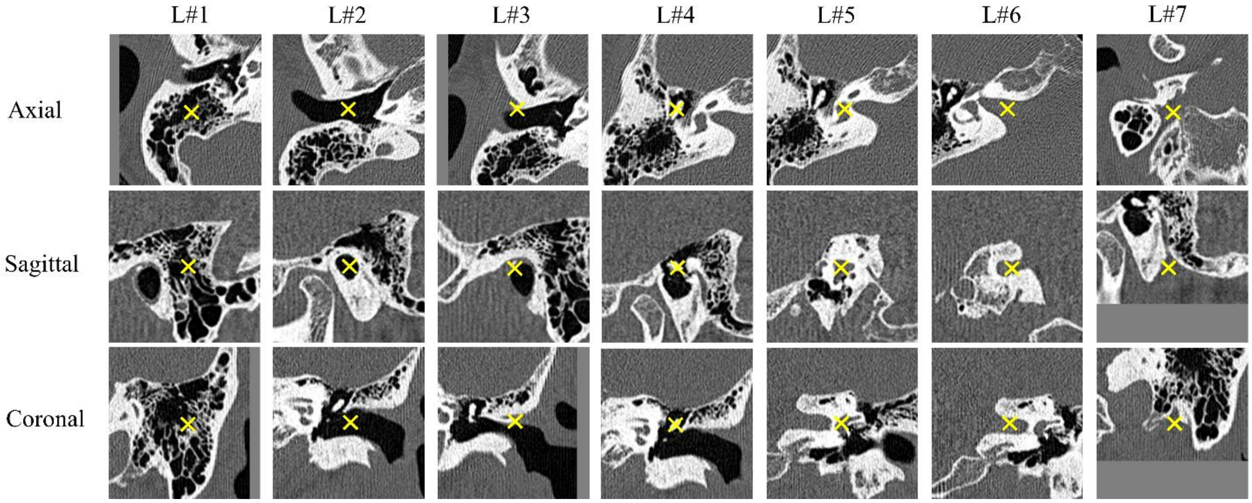

Fig. 4. 4 The seven landmarks we select for the study

directions and from 0.14 mm to 5.00 mm in the inferior-superior direction.

To develop and evaluate our proposed method, we randomly split the data into a training set and a testing set. We verify that image volumes pertaining to a single patient are not split between the two sets. The number of CT volumes in the training set and the testing set is listed in the upper part of **Table 4.1**. The images are categorized according to whether they are convCTs or lCTs and whether they contains an implant (w/ CI) or not (w/o CI).

For each volume, we visually check the presence of inner ears. For each visible inner ear, 7 pre-defined landmark points surrounding the cochlea are manually selected. They represent the positions of the mastoid, the external auditory canal, the spine of henle, the ossicles, the cochlear labyrinth, the internal auditory canal, and the stylomastoid foramen. They are shown in **Fig. 4.4**. The region of interest is shown in **Fig. 4.3**. As mentioned earlier, scans can include both inner ears, only one (left/right), or neither. However, the number of image volumes in each of these four categories is not balanced. Indeed, in our current data set about 80% of the volumes include both ears. About 20% include one inner ear. Image volumes that include neither inner ear exist but are rare. To tackle this issue, we augment each set by cropping sub-volumes from CT volumes that include both ears to create artificial samples for the other three categories. After cropping, substantial but reasonable deformation, including scaling, rotation and skewing are applied to increase data variance. The scaling, rotation and skewing values in each axis are randomly generated by uniformly sampling values in intervals [0.95, 1.05], [-5°, 5°] and [0, 0.05], respectively. All image volumes are resampled to 2.25×2.25×2.25 mm^3 per voxel. It is a convention in deep learning to map images' intensity to [-1, 1] (or [0, 1]) by doing a linear scaling that converts the maximum intensity to 1 and the minimum to -1. In our dataset, the intensity of metal implants in post-operative CTs is much higher than that of normal human tissue. Applying a linear scaling would compress the intensity range of the tissues and potentially affect the training of the network. To avoid this issue, we first apply an intensity cutoff, i.e., we set the intensity values of the 0.1% voxels with the highest intensities (an empirical estimation of the fraction of voxels occupied by the metal implant) to their lower bound. We subsequently apply the intensity remapping. The images are cropped or padded to 96×96×96 voxels. The number of volumes in each category in the training and testing sets after augmentation is shown in the lower part of **Table 4.1**. It shows that our data set does not contain pre-operative lCT volumes. This is because the flat-panel scanner is only used to acquire post-operative images. As a consequence, our lCT pre-operative images in the augmented training and testing sets only include volumes that contain a single ear. These have been generated from unilateral post-operative CT volumes. In the test set, there are 625, 625, 625 and 611 CTs that include both, left, right and neither ear(s), respectively.

4.3 Methods

4.3.1 Related work

Detecting landmarks in medical images has been a well-studied topic for a many years. The reader is referred to the work of Rohr for a review of earlier work [9]. In the past few years, learning-based methods have been successfully applied to this task starting with random forest−based methods [10]. An exhaustive review of this body of work would be outside the scope of this article but some representative papers include [11-17].

More recently, deep learning methods and more specifically deep convolutional neural networks (CNNs) have superseded random forest techniques. Originally developed for 2D applications, they have been expanded to 3D and the 3d U-Net architecture proposed by [18] has been widely used for a range of medical image segmentation and detection tasks. Again a complete coverage of this body of work would outside the scope of this article but examples that are germane to our work include the work of Payer *et al.* [19] who have proposed a “SpatialConfiguration-Net” to detect landmarks in MR volumes, of Zhang, Liu and Shen [20] who devised a system that consists of two 3D deep networks for both brain and prostate landmark detection, of Yang *et al*, [21] who used a volume-to-volume network to detect a set of vertebra points in 3D CT volumes, and of Liu *et al.* [22] who used a 3D network to detect landmarks in the brain for disease diagnosis.

Our own work has been focused on head CT images and more specifically on the documentation of images covering the ears and on the localization of landmarks in these images to initialize registration algorithms. We have proposed methods to perform image content documentation and registration initialization tasks separately in the last two chapters. They are also published as [23], and [24], respectively and presented in previous chapters of this dissertation. In the former work, the content of head CT volumes is documented using a 2D CNN.

The CT volume is processed slice by slice or in very thin 3D volumes (3 consecutive slices). This makes the algorithm computationally inefficient and full 3D information is not exploited. As will be shown in the results section, the solution we propose herein leads to better results. In the latter paper, we estimate a rigid-body transformation via a set of landmarks but the system was not designed to document image content. The evaluation is also done on a screened dataset in which each image includes the region of interest, i.e., a region that encompasses the ear. In the work reported herein, we present an end-to-end solution that performs both tasks. It takes as input a CT volume and outputs both the image content and a set of landmarks that surround ear/ears that are included in the volume. In our 2018 MICCAI paper we propose a 3D method that can both localize one single landmark and document the image content. In the following sections, we detail this approach, we present improvements we have brought to this early solution, and we compare the two approaches.

4.3.2 HeadLocNet-1: inner ear detection using the 3d U-Net with false positive suppression and a shape constraint

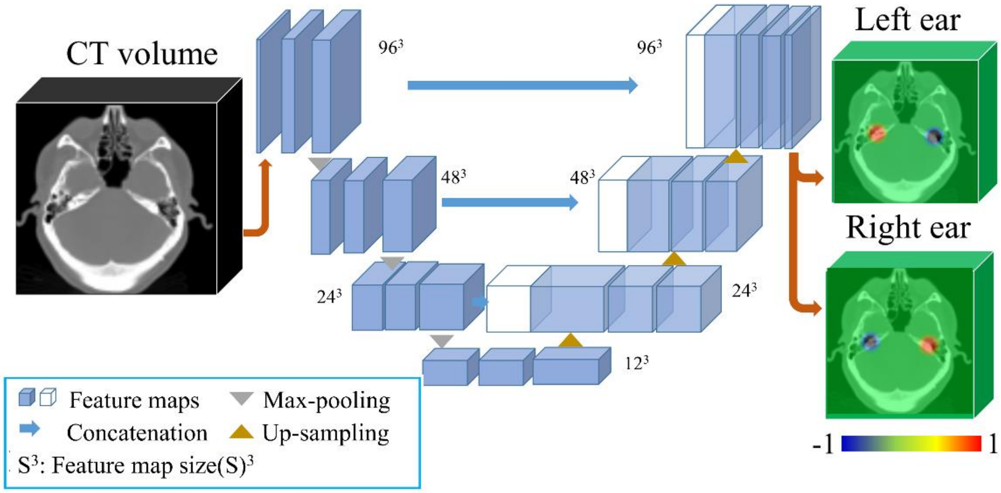

Fig. 4. 5 The HeadLocNet-1 architecture

In this first solution, we formulate the inner ear detection problem as a single landmark detection problem. We use the fifth landmark, i.e., the one representing the cochlear labyrinth, in the set of seven shown in **Fig. 4.4** because it is the closest to the cochlea and we adopted the 3d U-Net proposed by [18] to map a whole 3D image volume to two probability maps that have the same dimensionality as the input volumes. As is shown in **Fig. 4.5**, the 3d U-Net requires a 3D volume as input. The network consists of a sequence of convolution-pooling layers which compress the raw input volume into low-resolution, highly-abstracted feature maps. Following them are a sequence of convolution-upsampling layers, which process the abstracted feature maps into outputs with the same resolution as the input, in a way that is symmetrical to what is done in the compression layers. The channel numbers of the feature blocks are (1,32,64)-(64,64,128)-(128,128,256)-(256,256,512)-(768,256,256)-(384,128,128)-(192,64,64), from left to right. Each pair of parentheses contain channel numbers in three consecutive feature blocks at one level. In our first attempt, at the training stage, for each inner ear, we use a 3D Gaussian function centered at the manually labeled landmark position as a probability map. The standard deviation σ of the Gaussian is empirically set to 3 voxels in the resampled image. The probability values are multiplied by a constant to scale the maximum to 1. Any value below 0.05 is set to 0. If the inner ear is not included in the image, all values in the corresponding probability map are set to 0. We treat this volume-to-volume mapping as a voxel-wise regression problem. The weighted mean of voxel-wise squared errors between the output probability maps and those generated with the ground truth inner ear landmarks is used as the loss function. Larger weights are assigned to voxels with non-zero probabilities in the supervising maps. They are sparse but are very important features. Specifically, suppose the numbers of non-zero entries and zero entries are $N_{nonzero}$ and

N_{zero}, respectively, in the output probability map. The weights associated with non-zero entries and zero entries are $w_{nonzero}$ and w_{zero} defined as follows:

$$\begin{cases} w_{nonzero} = \frac{N_{zero}}{N_{nonzero}+N_{zero}} \\ w_{zero} = \frac{N_{nonzero}}{N_{nonzero}+N_{zero}} \end{cases}. \tag{4.1}$$

For a new CT volume, we preprocess it in the same way as we do for training images. Using the trained network, we generate two probability maps, one for the left ear and the other for the right ear. For each probability map, we find its maximum. If it is larger than p_{thres}= 0.5, we predict that the corresponding inner ear is present. Otherwise, we predict that it is absent.

Results we obtain with this approach are not satisfactory because it leads to a large number of false positives. We observe that the response map associated with one inner ear can have a very high response at the location of the other ear, possibly due to their similar intensity characteristics. In turn, this leads to a substantial number of wrong detections. To solve this problem, we incorporate a false positive suppression strategy during training. Specifically, for the probability map associated with one ear, if the ear on the other side of the head is included in the image, we force the values around this second inner ear to be negative rather than zero to penalize the detection of the erroneous ear. The negative values that are used are the same Gaussian-distributed values that are used for the correct ear but centered on the incorrect ear and multiplied by minus one. By penalizing the network in such a way, we effectively suppress the number of false positives, as will be shown in the results section. **Fig. 4.5** shows our final network architecture. The left ear response maps of a test image with both ears with and without false positive suppression are shown in **Fig. 4.6** as an example. This figure shows that the false positive caused by the right ear is effectively suppressed.

Even though the aforementioned method suppresses false positives caused by the contralateral ear, other false positives remain present at some random positions, e.g., the location of the CI transmitters in some post-operative CTs, as shown in **Fig. 4.7**. To solve this problem, we capture the spatial relationship between inner ear pairs using a low-dimension shape model and use this *a-priori* information to further evaluate the plausibility of the detected inner ear pairs. To do so we first collect the coordinates of the inner ear pairs in the training set. For the i^{th} pair, these are denoted as $l_{left}^{i} = (x_{left}^{i}, y_{left}^{i}, z_{left}^{i})$ and $l_{right}^{i} = (x_{right}^{i}, y_{right}^{i}, z_{right}^{i})$ for the left and right ear, respectively. We subtract from each point the center of the two and stack the coordinate vectors to create a 6-d shape vector $\boldsymbol{s}^{i}$. The mean shape is computed as

$$\bar{\boldsymbol{s}} = \frac{\sum_{i=1}^{N} \boldsymbol{s}^{i}}{N}. \tag{4.2}$$

Here, *N* is the number of inner ear pairs in the training set. The modes of variation of the shapes are computed as the k eigenvectors $\{\overrightarrow{u_j}, \mathrm{j} = 1,2, \dots, \mathrm{k}\}$ of the covariance matrix of $\{\boldsymbol{s}^{i}, i = 1,2, \dots, N\}$. The *k* (in our case, *k*=3 because ears come in pairs) non-zero eigenvalues associated with these are $\{\lambda_j, \mathrm{j} = 1,2, \dots, \mathrm{k}\}$. Suppose the projections of $\boldsymbol{s}^{i} - \bar{\boldsymbol{s}}$ onto $\{\overrightarrow{u_j}, \mathrm{j} = 1,2, \dots, \mathrm{k}\}$ are $\{b_j^{i}, \mathrm{j} = 1,2, \dots, \mathrm{k}\}$. The Mahalanobis distance between $\boldsymbol{s}^{i}$ and the mean shape $\bar{\boldsymbol{s}}$ is thus

$$M(\boldsymbol{s}^{i}, \bar{\boldsymbol{s}}) = \sqrt{\frac{\sum_{j=1}^{k} {b_j^{i}}^{2}}{\lambda_j}}\ . \tag{4.3}$$

It measures how much the spatial distribution of the ear pair deviates from the spatial distribution of ear pairs observed in the training set. We record the maximal Mahalanobis distance of the training shapes as M_{max}. For each test volume, when two inner ears are detected from the probability maps, the position vectors of the left and right inner ears, i.e., $l_{left}^{test} = (x_{left}^{test}, y_{left}^{test}, z_{left}^{test})$ and $l_{right}^{test} = (x_{right}^{test}, y_{right}^{test}, z_{right}^{test})$, are stacked and demeaned to create a shape vector $\boldsymbol{s}^{test}$. If $M(\boldsymbol{s}^{test}, \bar{\boldsymbol{s}}) > M_{max}$, we reject the detected inner ear with the lower

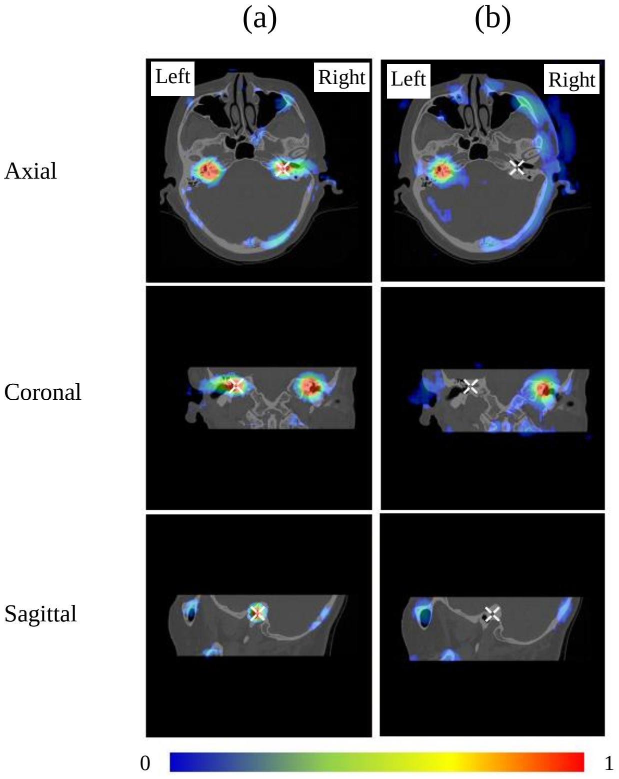

Fig. 4. 6 The left ear response maps of an input image containing both ears. Column (a): before applying false positive suppression: the response at the right ear is also very high, (b): after applying false positive suppression, the false positive is eliminated (the location of the right ear is marked).

response. We denote this network as HeadLocNet-1 (Head CT Localization Network for 1 landmark).

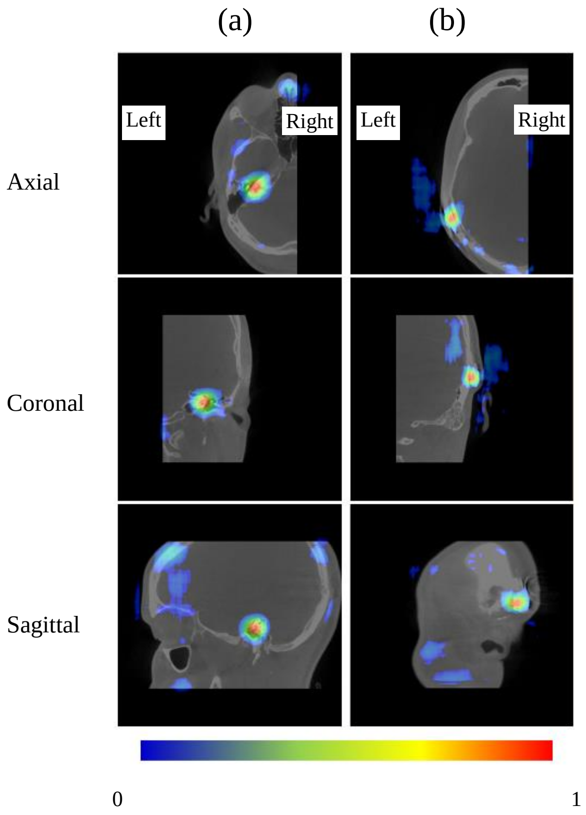

Fig. 4. 7 The response maps of an input image that includes the left half of the head. Column (a): the response map associated with the left ear, (b): the response map associated with the right ear. The response at the location of the CI transmitter is so high that it is detected as an ear.

4.3.3 HeadLocNet-MC: co-learning image volume classification and landmark set localization

As discussed previously, to automate our IGCIP pipeline we need to estimate transformations that are used to initialize intensity-based registration algorithms. The approach we follow in this work is to localize a set of landmarks that are used to compute a rigid-body transformation using a point-based registration method. To do so, we extend the solution we have presented above and we propose a network architecture that can co-task image content classification and landmark set detection. Based on the previous network architecture, we first set the number of output channels to 14; 7 for each side of the head. Second, instead of using the output probability maps as indicators of whether or not an inner ear exists, we add a classification branch to the main path of the 3d U-Net. As is shown in **Fig. 4.8**, in the upsampling stage of the 3d U-Net, we use the feature maps as input to the classifier. Because the dimensionality of the feature maps is too high, which could cause overfitting, we use global pooling operations to perform dimensionality reduction. For each feature map, we use a global max pooling and a global average pooling, reducing its dimensionality from M^3 with M =12, 24, 48 or 96 to 2. The hierarchical features we extract from the multi-level feature maps are used as input to a classifier. The classifier is designed to be a fully-connected network with one hidden layer (500 units). The number of output units is 4. We thus use a richer feature set than the one we used previously, which, as will be shown, improves performance. The network is trained to predict if the input volume includes both inner ears, only the left, only the right or neither inner ear. We use the categorical cross entropy between the ground truth and the prediction as the loss function.

We also modify the probability map regression loss function we used previously. In addition to assigning larger weights to non-zero entries in the ground truth probability maps, we also penalize more the entries for which the predicted values deviate too much from the ground truth maps. That is to say, in one training iteration, for each sample we define the loss as follows:

$$loss(\boldsymbol{P}, \widehat{\boldsymbol{P}}) = \frac{1}{N^3 L} \sum_{l=1}^{L} \sum_{k=1}^{N} \sum_{j=1}^{N} \sum_{i=1}^{N} w_{i,j,k,l} \left(P_{i,j,k,l} - \widehat{P}_{i,j,k,l}\right)^2. \tag{4.4}$$

Here, L is the number of landmarks, N = 96 is the dimensionality of each probability map. $\boldsymbol{P}$ and $\widehat{\boldsymbol{P}}$ are the ground truth probability maps and the predicted probability maps, respectively. $w_{i,j,k,l}$ is set to a large value if $|P_{i,j,k,l} - \hat{P}_{i,j,k,l}| > \Delta$ (Δ is empirically set to 0.2) or if $|P_{i,j,k,l}| > 0$. Otherwise, it is set to a small value. We replace $w_{nonzero}$ and w_{zero} introduced in Section *4.3.2* by the large and small weights, respectively. By doing so, we can penalize more the regions in which false negatives and false positives happen. An illustration of our weighting scheme in 2D is shown in **Fig. 4.9**. We call this new network HeadLocNet-MC (Head CT Localization Network for Multiple landmarks with Classification).

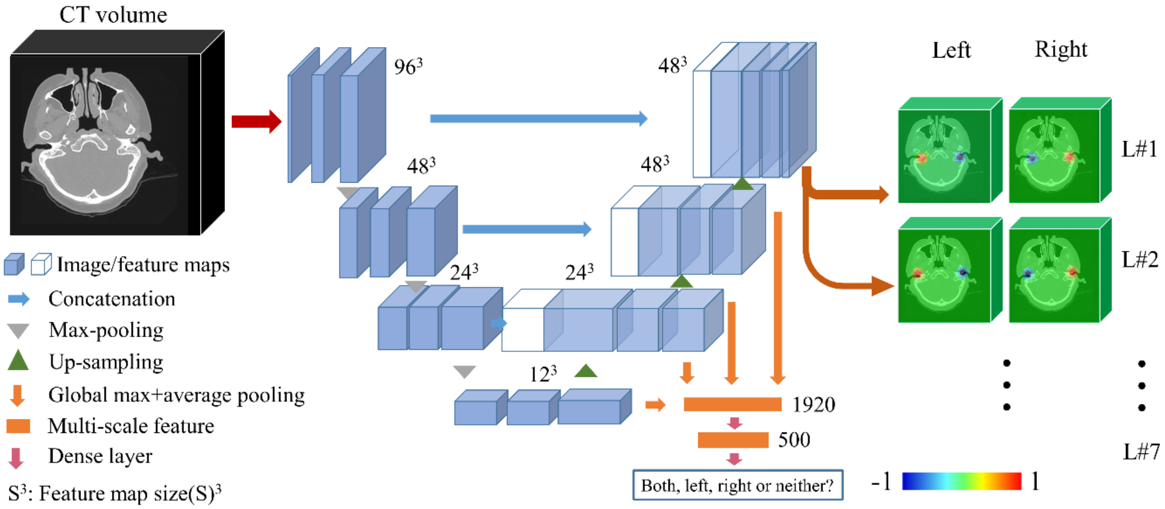

Fig. 4. 8 The HeadLocNet-MC architecture

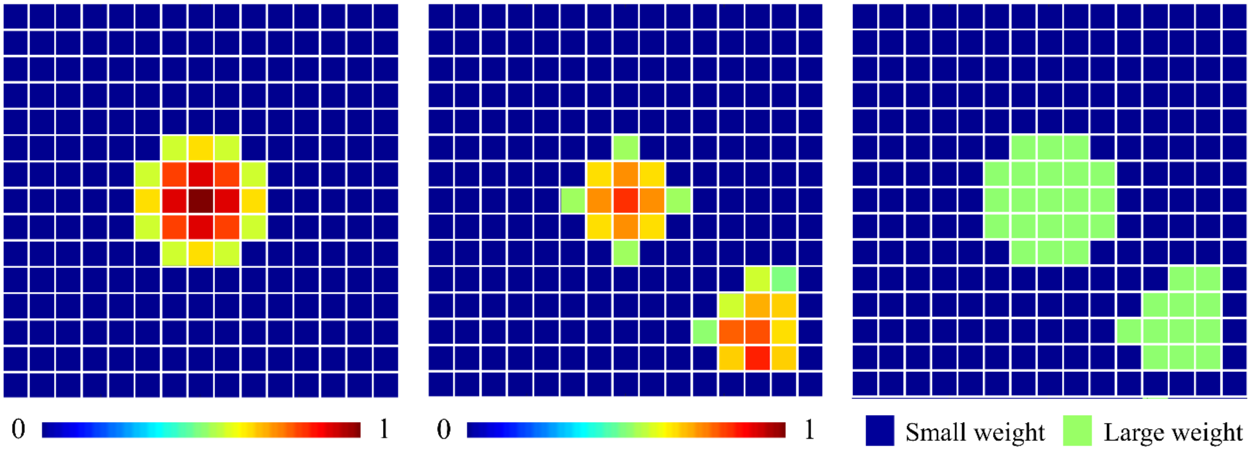

Fig. 4. 9 An example of the probability map ground truth (left), predicted probability map at one iteration (middle) and the generated weight matrix (right) in HeadLocNet-MC.

4.4 Experimental settings

Image preprocessing, including resampling, cropping, padding and intensity normalization is done using MATLAB. We train the neural networks using stochastic gradient descent (SGD) with 0.9 momentum and an initial learning rate of 0.0001. The batch size is set to 1. The code is written in Keras developed by Chollet [25] and runs on a Nvidia Titan X GPU. Training each neural network model takes ~2 days. A forward propagation to process one image using the model takes ~1.4 seconds on average. We also integrate this method (written in python) into our CI programming system. The total time to process one patient's image including the overhead to load Keras and the trained weights, and the forward propagation takes ~15s.

4.5 Results

4.5.1 Results obtained with HeadLocNet-1

Classification error rate					
	CTs Classified by presence of CI		CTs Classified by scanner		Overall
	w/ CI	w/o CI	convCT	lCT	
Before FP reduction	14.65%	7.39%	4.65%	22.14%	8.53%
After FP reduction	0.77%	1.53%	1.50%	1.09%	1.41%
Localization error (in mm)					
HeadLocNet-1	**2.32±2.34**	**2.48±2.35**	**2.41±1.13**	**2.57±4.49**	**2.45±2.35**
RF-based	6.80±18.14	5.39±11.80	5.01±11.14	8.17±19.61	5.87±13.57

Table 4. 2 Error rates before and after false positive suppression and shape-based constraint (denoted as FP reduction), and localization error using our HeadLocNet-1 solution and the baseline RF-based method

Truth \ Predict	Both	Left	Right	Neither
Both	618	5	1	1
Left	0	619	2	4
Right	0	4	613	8
Neither	1	4	5	601

Table 4. 3 Confusion matrix for each category obtained with HeadLocNet-1

In the upper part of **Table 4.2**, we show classification results obtained with and without the application of false positive suppression and the shape-based constraint. We successfully improve the error rate by ~7% when using the two strategies. In **Table 4.3**, we show the confusion matrix obtained with the test set, showing an overall accuracy of 98.6%. This is substantially higher than the 96% obtained with the slice-wise network that was proposed in [23].

For the test CT volumes that are correctly classified, we calculate the localization error, which is shown in the lower part of **Table 4.2**. The localization error is computed as the distance between the manually labeled inner ear position and the automatic localization. For comparison purpose, we use our previous Random Forest (RF)-based approach [26] to find the same landmark in our current dataset and we report the localization accuracy obtained with this method. A number in bold indicates that the localization error generated by the method in this row is lower and significantly different from the other method. The results show that the proposed method produces substantially lower localization error for all image groups. All differences for the five groups of comparisons are statistically significant using a paired t-test (*p*<0.01).

4.5.2 Results obtained with HeadLocNet-MC

Predict Truth	Both	Left	Right	Neither
Both	621	2	0	2
Left	1	623	1	0
Right	5	1	619	0
Neither	1	0	0	610

Table 4. 4 Confusion matrix for each category obtained with HeadLocNet-MC

Classification error rate					
	CTs Classified by presence of CI		CTs Classified by scanner		Overall
	w/ CI	w/o CI	convCT	lCT	
HeadLocNet-M	1.03%	1.47%	1.65%	0.54%	1.40%
HeadLocNet-MC	0.26%	0.57%	0.67%	0	0.52%
Localization error (in mm)					
Random Forest	**6.91±14.53**	**6.35±11.68**	**6.25±12.24**	**7.49±13.91**	**6.52±12.63**
HeadLocNet-MC	**3.31±1.32**	**3.49±2.10**	**3.45±1.96**	**3.48±1.98**	**3.45±1.97**

Table 4. 5 Error rates of the two models, and landmark set localization error of random forest & HeadLocNet-MC

We evaluate the classification performance of our trained HeadLocNet-MC model presented in Section *4.3.3* on the same test set we use to evaluate the HeadLocNet-1 solution. The classification performance is quantified and reported in **Table 4.4**. The classification error rate is 0.52%, showing a further substantial improvement over the 1.41%, produced by HeadLocNet-1. The two main differences between HeadLocNet-MC and HeadLocNet-1 are: (1) HeadLocNet-MC uses multi-scale feature maps in the intermediate layers of the neural network instead of only relying on the final output to classify images and (2) it uses information about 7 landmarks rather than one for supervision. To evaluate each factor's contribution to the classification performance improvement, we perform an ablation study: we train another model which is almost identical to the proposed HeadLocNet-MC except that it does not have the classification branch. We call this version HeadLocNet-M. As we do in HeadLocNet-1, in HeadLocNet-M, we threshold the maximum of each probability map on one side of the head and use a majority voting strategy to determine the presence of the inner ear. Specifically, we predict that an inner ear exists if and only if at least 4 landmarks are detected around it. With this approach we achieve a classification error rate of 1.40%, which is almost the same as what is obtained with HeadLocNet-1. This indicates that it is the utilization of multi-scale features in the intermediate layers rather than the additional landmarks' supervision information that contributes to the improvement. We check the cases which are incorrectly classified and show two examples in **Fig. 4.10**. The first case is manually labeled "both inner ears" because both cochlea are present. However, the right side of the head is only partially covered and most landmarks surrounding the right cochlea are not included. The network labels it as a "left ear" image. For the second image, the wrong prediction of "both ears" is likely due to an abnormally large FOV. It is worth noting that even though the right ear of CT#1 and the left ear of CT#2 are included, they cannot be used for IGCIP because: (1) in CT#1 the portion of the right ear is so small that image segmentation algorithms cannot be applied. (2) The quality of CT#2 is too low to get meaningful cochlear segmentation. Visual inspection of other failure cases reveals that they are caused by unusual acquisitions.

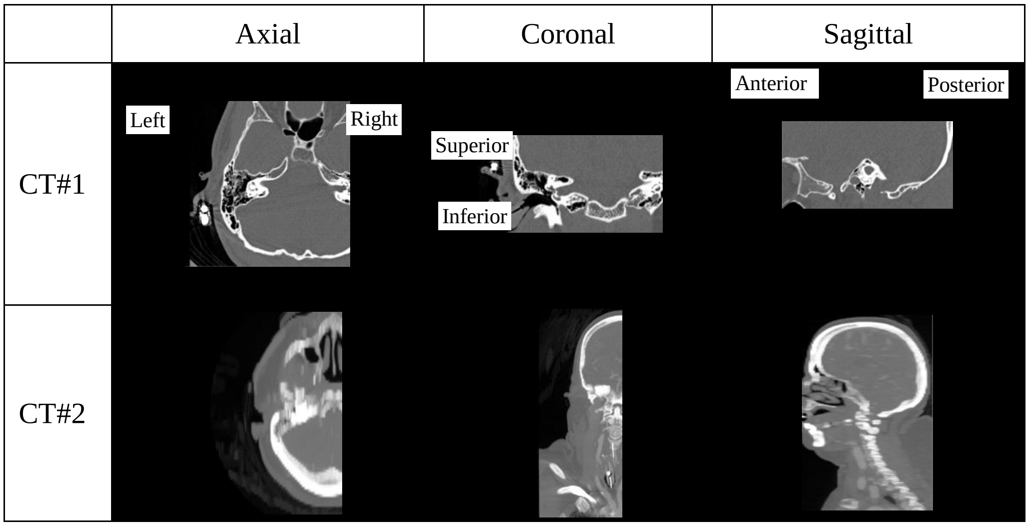

Fig. 4. 10 Two examples of wrong predictions

We then evaluate the landmark localization performance of HeadLocNet-MC. To do so, we use the maximum of each probability map as our prediction of the landmark position. We show the overall localization error in **Table 4.5**. For comparison purpose, we again list the mean overall localization error obtained with our random forest-based method [26] for all 7 landmarks on the same test set. **Table 4.5** shows that the proposed HeadLocNet-MC is substantially better (the localization error is reduced by about half). A paired t-test shows that the difference is statistically significant (p<0.01). We note that instead of training 7 separate models, as we did in the random forest-based approach, the proposed network produces a single model capable of localizing all landmarks in one pass. To test the effect of simultaneous localization on localization accuracy we compare the localization error of the same landmarks (Landmark #1 and Landmark #5) obtained with HeadLocNet-1 and HeadLocNet-MC. The HeadLocNet-1 results for Landmark #5 are readily available because this landmark has been used to produce the results obtained earlier. To produce results for Landmark #1 we train another HeadLocNet-1 model. Results we have obtained are shown in **Table 4.6**. These show that localization accuracy obtained with the two methods is comparable. A paired t-test shows that there is no statistically significant difference ($p > 0.01$) between the two methods.

Table 4.5 shows that, although satisfactorily small, the localization error of Landmark #1 is larger than that of landmark #5. We believe this is caused by the lack of features surrounding Landmark #1. As shown in **Fig. 4.3**, Landmark #5 is relatively easy to localize because it has very distinguishable local features. Landmark #1 however is located at the center of the temporal bone and the nearly random pattern of trabecular bones surrounding it over a large region makes pinpointing the landmark an actually ill-posed problem.

Localization error (in mm)		
	Landmark #1	Landmark #5
HeadLocNet-1	3.65±3.05	2.45±2.35
HeadLocNet-MC	3.77±1.83	2.58±1.25

Table 4. 6 Localization errors of Landmark #1 & #5, generated using HeadLocNet-1 and HeadLocNet-MC (in mm)

4.6 Conclusions and discussions

In this article, we begin by describing the work we have presented in our conference article. We call this early solution HeadLocNet-1. It relies on a 3d U-Net with false positive suppression and a shape-based constraint to document the content of head CTs and to localize a landmark around the inner ear. This deep-learning solution outperforms earlier methods we have proposed for content labeling of head CT images [23] and for landmark localization [24]. We expand our

HeadLocNet-1 solution to produce a new network architecture which we call HeadLocNet-MC. This new solution includes a one-hidden-layer classification branch that uses hierarchical features from the intermediate layers of 3d U-Net as input. The overall network is therefore trained using two loss functions. We show that HeadLocNet-MC (1) works in an end-to-end fashion, with no need for post-processing, (2) is able to document the content of head CTs better than HeadLocNet-1, and (3) is able to robustly localize the 7 landmarks (14 in total for both sides) in one pass with no loss of accuracy compared with HeadLocNet-1. This is a significant step toward full automation of our IGCIP process, thus facilitating its clinical use and deployment.

References

[1] NIDCD, "Fact Sheet: Cochlear Implants," NIH Publ. No. 11-4798, pp. 1–4, 2011.

[2] J. H. Noble, R. F. Labadie, R. H. Gifford, and B. M. Dawant, "Image-Guidance enables new methods for customizing cochlear implant stimulation strategies," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 21, no. 5, pp. 820–829, 2013.

[3] J. H. Noble, R. F. Labadie, O. Majdani, and B. M. Dawant, "Automatic segmentation of intracochlear anatomy in conventional CT," IEEE Trans. Biomed. Eng., vol. 58, no. 9, pp. 2625–2632, 2011.

[4] Y. Zhao, B. M. Dawant, R. F. Labadie, and J. H. Noble, "Automatic localization of closely-spaced cochlear implant electrode arrays in clinical CTs," Med. Phys., vol. 45, no. 11, pp. 5030–5040, 2018.

[5] Y. Zhao, S. Chakravorti, R. F. Labadie, B. M. Dawant, and J. H. Noble, "Automatic graph-based method for localization of cochlear implant electrode arrays in clinical CT with sub-voxel accuracy," Med. Image Anal., vol. 52, pp. 1–12, 2019.

[6] J. H. Noble, R. H. Gifford, A. J. Hedley-Williams, B. M. Dawant, and R. F. Labadie, "Clinical evaluation of an image-guided cochlear implant programming strategy," Audiol. Neurotol., vol. 19, no. 6, pp. 400–411, 2014.

[7] J. H. Noble, A. J. Hedley-Williams, L. Sunderhaus, B. M. Dawant, R. F. Labadie, S. M. Camarata and R. H. Gifford, "Initial Results With Image-guided Cochlear Implant Programming in Children," Otol. Neurotol., vol. 37, no. 2, pp. e63–e69, 2016.

[8] D. Zhang, J. Wang, J. H. Noble, and B. M. Dawant, "Accurate Detection of Inner Ears in Head CTs Using a Deep Volume-to-Volume Regression Network with False Positive Suppression and a Shape-Based Constraint," in Medical Image Computing and Computer-Assisted Intervention, 2018, pp. 703–711.

[9] K. Rohr, "Landmark-based image analysis: using geometric and intensity models," vol. 21. 2001.

[10] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[11] R. Donner, B. H. Menze, H. Bischof, and G. Langs, "Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization," Med. Image Anal., vol. 17, no. 8, pp. 1304–1314, 2013.

[12] Y. Zheng, M. John, R. Liao, A. Nottling, J. Boese, J. Kempfert, G. Brockmann, and D. Comaniciu, "Automatic aorta segmentation and valve landmark detection in C-Arm CT for transcatheter aortic valve implantation," IEEE Trans. Med. Imaging, vol. 31, no. 12, pp. 2307–2321, 2012.

[13] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui , "Regression forests for efficient anatomy detection and localization in computed tomography scans," Med. Image Anal., vol. 17, no. 8, pp. 1293–1303, 2013.

[14] D. Han, Y. Gao, G. Wu, P. T. Yap, and D. Shen, "Robust anatomical landmark detection with application to MR brain image registration," Comput. Med. Imaging Graph., vol. 46, pp. 277–290, 2015.

[15] T. Ebner, D. Stern, R. Donner, H. Bischof, and M. Urschler, "Towards automatic bone age estimation from MRI: Localization of 3D anatomical landmarks," in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016

[16] C. Lindner, C. W. Wang, C. T. Huang, C. H. Li, S. W. Chang, and T. F. Cootes, "Fully Automatic System for Accurate Localisation and Analysis of Cephalometric Landmarks in Lateral Cephalograms," Sci. Rep., vol. 6, no. September, 2016.

[17] J. Zhang, Y. Gao, Y. Gao, B. C. Munsell, and D. Shen, "Detecting anatomical landmarks for fast Alzheimer's disease diagnosis," IEEE Trans. Med. Imaging, vol. 35, no. 12, pp. 2524–2533, 2016.

[18] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016, pp. 424–432.

[19] C. Payer, D. Stern, H. Bischof, and M. Urschler, "Regressing Heatmaps for Multiple Landmark Localization Using CNNs," in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016, pp. 230–238.

[20] J. Zhang, M. Liu, and D. Shen, "Detecting Anatomical Landmarks from Limited Medical Imaging Data Using Two-Stage Task-Oriented Deep Neural Networks," IEEE Trans. Image Process., vol. 26, no. 10, pp. 4753–4764, 2017.

[21] D. Yang, T. Xiong, D. Xu, Q. Huang, D. Liu, S. K. Zhou, Z. Xu, J. Park, M. Chen, T. D. Tran, and S. P. Chin, "Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization," in Information Processing in Medical Imaging, 2017, pp. 633–644.

[22] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Landmark-based deep multi-instance learning for brain disease diagnosis," Med. Image Anal., vol. 43, pp. 157–168, 2018.

[23] D. Zhang, J. H. Noble, and B. M. Dawant, "Automatic detection of the inner ears in head CT images using deep convolutional neural networks," in Proceedings of SPIE conference on Medical Imaging, 2018, p. 10574.

[24] D. Zhang, Y. Liu, J. H. Noble, and B. M. Dawant, "Localizing landmark sets in head CTs using random forests and a heuristic search algorithm for registration initialization, " J. Med. Imaging, vol. 4, no. 4, p. 44007, 2017.

[25] F. Chollet, "Keras," 2015.

[26] D. Zhang, Y. Liu, J. H. Noble, and B. H. Dawant, "Localizing landmark sets in head CTs using random forests and a heuristic search algorithm for registration initialization," J. Med. Imaging, vol. 4, no. 4, p. 44007, 2017.

Chapter V

TWO-LEVEL TRAINING OF THE 3D U-NET FOR ACCURATE SEGMENTATION OF THE INTRA-COCHLEAR ANATOMY IN HEAD CTS WITH LIMITED GROUND TRUTH TRAINING DATA

Dongqing Zhang, Rueben Banalagay, Jianing Wang, Yiyuan Zhao, Jack H. Noble, Benoit M. Dawant

Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235, USA

This chapter was presented in the 2019 SPIE Medical Imaging Conference and is in publishing at the time of writing.

Abstract

Cochlear implants (CIs) use electrode arrays that are surgically inserted into the cochlea to treat patients with hearing loss. For CI recipients, sound bypasses the natural transduction mechanism and directly stimulates the neural regions, thus creating a sense of hearing. Post-operatively, CIs need to be programmed. Traditionally, this is done by an audiologist who is blind to the positions of the electrodes relative to the cochlea and only relies on the subjective response of the patient. Multiple programming sessions are usually needed, which can take a frustratingly long time. We have developed an image-guided cochlear implant programming (IGCIP) system to facilitate the process. In IGCIP, we segment the intra-cochlear anatomy and localize the electrode arrays in the patient's head CT image. By utilizing their spatial relationship, we can suggest programming settings that can significantly improve hearing outcomes. To segment the intra-cochlear anatomy, we use an active shape model (ASM)-based method. Though it produces satisfactory results in most cases, sub-optimal segmentation still happens. As an alternative, herein we explore using a deep learning method to perform the segmentation task. Large image sets with accurate ground truth (in our case manual delineation) are typically needed to train a deep learning model for segmentation but such a dataset does not exist for our application. To tackle this problem, we use segmentations generated by the ASM-based method to pre-train the model and fine-tune it on a small image set for which accurate manual delineation is available. Using this method, we achieve better results than the ASM-based method.

5.1 Introduction

Cochlear implants (CIs) are one of the most successful neural prosthetics that have been developed in the past few decades. They are used to treat patients with severe-to-profound hearing loss. In a

CI surgery, an electrode array is inserted into the cochlea. Post-operatively, the CI is programmed. The spatial location of the electrode array relative to the cochlea is important for CI programming. Traditionally, this process is done by an audiologist who is blind to this information. As a result, the programming is a trial-and-error process. Multiple sessions are usually needed and can take a frustratingly long time. In recent years, our group has developed an image-guided cochlear implant programming (IGCIP) system [1]. In IGCIP, we use image processing algorithms to automatically segment the intra-cochlear anatomy, typically in the patient’s pre-operative CT image [2], and to localize the electrode array in the post-operative CT image [3-5]. Using such knowledge derived from the CT images, we can suggest programming settings that have been shown to significantly improve hearing outcomes [6-8].

The intra-cochlear anatomy includes three main structures: the scala tympani (ST), the scala vestibuli (SV), and the modiolus (MD). Segmenting them in pre-operative CTs is very challenging because these structures are small and the resolution of clinical CTs is too low to clearly show their boundaries. Currently, we use an active shape model (ASM)-based method [2]. In this method, a deformable shape model is trained from a group of anatomy examples, each represented as a set of points on the anatomical surface. These shapes are obtained by manual delineation in a group of high-resolution microCTs (μCTs) of ear specimens. To segment the intra-cochlear anatomy in a new patient’s pre-operative CT, we fit the model to the external walls of the cochlea. Once the model is fitted it permits inferring the location of the intra-cochlear structures. Although this algorithm produces satisfactory results in most cases and we are using it on a regular basis to support clinical studies, sub-optimal segmentations can still be observed.

Deep learning-based methods, especially convolutional neural networks, have achieved impressive performances in a variety of image processing tasks. Notably, a 3d U-Net [9] has been proposed to produce 3d medical image dense segmentation when it is trained using a dataset in which only sparse delineation is available. Inspired by its successful application to image segmentation and landmark detection [10] since then, we propose to use it to segment the intra-cochlear anatomy. However, due to the large number of parameters in deep neural networks, training a model from scratch requires a large set of images for which the ground truth is known. For our application, such a dataset does not exist. This is because the intra-cochlear anatomy cannot be seen in clinical CTs. It can thus not be manually localized in these images. The only way to build a ground truth data set is to acquire μCT images of temporal bone specimens and to delineate structures in these images. It is an extremely tedious process that cannot be done on a large scale.

To circumvent difficulties caused by the size of the training set required, we propose to first use the automatic segmentations generated by our ASM-based method on a large-scale image set as weak supervision to pre-train the deep learning model. Second, we fine-tune it using a small dataset for which accurate manual segmentation of the intra-cochlear anatomy is available. Similar concepts can be found in [11,12].

5.2 Data

We have collected 346 clinical head CTs, each from one unique patient, obtained prior to surgery. The ST, SV and MD of each ear which is not implanted with a CI are segmented using the ASM-based method. For each ear, we crop a region of interest using a bounding box which includes the three structures with an additional 5mm-margin at each of the anterior, posterior, superior, inferior, left and right directions. We denote this set as dataset #1. We also have another set that includes 11 clinical CT-μCT image pairs. Each pair is obtained from the same ear specimen. The voxel dimensions are 0.3mm isotropic in the clinical CTs and are 36μm or 37.6μm isotropic in the μCTs. In the μCTs, one can visually identify the boundaries of the intra-cochlear anatomy. To create the ground truth manual segmentations from μCT images, we proceed as follows: first, in each μCT image, the three anatomical structures of interest are manually drawn as binary masks. Second, we generate a mesh for each structure from the segmentation masks, using the marching cubes algorithm [13]. Third, since the meshes created directly from binary masks do not have the same number of vertices across specimens, we establish point correspondence between different specimens. The point correspondence is important for our application because (1) creating the ASM requires point correspondence. And (2) for every new patient, frequency values need to be assigned to the neural regions of the segmentation meshes for CI programming purposes. Currently this is done by direct transfer from a reference specimen's segmentation meshes where the frequencies have already been defined and such transfer also requires point correspondence. Point correspondence is established in the following way: (1) a specimen and its manual segmentation meshes are chosen as the reference specimen and reference meshes, respectively. (2) For a new specimen (target specimen), we take its μCT, perform a sequence of rigid and non-rigid registrations [14] from the reference specimen's μCT to it and deform the reference meshes accordingly. To compensate for the imperfection of automatic registration, the deformed reference meshes are manually adjusted until they are well aligned with the target specimen meshes. (3) For each point on the reference meshes, the closest point on the target specimen's meshes is found and used as a new point to re-generate the target meshes. An example of the target meshes and the reference meshes before and after manual adjustment is shown in **Fig. 5.1**.

After the ground truth segmentation masks and meshes are created in the μCT image of the specimen, the μCT is rigidly registered to its clinical CT counterpart and the manual segmentations are projected onto the clinical CT space via the same transformation matrix. We denote this dataset as dataset #2. An example of registered CT-μCT image pair is shown in **Fig. 5.2**. Dataset #2 consists of dataset #2A (5 specimens) and dataset #2B (6 specimens). The manual segmentation meshes in dataset #2A are used to build the ASM model which we use to do coarse segmentation in dataset #1, as well as to fine-tune the deep learning model. Dataset #2B is used to test our algorithm and to fine-tune the deep learning model in a leave-one-out way. Since the ASM-based method only generates meshes, for each image in both datasets, the meshes are converted to binary segmentation masks. All images and their segmentation masks are resampled to 0.1 × 0.1 × 0.1 $mm^3/voxel$.

We have another dataset to do a relatively large scale validation. We denote it as dataset #3. In dataset #3, there are 147 ear images located from clinical head CTs. For each ear image, the segmentation is generated by the ASM-based method. Because dataset #3 is a clinical head CT set from patients, there is no μCT for each of them. It means no ground truth segmentation is available for quantitative evaluation. We use this dataset to do qualitative validation. A specialist will be asked to judge the quality of the segmentations generated by the proposed method. We will talk about the details on how the evaluation is done in section 5.4.

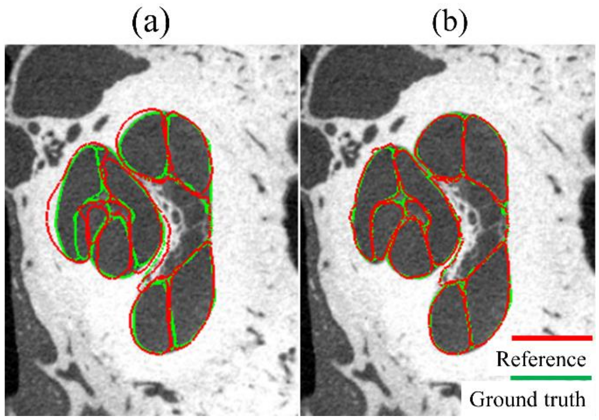

Fig. 5. 1 Target ground truth meshes and reference meshes: (a) after rigid and non-rigid registration and (b) after manual adjustment

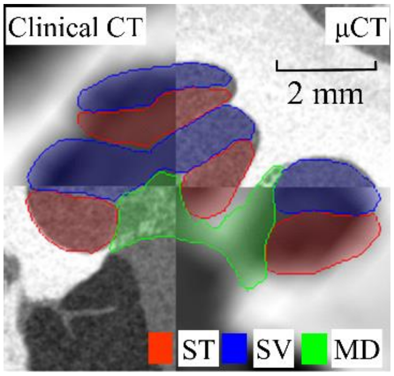

Fig. 5. 2 A pair of registered clinical CT image and µCT image in axial view. Manual delineation is shown.

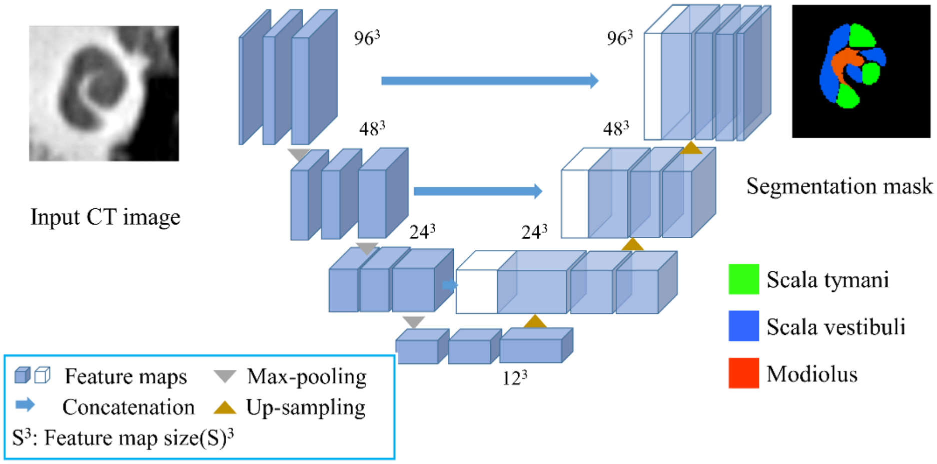

Fig. 5. 3 The neural network architecture we employ

5.3 Methods

In this work, we use the 3d U-Net as the base network architecture. It is shown in **Fig. 5.3**. The input volume size is set to 96×96×96. The channel numbers of the feature blocks are (1,32,64)-(64,64,128)-(128,128,256)-(256,256,512)-(768,256,256)-(384,128,128)-(192,64,64), from left to right. Each pair of parentheses contain channel numbers in three consecutive feature blocks at one level. The output has 4 channels, corresponding to the ST, SV, MD and the background. Suppose that $S^g \in \boldsymbol{B}^{96\times96\times96\times4}$ ($\boldsymbol{B}$ = {0, 1}) is the binary segmentation mask generated either by the ASM-based method or manual delineation for training. $S^g_{i,j,k,n}$ = 1 if voxel (i, j, k) is in the n^{th} structure and $S^g_{i,j,k,n}$ = 0 if voxel (i, j, k) is not in the n^{th} structure. Here, $n = 1{,}2{,}3{,}4$ corresponds to the ST, SV, MD and background, respectively. $\hat{S} \in \boldsymbol{R}^{96\times96\times96\times4}$ is the probability map generated by the 3d U-Net. We use the following loss function [15]:

$$loss\left(S^g, \hat{S}\right) = 1 - \frac{2\sum_{n=1}^{4} w_n \sum_{i=1}^{96}\sum_{j=1}^{96}\sum_{k=1}^{96}(S^g_{i,j,k,n} * \hat{S}_{i,j,k,n})}{\sum_{n=1}^{4} w_n \sum_{i=1}^{96}\sum_{j=1}^{96}\sum_{k=1}^{96}(S^g_{i,j,k,n} + \hat{S}_{i,j,k,n})} \tag{5.1}$$

It measures the disagreement between two segmentation masks. Here, since the numbers of voxels in each structure and the background are highly imbalanced, we use a weight w_n(n = 1,2,3,4) to ensure that the four classes contribute equally to the loss function. It is defined as

$$w_n = \frac{1}{\sum_{i=1}^{96}\sum_{j=1}^{96}\sum_{k=1}^{96} S^g_{i,j,k,n}}, (n = 1{,}2{,}3{,}4) \tag{5.2}$$

Since the image volumes vary in size, to fit them in the neural network and for data augmentation purpose, we crop 10 image cubes with a size of 96×96×96 from each located region of interest at different positions as the input. We use a two-step training strategy. In step one, we use the images and segmentation masks in dataset #1 to pre-train the model; the convolutional kernels of this network are randomly initialized. In this step, 80% of the volumes are randomly selected as the training set and the remaining 20% as the validation set. In step two, we fine-tune the model using images and the corresponding manual segmentations in dataset #2A. Since dataset #2A is small, to leverage our ground truth segmentation data, as we have mentioned, we also use images in data set #2B for fine-tuning, in a leave-one-out way. That is to say, to test each image in dataset #2B, we fine-tune the coarse segmentation model using dataset #2A and all the other images in dataset #2B. During fine-tuning, only the last two convolutional layers are set to trainable. We split the volumes into a training set (80 volumes) and a validation set (20 volumes) in the same fashion as we do in step one. In both training steps, we set the batch size to 1 and use the Adam [16] optimizer with an initial learning rate of 0.0001. A flowchart of our two-level training scheme is shown in **Fig. 5.4**.

In the testing phase, since the test volumes do not necessarily fit in the 3d U-Net, we use a sliding window with a stride of 10 voxels in each of the three orthogonal directions to crop the original test volumes and feed them to the network. The outputs are concatenated according to the spatial positions of the corresponding input cropped cubes. In overlapping regions, the responses are averaged.

In the segmentations produced by the network, noticeable false segmentations can happen. Specifically, random regions that are not part of the three anatomical structures can be classified

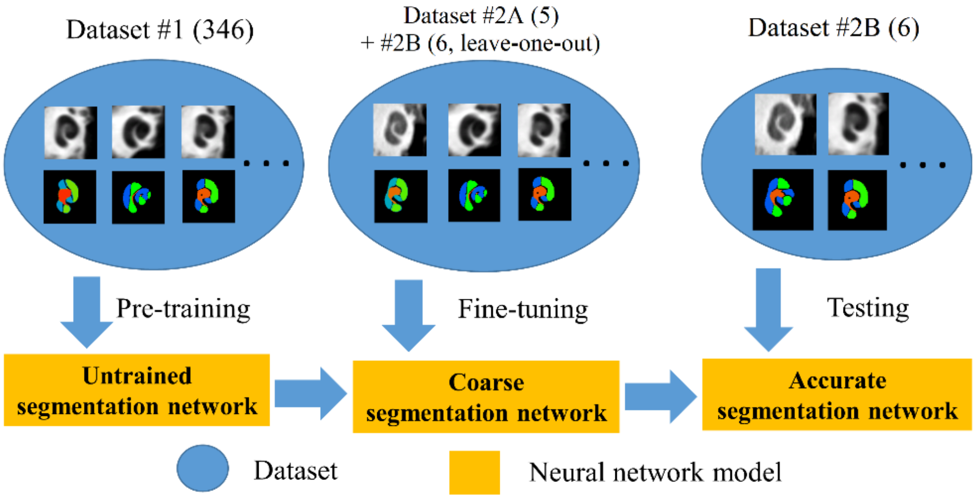

Fig. 5. 4 A flowchart of our proposed algorithm

as one of them. An example is shown in **Fig. 5.5**. Since such regions are typically small, they are simply removed by keeping only the largest connected component, i.e., the true anatomy in the segmentation mask.

For comparison purpose, we segment each image in dataset#2B using the ASM-based method (the model is created with 9 training examples that include the 5 dataset #2A). We also do segmentations using the pre-trained (1-level) model to show the effectiveness of the 2-level training scheme. For evaluation, we use three criteria. First, we compute the DICE similarity coefficients (DSC) between the automatic segmentations and the ground truth. Second, we compute the average surface distance (ASD) between the automatic segmentations and the ground truth segmentations. Specifically, suppose there are two meshes M_1 and M_2. For each point $\boldsymbol{p}$ on M_1, its distance to M_2 is

$$d(\boldsymbol{p}, M_2) = inf_{\boldsymbol{q} \in M_2} ||\boldsymbol{p} - \boldsymbol{q}||_2 \tag{5.3}$$

The mean distance from M_1 to M_2 is thus,

$$d(M_1, M_2) = mean_{\boldsymbol{p} \in M_1} d(\boldsymbol{p}, M_2) \tag{5.4}$$

We define the ASD between M_1 and M_2 as

$$ASD(M_1, M_2) = \frac{d(M_1, M_2) + d(M_2, M_1)}{2} \tag{5.5}$$

The smaller the ASD is with the ground truth, the better the segmentation quality is. Since the 3d U-Net generates image masks, we convert them into meshes to compute ASDs. Third, as we have discussed in section 5.2, because point correspondence between segmentation meshes of different subjects is important for frequency assignment and CI programming, we compute the point-to-point (P2P) errors. For the ASM-based method, we use the point correspondence encoded in the meshes. This cannot be done with the deep learning-generated meshes because point correspondence with the reference specimen mesh is not established. To establish point correspondence for these meshes, we use the ASM-generated meshes of the same specimen. We first do an iterative closest point (ICP) registration between the two meshes. For each point on the transformed ASM-generated meshes, we find the closest point on the deep learning-generated meshes. Finally, the P2P distance is computed between the ground truth meshes and the automatic segmentation meshes for each specimen, i.e.,

$$P2P(M_a, M_g) = mean_{\mathbf{0}<\boldsymbol{i}\leq V}||\boldsymbol{p}_a^i - \boldsymbol{p}_g^i|| \tag{5.6}$$

Here, M_a and M_g represent the automatic segmentation mesh and the ground truth segmentation mesh, respectively. V is the number of vertices. $\boldsymbol{p}_a^i$ and $\boldsymbol{p}_g^i$ are the i^{th} point of M_a and M_g, respectively.

5.4 Results

The DSCs, ASDs and P2P distances are shown in **Table 5.1**. As can be seen, in terms of DSC and ASD, the proposed 2-level trained model yields the best results and is substantially better than the ASM-based method and the l-level trained model. For each method, the P2P errors are substantially larger than ASDs. The three methods have close P2P distances and the proposed method does better than the other two only by a narrow margin according to this measure. We provide an explanation for this using **Fig. 5.6**. For a point P_g on the ground truth mesh, its corresponding points on the ASM-generated mesh and the deep learning-generated mesh are denoted as P_{asm} and P_{dl}, respectively. The P2P error can be seen as an aggregation of the error along the normal direction, denoted as the "normal error" and the error in the tangent direction, denoted as "tangential error". The tangential error is due to the shape rotation happening in the ASM algorithm, making the P2P error substantially larger than APD, even if the rotation is slight. Since point correspondence for the deep learning-generated meshes is established using the ASM-generated meshes, they have similar tangential errors. P2P errors are substantially larger than ASD errors for the ASM method, which indicates that the tangential error is the dominant error component for this approach. Tangential error caused by the ASM cannot be eliminated from the P2P deep learning method, which explains why the P2P error difference between ASM and the proposed method is not as large as the difference observed with DSC and APD. In **Table 5.2**, we show the *p* values of paired t tests when comparing different methods in segmenting each structure. For the DSCs and APDs the proposed method is statistically significantly different from the ASM method and the 1-level model.

	ST			SV			MD		
	ASM	1-level	Proposed	ASM	1-level	Proposed	ASM	1-level	Proposed
Mean DSC	0.76	0.77	0.87	0.75	0.71	0.86	0.60	0.63	0.83
Std of DSC	0.012	0.025	0.016	0.019	0.058	0.017	0.049	0.026	0.017
Mean ASD*	0.15	0.14	0.077	0.14	0.15	0.084	0.21	0.18	0.09
Std of ASD	0.027	0.017	0.016	0.021	0.028	0.017	0.039	0.01	0.027
Mean of P2P	0.55	0.58	0.52	0.56	0.59	0.54	0.58	0.57	0.53
Std of P2P	0.2	0.22	0.23	0.22	0.25	0.25	0.13	0.13	0.14

*: All distance values in this table are in *mm*

Table 5. 1 DSCs, ASDs and P2P distances of the segmentations using the three methods

	DSC			ASD			P2P		
	ST	SV	MD	ST	SV	MD	ST	SV	MD
Proposed vs ASM	**0.0002**	**0.0002**	**0.0002**	**0.014**	**0.012**	**0**	0.052	0.083	0.059
1-level vs ASM	0.13	0.077	0.17	0.68	0.12	0.26	**0.013**	0.055	0.078
1-level vs proposed	**0***	**0.0001**	**0.0001**	**0**	**0.0001**	**0.0001**	**0.0003**	**0.0003**	0.12

*: Any value that is below 0.0001 is shown as 0. A value in bold indicates statistical significance (<0.05)

Table 5. 2 Values of *p* using paired t tests for different comparisons

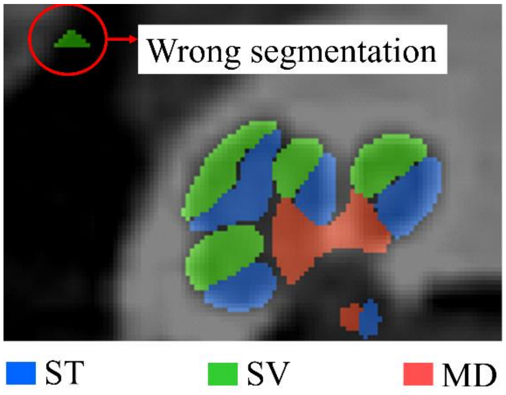

Fig. 5. 5 An example of wrong segmentation

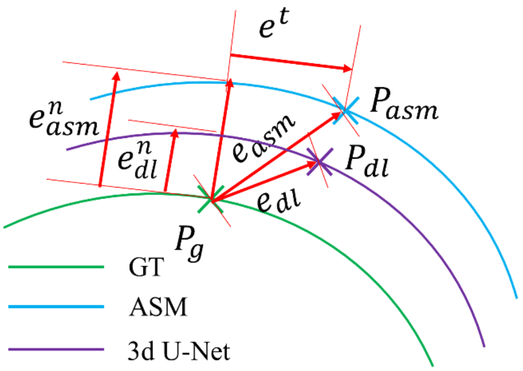

Fig. 5. 6 A sketch of the surface distances and P2P distances between the ground truth mesh and the automatically-generated meshes: $\boldsymbol{e}^{\boldsymbol{n}}_{\boldsymbol{asm}}$and $\boldsymbol{e}^{\boldsymbol{n}}_{\boldsymbol{dl}}$ are the normal errors for the ASM and the deep learning method, respectively. $\boldsymbol{e}^{\boldsymbol{t}}$ is the tangential error. $\boldsymbol{e}_{\boldsymbol{asm}}$ and $\boldsymbol{e}_{\boldsymbol{dl}}$ are the P2P errors of the ASM and the deep learning method, respectively.

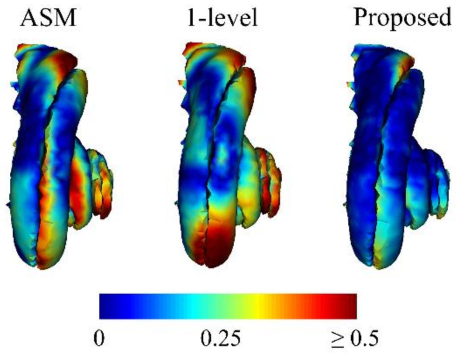

Fig. 5. 7 An example of surface distances from the ground truth meshes to automatic segmentation meshes (in *mm*), displayed on the ground truth

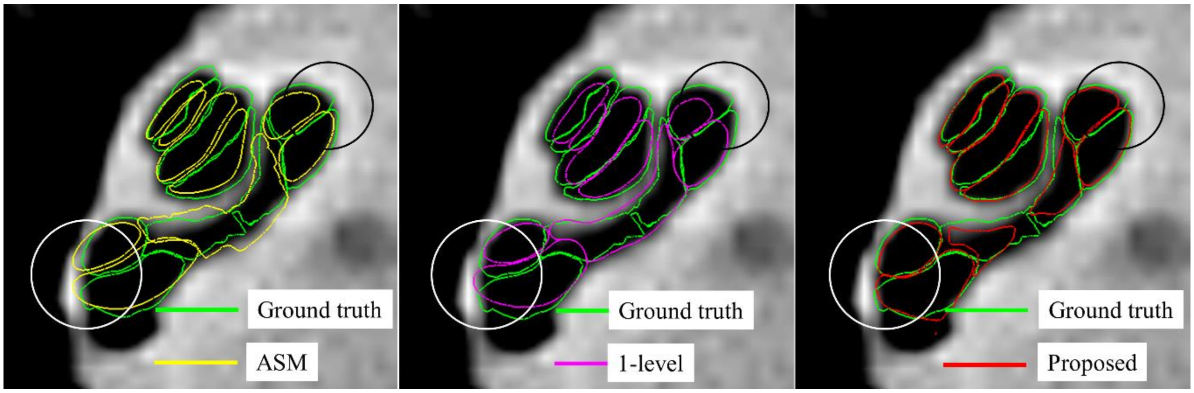

Fig. 5. 8 An example of the results obtained with the three methods. Areas with noticeable difference are encircled

A representative example illustrating the point-wise distances from the ground truth mesh to the automatic segmentations is shown in **Fig. 5.7**. An example showing the segmentations superimposed on the image is shown in **Fig. 5.8**. As can be seen, the proposed method produces a segmentation that is closer to the ground truth than the ASM-based method and the 1-level model. Areas in which the proposed method produces noticeably better results are encircled.

We also do a large scale qualitative evaluation of our proposed method in dataset #3. Specifically, for each ear image, we segment the intra-cochlear anatomy using our proposed method and the result is compared with the ASM-generated segmentation. An expert is asked to rank the quality of the two segmentations for each image. The rater judges which segmentation is better or they are equally good. To avoid evaluation bias, for each image, the order of the two segmentations is randomized and anonymized to the rater. We show the evaluation result in **Table 5.3**. As we can see, the proposed method produces better segmentations in more than 70% of the ear images and gives segmentations that are at least as good as the ASM-based method in ~80% of the images.

The method that produces better segmentation	ASM	Equally good	Proposed method
Number of ear images	29	11	107

Table 5. 3 The result of the qualitative evaluation

We do an analysis of the images in which the deep learning does not generate segmentations that are as good as the ASM-based method. It shows that the majority of them are either of lower quality, including being noisy and blurry, or of abnormal orientation. The detailed analysis is shown in **Table 5.4** and two examples are shown in **Fig. 5.9**. This work is still ongoing and the ear images that have abnormal orientation will be corrected and the segmentation will be re-generated. Further, to improve the robustness of our method, we propose (1) to re-train our model by adding simulated degraded images in the training set, and (2) to explore other deep architecture, for example, the conditional generative adversarial networks (cGAN), for the purpose of introducing anatomical shape constraint on the generated segmentation.

Image type	Number
Blurry image	12
Noisy image	3
Blurry+noisy image	1
Abnormally oriented image	5
Blurry+abnormally orientated image	3
Image with prosthetic stapes	1
Normal image	4

Table 5. 4 Categorization of the images in which the proposed method does not produce segmentations that are as good as the ASM-based method.

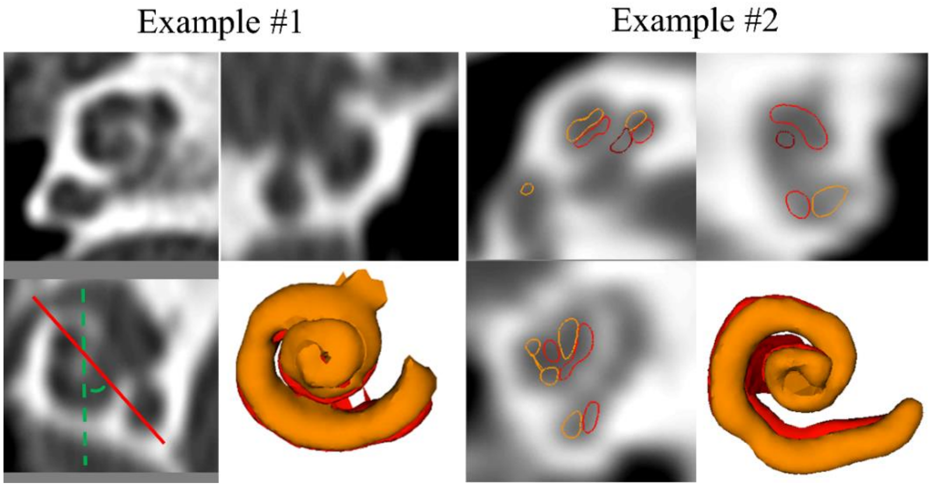

Fig. 5. 9 Two examples of bad segmentation produced by the proposed method. The ear in the first image has an abnormal orientation, which has an angle of ~45 degree with the standard orientation. The second image is blurry than normal images.

5.5 Conclusions and discussions

In this work, we propose to use a two-level training strategy to accurately segment the intra-cochlear anatomy in ear specimen clinical CTs using a very limited set of training images for which manually delineated ground truth segmentations are available. To circumvent the training data paucity problem, we use automatically generated segmentations obtained with our previous ASM-based method to train a weak segmentation model and use a small dataset for which ground truth segmentations are available to fine-tune it. The results show that the proposed method is more accurate than the ASM-based method in terms of DICE and ASD evaluated in this data set. To assess the robustness of the proposed method, we conduct an evaluation study on a relatively large dataset (147 ear images). Segmentations produced by the ASM-based approach and the proposed method are compared by an expert. The result shows that the proposed method have better performance in the majority of the images. An analysis of the images in which the proposed method does not produce segmentations that are as good as the ASM-based method shows that the abnormal image orientation and image degradation could be the main reasons. The images which have abnormal orientations will be corrected and the segmentations will be re-generated and re-evaluated. As our future work, we aim to improve the robustness of the proposed method by (1) simulating degraded images using images of high quality in the training stage and (2) exploring other deep learning architectures, for example, cGAN to introduce discrimination loss as an atomical shape constraint.

References

[1]. J. H. Noble, R. F. Labadie, R. H. Gifford and B. M. Dawant, "Image-guidance enables new methods for customizing cochlear implant stimulation strategies." IEEE Transactions on Neural Systems and Rehabilitation Engineering 21.5 (2013): 820-829.

[2]. J. H. Noble, R. F. Labadie, O. Majdani, and B. M. Dawant, "Automatic segmentation of intracochlear anatomy in conventional CT," IEEE Transections on Biomedical Engineering, vol. 58, pp. 2625-32, 2011.

[3]. J. H. Noble and B. M. Dawant, "Automatic graph-based localization of cochlear implant electrodes in CT," Medical Image Computing and Computer Assisted Intervention, pp. 152-159, 2015.

[4]. Y. Zhao, B. M. Dawant, R. F. Labadie, and J. H. Noble, "Automatic localization of cochlear implant electrodes in CT," Medical Image Computing and Computer Assisted Intervention, pp. 331-338, 2014.

[5]. Y. Zhao, B. M. Dawant, and J. H. Noble, "Automatic localization of cochlear implant electrodes in CTs with a limited intensity range", Proc. SPIE 10133, Medical Imaging 2017: Image Processing, 101330T, 2017

[6]. J. H. Noble, R. H. Gifford, A. J. Hedley-Williams, B. M. Dawant and R. F. Labadie. "Clinical evaluation of an image-guided cochlear implant programming strategy." Audiology and Neurotology 19.6 (2014): 400-411

[7]. J. H. Noble, A. J. Hedley-Williams, L. Sunderhaus, B. M. Dawant, R. F. Labadie, S. M. Camarata and R. H. Gifford. "Initial results with image-guided cochlear implant programming in children." Otology & Neurotology, 37.2 (2016): e63.

[8]. R. F. Labadie, J. H. Noble, A. J. Hedley-Williams, L. W. Sunderhaus, B. M. Dawant, and R. H. Gifford, "Results of Postoperative, CT-based, Electrode Deactivation on Hearing in Prelingually Deafened Adult Cochlear Implant Recipients," Otology & Neurotology, vol. 37, pp. 137-45, 2016.

[9]. O. Çicek, A. Abdulkadir, S. S. Lienkamp, T. Bronx and O. Ronneberger. "3D U-Net: learning dense volumetric segmentation from sparse annotation." Medical Image Computing and Computer Assisted Intervention, pp. 424-432, 2016.

[10]. D. Zhang, J. Wang, J.H. Noble, B.M. Dawant. "Accurate Detection of Inner Ears in Head CTs Using a Deep Volume-to-Volume Regression Network with False Positive Suppression and a Shape-Based Constraint". Medical Image Computing and Computer Assisted Intervention, pp. 703-711, 2018

[11]. D. Mahajan, R. Girshick, V. Ramanathan and K. He. "Exploring the Limits of Weakly Supervised Pretraining." arXiv preprint arXiv:1805.00932(2018).

[12]. Y. Huo. Z. Xu, K. Abound, P. Parvathaneni, S. Bao, C. Bermudez, S.M. Resnick, L.E. Cutting, B.A. Landman, et al. "Spatially Localized Atlas Network Tiles Enables 3D Whole Brain Segmentation from Limited Data". Medical Image Computing and Computer Assisted Intervention, pp. 698-705. 2018

[13]. W. E. Lorensen and H. E. Cline. "Marching cubes: A high resolution 3D surface construction algorithm." ACM siggraph computer graphics. Vol. 21. No. 4. ACM, 1987

[14]. G. K. Rohde, A. Aldroubi, and B. M. Dawant. "The adaptive bases algorithm for intensity-based non-rigid image registration." IEEE transactions on medical imaging 22.11 (2003): 1470-1479.

[15]. C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M.J. Cardoso. "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations". In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (pp. 240-248). Springer, Cham. 2017

[16]. D. P. Kingma, and J. Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980(2014).

Chapter VI

SELECTING ELECTRODE CONFIGURATION FOR IMAGE-GUIDED COCHLEAR IMPLANT PROGRAMMING USING TEMPLATE MATCHING

Dongqing Zhang, Yiyuan Zhao, Jack H. Noble, Benoit M. Dawant

Department of Electrical Engineering and Computer Science, Vanderbilt University, 2201 West End Ave, Nashville, USA, 37235

This chapter has been published in Journal of Medical Imaging 5.2 (2017): 021202.

Abstract

Cochlear implants (CIs) are neural prostheses that restore hearing using an electrode array implanted in the cochlea. After implantation, the CI processor is programmed by an audiologist. One factor that negatively impacts outcomes and can be addressed by programming is cross-electrode neural stimulation overlap (NSO). In the recent past, we have proposed a system to assist the audiologist in programming the CI that we call Image-Guided CI Programming (IGCIP). IGCIP permits using CT images to detect NSO and recommend deactivation of a subset of electrodes to avoid NSO. In an ongoing clinical study, we have shown that IGCIP significantly improves hearing outcomes. Most of the IGCIP steps are robustly automated but electrode configuration selection still sometimes requires expert intervention. With expertise, Distance-Vs-Frequency (DVF) curves, which are a way to visualize the spatial relationship learned from CT between the electrodes and the nerves they stimulate, can be used to select the electrode configuration. In this work, we propose an automated technique for electrode configuration selection. A comparison between this approach and one we have previously proposed shows that our new method produces results that are as good as those obtained with our previous method while being generic and requiring fewer parameters.

6.1 Introduction

Cochlear implants (CIs) are one of the most successful neural prosthetics that have been developed over the past few decades [1]. CIs are used to treat patients with severe-to-profound hearing loss. In surgery, an electrode array with 12 to 22 contacts is implanted in the cochlea. A processor worn behind the ear is connected to the array and sends signals to individual electrodes to stimulate the spiral ganglion (SG) nerves, i.e., the auditory nerves that are frequency-mapped within the cochlea. After implantation, the CI is programmed by an audiologist. Programming includes selecting contacts to deactivate and assigning sound frequency ranges and stimulation levels to each contact. Recent studies have suggested that hearing outcomes with CIs are correlated with the spatial relationship between the electrode array and the nerves it stimulates [2-6]. But, because the electrode array is blindly inserted into cochlea, its final position is generally not known. Recently, we have developed a series of algorithms that permit localizing both inner ear structures and the electrode array in CT images [7-14]. We have shown that this information can be used to estimate how much overlap exists between neural areas each contact activates. **Fig. 6.1** illustrates the channel interaction or cross-electrode neural stimulation overlap phenomenon which is known to negatively affect hearing outcomes. We have shown in clinical studies that for both adults and pediatric populations, hearing outcomes improve significantly when stimulation overlap is detected and the *configuration*, i.e., the set of active electrodes, is adjusted to try to reduce it [8,9].

To estimate the amount of cross-electrode neural stimulation overlap, we have developed what we refer to as distance-vs-frequency curves (DVFs). These are 2-D plots as shown in **Fig. 6.2** that capture the patient-specific spatial relationship between electrodes and SG nerves. In this figure, the *x* axis is the position along the length of the SG nerves in terms of characteristic frequency, i.e., the sound frequency at which the local auditory nerves are activated in natural hearing, and the *y* axis is the distance to the SG nerves. The number of 2-D roughly parabolic curves in this plot is equal to the number of contacts in the electrode array. Each curve is labeled with the electrode number. The height of each individual DVF curve thus represents the distance from the corresponding electrode to the SG nerve pathways. An example is shown in **Fig. 6.2**. Here, the distance from electrode #2 to the SG nerves is larger than that from electrodes #1 and

#3. We assume that electrodes that are farther away from the SG nerves activate a broader region than the electrodes that are close to the nerves. Here, this means that electrode #2 will interfere with electrodes #1 and #3. Electrodes #5, #6, and #7 are all far from the nerves and close to each other. It is thus also likely that they interfere with each other. Looking at the plot, an experienced user would deactivate electrodes #2 and #6. In [15] and its extension [16], Zhao et al. have proposed a method that captures the heuristics used by experts and translates them into a set of rules. Weights associated with each rule are estimated from a training set and we have demonstrated that this method leads to satisfactory results but it requires estimating the values of more than 10 parameters.

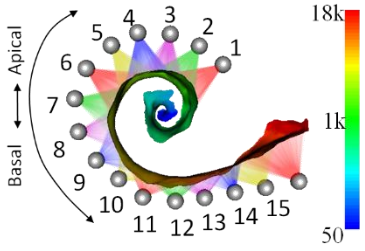

Fig. 6. 1 An illustration of the electrode interaction effect: gray spheres are electrodes and the spiral-shaped surface is the neural region. Colored cones indicate neural areas the corresponding electrodes are stimulating.

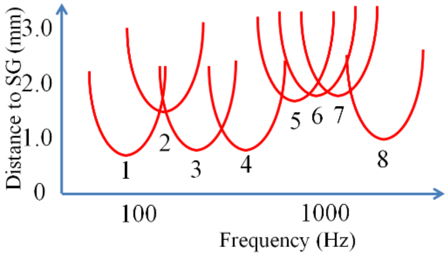

Fig. 6. 2 A synthetic 8-electrode DVF example

As the size of our DVF library increase, we hypothesizes that it would become possible to find DVFs that are highly similar to each other. If this sis the case, configurations that have been selected by an expert could be applied to new cases for which the configuration is unknown. But, with experience, we observe that matching complete DVFs is difficult and would require a very large library to be successful. We addresss this issue by decomposing entire DVFs into what we call patches and matching patches rather than entire DVFs. We then impose constraints to make the electrode configuration patterns compatible between subsequent patches. Our approach is tested on 20 cases for each of the three main CI manufacturers and we compare the configurations generated by our method with those generated by Zhao's method as well as with configurations produced manually. The results we have obtained show that our method produces configurations that are generally as good as those obtained by Zhao's method but our method requires far fewer parameters and could potentially improve further with larger library sizes.

6.2 Methods

CIs are manufactured by three main companies: Med-El (MD) (Innsbruck, Austria), Advanced Bionics (AB) (Valencia, California), and Cochlear (CO) (New South Wales, Australia). Implants distributed by these companies differ mainly by the number of electrodes in the array. To accommodate these differences, we create three libraries, one for each array type. In the study presented herein, the data set contains 58, 43 and 152 DVFs for MD, AB and CO arrays, respectively. For each DVF in the libraries, the manual configuration, which we consider to be the ground truth, has been defined by JHN who is primarily responsible for creating configurations in the clinical studies we are conducting to evaluate our IGCIP approach. Each of these ground truth configurations have been used clinically. Both DZ and YZ have been trained by JHN to produce and evaluate deactivation solutions but did not produce deactivation solutions for the DVFs used in this study. They will be referred to as rater 1 and rater 2 in the remainder of this article. Each DVF in the libraries is decomposed into a series of patches that are defined as 5 consecutive electrode curves. Two consecutive patches overlap by 4 electrodes. Since the MD, AB and CO implants have 12, 16 and 22 electrodes, respectively, this leads to 8, 12 and 18 patches per implant for the first, second, and third implant type. When computing the configuration automatically we use the same method for each CI type but we adjust the number of electrodes to match the number of electrodes in the array of interest. When matching a new patient's DVFs to our libraries we also decompose them into patches. The matching process involves two steps. First, individual patches are matched and then patch-to-patch consistency is enforced. A flowchart of our algorithm is shown in **Fig. 6.3**. Without loss of generality, the description of our method is based on DVFs produced for CO implants.

To match individual patches we define a feature vector as follows. For each electrode, eight features illustrated in **Fig. 6.4** are computed. These are the minimum distance to the SG (1), the

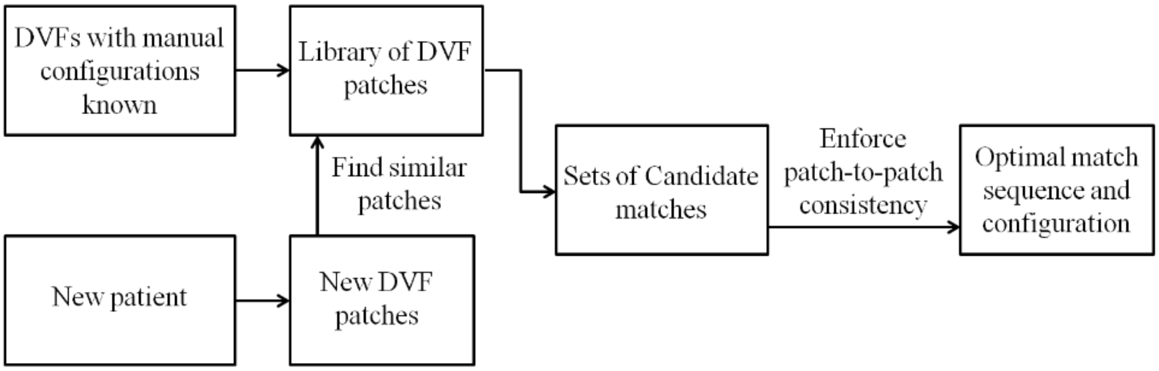
DVFs with manual configurations known
Library of DVF patches
Find similar patches
Sets of Candidate matches
Enforce patch-to-patch consistency
Optimal match sequence and configuration
New patient
New DVF patches

Fig. 6. 3 The flowchart of our automatic electrode configuration algorithm

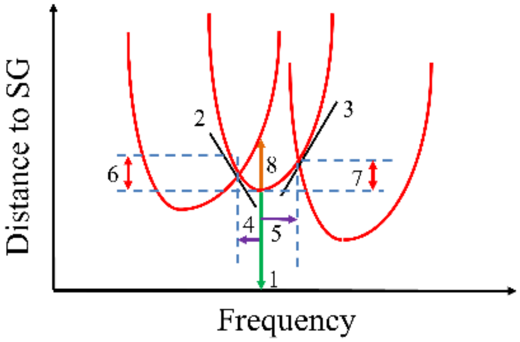

Fig. 6. 4 Illustration of features. Feature (1) is shown in green arrow, (2): black lines, (3): purple arrows, (4): red arrows, (5): brown arrow

slopes of the curve at the intersection points with its left (2) and right (3) neighboring curves, the frequency ranges from the minimum to the two intersection points (4-5), the difference in the distance to the SG at the two intersection points (6-7), and the distance along a vertical line passing through the minimum between this minimum and the first intersection with one of the neighboring curves (8). For each individual feature, first and second-order statistics are computed for all electrodes in the library. Each feature is subsequently normalized by subtracting its mean and dividing by its standard deviation. The feature vector for a patch is obtained by stacking the five individual feature vectors, leading to a 40-D feature vector. When determining the feature vector for new test patches not in the library, the new feature vector is normalized using the same mean and standard deviation computed from the library. The similarity score *s* between patches P_i and P_j with feature vectors $\boldsymbol{F_i}$ and $\boldsymbol{F_j}$ is then computed as

$$s(P_i, P_j) = \frac{\boldsymbol{F_i} \cdot \boldsymbol{F_j}}{|\boldsymbol{F_i}||\boldsymbol{F_j}|} \tag{6.1}$$

Fig. 6.5 shows two exemplar patient patches and a series of library patches with their similarity score *s*. This figure shows that our feature vector leads to a similarity metric that is reasonable, i.e., high scores correspond to patches that are visually similar.

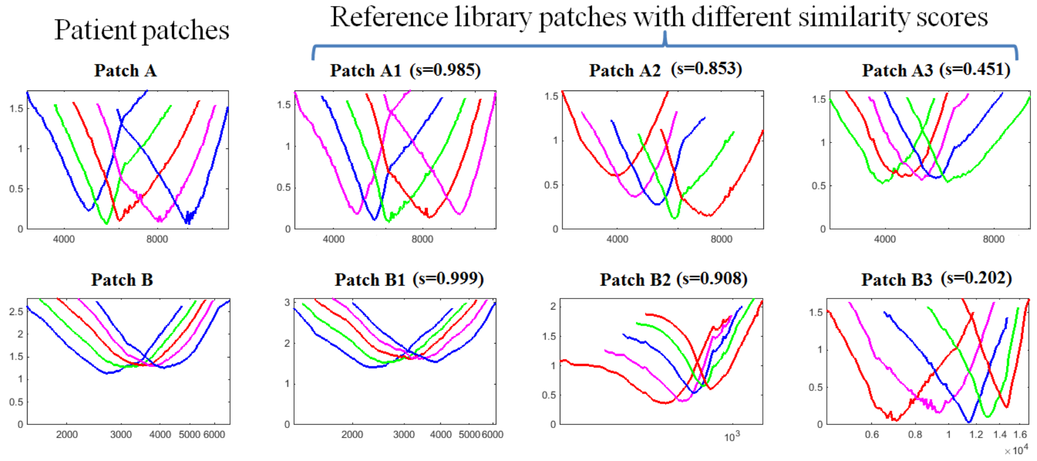

Fig. 6. 5 Two patient patches, A and B, each with 3 library patches shown on the same row. The value of the similarity measure *s* increases as the visual similarity between patches increases.

With this similarity metric, a straightforward way to deactivate electrodes would be to find the best match for each test patch and to assign its central electrode's deactivation state to the test patch's central electrode. However, this oversimplifies the problem. When experts manually select the 'on/off' state of one electrode, they not only consider the state of the central electrode, they also consider its consistency with the state of neighboring electrodes. An example can be seen in **Fig. 6.2**. Here, as discussed earlier, since electrode #6 is close to electrodes #5 and #7 but far away from the nerves, it is likely to interfere with electrodes #5 and #7. A good choice to reduce interaction would be to deactivate it, but only if #5 and #7 are kept active. An alternative would be to deactivate #5 and #7. In that case, electrode #6 should be left on to avoid a big stimulation gap. However, it is possible that the best matches for each of #5, #6 and #7 all suggest turning off their middle electrodes but leaving neighbors on. This leads to deactivating all three electrodes. If this happens, electrode interaction is eliminated but too many members are turned off from the already limited number of electrodes in this region and a stimulation gap is left between electrodes #4 and #8. This will result in excessive compression of the sound frequency spectrum, i.e., fewer independent frequency channels are used than could be used and some neural areas may not be stimulated at all. The patient will thus have sub-optimal signal quality. So, without enforcing deactivation pattern consistency between patches, the strategy mentioned above would likely lead to bad configurations.

To tackle this problem we impose patch-to-patch consistency as follows. We denote the i^{th} (i = 1, 2, 3,.., 18) test patch as T_i, For T_i, instead of only selecting the most similar match, we store *k* matches with the highest similarity scores, forming a candidate set, denoted as $C_{i,j}$ with similarity scores $s_{i,j}$ (*j* = 1, 2, 3,…, *k*). For each test patch our final goal is to find a match in the candidate set that is both similar to test patch and leads to good patch-to-patch consistency. Specifically, for two neighboring patches, we define the inconsistency D using the number of ‘on/off’ state disagreements in their two middle overlapping electrodes. With ‘1’ meaning ‘electrode on’ and ‘0’ meaning ‘electrode off’, suppose patches T_1 and T_2 in one set of DVFs consist of electrodes #1-#5 and electrodes #2-#6, respectively. Suppose also that for each patch we found a match, $C_{1,1}$ and $C_{2,1}$, respectively. Finally, suppose that the deactivation patterns of these matches are $a_{1,1}$ = [1, 1, 0, 0, 1] and $a_{2,1}$ = [1, 0, 1, 1, 0]. The electrodes they both include are #2-#5 and the two middle electrodes are #3 and #4. We define the disagreement as the Hamming distance [17] between the two state vectors for electrodes #3 and #4. In our example it would be the Hamming distance between [0, 0] and [0, 1], i.e., $D(a_{1,1}, a_{2,1}) = 1$. The bigger the disagreement, the higher the inconsistency. Searching for the best global deactivation strategy involves minimizing a cost function containing both patch-match dissimilarity and patch-to-patch inconsistency terms. The simplest solution would be to do an exhaustive search but that would be time-consuming. Instead, we have formulated it as searching for the shortest path in a directed graph. As is shown in **Fig. 6.6**, every node in the graph is a match. Matches in the same column are the candidates for the same test patch. We put two extra nodes in the graph as the starting and ending nodes. Arbitrary similarity scores are assigned to them and their disagreements with neighboring matches are set to 0. Edges only exist from patch $C_{i,j}$ to $C_{i+1,l}$, $(i = 0, 1, 2, \dots, 18$ and $j, l = 1, 2, \dots, k)$ and the cost along an edge is defined as

$$\mathrm{cost}(C_{i,j}, C_{i+1,l}) = \lambda[1 - s(T_{i+1}, C_{i+1,l})] + (1 - \lambda)D(a_{i,j}, a_{i+1,l}) \tag{6.2}$$

Here, we use the first term $1 - s(T_{i+1}, C_{i+1,l})$ to represent the patch-match dissimilarity between T_{i+1} and $C_{i+1,l}$. $a_{i,j}$ and $a_{i+1,l}$ are the deactivation patterns of patch $C_{i,j}$ and $C_{i+1,l}$. The second term $D(a_{i,j}, a_{i+1,l})$ is the deactivation pattern inconsistency. λ $(0 < \lambda < 1)$ balances these two terms. The shortest path from the starting node to the ending node gives the optimal match sequence. To find the shortest path, we used Dijkstra's algorithm [18].

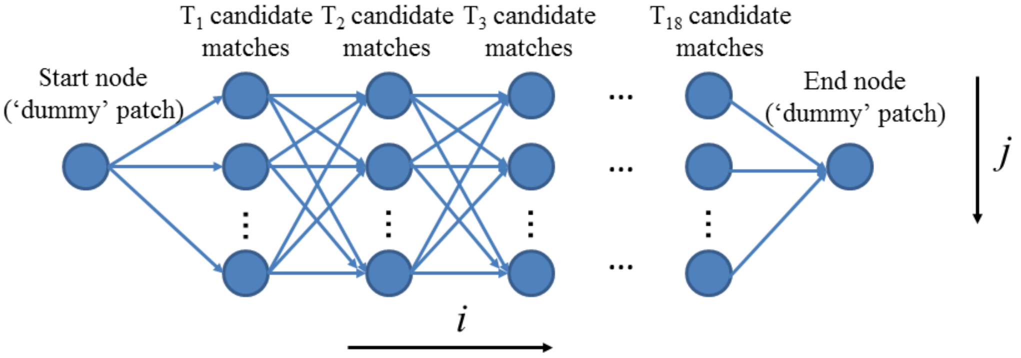

Fig. 6. 6 The directed graph we construct in order to find the optimal match sequence

After the optimal match sequence is found, the 'on/off' state of each optimal match's central electrode is used for the corresponding contact in the new patient's DVFs. The states of the first two and of the last two electrodes are assigned to be the same as the states of these electrodes in the first and last matches of the sequence, respectively.

Note that with the proposed algorithm, the final electrode configurations depend on the value of two parameters, k and λ. The next section will present the results we have obtained. In Section 4 we discuss how these parameters were chosen as well as the sensitivity of our results to these parameter values.

6.3 Results

At the time of writing, we have evaluated our technique on DVFs generated from images of 20 patients/brand for a total of 60 patients for whom manual solutions are available. We automatically generated the electrode configurations using the method proposed herein and Zhao's method. To evaluate the solutions, human expert 1 (DZ) and expert 2 (YZ) were asked to compare automatic and manual configurations as well as automatic and control configurations. The control configurations were generated by manually producing a configuration that is not "acceptable" but "close" to acceptable for all test subjects. This was done by DZ who changed the on/off state of 1-3 electrodes. All the validation solutions were created before the comparative study started to minimize the chance that particular solutions could be remembered. This control configuration was added to avoid evaluation bias, i.e., to avoid having the human expert biased towards rating solutions as acceptable/comparable if they knew that they were always presented with a human-generated and a computer-generated solution. Specifically, for each subject, three comparisons were performed: the manual configuration versus automatic configuration #1 (generated by our proposed method), the manual configuration versus automatic configuration #2 (generated by Zhao's method) and the manual configuration versus the control configuration. For each comparison the experts were presented with the two configurations blind to their origin and the ordering was randomized. For each comparison the two experts were asked to rank the solutions in terms of quality and to rate whether each was acceptable or not. "Acceptable" means that in the expert's opinion, the configuration can be used for CI programming and is likely to lead to hearing outcomes that are comparable to those achieved using the best possible configuration.

The results are shown in **Fig. 6.7**. According to expert 1 (DZ), across all 60 subjects, Zhao's method generated "better than the manual" configurations for 5 subjects, "at least equivalent as the manual" ("Better than the manual" + "Replicate the manual" + "Equally good as the manual") configurations for 35 subjects and unacceptable configurations for 2 subjects. Our proposed method generated "better than the manual" configurations for 4 subjects, which was one less than Zhao's method, "at least equivalent as the manual" configurations for 40 subjects, 5 more than Zhao's method, and unacceptable configurations for the same number of subjects as Zhao's method. According to expert 2 (YZ), Zhao's method generated "better than the manual" configurations for 10 subjects, "at least equivalent as the manual" configurations for 40 subjects, and unacceptable configurations for 5 subjects. Our proposed method generated "better than the manual" configurations for 6 subjects, which was 4 less than Zhao's method, "at least equivalent as the manual" configurations for 46 subjects, 6 more than Zhao's method and unacceptable configurations for only 3 subjects, 2 less than Zhao's method. For statistical analysis, we performed a McNemar mid-p test between groups of configurations. The McNemar mid-p test is a statistical test used for binary matched-pairs data [19]. We found that the differences between acceptance rates of our method and control configurations are statistically significant for both expert 1 ($p = 1.8 \times 10^{-13}$) and expert 2 ($p = 4.4 \times 10^{-16}$). We did not find statistically significant difference when comparing the acceptance rate of our method across the two raters ($p = 0.63$). Neither did we find statistically significant difference when comparing the acceptance rate of Zhao's method across the two raters ($p = 0.22$) and acceptance rate of control configurations across the two raters ($p = 0.21$). To summarize, our proposed method generated more configurations of high quality (at least equivalent as the manual configurations) and has a higher acceptance rate than Zhao's method, although it does not outperform the manual configurations as often as Zhao's method. McNemar mid-p tests between our method and Zhao's method in terms of the configuration numbers reaching different ranks, e.g., differences in the number of configurations deemed “better than the manual configuration” between the two methods by each rater are shown in **Table 6.1**. None of them are found to be significantly different.

p-value	"Better than manual" rate	"At least equivalent as manual" rate	Acceptance rate
Rater #1	0.73	0.5	1
Rater #2	0.3	0.23	0.51

Table 6. 1 *p*-values of McNemar mid-*p* tests between our method and Zhao’s method at different ranks

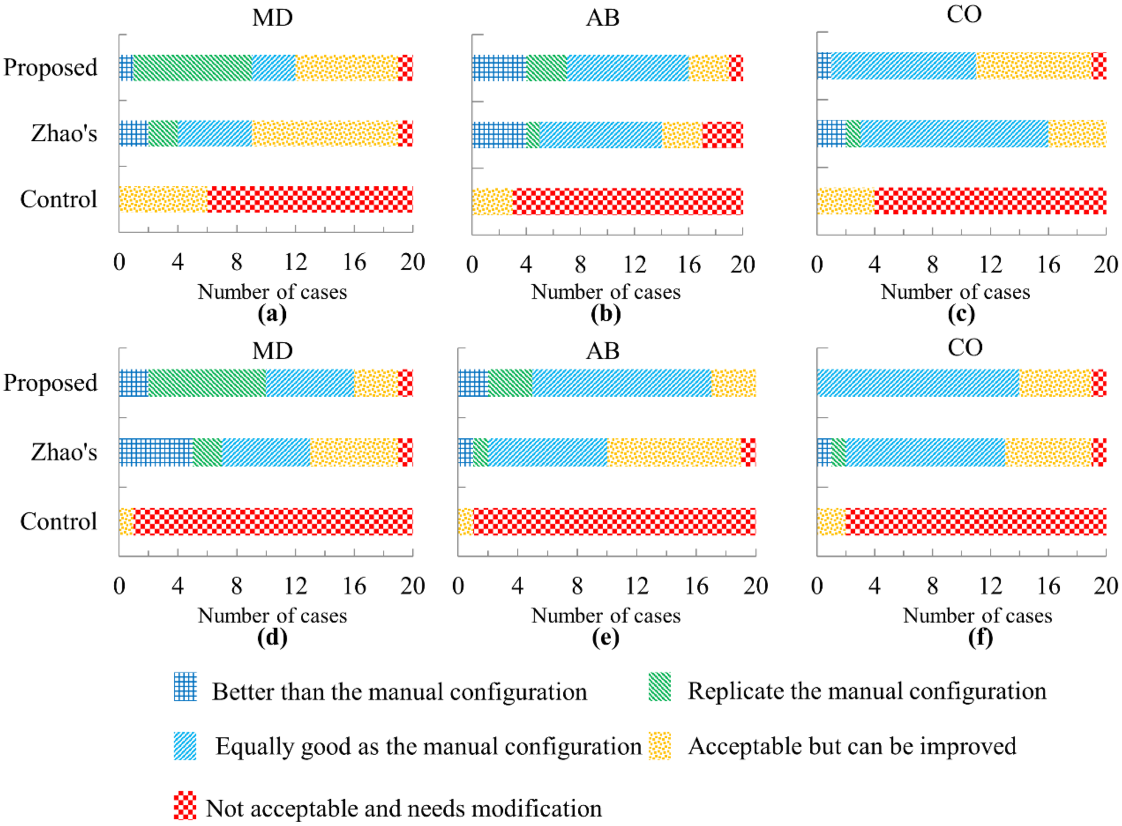

Fig. 6. 7 Validation study results. (a)-(c) The results of validation studies performed by expert 1 (DZ) and (d)-(f) expert 2 (YZ) on our proposed automatic, Zhao’s automatic and control electrode configurations

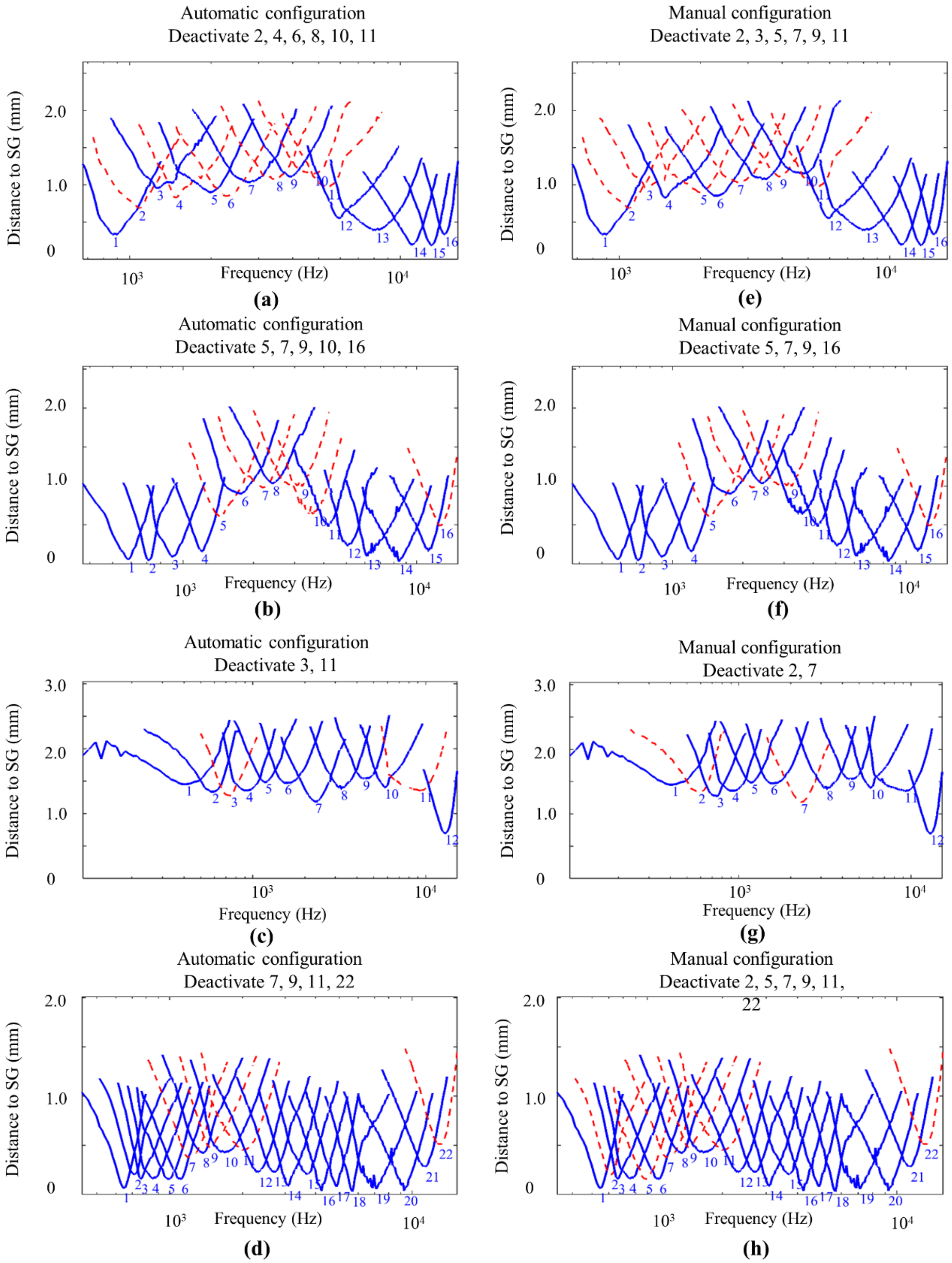

Fig. 6. 8 Visualization of automatically selected (a–d) using our proposed method and corresponding manual (e–h) electrode configurations for several cases. An automatic AB plan that was judged as better than the manual plan is shown in (a). An automatic AB plan judged to be equivalently good is shown in (b). An automatic MD plan judged as acceptable is shown in (c). An automatic CO plan that was judged as not acceptable is shown in (d)

In **Fig. 6.8**, we show the DVFs for automatically determined electrode configurations for four example cases. The blue curves represent DVFs for electrodes that are kept active in the configuration and the red dashed curves represent DVFs for electrodes that are deactivated. In **Fig. 6.8**(a), a result for an AB case is identified as better than the manual configuration shown in **Fig. 6.8(**e) because electrode #10 is likely to interfere with electrode #12 in the manual configuration and it is turned off in the automatic one. In **Fig. 6.8**(b), the automatic configuration for an AB case is rated as equally good as the manual configuration. The automatic configuration deactivates electrode #10 but the manual configuration keeps it. Though deactivating electrode #10 reduces some interaction, because the gap between electrode #8 and #11 is fairly big, it will be offset by an increase in frequency spectrum compression artifacts we mentioned above. In **Fig. 6.8**(c), the automatic configuration for an MD case is identified as only acceptable because turning off electrode #11 will probably result in slight frequency spectrum compression artifacts but it is not a serious issue here. In **Fig. 6.8**(d), the automatic configuration for a CO case is not acceptable because there will be serious electrode interaction between electrode #1, #2 and #3 if electrode #2 is not deactivated.

We also conducted an experiment to evaluate the sensitivity of our proposed method to the size of the library. The CO library is the only one that is large enough to conduct such a study. For this array type, we used 10, 20, 40, 80 DVFs randomly selected from the complete set of 152 DVFs. We then produced four different configurations, one for each library size and asked expert 1 to evaluate them. In this experiment, we did not compare our method to Zhao's method but a control configuration was included for each subject. This results in 5 comparisons. Each comparison involves the manual configuration versus one of the 4 automatic or the control configurations. Again, the trials are presented in a blinded fashion with the ordering randomized.

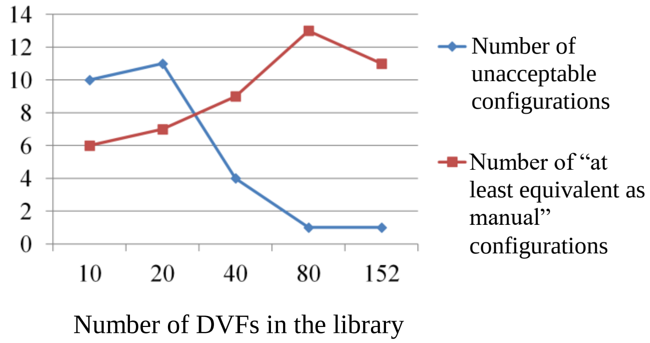

Fig. 6. 9 Evaluation results when different library sizes are used

Fig. 6.9 shows a plot of the number of "at least equivalent as the manual" configurations and "unacceptable" configurations for different library size. The result obtained with the entire library (152 DVFs) is also shown. This figure shows that results tend to improve as the library size increases, at least up to 80 DVFs. With the number of DVFs currently at our disposal we cannot assess whether we have reached a plateau or whether results could improve further with a larger library.

6.4 Parameter selection

The algorithm we present herein involves two parameters: k, which is the number of candidate matches for each test patch, and λ, which is the weight of the patch-match dissimilarity term in the cost function. To find the best values for these parameters we created a validation test that contained 10 randomly selected CO subjects. The remaining 142 DVFs were used as library for these validation cases. We generated the configurations for this validation set and asked one of the experts to do the evaluation. The evaluation was done as described in the library size sensitivity study discussed in Section 3. Since the evaluation is a laborious process, instead of doing a grid search for the two parameters, we adopted the following strategy: we first heuristically chose a few values of *k* and evaluated the results when different λ's were used. Following this experiment, we fixed the value of λ and evaluated the method on a larger range of *k* values.

Specifically, we started with *k* set to 5, 10, and 15 and we varied λ from 0 to 1 and counted both the number of configurations that were at least equivalent to manual configurations and the number of unacceptable configurations. The results are shown in **Fig. 6.10**. As can be seen, the results are largely insensitive to the value of λ but combining patch-match similarity and patch-to-patch consistency is nevertheless important.

Based on the results discussed above we chose $\lambda = 0.5$, which is roughly in the middle of the [0.0001, 0.99] interval, and varied *k* between 1 and 80, starting at one and iteratively doubling its value. **Fig. 6.11** shows that the results improve as *k* increases from 1 to 10 and reaches a plateau when *k* = 10. Based on these observations, we chose $\lambda = 0.5$ and *k* = 10 to run our experiments on the testing set that was used to generate the results presented in section 3.

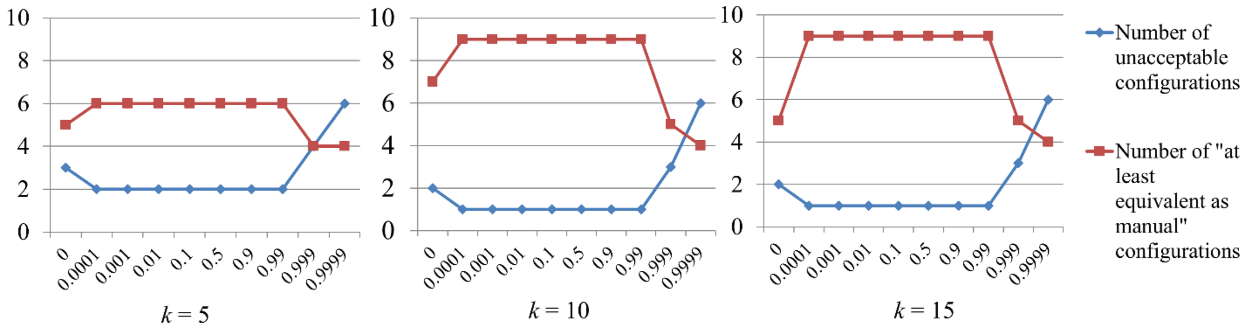

Fig. 6. 10 Evaluation results when different λ's are used (k = 5, 10 and 15, respectively)

6.5 Conclusions

This paper presents a generic algorithm for automatic electrode configuration selection for cochlear implants. This method produces results that are comparable to those obtained by Zhao's

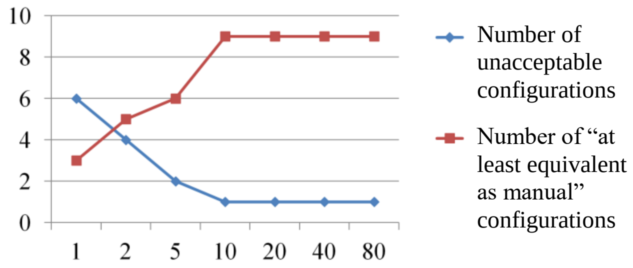

Fig. 6. 11 Evaluation results when different k's are used (λ = 0.5)

method described in [15] and [16]. Specifically, in the large scale study we have conducted, this new method generates more configurations of high quality (at least equivalent to manual configurations) than Zhao's. It also generates fewer unacceptable solutions but it does not outperform the manual configurations as often as Zhao's method. We have also observed that performance tends to increase with library size. As more data becomes available, we will assess whether our technique reaches a plateau or keeps improving. Finally, this method only requires a few parameters (λ and k, i.e., the number of patches kept when matching the library to a new patch). In contrast, the previous method relies on heuristics used by experts to come up with a deactivation strategy. Capturing a complete set of heuristics, translating them into equations, assigning weights to each rule, and dealing with unusual cases is difficult. The weights would also need to be re-estimated when the training set is modified to, for instance, include new electrode arrays. Here we rely on deactivation plans that have been vetted by experts, and we match new cases to known ones. In terms of computation, the proposed algorithm takes around 1s/case. Zhao's method takes 2-3s for MD and AB implants and ~35s for CO implants, which have more contacts. We note however that computation time is not a critical issue at this point.

References

[1] NIDCD, "Fact Sheet: Cochlear Implants, " NIH Publ. No. 11-4798, pp. 1–4, 2011.

[2] A. Aschendorff, R. Kubalek, B. Turowski, F. Zanella, A. Hochmuth, M. Schumacher and R. Laszig, "Quality Control after Cochlear Implant Surgery by means of Rotational Tomography," Otol. Neurotol., vol. 26, no. 1, pp. 34–37, 2005.

[3] C. C. Finley and M.W. Skinner, "Role of electrode placement as a contributor to variability in cochlear implant outcomes, " Otol Neurotol, vol. 29, pp. 920–8, 2008.

[4] L. K. Holden, C. C. Finley, J. B. Firszt, T. A. Holden, C. Brenner, L. G. Potts and M. W. Skinner, "Factors affecting open-set word recognition in adults with cochlear implants, " Ear Hear., vol. 34, no. 3, p. 342, 2013.

[5] G. B. Wanna, J. H. Noble, R. H. Gifford, M. S. Dietrich, A. D. Sweeney, D. Zhang and R. F. Labadie, "Impact of Intrascalar Electrode Location, Electrode Type, and Angular Insertion Depth on Residual Hearing in Cochlear Implant Patients: Preliminary Results, " Otol Neurotol, vol. 36, pp. 1343–8, 2015.

[6] K. F. Nordfalk, K. Rasmussen, E. Hopp, R. Greisiger, and G. E. Jablonski, "Scalar position in cochlear implant surgery and outcome in residual hearing and the vestibular system, " Int. J. Audiol., vol. 53, no. 2, pp. 121–127, 2014.

[7] G. B. Wanna, J. H. Noble, M. L. Carlson, R. H. Gifford, M. S. Dietrich, D. S. Haynes, B. M. Dawant and R. F. Labadie, "Impact of electrode design and surgical approach on scalar location and cochlear implant outcomes," Laryngoscope, vol. 124, no. S6, pp. S1–S7, 2014.

[8] J. H. Noble, R. H. Gifford, A. J. Hedley-Williams, B. M. Dawant, and R. F. Labadie, "Clinical evaluation of an image-guided cochlear implant programming strategy, " Audiol. Neurotol., vol. 19, no. 6, pp. 400–411, 2014.

[9] J. H. Noble, A. J. Hedley-Williams, L. Sunderhaus, B. M. Dawant, R. F. Labadie, S. M. Camarata and R. H. Gifford, "Initial Results With Image-guided Cochlear Implant Programming in Children," Otol. Neurotol., vol. 37, no. 2, pp. e63–e69, 2016.

[10] Y. Zhao, B. M. Dawant, R. F. Labadie, and J. H. Noble, "Automatic Localization of Cochlear Implant Electrodes in CT," in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2014, pp. 331–338.

[11] Y. Zhao, B. M. Dawant, and J. H. Noble, "Automatic localization of cochlear implant electrodes in CTs with a limited intensity range," SPIE Med. Imaging. Int. Soc. Opt. Photonics, vol. 10133, 2017.

[12] Y. Zhao, B. M. Dawant, R. F. Labadie, and J. H. Noble, "Automatic localization of closely-spaced cochlear implant electrode arrays in clinical CTs," Med. Phys., vol. 45, no. 11, pp. 5030–5040, 2018.

[13] J. H. Noble and B. M. Dawant, "Automatic graph-based localization of cochlear implant electrodes in CT," Int. Conf. Med. Image Comput. Comput. Interv., pp. 152–159, 2015.

[14] Y. Zhao, S. Chakravorti, R. F. Labadie, B. M. Dawant, and J. H. Noble, "Automatic graph-based method for localization of cochlear implant electrode arrays in clinical CT with sub-voxel accuracy, " Med. Image Anal., vol. 52, pp. 1–12, 2019.

[15] Y. Zhao, B. M. Dawant, and J. H. Noble, "Automatic electrode configuration selection for image-guided cochlear implant programming," vol. 9415, p. 94150K, 2015.

[16] Y. Zhao, B. M. Dawant, and J. H. Noble, "Automatic selection of the active electrode set for image-guided cochlear implant programming," J. Med. Imaging, vol. 3, no. 3, pp. 035001–035001, 2016.

[17] R. W. Hamming, "Error detecting and error correcting codes, " Bell Labs Tech. J., vol. 29, no. 2, pp. 147–160, 1950.

[18] E. W. Dijkstra, "A note on two problems in connexion with graphs," Numer. Math., vol. 1, no. 1, pp. 269–271, 1959.

[19] M. W. Fagerland, S. Lydersen, and P. Laake, "The McNemar test for binary matched pairs data: mid-p and asymptotic are better than exact conditional." BMC Med. Res. Methodol., vol. 13, no. July, p. 91, 2013.

Chapter VII

SUMMARY AND FUTURE WORK

7.1 Summary

In Chapter I, we give an introduction to natural hearing mechanism, cochlear implants (CIs) and an image-guided cochlear implant programming (IGCIP) system. The full automation of IGCIP has not been achieved which hinders its large-scale evaluation and deployment. Also, it is worth improving the accuracy of the intra-cochlear anatomy segmentation algorithm we currently use in pre-operative CTs. To tackle these issues, in the remaining chapters of this dissertation, several learning-based techniques are proposed to automate the various steps in the IGCIP system which needed manual intervention and to provide alternatives to existing techniques.

In Chapter II, to initialize the mutual information-based image registration algorithm, we propose to localize a set of pre-defined landmarks surrounding the inner ear in head CTs and use a point-based initial registration. To achieve this goal, we employ random forest regression to map image intensity values to heat maps that indicate positions of each landmark. However, false positive detections could happen in heat maps. To further distinguish the true positive from multiple false positives in each heat map, we incorporate the spatial relationship between landmarks as *a-priori* knowledge. We use a heuristic search algorithm to find the candidate landmark set that has the minimum registration error with the same landmark set on the atlas as the final solution [1,2]. We achieve a localization error of less than two voxels in the resampled images that are directly processed by the random forests. We further validate the initialization scheme and compare it with manual initialization and find that it has similar performance with manual initialization.

In Chapter III, to document a head CT image content in terms of the inner ears it includes, we use a 2D convolutional neural network (CNN) model to process image volumes in a slice-wise manner [3]. For a test CT volume, each axial slice is fed to the deep neural network model. The responses of all slices are combined to predict the content of the whole volume in terms of whether it includes both inner ears, only the left/right inner ear or neither inner ear. We test the algorithm on a large-scale, unscreened CT dataset and achieve a prediction accuracy of 96%.

In Chapter IV, to further improve the performance of the 2D CNN and to find the position of the inner ear, we use the 3d U-Net as a regressor to map a whole CT volume to heat maps that indicate the positions of the inner ear. We denote it as HeadLocNet-1 (Head CT Localization Network for 1 landmark) [4]. Using HeadLocNet-1, we improve the classification accuracy over our 2D CNN model from 96% to 98.6%, and achieve a localization error of only 2.45mm. To enable our deep learning model to localize multiple landmarks and make it more reliable, we devise the HeadLocNet-MC (Head CT Localization Network for Multi-landmarks with Classification) by adding six more landmarks' heat maps in the output and using multi-level features to directly do classification. We achieved an accuracy of 99.5% in head CT classification and a mean localization error of 3.45mm, averaged across all landmarks. The classification error rate is only ~1/3 of HeadLocNet-1 and the localization accuracy is significantly better than our random forest-based method. Further ablation studies show that the improvement of the classification accuracy is mainly due to the hierarchical features used in the classification branch. Also, we do not find an accuracy loss in localizing individual landmarks when we use HeadLocNet-MC compared with when we use HeadLocNet-1.

In Chapter V, to improve the accuracy of our current intra-cochlear anatomy segmentation algorithm, the active shape model (ASM)-based method, we propose a 2-level training scheme of the 3d U-Net [5] to perform the same task. To combat the ground truth data paucity problem in training deep networks for medical image applications, in this scheme, the 3d U-Net is first trained from scratch using more than 300 clinical CT images using the ASM-generated segmentations. That is to say, we use weak supervision at this stage. Second, the model is fine-tuned using a small set of specimens' clinical CTs of which the ground truth segmentations are generated by manual delineation in μCT images, i.e., strong supervision. Using this method, we achieve significant improvement over the ASM-generated segmentations when measuring DICE similarity coefficients and average surface distances in a leave-one-out training-testing study. A large scale qualitative evaluation on another 147 ear images also shows that the proposed method produces segmentations of visually better quality in the vast majority of the images, compared with ASM.

In Chapter VI, to automate the electrode configuration selection process for CI patients, we propose a template matching approach [6]. In this approach, a library which consists of past patients' distance-vs-frequency curves (DVFs) and their electrode configurations selected by experts is built. Each patient's DVFs are decomposed into patches. For a new patient, his/her DVFs are first created and decomposed into overlapping test patches. A set of similar candidate matches are found from the library for each test patch. We enforce deactivation pattern consistency on the candidate matches and find the optimal DVF match of each test patch. They are used to automatically determine the deactivation state of each electrode for the new patient. The advantage of our method over the previous method proposed by Zhao et al in [7] is that, it does not require a long training process to estimate more than 10 parameters. Once a new patient's electrode configuration data from experts is added to the library, it can be immediately used for testing without a re-training process. A parameter sensitivity study shows that the algorithm is not sensitive to parameter settings.

7.2 Some future work on intra-cochlear anatomy segmentation

7.2.1 Improving the robustness of the two-level trained model for intra-cochlear anatomy segmentation

From the large-scale qualitative evaluation result of the 2-level trained 3d U-Net model to segment the intra-cochlear anatomy, it is found that the model lacks robustness when it processes images degraded by noise, blurring and prosthetics. In the future, we aim to improve the robustness of this model in dealing with such images. Possible approaches include (1) designing data augmentation strategies by adding artificially degraded images to the training set, and (2) exploring other architectures, for example, conditional Generative Adversarial Networks (cGANs) [8] to introduce anatomical shape constraint to the segmentations generated by the deep neural networks.

7.2.2 Validation of the segmentation network model in predicting CI electrode positions

Studies have shown that each CI electrode is positioned in one of the cavities of the cochlea. It could be in the ST or the SV. The locations of the electrodes are significantly associated with hearing outcomes. So a critical criterion to judge the segmentation quality of the ST and the SV produced by the proposed segmentation model is to test the accuracy of classifying the electrodes in terms of the cavities they lie in, using the automatic segmentation results. For this purpose, we plan to do another validation study. In this validation study, a new dataset will be used. In this dataset, imaging data of a group of ear specimens will be prepared. Since the boundary of the ST and the SV, i.e., the basilar membrane (BM) cannot be seen in clinical CTs but can be visualized in μCTs, the validation requires μCTs of each ear specimen. So for each specimen, three CT scans are needed, i.e., (1) a clinical CT scan prior to implanting a CI, denoted as ‘Pre-C’, (2) a clinical CT scan after implanting a CI, denoted as ‘Post-C’ and (3) a μCT scan after implanting the CI, denoted as ‘Post-μ’. For each ear specimen, two steps that are typically done in IGCIP, i.e., automatic segmentation of the intra-cochlear anatomy and automatic localization of CI electrodes are done using Pre-C and Post-C, respectively. Positions of the electrodes can thus be predicted by registering the Pre-C and the Post-C and performing the prediction method proposed in [9]. The ground truth positions of the electrode set in the same specimen are obtained from Post-μ. The predicted positions of the electrodes are compared with the ground truth positions.

7.2.3 Reducing P2P errors of the segmentation network model

Currently as we have observed in the quantitative evaluation, the segmentations generated by our proposed segmentation model have P2P errors that are similar to those obtained with the ASM method despite the fact that our proposed method leads to better DSC and ASD results. The main reason, as we have explained in Chapter V is that point correspondence between the deep learning-generated meshes and the reference subject mesh has to be done through ASM-generated meshes. As also discussed, the tangential errors of the ASM-generated meshes are thus propagated to the deep learning-generated meshes. To avoid the propagation of the tangential errors, we will explore other methods which can help establish point correspondence for deep learning-generated meshes and potentially have less tangential errors. One possible method would be (1) first to perform non-rigid registration between the reference subject’s meshes and the deep learning-generated meshes of the target specimen, using a non-rigid ICP [10] algorithm, and (2) find the closest point from the deformed reference meshes to the target automatically obtained meshes.

References

[1] D. Zhang, Y. Liu, J. H. Noble, and B. M. Dawant, "Automatic localization of landmark sets in head CT images with regression forests for image registration initialization," in Proceedings of SPIE Medical Imaging Conference, 2016.

[2] D. Zhang, Y. Liu, J. H. Noble, and B. M. Dawant, “Localizing landmark sets in head CTs using random forests and a heuristic search algorithm for registration initialization,” Journal of Medical Imaging, vol. 4, no. 4, p. 44007, 2017.

[3] D. Zhang, J. H. Noble, and B. M. Dawant, "Automatic detection of the inner ears in head CT images using deep convolutional neural networks, " in Proceedings of SPIE conference on Medical Imaging Conference, 2018, p. 10574.

[4] D. Zhang, J. Wang, J. H. Noble, and B. M. Dawant, "Accurate Detection of Inner Ears in Head CTs Using a Deep Volume-to-Volume Regression Network with False Positive Suppression and a Shape-Based Constraint", in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 703–711, 2018.

[5] D. Zhang, R. Banalagay, J. Wang, Y. Zhao, J. H. Noble and B. M. Dawant, "Two-level Training of a 3d U-Net for Accurate Segmentation of the Intra-cochlear Anatomy in Head CT with Limited Ground Truth Training Data", In Proceedings of SPIE Medical Imaging Conference 2019.

[6] D. Zhang, Y. Zhao, J. H. Noble and B. M. Dawant. "Selecting electrode configurations for image-guided cochlear implant programming using template matching. Journal of Medical Imaging", *5*(2), 021202, 2017.

[7] Y. Zhao, B. M. Dawant, and J. H. Noble, “Automatic selection of the active electrode set for image-guided cochlear implant programming,” J. Med. Imaging, vol. 3, no. 3, pp. 035001–035001, 2016.

[8] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, "Image-to-image translation with conditional adversarial networks". arXiv:1611.07004 (2017)

[0] J. H. Noble, R. F. Labadie and B. M. Dawant. "Automatic Classification of Cochlear Implant Electrode Cavity Positioning." In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2018.

[10] B. Amberg, S. Romdhani and T. Vetter. "Optimal step nonrigid ICP algorithms for surface registration". In IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). 2007.