This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2697916, IEEE Transactions on Biomedical Engineering

Highly accurate Facial Nerve Segmentation Refinement from CBCT/CT Imaging using a Super Resolution Classification Approach

Ping Lu[1], *Member, IEEE,* Livia Barazzetti[1], Vimal Chandran[1], *Member, IEEE,* Kate Gavaghan[2], Stefan Weber[2], *Member, IEEE,* Nicolas Gerber[2], and Mauricio Reyes[1], *Member, IEEE*

***Abstract*—Facial nerve segmentation is of considerable importance for pre-operative planning of cochlear implantation. However, it is strongly influenced by the relatively low resolution of the cone-beam computed tomography (CBCT) images used in clinical practice. In this paper, we propose a super-resolution classification method, which refines a given initial segmentation of the facial nerve to a sub-voxel classification level from CBCT/CT images. The super-resolution classification method learns the mapping from low-resolution CBCT/CT images to high-resolution facial nerve label images, obtained from manual segmentation on micro-CT images. We present preliminary results on dataset, 15 ex-vivo samples scanned including pairs of CBCT/CT scans and high-resolution micro-CT scans, with a Leave-One-Out (LOO) evaluation, and manual segmentations on micro-CT images as ground truth. Our experiments achieved a segmentation accuracy with a Dice coefficient of 0.818 ± 0.052, surface-to-surface distance of $0.121 \pm 0.030mm$ and Hausdorff distance of $0.715 \pm 0.169mm$. We compared the proposed technique to two other semi-automated segmentation software tools, ITK-SNAP and GeoS, and show the ability of the proposed approach to yield sub-voxel levels of accuracy in delineating the facial nerve.**

***Index Terms*—Facial nerve, Segmentation, Cochlear implantation, Supervised Learning, Super-Resolution, CBCT, micro-CT.**

I. Introduction

Cochlear implantation is a conventional treatment that helps patients with severe to profound hearing loss. The surgical procedure requires drilling of the temporal bone to access the cochlea. In the traditional surgical approach, a wide mastoidectomy is performed in the skull to allow the surgeon to identify and avoid the facial nerve, whose damage can cause temporal or permanent ipsilateral facial paralysis. In order to minimize invasiveness, a surgical robot system has been developed to perform highly accurate and minimally invasive drilling for direct cochlear access [1]. The associated planning software tool, OtoPlan [2], allows the user to semiautomatically segment structures of interest and define a safe drilling trajectory. The software incorporates a semiautomatic and dedicated method for facial nerve segmentation using interactive centerline delineation and curved planar reformation [2].

The surgical planning for minimally invasive cochlear implantation is affected by the relatively low resolution of the patient images. Imaging of the facial nerve is typically performed using CT or CBCT imaging with a resolution in the range of $0.15 - 0.3mm$ slice thickness, and a small field of view $80 - 100mm$ temporal bone protocol. This resolution is comparatively low in regards to the diameter of the facial nerve, which lies in the range of $0.8 - 1.7mm$.

Atlas-based approaches combined with level-set segmentation have been proposed before to segment the facial nerve in adults [3] and pediatric patients [4]. These methods automatically segment the facial nerve, with reported average and hausdorff accuracies in the ranges of $0.13 - 0.24mm$ and $0.8 - 1.2mm$, respectively. This reported accuracy is similar to other approaches, such as OtoPlan [2] or NerveClick [5], where a semi-automatic statistical model of shape and intensity patterns was developed with a reported RMSE accuracy of $0.28 \pm 0.17mm$. Since for the facial nerve a margin of up to $1.0mm$ is available and an accuracy of at least $0.3mm$ (depending on the accuracy of the navigation system) is required [6], an accurate facial nerve segmentation is crucial for an effective cochlear implantation surgical plan.

Super-resolution methods have been presented in computer vision related tasks to reach sub-voxel accuracy in regression problems, where the goal is to reconstruct a high-resolution image from low-resolution imaging information. Most of such methods employ linear or cubic interpolation [7], but are sub optimal for CBCT/CT images of the facial nerve, due to their SNR and local structural variability. In a recent study [8], a Random Forest based regression model was used to perform upsampling of natural images. Similarly, in [9] a supervised learning algorithm was used to generate diffusion tensor images at super-resolution (i.e. upscaling from $2 \times 2 \times 2mm$ to $1.25 \times 1.25 \times 1.25mm$ resolution). Recently, in [10], a super-resolution convolutional neural network (SRCNN) learns an end-to-end mapping between the low- and high-resolution images. In [11] a super-resolution (SR) approach reconstructs high resolution 3D images from 2D image stacks for cardiac MR imaging, based on a convolutional neural network (CNN) model.

In the present clinical problem, we are concerned with the delineation of the facial nerve for cochlear implantation planning. Hence, as opposed to other super-resolution schemes, here we propose a super-resolution *classification* method for accurate segmentation refinment of the facial nerve.

[1] Ping Lu, Livia Barazzetti, Vimal Chandran and Mauricio Reyes are with Institute for Surgical Technology & Biomechanics, University of Bern, CH-3014 Bern, Switzerland. e-mail: ping.lu@istb.unibe.ch

[2] Nicolas Gerber, Kate Gavaghan and Stefan Weber are with the ARTORG Center for Biomedical Engineering Research, University of Bern, CH-3010 Bern, Switzerland.

We adopted a supervised learning scheme to learn the mapping between CBCT/CT images to high-resolution facial nerve label images, obtained from manual segmentations on micro-CT images. Here we coin the method super-resolution classification (SRC). The proposed approach then employs SRC to refine an initial segmentation provided by OtoPlan [2] to generate accurate facial nerve delineations.

In the following sections we present a description of the image data and the proposed algorithm, followed by segmentation results on test cases, and a comparison with two other general-purpose segmentation (ITK-SNAP, GeoS) software tools used to perform segmentation refinement.

II. Materials and Methods

This section describes the image data used to train and test the proposed SRC algorithm. Figure 1 shows an overview of the complete pipeline, composed of two phases for training and testing. During training, image upsampling, image preprocessing, registration to micro-CT images, and building of a classification model based on extracted features, are performed. During testing, the input CBCT image is upsampled, preprocessed, and features are extracted to perform super-resolution classification.

A. Image Data

We developed and tested our approach on a database of 15 patient cases, comprising 7 pairs of CBCT and micro-CT images, and 8 pairs of CT and micro-CT images of temporal bones. The CBCT temporal bones were extracted from four cadaver heads in the context of an approved clinical study on cochlear implantation [12]. The CBCT were obtained with a Planmeca ProMax 3D Max with $100 \times 90mm^2 FOV$. Micro-CT was performed with a Scanco Medical μCT 40 with $36.9 \times 80mm^2 FOV$. For each sample, a CBCT ($0.15 \times 0.15 \times 0.15mm^3$) and micro-CT ($0.018 \times 0.018 \times 0.018mm^3$) scan was performed. The set of 8 pairs of CT and Xtreme CT images of temporal bones were obtained with a CT imaging (Siemens SOMATOM Definition Edge) and a temporal bone imaging protocol with parameters: $120kVp$, $26mA$, $80mmFOV$. The spatial resolution of the scanned CT images was $0.156 \times 0.156 \times 0.2mm^3$. The Xtreme CT ($0.0607 \times 0.0607 \times 0.0607mm^3$) scans were obtained with a Xtreme CT imaging (SCANCO Medical). We note that cadaver images are similar to clinical images of the facial nerve, enabling the evaluation of our method with cadaver images. The image volume size ranges around $70 \times 80 \times 110$ voxels from CBCT and $60 \times 55 \times 60$ voxels from CT .

To create ground truth datasets, manual segmentations of the facial nerve on micro-CT images was performed following the segmentation protocol presented in [2], and were verified by experts using Amira 3D Software for Life Sciences version 5.4.4 (FEI, USA) [13]. The experts verifying the manual ground-truth are two senior biomedical engineers trained in the cochlear anatomy and with years of experience in manual segmentation of the cochlear structures.

B. Preprocessing

To create the supervised based machine learning model and to evaluate the approach on ground truth data, derived from micro-CT images, we rigidly aligned the pairs of CBCT and CT and micro-CT images using Amira (version 5.4.4) via the normalized mutual information method [14] and elastix [15] [16] with the advanced normalized correlation metric (see Appendix A). For the sake of clarity, we refer for the rest of the paper to both CBCT and CT as CBCT/CT to indicate that the operations are applied to both.

1) Intensity normalization: We normalize the intensities of the CBCT/CT images by histogram matching, with a common histogram as a reference. Since we are computing the histogram only on a ROI (described below in section II-B3) to match the range of intensities being targeted for the facial nerve, we avoid the effect of background voxels and hence, there is no need to set an intensity threshold for the histogram matching.

2) CBCT/CT and micro-CT image alignment (only training phase): In order to learn the mapping between CBCT/CT and micro-CT images we rigidly aligned the pairs of CBCT/CT and micro-CT images. Due to the fact that the diameter of the facial nerve lies in the range of $0.8 - 1.7mm$ and the facial nerve is only imaged across approximately at $5 - 11$ slices of CBCT/CT ($0.15 \times 0.15 \times 0.15mm^3$), we manually initialized a rigid registration based on landmarks defined by screws implanted in the specimens for patient-to-image registration, as presented in [12], followed by an automated rigid registration in Amira (version 5.4.4) using normalized mutual information metric. A second rigid registration was performed between the transformed micro-CT image and the CBCT image, using elastix with advanced normalized correlation metric. We observed that in practice this pipeline resulted in an improved robustness and accuracy, as opposed to performing a single registration. We also remark that no change of resolution is performed when registering the micro-CT image to the CBCT image (as typically is the case for image registration tasks). The set of sought transformations are then applied to the ground truth image in order to map them onto the CBCT image space.

3) Region of interest selection (ROI): Since the main focus of the method is to obtain sub-voxel accuracy of the facial nerve border, and to reduce computational costs, we adopted a band-based region of interest selection strategy. Here we use the segmentation results from OtoPlan as initial segmentation to be refined through Super Resolution Classification. From the preliminary OtoPlan segmentation of the CBCT/CT image, a region-of-interest is created via a combination of erosion and dilation morphological operations. The region of interest, on which the super-resolution classification takes place, corresponds to the arithmetic difference between the dilated and eroded label images. In practice, a 16 and 24 voxel structuring element ($0.3mm$ and $0.4mm$ respectively on each side, which effectively translates as an additional two times magnitude of the accuracy error reported by other approaches)) was tested on the upsampled CBCT/CT images.

C. Super-Resolution Classification (SRC)

This section describes the steps for CBCT/CT upsampling, the feature extraction and the classification model building.

1) CBCT/CT upsampling: Similar to [8], we perform an upsampling of the CBCT/CT image to the target resolution in order to combine features extracted from the upsampled and the original image. In this study we employed a B-spline upsampling scheme. However, other interpolation schemes, such as linear or cubic, can be used since as classification results were not sensitive to this choice.

2) Feature Extraction: We employ texture-based features derived from first-order statistics, percentiles and Grey Level Co-occurrences Matrix (GLCM) [17] [18] [19], which are only extracted on the computed region of interest. First-order statistics [20] and percentile features [21] [22] are computed at original and upsampled resolutions, while GLCM features are computed only on patches from the upsampled image. This is supported by direct testing of GLCM features derived from both the original and upsampled images with poorer results (in terms of all evaluated metrics) than using only GCLM features extracted from the upsampled image. This can also be explained by the fact that the much larger size of voxel-wise GLCM features (in comparison to the other imaging features). We remark that through direct testing of GLCM features derived from both original and upsampled CBCT/CT images, GLCM features extracted from the original CBCT/CT image do not contribute as much as those extracted from the upsampled image.

a) ***First-order statistics**:* Mean, standard deviation, minimum, maximum, skewness and kurtosis of voxel intensities are computed for each image patch of the CBCT/CT and upsampled CBCT/CT image.

b) ***Percentiles**:* From each image patch of the CBCT/CT and upsampled CBCT/CT image, the 10th percentile, 25th percentile, 50th percentile, 75th percentile, 80th percentile, 95th percentile of the intensity distribution, are used as features.

c) ***The Grey Level Co-occurrences Matrix (GLCM)**:* The Grey Level Co-occurrences Matrix (GLCM) is a second-order statistical texture that considers two-voxels relationship in an image. Following [19], we adopted 8 GLCM features: inertia, correlation, energy, entropy, inverse difference moment, cluster shade, cluster prominence, haralick correlation. Mean and variance of each feature with 13 independent directions in the center voxel of each image patch are calculated. Hence, 16 features of GLCM were calculated in the upsampled CBCT/CT image.

3) Classification model – Training phase: Given a training set $\{\langle \boldsymbol{X_i}, \boldsymbol{Y_i}\rangle | i = 1, ..., N\}$ of CBCT/CT and micro-CT aligned pairs of images, we extract from each i_{th} image patch, a feature vector $\boldsymbol{X_i} = (\boldsymbol{v}_1, ...\boldsymbol{v}_n) \in \mathbb{X}$ and responses $\boldsymbol{y} \in \{0, 1\}$, which describes the background/foreground label of the center voxel over a grid of C voxels. Then, a function $\hat{\boldsymbol{y}} : \mathbb{X} \mapsto \boldsymbol{y}$ from a space of features $\mathbb{X}$ to a space of responses $\boldsymbol{y}$ is constructed. The mapping is cast as a *classification* problem.

As classification model, we adopted extremely randomized trees (Extra-Trees) [23], which is an ensemble method that combines the predictions of several randomized decision trees to improve robustness over a single estimator. Extra-Trees have shown to be slightly more accurate than Random Forests (RF) and other tree-based ensemble methods [23]. During the training phase of Extra-Trees, multiple trees are trained and each tree is trained on all training data. Extra-Trees randomly selects without replacement, K input variables $\{v_1, ...v_k\}$ from the training data. Then, a cutpoint s_i is randomly selected, ruled by a splitting criteria $[v_i < s_i]$, for each selected feature within the interval $[v_i^{min}, v_i^{max}]$. Among the K candidate splits, the best split is chosen via normalization of the information gain [24]. We note that in our experiments, and in order to reduce irrelevant features [23], the number of input variables K is set to the size of the input feature vector n.

4) Classification model – Prediction phase: During testing, the CBCT/CT image is pre-processed through image intensity normalization (using the same reference image as for the training phase). Image features in a band of interest (results reported using a band size of 16 and 24 voxels) are extracted from the original and upsampled CBCT/CT image, and passed through the Extra-Trees classification model. The computed output corresponds to the label of the central voxel from the extracted patch.

5) Postprocessing: The refined segmentation is regularized in order to remove spurious and isolated segmented regions. In this study we adopted a basic regularization scheme based on erosion (kernel size=16 or 24) and dilation (kernel size=16 or 24) morphological operations.

III. Experimental Design

A Leave-One-Out (LOO) cross-validation study was carried out to evaluate the accuracy of the proposed super-resolution segmentation approach. The idea of LOO is to split data into train and test sets. One image data is chosen as a test set while the remaining data are used for training. This method is repeated until every image data has been tested and evaluated using the following evaluation metrics.

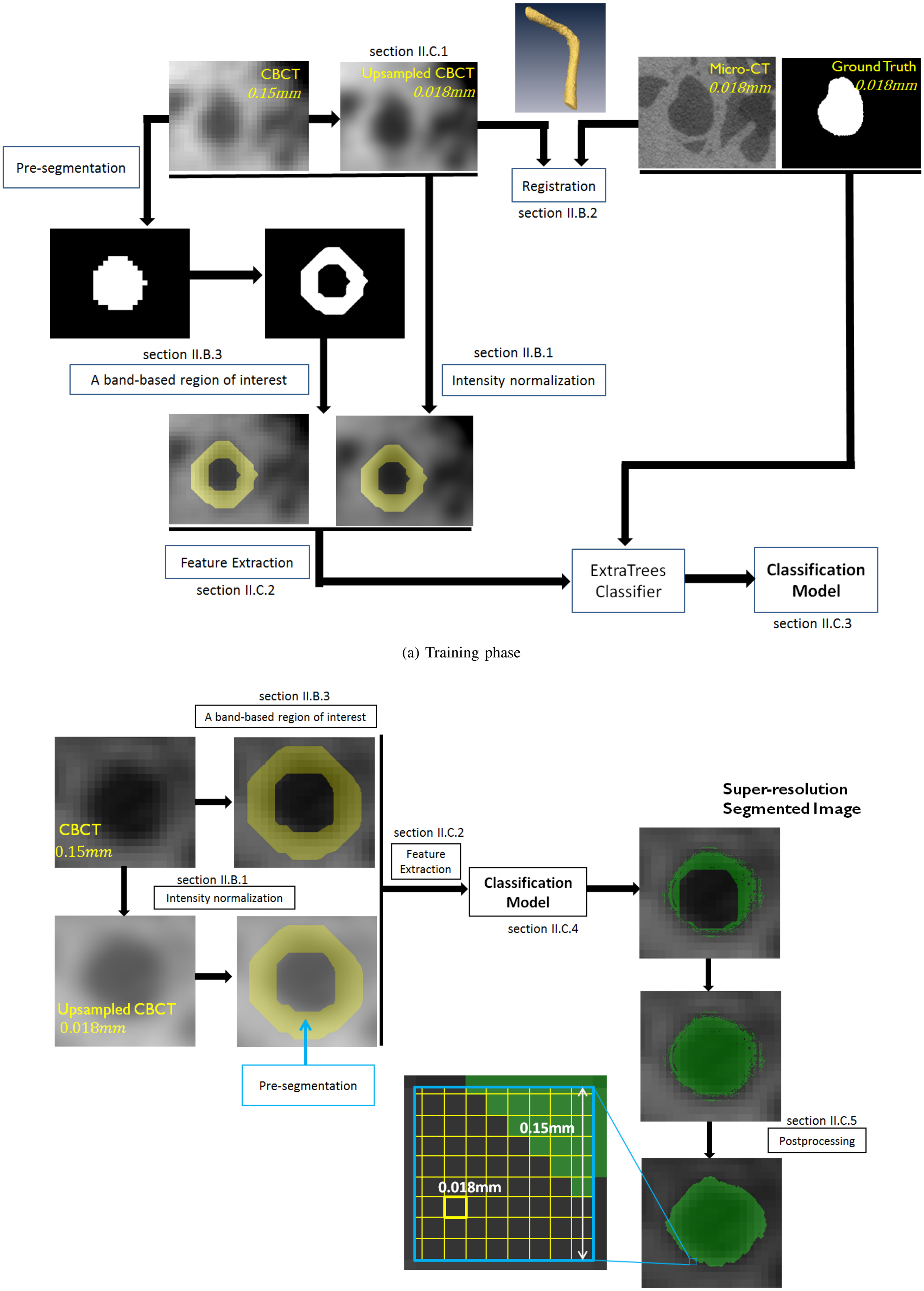
A. Experimental detail

The upsampling of the CBCT/CT images was performed in Amira with a B-spline interpolation kernel. For computation of features, patches of size $5 \times 5 \times 5$ were extracted on the original and upsampled CBCT/CT images. Feature extraction, morphological operations to create the ROI, and intensity normalization was performed with the Insight-Toolkit version 4.4.1 [25], and classification was completed with Scikit-learn: Machine Learning in Python [26]. Default parameters were used for the Extra-Trees classifier.

B. Segmentation initialization with OtoPlan

As input to the segmentation refinement step with SRC we utilized OtoPlan [2] to obtain an initial segmentation from where the ROI bands are extracted, and then classified.

In OtoPlan, the centerline of the facial nerve is manually drawn and the borders are automatically defined by the tool. After this step, the facial nerve border can be manually modified by dragging contours of the facial nerve.



(a) Training phase

(b) Testing phase

Fig. 1: Proposed super-resolution classification (SRC) approach, described for the training (a), and testing phase (b). During training, the original CBCT/CT image is aligned to its corresponding micro-CT image. OtoPlan [2] is used to create an initial segmentation, from where a region-of-interest (ROI) band is created. From the original and upsampled CBCT/CT images, features are extracted from the ROI-band to build a classification model, which is used during testing to produce a final super-resolution segmented image. The zoomed square on the segmented super-resolution image shows on one voxel of the CBCT/CT image, the more accurate segmentation yielded by SRC.

Based on this initial segmentation we extracted a ROI with two different sizes (section II-B3), referred as to band 16 and band 24, to indicate 16 and 24 voxels band size. The rationale behind is to analyze the sensitivity of SRC to different band sizes, as well as to analyze a potential dependency between the accuracy of the initial segmentation and the performance of SRC.

C. Evaluation metrics

(1) Hausdorff Distance (HSD). This metric measures the Hausdorff distance [27] from the ground truth surface to its nearest neighbor in the segmented surface.

(2) Root Mean Squared Error (RMSE). The RMSE is calculated by the square root of the Mean Squared Error (MSE). $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(a-b)^2}{m}}$, where $a \in A$, $b = \min_{b^* \in B} ||a-b^*||$, and m corresponds to the number of surface points used to compute RMSE.

(3) Average Distance (AveDist). $AveDist(A,B) = max(d(A,B), d(B,A))$, where $d(A,B) = \frac{1}{m}\sum_{a \in A} \min_{b \in B} ||a-b||$. The smaller the value of the average distance the better the accuracy of the facial nerve segmentation is.

(4) Positive Predictive Value (PPV). $PPV = \frac{TP}{(TP+FP)}$, where TP stands for true positive — the number of correctly segmented facial nerve voxels— and FP stands for false positive, the number of wrong segmented voxels.

(5) Sensitivity (SEN). $SEN = \frac{TP}{(TP+FN)}$, where FN stands for the number of wrong labeled background voxels (i.e., segmenting facial nerve voxels as background).

(6) Specificity (SPC). $SPC = \frac{TN}{(TN+FP)}$, where TN stands for the number of correctly segmented background voxels.

(7) Dice Similarity Coefficients (DSC). A DSC value of 1 indicates the segmentation fully overlaps with the ground truth (i.e. perfect segmentation), while a DSC value of 0 indicates no overlap beween the segmentation and the ground truth segmentation.

D. Evaluation

In this section we present segmentation results separately for CBCT and CT images with the intention to show the performance of the proposed approach on two different scan types. We remark that the adopted LOO evaluation strategy employs the entire set of clinical CBCT/CT scans for the training phase.

1) Experiment 1: Segmentation results on CBCT(training and testing with LOO): We compared the proposed SRC method with the segmentation software GeoS and ITK-SNAP. We employed ITK-SNAP (version 3.4.0) [28] and its Random-Forest-based generation of speed images, which relies on defining brushes on the foreground and background areas of the facial nerve. The number of brushes was found empirically via trial-and-error with the main criteria of yielding robust segmentation results. In practice this resulted in approximately four brush strokes per image, and eight bubbles for contour initialization. No extensive search of optimal placement of brushes was conducted in order to keep the experiments to the typical usage scenario of the tool. Similar procedure was conducted for the GeoS tool (version 2.3.6) [29], a semi-automatic tool based on brush strokes and Random Forest supervised learning. On average over fifteen brush strokes were used for GeoS, with no further improvements observed beyond this number.

For both software tools, brush strokes were defined on the ROI-band (16 or 24 voxels) on background and foreground areas (c.f. section II-B3).

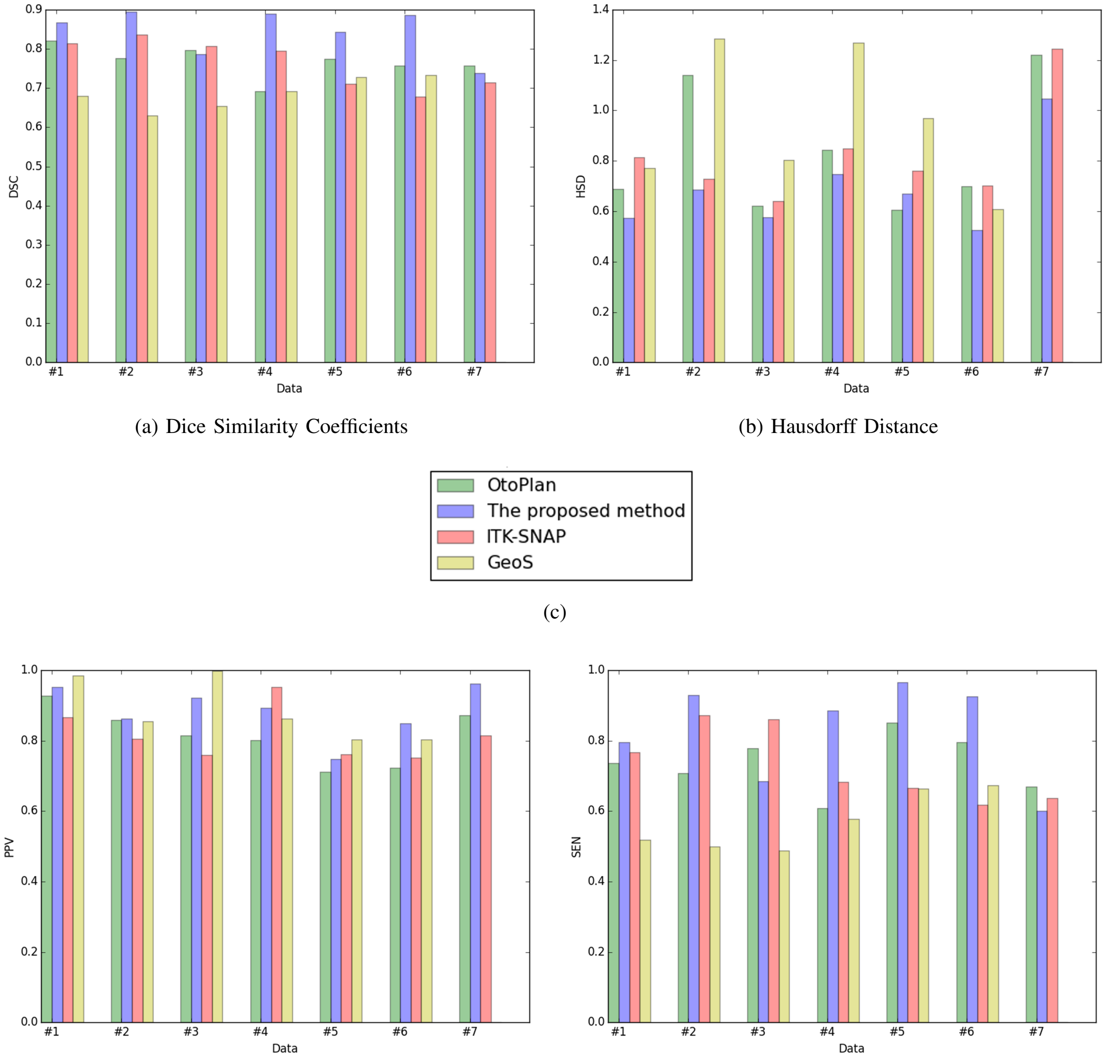
In order to compare DSC values among segmentation results and the ground truth (produced at micro-CT resolution), we resampled the results from the tools to the resolution of the ground truth using nearest interpolation. Second, we converted the segmentation results to surfaces [30] and computed average and Hausdorff distances.

Figure 2 shows the facial nerve segmentation results for each sample of the CBCT dataset for the proposed SRC method, and ITK-Snap and GeoS. From the DSC values and the Hausdorff distances (Figures 2a and 2b), it can be observed that the proposed method is robust and provides a higher average DSC and a lower Hausdorff distance than the other methods. From Figure 2d and Figure 2e it can be observed that that overall ITK-SNAP and GeoS tend to undersegment the CBCT cases. Conversely, SRC did not show a potential bias towards over-, or under-segmentation.

Table I summarizes the comparative results between the proposed approach and GeoS and ITK-SNAP, with two different band-based region of interest. It can be observed that in comparison to ITK-SNAP and GeoS, the proposed SRC method is more accurate and robust to an increase of the band size. Particularly, GeoS resulted to be less robust to an increase of the band size, as described by the increase variance of the metrics. The proposed SRC method achieved an average DSC value of 0.843, a mean Hausdorff distance of $0.689mm$, and a sub-voxel average distance accuracy of $0.156mm$. Regarding the tested segmentation tools, GeoS yielded the lowest DSC value among the evaluated approaches (average DSC of 0.686), followed by ITK-SNAP with an average DSC value of 0.765. In terms of distance metrics, the average Hausdorff metric for GeoS and ITK-SNAP was $0.951mm$ and $0.819mm$, respectively. Using a two-tailed t-test and Wilcoxon signed ranks tests, statistically significantly greater results than GeoS and ITK-SNAP were obtained ($p < 0.05$, Bonferroni corrected) for the dice, average distance and RMSE metrics. Figure 3 shows an example result, put in the context of the original and high-resolution ground-truth, while Figure 4 shows example results for all tested approaches. It can be observed that the proposed approach yields a more precise delineation than the other tested methods. Particularly, the postprocessing step based on simple morphological removes any potential holes and isolated small regions.

2) Experiment 2: Segmentation results on CT: Figure 5 and Table II summarize the comparative results on the CT database, between the proposed approach and GeoS and ITK-SNAP for two different band-based sizes. Similar to the results on CBCT cases, it is observed that the proposed SRC method is superior to ITK-SNAP and GeoS, and is more robust to the band size. The proposed method achieved an average DSC

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2697916, IEEE Transactions on Biomedical Engineering



(a) Dice Similarity Coefficients

(b) Hausdorff Distance

(c)

(d) Positive Predictive Value

(e) Sensitivity

Fig. 2: Evaluation on CBCT cases between the proposed super-resolution segmentation method and GeoS (version 2.3.6) and ITK-SNAP (version 3.4.0). Average values for DSC (a), Hausdorff (b), Positive Predictive Value (d), and Sensitivity (e). Case number 7 could not be segmented via GeoS. Note: best seen in colors.

	CBCT dataset with band16			CBCT dataset with band24		
Method	Proposed method (SRC)	GeoS	ITK-SNAP	Proposed method (SRC)	GeoS	ITK-SNAP
Dice	**0.843±0.055(0.866)**$*$	0.686±0.037(0.686)	0.765±0.058(0.795)	**0.822±0.062(0.847)**	0.578±0.123(0.547)	0.732±0.099(0.777)
AveDist	**0.112±0.034(0.100)**$*$	0.296±0.058(0.193)	0.196±0.052(0.205)	**0.131±0.032(0.121)**	0.436±0.137(0.461)	0.197±0.064(0.164)
RMSE	**0.156±0.038(0.144)**$*$	0.345±0.062(0.367)	0.247±0.063(0.248)	**0.186±0.034(0.180)**	0.493±0.134(0.526)	0.237±0.062(0.212)
Hausdorff	**0.689±0.163(0.670)**$*$	0.951±0.253(0.886)	0.819±0.185(0.760)	**0.747±0.117(0.736)**	1.309±0.207(1.364)	0.744±0.090(0.708)

TABLE I: Quantitative comparison on CBCT cases between our method and GeoS and ITK-SNAP, for band sizes 16 (left) and 24 (right). Dice and surface distance errors (in mm). The measurements are given as mean ± standard deviation (median). The best performance is indicated in boldface. The $*$ indicates that SRC results are statistically significantly greater ($p < 0.05$) than GeoS and ITK-SNAP using a two-tailed t-test and Wilcoxon signed ranks tests.

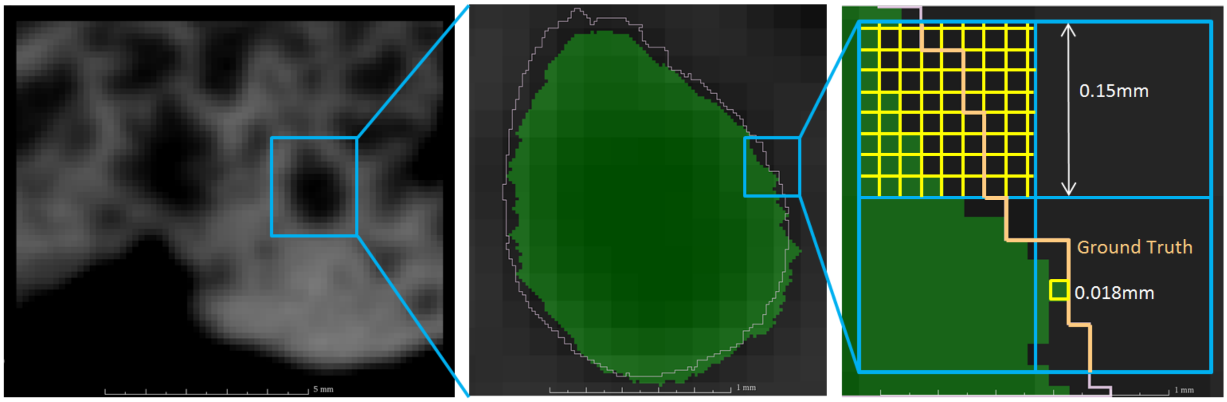


Fig. 3: Example results for the proposed super-resolution segmentation approach. From left to right: Original CBCT image with highlighted (in blue) facial nerve, resulting segmentation and ground truth delineation (orange contour), and zoomed area describing SRC results on four corresponding CBCT voxels.

value of 0.797, a mean Hausdorff distance of $0.739mm$, and a sub-voxel average distance accuracy of $0.129mm$. GeoS yielded the lowest DSC value among the evaluated approaches, followed by ITK-SNAP with an average DSC value of 0.731. In terms of distance metrics, the average Hausdorff metric for GeoS and ITK-SNAP was $0.954mm$ and $0.829mm$, respectively. Using a two-tailed t-test and Wilcoxon signed ranks tests, statistically significantly greater results than GeoS and ITK-SNAP were obtained ($p < 0.05$, Bonferroni corrected) for the dice, average distance and RMSE metrics, see Table II.

IV. Discussions

In this study we developed an automatic super-resolution facial nerve segmentation via a random forest Extra-Trees based classification framework that refines an initial segmentation of the facial nerve to sub-voxel accuracy. To our knowledge, this is the first attempt to perform super-resolution classification for facial nerve segmentation in CBCT/CT images by exploting imaging modalities featuring different resolution levels. Preliminary results, based on a leave-one-out evaluation on fifteen ex-vivo cases, suggest that the proposed method is able to classify the facial nerve with high accuracy and robustness. On a standard desktop computer, the learning phase is the most time-consuming part, requiring for our set-up around 2 hours. The testing phase (running on a new case) takes only 9 minutes. Given an input CBCT or CT image, the proposed pipeline start with an initial segmentation of the facial nerve region, which in this study was obtained via OtoPlan [2]. However, we remark that other approaches can be used to yield the initial segmentation (e.g. [3], ITK-SNAP [28]). A band ROI is then created from this initial segmentation and used by SRC to attain a highly accurate segmentation of the facial nerve in an automated fashion. Comparison with other available segmentation tools, ITK-SNAP and GeoS, confirms the higher accuracy and robustness of the proposed SRC approach.

According to our experiments, better segmentation results are obtained when the features computed on both the original and upsampled CBCT images than with features extracted only from the original CBCT image. This is in agreement with recent findings in semi-supervised regression based image upscaling where features extracted from an initial upsampling has shown to yield better estimates of the sought high-resolution image [8]. This is motivated by the fact that the training phase is enriched by including model samples that stem from micro-CT labels (i.e. from the micro-CT ground-truth image) and corresponding imaging features approximated at micro-CT level by the upsampling step on the CBCT/CT images.

As described, GLCM features extracted from the original CBCT/CT image do not contribute as much as those extracted from the upsampled image. This behavior can be conceptually explained since GLCM features computed on the upsampled image describe textural patterns on a much localized 5^3 patch size that better correlates to the label of the central voxel, extracted from the micro-CT image. Conversely, GLCM features computed on the original CBCT/CT image covers a much larger spatial extent, and hence the described textural information correlates less to the label of the central voxel at micro-CT resolution. Interestingly, the role of features from first-order statistics and percentiles provide benefits on both original and upsampled CBCT/CT images. First-order statistics and percentiles computed on the original CBCT/CT image improve the positive predictive value, but yields to a blocky effect in the segmentation result when not used in combination with first-order statistics and percentiles computed on the upsampled CBCT/CT image. We also checked (not reported here) the accuracy of the general-purpose segmentation tools on the upsampled CBCT images. Obtained results suggests that these general-purpose tools do not benefit from an upsampling of the CBCT image. On the contrary, worse results were obtained, with an average worsening on the dice scores of 80.1% and 62.3% for ITK-SNAP and GeoS, respectively. We refrained from further investigating the reasons as to why of this behavior due to the lack of implementation details of the tools.

The proposed SRC approach can also be used on pathological anatomies as it does not rely on shape priors. For instance, in case of bony dehiscence of the fallopian canal. In facial nerve dehiscence the nerve is uncovered in the middle ear cavity, leading to proximity of air voxels to the facial nerve.

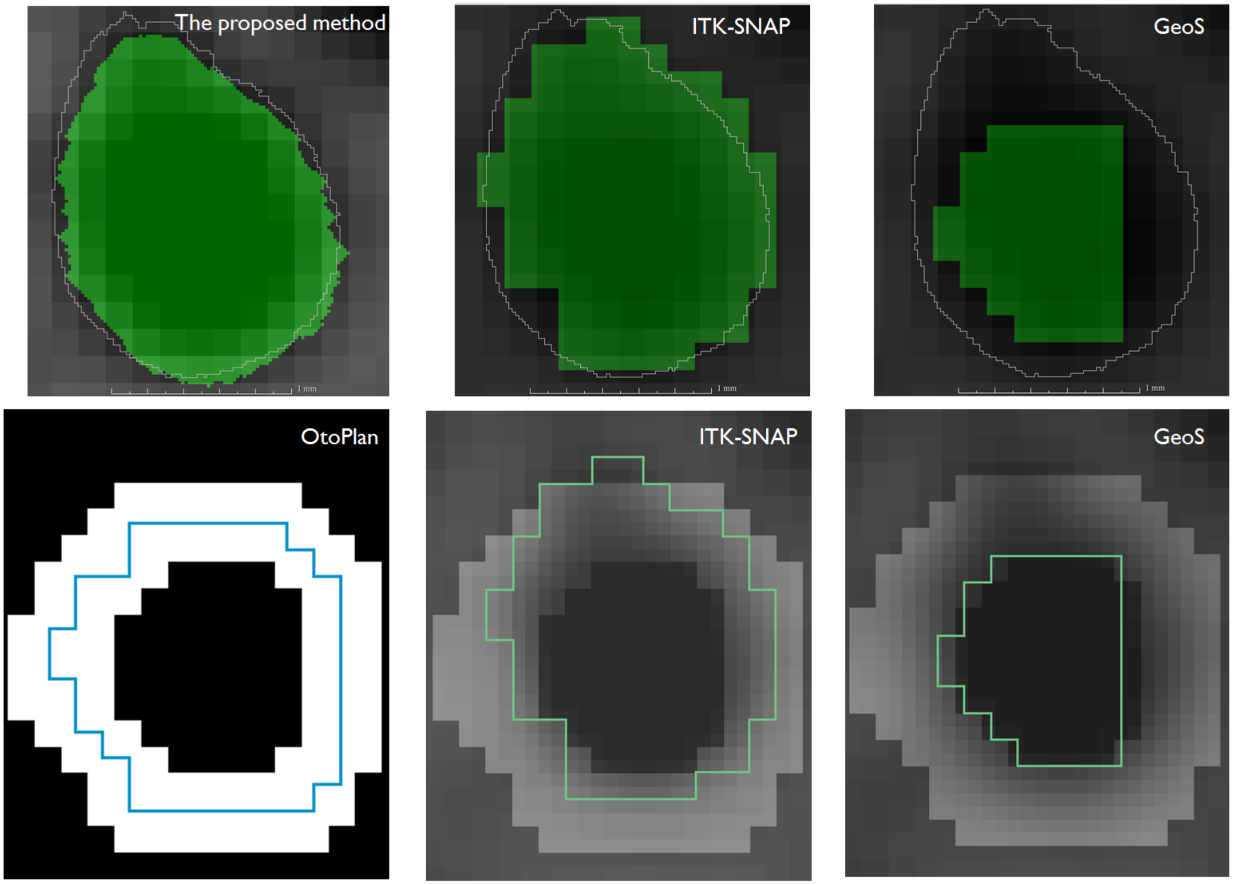


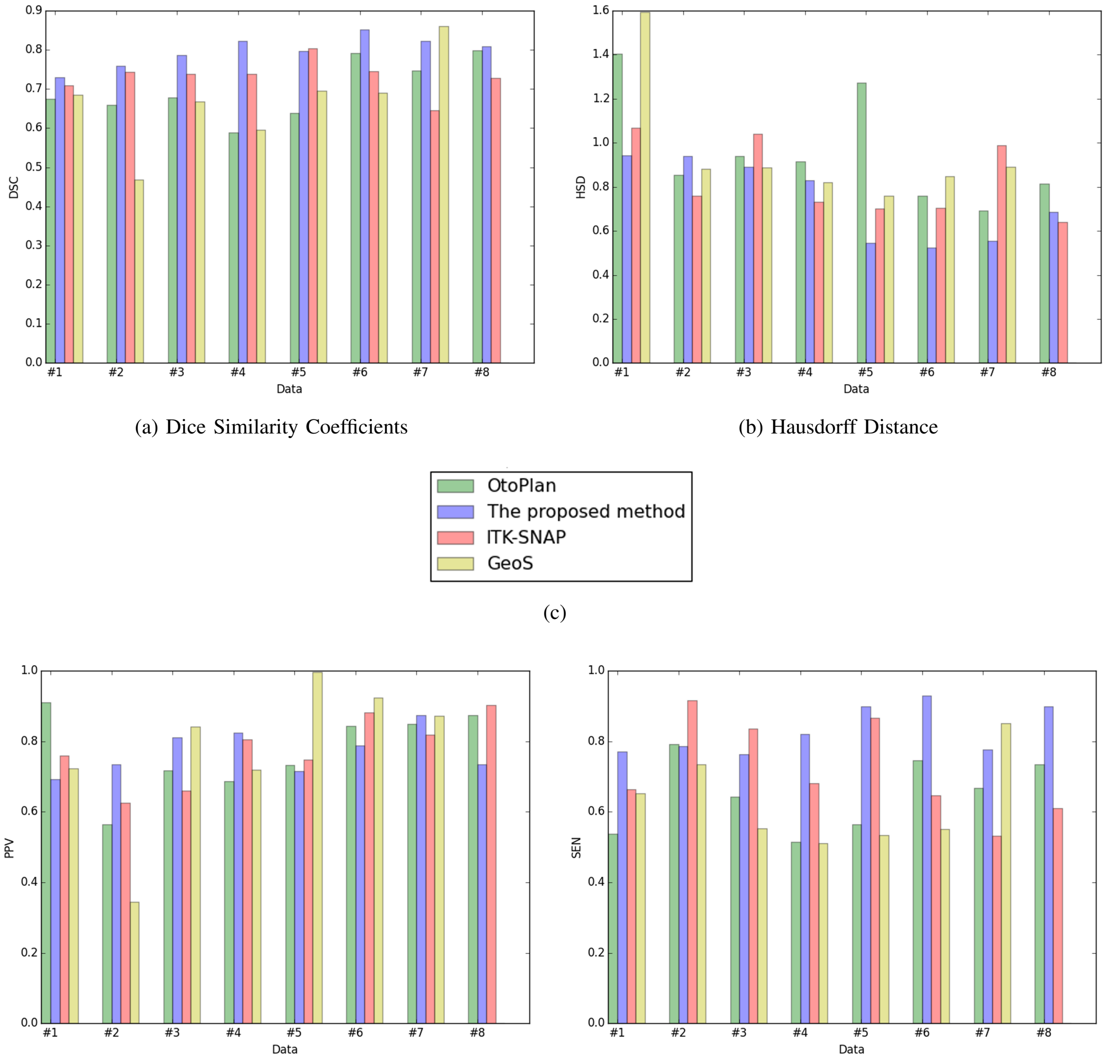
Fig. 4: The facial nerve segmentation comparison on the original CBCT image between the proposed SRC method and other segmentation software — ITK-SNAP and GeoS. The ROI selection via band 16 from OtoPlan initial segmentation.

	CT dataset with band16			CT dataset with band24		
Method	Proposed method (SRC)	GeoS	ITK-SNAP	Proposed method (SRC)	GeoS	ITK-SNAP
Dice	**0.797±0.036(0.802)**∗	0.666±0.109(0.685)	0.731±0.041(0.738)	**0.749±0.048(0.759)**	0.549±0.084(0.545)	0.744±0.049(0.734)
AveDist	**0.129±0.023(0.125)**∗	0.292±0.021(0.292)	0.194±0.036(0.204)	**0.156±0.027(0.155)**	0.484±0.181(0.433)	0.176±0.061(0.163)
RMSE	**0.177±0.033(0.178)**∗	0.353±0.030(0.341)	0.243±0.045(0.251)	**0.216±0.037(0.226)**	0.593±0.247(0.477)	0.229±0.073(0.229)
Hausdorff	**0.739±0.171(0.758)**∗	0.954±0.264(0.883)	0.829±0.162(0.746)	**0.854±0.168(0.901)**	1.524±0.732(1.087)	0.812±0.221(0.848)

TABLE II: Quantitative comparison on CT cases between our method and GeoS and ITK-SNAP, for band sizes 16 (left) and 24 (right). Dice and surface distance errors (in mm). The measurements are given as mean ± standard deviation (median). The best performance is indicated in boldface. The $*$ indicates that SRC results are statistically significantly greater ($p < 0.05$) than GeoS and ITK-SNAP using a two-tailed t-test and Wilcoxon signed ranks tests.

As the proposed approach uses a band surrounding the facial nerve, it already includes air voxels labeled as background to train the model. Therefore, it is expected that the proposed method can handle these cases. However, due to the absence of this type of cases in our database, we were not able to test this point in this study. In the context of the required accuracy for an effective and safe cochlear implantation planning of at least $0.3mm$ [6], analysis of the RMSE error (suitable for this clinical scenario as large errors are to be penalized), the proposed SRC approach is the only one yielding RMSE errors with ranges not surpassing the required accuracy for the tested CBCT/CT cases (Table I & II).

There are some limitations in this study. First, the approach relies on aligned pairs of CBCT/CT and micro-CT images, which are not readily available on all centers. A potential solution to this limitation, is the use of synthetically-generated images from a phantom of known geometry. Similarly, our short-term goal is to prepare a data descriptor in order to make the datasets in this study available for research purposes. Secondly, the learned mapping between clinical and high resolution imaging is specific for the corresponding imaging devices used to generate training data. However, as technical specifications of CBCT/CT imaging devices among different vendors do not differ substantially for facial nerve imaging of cochlear patients, we hypothesize that utilization of an existing super-resolution classification model to a different CBCT/CT vendor might require slight adaptations related to straightforward intensity normalization and histogram matching operations. In this direction, future work includes evaluation of the approach on a large dataset including images from different



(a) Dice Similarity Coefficients

(b) Hausdorff Distance

(c)

(d) Positive Predictive Value

(e) Sensitivity

Fig. 5: Evaluation on CT cases between the proposed super-resolution segmentation method and GeoS (version 2.3.6) and ITK-SNAP (version 3.4.0). Average values for DSC (a), Hausdorff (b), Positive Predictive Value (d), and Sensitivity (e). Case number 8 could not be segmented via GeoS. Note: best seen in colors.

CBCT/CT devices in order to produce a more generally applicable algorithm. Future work also consider a larger dataset of cochlear image datasets including pathological cases to further validate the approach. The segmentation method was made specific to the task of super-resolution segmentation of the facial nerve. Our next step is to extend it to the segmentation of the chorda tympani by creating a dedicated model for it. In order to share the data with the scientific community and to foster future research in this and other related research lines, a data descriptor and open repository will be released.

Another limitation is the computational cost needed to extract features on the upsampled CBCT/CT image (hour range depending on the length of the facial nerve), which is expected to be improved through a pyramidal upsampling scheme, on which features are progressively extracted on each resolution level and concatenated, similar to the pyramid approach recently proposed in [31].

In this study we employed an ad-hoc regularization postprocessing of the resulting segmentation based on morphological operations, aiming at removing isolated small regions and holes in the segmentation. Future work includes the use of a regularization component based on a conditional random field, similar to [32]. In practice, the postprocessing step had a larger impact on the Hausdorff distance metric, as single and isolated voxels outside of the facial nerve region would be used to compute it.

We anticipate that the proposed approach can be seamlessly applied as well to pediatric cases, because it does not rely on shape priors as it is the case of atlas-based methods. Moreover, as demonstrated, this approach can be applied to other image modalities for super-resolution image segmentation, particularly for CT, which is an imaging modality often employed for bone imaging.

V. Conclusions

We have presented an automatic random forest based super-resolution classification (SRC) framework for facial nerve segmentation from CBCT/CT images, which refines a given initial segmentation of the facial nerve to sub-voxel accuracy. Preliminary results on seven 3D CBCT and eight 3D CT ex-vivo datasets suggests that the proposed method achieves accurate segmentations at sub-voxel accuracy.

VI. Acknowledgments

This research is funded by Nano-Tera.ch, scientifically evaluated by the SNSF and financed by the Swiss confederation.

VII. Appendix A

A protocol description of the registration pipeline

1) Registration with Amira:
First, manual alignment of the CBCT/CT image and the corresponding micro-CT image based on manually placed landmarks (from 4 landmarks). Second, first rigid registration with normalized mutual information in Amira (version 5.4.4).
2) Registration with elastix:
First, the rigidly transformed micro-CT image is defined as the moving image, and the CBCT/CT image is defined as the fixed image. In order to preserve the resolution of the micro-CT image and transform it to the CBCT/CT image space, the CBCT/CT image is resampled to micro-CT resolution. Default registration parameters taken from **http://elastix.bigr.nl/wiki/index.php/Default0**.
The resulting non-rigid trasnform parameters is used to transform the ground-truth label image using nearest interpolation. The resulting trasnformed ground-truth image is then used during training of the SRC approach.

VIII. Appendix B

Parameters		
n_estimators	10	the number of tress
criterion	default='gini'	the Gini impurity
max_feature	default='auto'	sqrt(the number of features)
max_depth	default='None'	nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

TABLE III: Employed parameters of the ExtraTreesClassifier in sklearn.

References

[1] B. Bell, N. Gerber, T. Williamson, K. Gavaghan, W. Wimmer, M. Caversaccio, and S. Weber, "In vitro accuracy evaluation of image-guided robot system for direct cochlear access," *Otology Neurotology*, 2013.

[2] N. Gerber, B. Bell, K. Gavaghan, C. Weisstanner, M. Caversaccio, and S. Weber, "Surgical planning tool for robotically assisted hearing aid implantation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, pp. 11–20, 2014.

[3] J. H. Noble, F. M. Warren, R. F. Labadie, and B. M. Dawant, "Automatic segmentation of the facial nerve and chorda tympani in CT images using spatially dependent feature values," *Medical physics*, vol. 35, no. 12, pp. 5375–5384, 2008.

[4] F. A. Reda, J. H. Noble, A. Rivas, T. R. McRackan, R. F. Labadie, and B. M. Dawant, "Automatic segmentation of the facial nerve and chorda tympani in pediatric CT scans," *Medical physics*, vol. 38, no. 10, pp. 5590–5600, 2011.

[5] E. H. Voormolen, M. van Stralen, P. A. Woerdeman, J. P. Pluim, H. J. Noordmans, M. A. Viergever, L. Regli, and J. W. B. van der Sprenkel, "Determination of a facial nerve safety zone for navigated temporal bone surgery," *Operative Neurosurgery*, vol. 70, pp. ons50–ons60, 2012.

[6] J. Schipper, A. Aschendorff, I. Arapakis, T. Klenzner, C. B. Teszler, G. J. Ridder, and R. Laszig, "Navigation as a quality management tool in cochlear implant surgery," *The Journal of Laryngology & Otology*, vol. 118, no. 10, pp. 764–770, 2004.

[7] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[8] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3791–3799.

[9] D. C. Alexander, D. Zikic, J. Zhang, H. Zhang, and A. Criminisi, "Image Quality Transfer via Random Forest Regression : Applications in Diffusion MRI," *Medical Image Computing and Computer-Assisted Intervention MICCAI 2014*, vol. 8675, pp. 225–232, 2014.

[10] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[11] O. Oktay, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, A. de Marvao, S. Cook, D. ORegan, and D. Rueckert, "Multi-input cardiac image super-resolution using convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 246–254.

[12] W. Wimmer, B. Bell, M. E. Huth, C. Weisstanner, N. Gerber, M. Kompis, S. Weber, and M. Caversaccio, "Cone beam and micro-computed tomography validation of manual array insertion for minimally invasive cochlear implantation," *Audiology and Neurotology*, vol. 19, no. 1, pp. 22–30, 2013.

[13] D. Stalling, M. Westerhoff, and H.-C. Hege, "Amira: a highly interactive system for visual data analysis," 2005.

[14] C. Studholme, D. L. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3d medical image alignment," *Pattern recognition*, vol. 32, no. 1, pp. 71–86, 1999.

[15] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: a toolbox for intensity-based

medical image registration," *Medical Imaging, IEEE Transactions on*, vol. 29, no. 1, pp. 196–205, 2010.

[16] D. P. Shamonin, E. E. Bron, B. P. Lelieveldt, M. Smits, S. Klein, and M. Staring, "Fast parallel image registration on CPU and GPU for diagnostic classification of alzheimer's disease," *Frontiers in Neuroinformatics, 7, 2014*, 2014.

[17] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973.

[18] A. Ortiz, A. A. Palacio, J. M. Górriz, J. Ramírez, and D. Salas-González, "Segmentation of brain MRI using SOM-FCM-based method and 3D statistical descriptors," *Computational and mathematical methods in medicine*, vol. 2013, 2013.

[19] V. Chandran, P. Zysset, and M. Reyes, "Prediction of trabecular bone anisotropy from quantitative computed tomography using supervised learning and a novel morphometric feature descriptor," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. Springer, 2015, pp. 621–628.

[20] G. Srinivasan and G. Shobha, "Statistical texture analysis," in *Proceedings of world academy of science, engineering and technology*, vol. 36, 2008, pp. 1264–1269.

[21] P. Elbischger, S. Geerts, K. Sander, G. Ziervogel-Lukas, and P. Sinah, "Algorithmic framework for hep-2 fluorescence pattern classification to aid auto-immune diseases diagnosis," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2009, pp. 562–565.

[22] S. Ghosh and V. Chaudhary, "Feature analysis for automatic classification of hep-2 florescence patterns: Computer-aided diagnosis of auto-immune diseases," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 174–177.

[23] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[24] R. Marée, L. Wehenkel, and P. Geurts, "Extremely randomized trees and random subwindows for image classification, annotation, and retrieval," *Decision Forests for Computer Vision and Medical Image Analysis*, pp. 125–134.

[25] H. J. Johnson, M. M. McCormick, and L. Ibanez, *The ITK Software Guide Book 1: Introduction and Development Guidelines - Volume 1*. USA: Kitware, Inc., 2015.

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

[27] D. P. Huttenlocher, G. Klanderman, W. J. Rucklidge *et al.*, "Comparing images using the hausdorff distance," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 15, no. 9, pp. 850–863, 1993.

[28] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[29] A. Criminisi, T. Sharp, and A. Blake, "Geos: Geodesic image segmentation," in *Computer Vision–ECCV 2008*. Springer, 2008, pp. 99–112.

[30] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *ACM siggraph computer graphics*, vol. 21, no. 4. ACM, 1987, pp. 163–169.

[31] Q. Zhang, A. Bhalerao, E. Dickenson, and C. Hutchinson, "Active appearance pyramids for object parametrisation and fitting," *Medical Image Analysis*, 2016.

[32] R. Meier, V. Karamitsou, S. Habegger, R. Wiest, and M. Reyes, "Parameter learning for crf-based tissue segmentation of brain tumors," in *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer, 2015, pp. 156–167.