IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 29, 2020, accepted August 13, 2020, date of publication August 26, 2020, date of current version September 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019532

Online Learning for the Hyoid Bone Tracking During Swallowing With Neck Movement Adjustment Using Semantic Segmentation

DONGHEON LEE[1], WOO HYUNG LEE[2], (Member, IEEE), HAN GIL SEO[2], BYUNG-MO OH[2,3,4,5], JUNG CHAN LEE[6,7,8], (Member, IEEE), AND HEE CHAN KIM[1,6,8], (Member, IEEE)

[1]Interdisciplinary Program, Bioengineering Major, Graduate School, Seoul National University, Seoul 03080, South Korea
[2]Department of Rehabilitation Medicine, Seoul National University College of Medicine, Seoul National University Hospital, Seoul 03080, South Korea
[3]Institute of Aging, Seoul National University, Seoul 08826, South Korea
[4]Neuroscience Research Institute, Seoul National University College of Medicine, Seoul 03080, South Korea
[5]National Traffic Injury Rehabilitation Hospital, Yangpyeong 12564, South Korea
[6]Department of Biomedical Engineering, Seoul National University College of Medicine, Seoul 03080, South Korea
[7]Department of Biomedical Engineering, Seoul National University Hospital, Seoul 03080, South Korea
[8]Institute of Medical and Biological Engineering, Medical Research Center, Seoul National University, Seoul 03080, South Korea

Corresponding authors: Byung-Mo Oh (keepwiz@gmail.com) and Jung Chan Lee (ljch@snu.ac.kr)

This work was supported by the Seoul National University Hospital Research Fund under Grant 04-2017-0660.

ABSTRACT Swallowing difficulty is a major health concern of the elderly population. The gold standard examination to assess swallowing function is videofluoroscopic swallowing study (VFSS). Hyoid kinematic parameters extracted from VFSS images can be quantitative indicators of swallowing difficulty. In previous studies, its tracking failures are still not resolved when passing through the mandible. Furthermore, it is difficult to be applied in kinematic analysis because the hyoid trajectories can be susceptible to irrelevant neck movements during swallowing. The aim of this study is to develop a robust algorithm for obtaining high-accuracy trajectories of the hyoid bone during swallowing with adjustment of the neck movements. We propose a CNN-based hyoid tracking algorithm which consists of single-domain networks for hyoid tracking and an attention U-Net with conditional random fields for semantic segmentation of the hyoid bone and the cervical vertebrae. The results show that the proposed method can track the hyoid bone robustly compared to the previous methods as measured by a success plot of one-pass evaluation. In addition, the proposed semantic segmentation method achieved the highest dice coefficient for the hyoid bone and the cervical vertebrae. Finally, the obtained hyoid trajectories were evaluated by a root mean squared error, relative error of range of motion, and Pearson's correlation analysis. The proposed algorithm can provide ability to automatically analyze the hyoid motions during swallowing in clinical practice and will potentially enable physician's decision making on diagnostic and therapeutic modalities based on quantitative swallowing assessments.

INDEX TERMS Swallowing difficulty, hyoid bone, cervical vertebrae, online learning, semantic segmentation, convolutional neural networks, videofluoroscopic swallowing study.

I. INTRODUCTION

Swallowing difficulty is a common and major health concern with a 15% to 22% prevalence in the elderly population [1]. It can develop in patients with several neurologic disorders such as dementia, stroke and Parkinson's disease, but can also occur as a normal age-related changes [1]–[3]. Swallowing difficulty in the elderly can result in serious complications such as dehydration, malnutrition, and aspiration pneumonia which may increase hospitalization and mortality rates [1].

The gold standard examination to assess swallowing function is a videofluoroscopic swallowing study (VFSS). In VFSS, swallowing-related anatomic structures and dynamics can be visualized using X-rays during the entire swallowing process including oral, pharyngeal, and esophageal phases [4]. Quantitative kinematic analysis for swallowing

The associate editor coordinating the review of this manuscript and approving it for publication was Chulhong Kim.

difficulty is carried out based on the VFSS images containing information of the anatomic and dynamic properties of swallowing [5], [6]. Unfortunately, the swallowing kinematic analysis demands labor-intensive and time-consuming processes for the manual marking of the swallowing structures, hence limiting its clinical utility and applicability [7].

Previously, various methods have been proposed to automatically track the swallowing structures such as the hyoid bone for kinematic analysis through image processing and machine learning algorithms. These algorithms include sobel edge detection [7], [8] and active shape matching [9], Haar classifier matching [10] and local binary patterns in the image processing, and recently, convolutional neural networks (CNN) in machine learning [11]. However, in these previously proposed methods, manual corrections of the swallowing kinematic analyses was still required because tracking the hyoid bone is frequently failed when the image contrast is abruptly changed in the duration of passing through the mandible.

In this study, we propose a robust algorithm to automatically track the hyoid bone frame by frame based on a CNN model. The proposed algorithm can update the location of the hyoid bone through online learning during the inference phase as well as the training phase, which enables tracking of the hyoid bone even when it passes through the mandible. Additionally, a segmentation method is applied to automatically segment the cervical vertebral bones that allows us to adjust for neck movements during swallowing by setting up a local coordinate system for every time frame.

This paper is organized as follows. An overview of the previous works on hyoid tracking algorithms is presented in Sec. II. The details of the proposed methods including the hyoid tracking algorithm and semantic segmentation algorithm for the hyoid bone and cervical vertebral bones are given in Sec. III including clinical information and swallowing assessments. A performance comparison between the proposed methods and those of previous studies are presented in Sec. IV. Finally, the discussion and conclusion of this study are presented in Sec. VI and Sec. VII, respectively.

II. RELATED WORK

The usual method to obtain hyoid trajectories in VFSS images has been manual marking the coordinates of the hyoid bone in each frame. However, manual process may accompany with measurement errors as well as inter-and intra-rater variations [12]. It can also be time-consuming and labor-intensive especially in cases with prolonged swallowing duration or high frame rates [7]. These factors have led to limited application of swallowing kinematic analysis in clinical practice. Development of automatic tracking systems for the hyoid bone is important to reduce human errors and workload.

There have been various vision-based approaches to track the hyoid bone including image processing methods. Patrick *et al.* proposed a method to identify the hyoid region of interest manually with sobel edge detection, they applied this to 9 subjects (some healthy and some with swallowing difficulty) [8]. Kim *et al.* also applied the same method to 17 patients with swallowing difficulty [7]. Aung *et al* proposed a 16-point active shape model to set anatomical boundaries with minimal user input. [9]. However, these approaches have limitations in that the number of experimental subjects was not large enough to validate the methods, in addition, these approaches require manual corrections. Hossain *et al.* proposed Haar classifier matching with manual identification of regions, this approach was applied to 350 randomly selected frames from multiple videos [10]. Lee *et al.* developed a software platform for acquiring the trajectory of the hyoid bone using histogram of local binary patterns and of multiscale local binary patterns. It was applied to 10 healthy people in 19 videos and to 8 patients with swallowing difficulty in 50 videos [13]. However, these approaches have limitations in that the number of subjects was again not enough to validate the method, in addition, the recognition performance for the hyoid bone was relatively low due to the diverse shapes of the patients' hyoid bone features obtained by the descriptor.

Recently, Zhang *et al.* proposed a CNN-based model to detect coordinates of the hyoid bone in VFSS as an attempt to overcome these low recognition issues. The architecture used in this previous study was a Single Shot MultiBox Detector (SSD) and it was applied to 256 patients with swallowing difficulty [11]. However, this detection method using a SSD still has a limitation in that tracking failures can occur when the hyoid bone is overlapped by the mandible during swallowing in the fluoroscopic images. The contrast of the hyoid bone in fluoroscopic images can be changed abruptly due to overlaps with the mandible, at which there is no representation in the fluoroscopy of the hyoid that matches the previously trained features and tracking fails.

Acquisition of accurate positional information for the hyoid bone is important in swallowing kinematic analysis. Notably, the positions of the hyoid bone can be substantially affected by neck movements that are irrelevant to the swallowing process. Adjustment of the neck movements is therefore needed to obtain the hyoid trajectories during swallowing, which has not been considered in previously developed the hyoid tracking algorithms. Additionally, the previously proposed methods have usually adopted object detection algorithms that are limited in obtaining accurate the hyoid trajectories because the estimated values are not specific coordinates of the hyoid bone that are just its regions of interest.

III. MATERIAL AND METHODS

A. CLINICAL INFORMATION

The VFSS data were retrospectively acquired from an anonymized VFSS data repository used in previous clinical studies [14]–[16]. The VFSS data were obtained from 77 individuals: healthy individuals (n=26), patients with Parkinson's disease (n=32), and patients with stroke (n=19). The mean age was 64.8 $\pm$ 13.6, ranging from 19 to 94. The research protocol for this study was approved by the

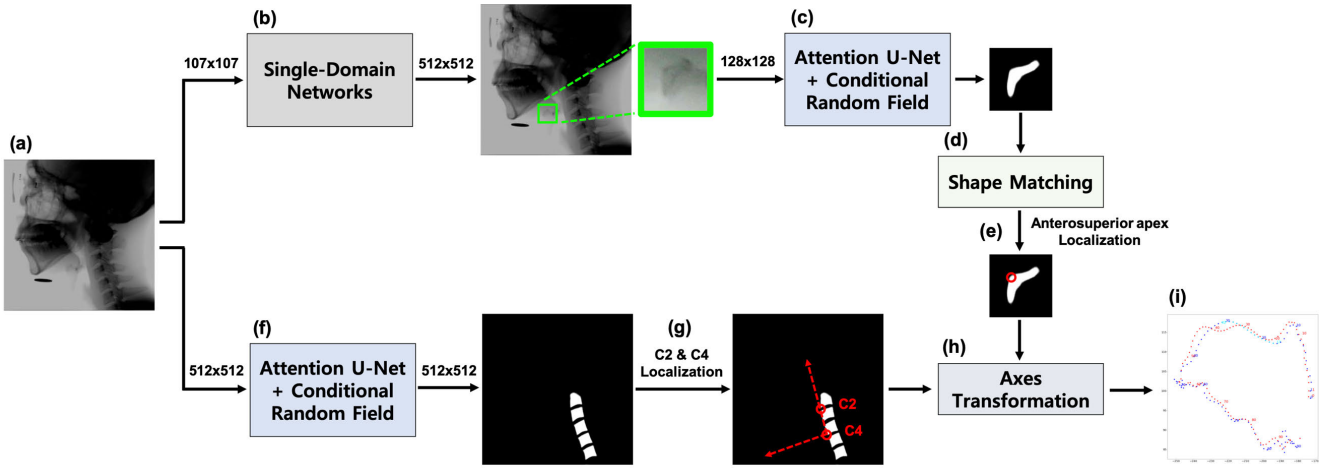


FIGURE 1. Overview of the hyoid tracking system. (a) Original images in videofluoroscopic swallowing study (b) Single-domain networks for hyoid tracking. (c) Attention U-Net with conditional random field (CRF) for automatic hyoid bone segmentation. (d) Shape matching between reference image from initial frame and all frames of the hyoid bone image. (e) Localizing center point of the hyoid bone. (f) Attention U-Net with CRF for automatic cervical vertebrae segmentation. (g) Localizing c2 and c4 in cervical vertebrae. (h) Axes transformation to calibrate a trajectory of the hyoid bone. (i) The result of calibrated hyoid trajectory and ground truth.

Institutional Review Board (IRB No. 1707-178-875) and informed consent was exempted from this retrospective study. All experiments were performed in accordance with relevant guidelines and regulations.

B. SWALLOWING ASSESSMENT

All VFSS were performed by physiatrists with the assistance of radiologic technologists. The individuals were seated in upright position and were viewed in the lateral projection. The volume of the administered liquid bolus was 2mL of a 35% w/v diluted barium solution (Solutop Suspension, Tae Joon Pharm Corp., Ltd., Seoul, Korea) for all subjects. To control the quality of the VFSS images, only images acquired from the identical fluoroscopy equipment (SONIALVISION G4R ®, Shimadzu Corporation, Kyoto, Japan) were used in this study. VFSS image files were obtained by a frame grabber board and image processing software (Pinnacle System Inc., Mountain View, CA, USA). The VFSS images were recorded at a rate of 30 images per second at the set 1280×1024 resolution. The length of the VFSS videos ranged from approximately 2 to 5 seconds. Swallowing motion analysis software, called the spatio-temporal analyzer for motion and physiologic study (STAMPS; https://github.com/cmookj/stamps), was used to obtain positional data of the hyoid bone [17]. The position of the hyoid bone was determined by two experienced examiners who were blinded to the results of the developed algorithm.

C. ONLINE LEARNING FOR TRACKING THE HYOID BONE

The overview of the hyoid tracking system is represented in Fig. 1. The hyoid tracking algorithm used in this study was from Multi-Domain Networks (MDNet), which is a CNN-based online learning model [18]. In the training phase, MDNet is pretrained with ImageNet datasets to learn common representations [19], then it is trained for hyoid datasets as a specific domain. It was originally designed to fit multi-domains by dividing the fully connected layer ends into k numbers of branches. In this study, MDNet was used as a single-domain networks (SDNet), so k is 1, tracking only the hyoid bone. In the inference phase, the weights were updated in the fully connected layers using online learning. SDNet performed complementary functions tracking the hyoid bone with robustness and adaptiveness through long-term and short-term updates. Long-term updates use positive samples collected at regular and long periods. On the other hand, the short-term updates use positive samples collected over a short period. This is carried out whenever potential tracking failures occur. It is triggered when a positive score of the estimated target which becomes less than 0.5, is detected. Additionally, in both training and inference phases, negative samples are collected and used during only the short-term periods. Negative samples obtained over the short-term are screened using hard negative mining [20], which means that the high scores are collected from the obtained positive samples. Fig. 2 shows the architecture of single-domain networks in training and inference phases.

D. SEMANTIC SEGMENTATION FOR THE HYOID BONE AND CERVICAL VERTEBRAE

In this study, a U-Net based CNN architecture was implemented for segmentation of both the cervical vertebrae and the hyoid bone. It is a variant of an encoder-decoder architecture, therefore coarse and fine feature maps can be obtained via skip connections [21]. We added an attention gate to the U-Net architecture that trains specific layers or nodes to find the context of the local region that should be focused on [22]. We also used batch normalization [23] and the parametric ReLU (PReLU) activation function [24]. Finally, noise reduction was achieved by applying a fully connected

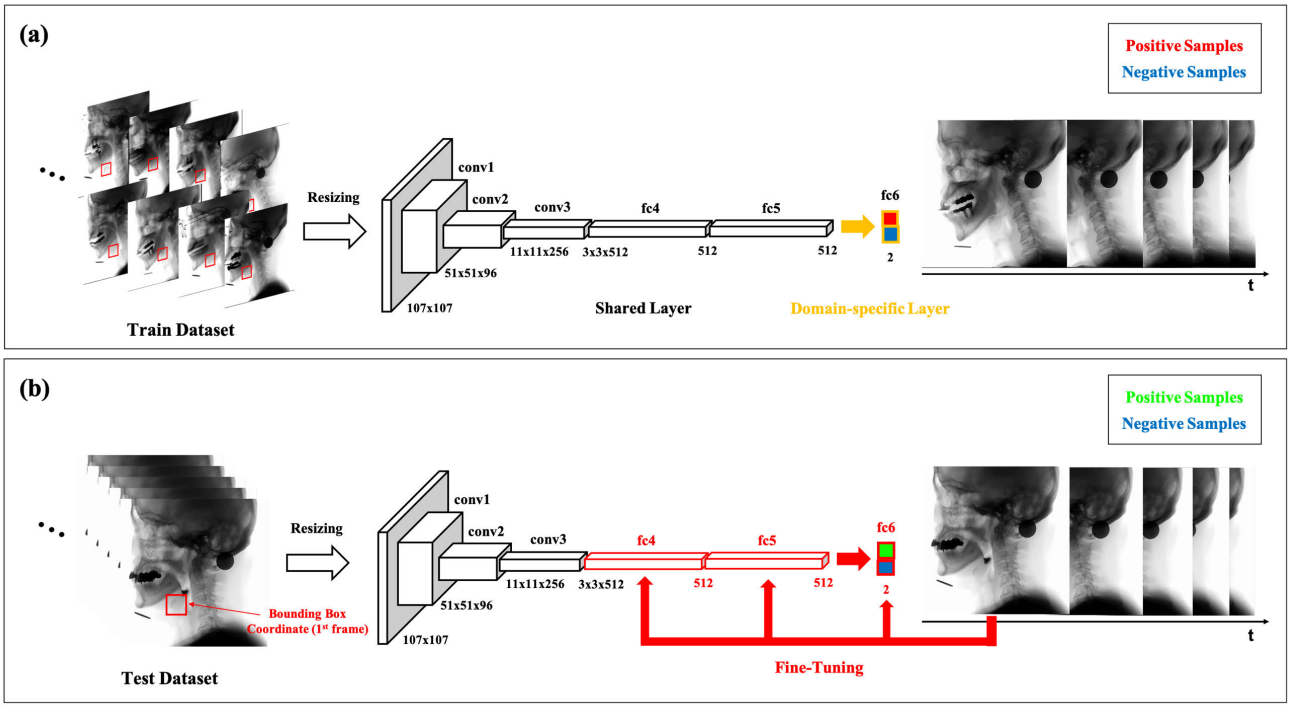


FIGURE 2. The architecture of single-domain networks. (a) Training phase: red and blue bounding boxes denote the positive and negative samples in x-ray domain. (b) Inference phase: green and blue bounding boxes denote the positive and negative samples, and fine-tune the fully connected layers.

Conditional Random Field (CRF) to the segmentation results of the proposed network [25]. The architecture of U-Net has some limitations in that feature maps are reduced in the process of repeating max pooling, which leads to the loss of detailed features. The loss of detailed information on the cervical vertebrae affects the axes transform, therefore, a CRF method is applied as a post-processing to prevent the loss of fine features. The CRF equation consists of appearance and smoothness kernels as follows.

$$k\left(f_i, f_j\right) = w^{(1)} e\left(-\frac{\left|p_i - p_j\right|^2}{2\theta_\alpha^2} - \frac{\left|I_i - I_j\right|^2}{2\theta_\beta^2}\right) + w^{(2)} e\left(-\frac{\left|p_i - p_j\right|^2}{2\theta_\gamma^2}\right) \quad (1)$$

In this equation, p_i and p_j are the positions of the pixel, and I_i and I_j are the intensity. The first gaussian appearance kernel that close-by pixels with similar color are likely to be in the same class, which is controlled by the degrees of nearness θ_α, and the similarity. The second gaussian smoothness kernel determines a smooth level based on the proximity of the pixels controlled by the smoothness θ_γ. Supplementary Fig. 1 shows the architecture of the proposed semantic segmentation method.

E. SHAPE MATCHING FOR REMOVAL OF THE DEFORMED HYOID BONE

Estimation of the hyoid trajectories needs semantic segmentation of the hyoid bone since the positional data is obtained from the anterosuperior apex of the hyoid bone (Fig. 1e). However, abrupt changes of the image contrast when the hyoid bone is overlapped by the mandible during swallowing may lead to deformation of the object shape, resulting in high prediction error (Supplementary Fig. 2). To mitigate this error effectively, we removed the coordinates of the hyoid bone while passing the mandible and instead interpolation was applied on the x- and y-axis to approximate the correct trajectory. A shape matching method that uses the image moments, including the area, centroid, and information about its orientation was applied to identify the swallowing process during which the hyoid bone passes through the mandible [26]. A reference frame was acquired in the first frame for each patient, then the method was applied 15 frames after VFSS starts, and it was determined as a deformed shape when the result of the image moment was not exceeded less than 2 standard deviations of the average image moment during the first 15 frames. In addition, it must be maintained that the image moment does not exceed the preceding conditions continuously for 5 frames.

IV. EXPERIMENTS AND RESULTS

We used NVIDIA Volta GPU and used PyTorch (v1.2) for SDNet and Keras (v.2.2.4) for an attention U-Net in this study. The algorithms developed for this experiment are as uploaded in https://github.com/dhlee-jubilee/dysphagia

A. PERFORMANCE COMPARISON OF THE HYOID BONE TRACKING

The proposed tracking algorithm based on single-domain networks was pretrained with the ImageNet dataset [19] and

TABLE 1. Performance comparison of semantic segmentation applied on cervical vertebrae and hyoid bone.

Subjects		U-Net [21]	U-Net-S [35]	U-Net++ [36]	Attention U-Net [37]	**Proposed Method**
26 Healthy (DSC)	CV	0.83	0.86	0.91	0.91	**0.93**
	HB	0.81	0.83	0.86	0.87	**0.87**
32 Parkinson (DSC)	CV	0.83	0.85	0.88	0.89	**0.91**
	HB	0.82	0.85	0.88	0.87	**0.88**
19 Stroke (DSC)	CV	0.84	0.87	0.89	0.91	**0.91**
	HB	0.8	0.83	0.84	0.84	**0.85**
77 Total (DSC)	CV	0.83	0.86	0.89	0.9	**0.92**
	HB	0.81	0.84	0.86	0.86	**0.87**

DSC, Dice Similarity Coefficient; CV, Cervical Vertebrae; HB, Hyoid Bone

trained with 845 images from 5 subjects including images with the hyoid bone when passing the mandible. We resized the images to 107 × 107 pixels to match the input size of SDNet that is designed for a real time tracking.

For the online learning in the inference step, the user manually selects a bounding box around the hyoid bone location and the box size is fixed at 80 × 80 in the first frame. It was confirmed through experiments that the periods of long-term and short-term periods showed the highest tracking performance at 30 frames and 5 frames, respectively. In this study, 50 positive and 200 negative samples were used in the training phase and have ≥ 0.7 and ≤ 0.5 Intersection over Union (IoU) overlap ratios with the bounding box of the ground truth, respectively. Also, in the inference phase, the same number of positive and negative samples were used as in the training phase and have ≥ 0.7 and ≤ 0.3 IoU overlap ratios with the bounding box of the ground truth, respectively.

The optimizer used in this study was stochastic gradient descent, the learning rate was 0.0001, and the number of positive and negative samples were 500 and 5,000, respectively. The model was trained for 5,000 epochs until the loss saturation, and the measured precision was 0.719 for the training dataset. The resulting images were restored to the original image size to perform the next step.

The performance of the proposed method were compared to that of the previous CNN-based the hyoid bone detection algorithm and the Single Shot MultiBox Detector (SSD) 500-VGG [11], [27] after one-pass evaluation (OPE) [28]. The results show that the proposed method has an average area under curve score of 0.774 in the success plot of the OPE, as represented in Fig. 3. Qualitative results of the hyoid bone tracking are shown in Fig.4 and in Supplementary Video 1. In Fig.4, the images in which any part of the hyoid bone passes the mandible are marked with a red box.

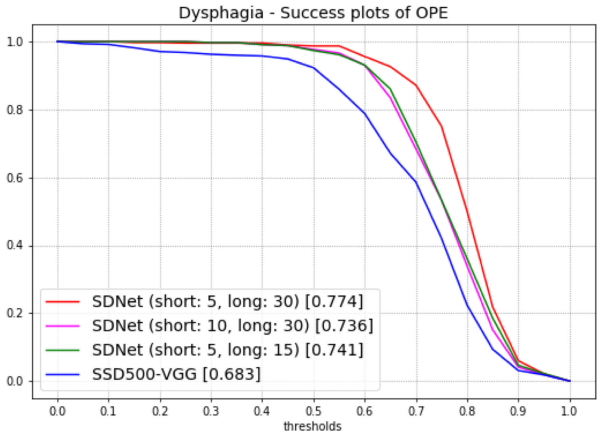


FIGURE 3. Success plots of one-pass evaluations (OPEs) on Single Domain Network (SDNet) and Single-Shot MultiBox Detector (SSD) 500-VGG. The performance of SDNet was compared according to short-term and long-term.

B. PERFORMANCE COMPARISON OF THE SEMANTIC SEGMENTATION

In this study, we compared the performance of the proposed method with various segmentation models based on U-Net [21]. Previous studies on semantic segmentation in the medical domain using CNNs have used U-Net based architectures [29]–[34]. AI Arif *et al.* proposed a shape-aware loss term with U-Net and applied it to the cervical vertebrae [35]. In addition, we compared the performance of U-Net ++ [36], which further enriches the connections between layers using skip connections. Finally, attention U-Net [37] was the structure used for adding an attention gate [22] to the decoder part of the basic U-Net structure. The proposed method was trained for 500 epochs until the loss saturation, the optimizer used was Adam, and 5-folds cross validation was then performed. The measured dice similarity coefficient (DSC) was 0.92 for the training dataset. The total number of test subjects was 77 and the applied algorithms were evaluated quantitatively using DSC. The training dataset and performance results are shown in Fig. 5 and Table 1. The images that were resized to fit the network input size were restored to their original image size to perform the next task.

1) CERVICAL VERTEBRAE

Training datasets for the cervical vertebrae segmentation were obtained from 53 subjects with one frame per subject. The images were resized to 512 × 512 pixels and the uninformative background was removed. The DSC was compared with the ground truth once per 15 frames. The results of the proposed method were 0.93 for healthy individuals,

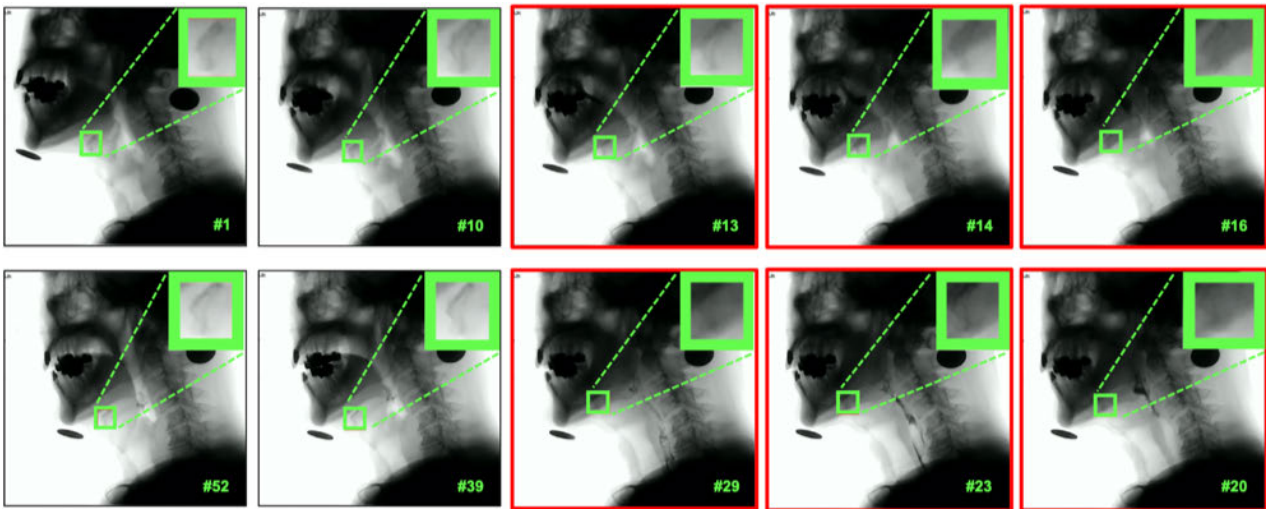


FIGURE 4. Qualitative result of the hyoid bone tracking procedure. Red boxes represent the moment hyoid passes through the mandible.

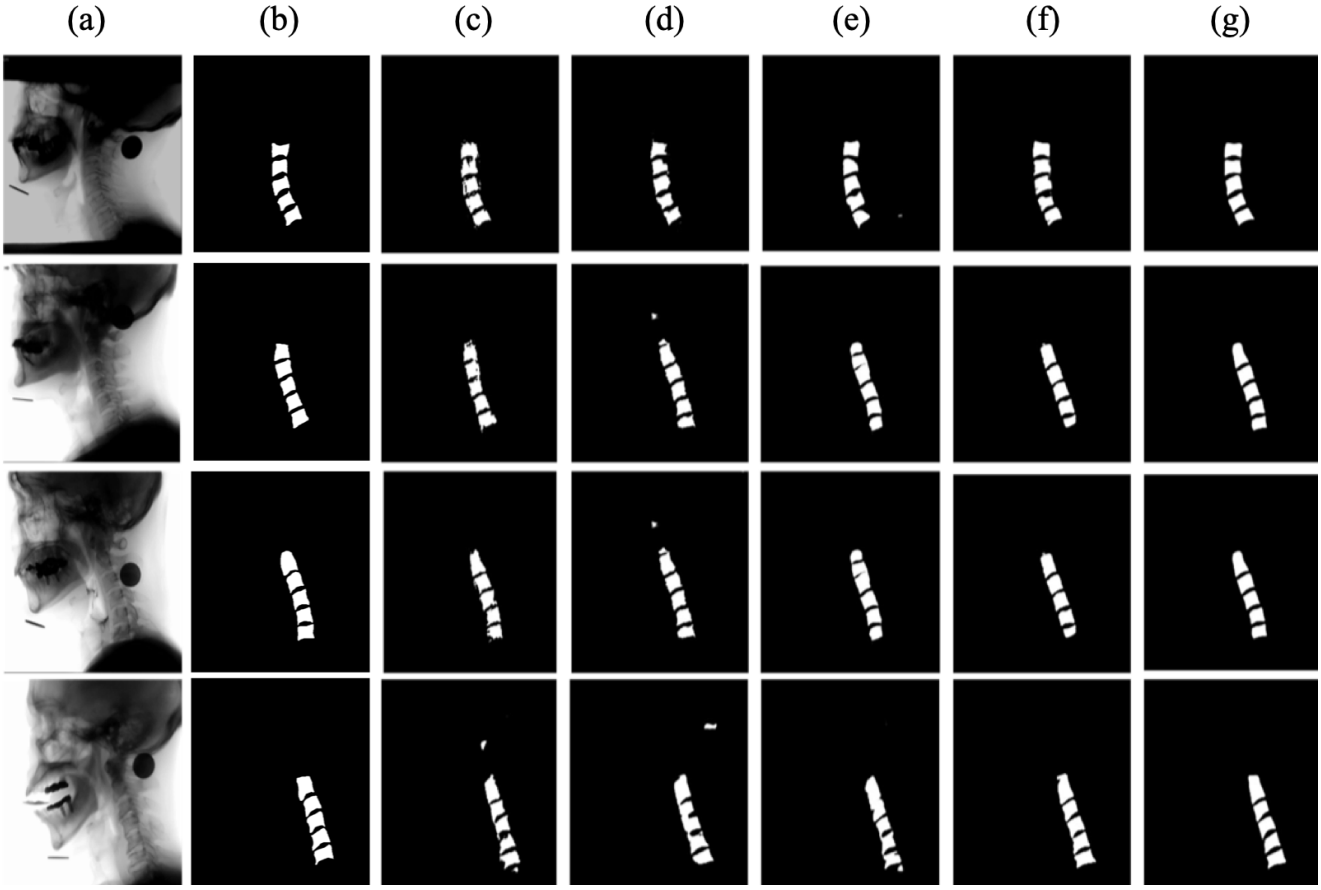


FIGURE 5. Qualitative performance comparison of semantic segmentation methods. (a) Original image (b) Ground truth (c) Result of U-Net (d) Result of U-Net-S (e) Result of U-Net++ (f) Result of Attention U-Net (g) Result of the proposed method.

0.91 for patients with Parkinson's disease, 0.91 for patients with stroke, and a total of 0.92.

2) HYOID BONE

Training datasets for the hyoid bone and tracking algorithm were made up of 845 images from 5 subjects. The images were resized to 128 × 128 pixels. The DSC was compared with the ground truth once per 5 frames, and frames where the hyoid bone was determined as deformed through shape matching were excluded from the evaluation. The DSC results for the proposed method were 0.87 for healthy individuals, 0.88 for patients with

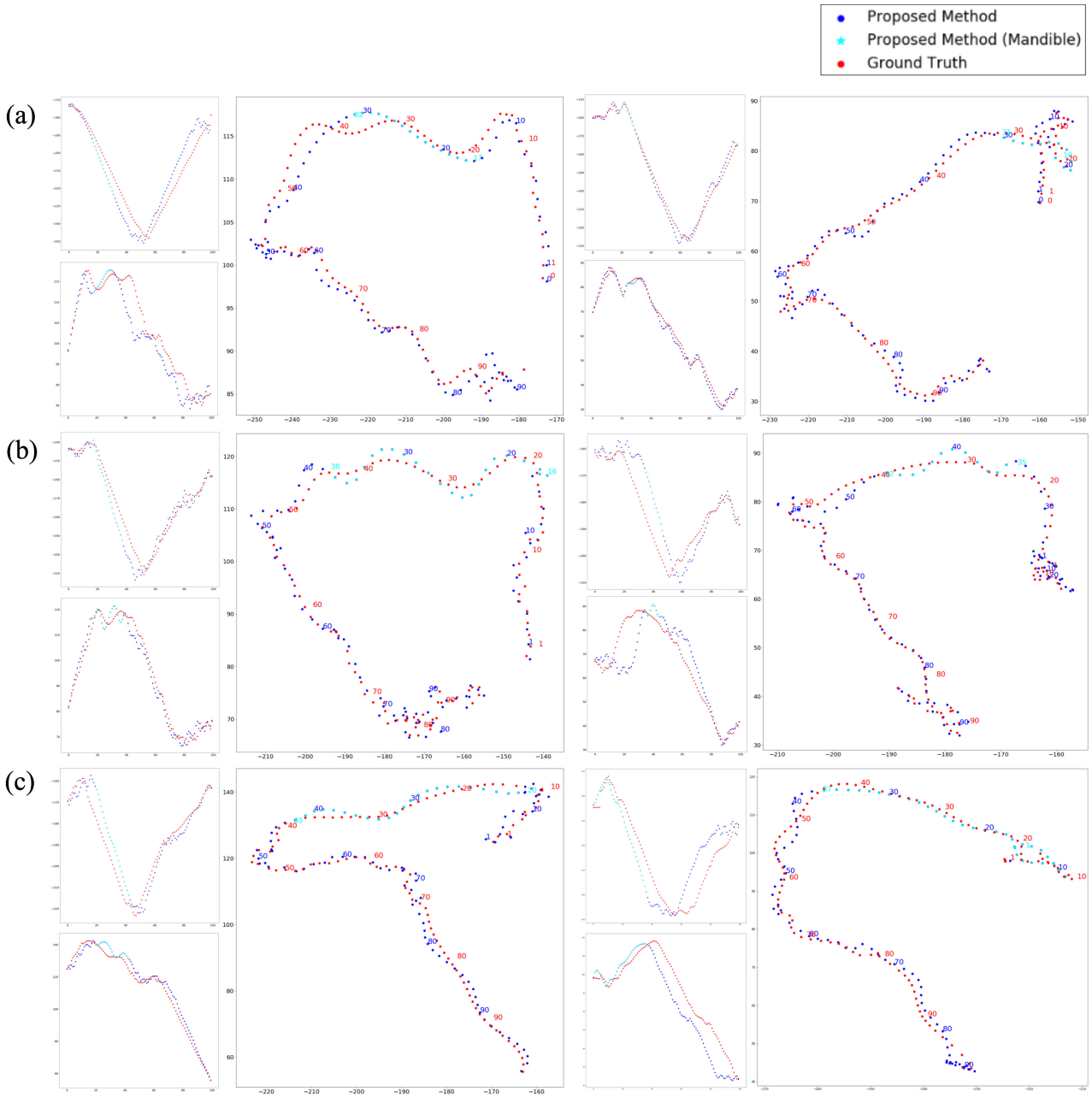


FIGURE 6. Results of calibrated trajectory of the hyoid bone. The sky points represent the moments passing through the mandible and the blue points represent those not passing through the mandible. They were then compared with the ground truth. Left Top: Hyoid trajectory (x-axis). Left Bottom: Hyoid trajectory (y-axis). Right: Hyoid trajectory (2D). (a) Healthy individuals (b) Patients with Parkinson's disease (c) Patients with stroke.

Parkinson's disease, 0.85 for patients with stroke, and a total of 0.87.

C. PERFORMANCE EVALUATION OF THE SHAPE MATCHING

Shape matching [26] was applied to the segmented hyoid bone image, obtained as a result of the semantic segmentation method applies 15 frames after VFSS starts. When the shape matching result was less than 2 standard deviations of average image moment during the first 15 frames and maintained continuously in 5 frames, it was determined as a deformed hyoid bone and removed. The ground truth of the deformed hyoid was determined if the average root mean squared error (RMSE) was increased when 5 consecutive hyoid images had been removed. Out of 77 subjects, 52 had cases of their hyoid bones passing through the mandible. Therefore, the performance evaluation for the shape matching was conducted only for these cases, and the accuracy was calculated by measuring the success rate among the frames when the hyoid passes through the mandible. As a result, the accuracy of the shape matching showed an average of 0.92.

TABLE 2. Evaluation of calibrated hyoid trajectory results.

Subjects	Average RMSE	Relative errors of ROM (x-axis)	Relative errors of ROM (y-axis)	Relative errors of ROM (2D)	Pearson r (x-axis)	Pearson r (y-axis)
26 Healthy (pixel)	7.21 ± 0.93	3.77	4.42	3.32	0.985	0.974
32 Parkinson (pixel)	9.1 ± 1.16	4.71	3.96	3.66	0.98	0.971
19 Stroke (pixel)	6.53 ± 1.88	3.21	5.1	4.17	0.986	0.98
77 Total (pixel)	7.83 ± 1.7	4.02	4.4	3.67	0.983	0.974

One pixel is 0.44 mm on average.
RMSE, Root Mean Squared Error; ROM, Range of Motion

D. AXES TRANSFORMATION OF THE HYOID BONE AND EVALUATION

In this study, the coordinates of the c2 and c4 were defined as the bottom left position of the corresponding bone. The method for localizing c2 and c4 positions is as follows: the uppermost object of the segmented cervical vertebrae is the c2 area, and the c3 and c4 areas are below it. Therefore, by selecting the first and third objects on the y-axis and obtaining the bottom left coordinates, the c2 and c4 can be localized as shown in Fig 1g.

By acquiring the c2 and c4 coordinates, we transformed to a new coordinate axes where the line connecting the c2 and c4 becomes the y-axis. We then measured the coordinate differences between the proposed algorithm and the ground truth. The error measured in the fluoroscopy is in pixel units, this was also converted into mm units by measuring the length of the marker attached to the patient's jaw in advance. The size of the marker before fluoroscopy was 24 mm.

First, we evaluated the pixel difference between the hyoid trajectory coordinates obtained by the proposed method and ground truth coordinates through RMSE. The RMSE of the average trajectory coordinates was 7.83 pixels. In addition, the range of motion (ROM) that is the maximum distance of the trajectory was measured. The relative error, which is the difference between the predicted ROMs and their ground truth, was normalized by the ROM of the algorithm. As a result, the average ROM was 4.02 pixels along the x-axis, 4.4 pixels along the y-axis, and 3.67 pixels in 2D, respectively. Finally, Pearson's correlation analysis was used to measure the similarity between the predicted trajectories and their ground truth. The average Pearson's correlation coefficients were 0.983 along the x-axis and 0.974 along the y-axis. Results for the calibrated hyoid trajectories and evaluation were shown in Fig. 6 and Table 2.

V. DISCUSSION

In this study, we proposed two types of deep neural networks as the main methods for tracking the hyoid bone in VFSS images. The key for training a single-domain network is to set parameters for short-term and long-term updates during online learning. Unlike tracking an object in general domains, the period terms was set relatively small because the movement of the hyoid bone should be tracked for a short duration due to the fluoroscopic video recording being only 2 to 5 seconds long. Parameter optimization for the VFSS images was performed through various experiments, we showed that the hyoid bone can be traced stably even when it passes through the mandible as shown in Fig. 3-4 and Supplementary Video 1.

Furthermore, this study applied a semantic segmentation algorithm to the hyoid bone and the cervical vertebrae. Using the proposed semantic segmentation method, both anatomical structures are automatically segmented with high performance in all frames. The previous study that we compared the proposed method to used U-Net based architectures that had been gradually developed in various ways. The method using a modified loss function is known as U-Net-S [35], an advanced methods from this architecture is called U-Net ++ [36], and another method was the attention U-Net [37]. By localizing the c2 and c4 coordinates in the segmented cervical vertebrae using the proposed method, converting them to the new y-axis coordinates, the noise from the patient's neck movements could be minimized. In addition, it is possible to obtain accurate trajectories by localizing the exact coordinates of the anterosuperior apex of the hyoid bone (Fig. 5 and Table 1). As shown in the Table 1, the proposed algorithm has the best performance, even though it does not differ significantly from the results of the other U-Net based methods. This is because the cervical vertebrae usually has a relatively clear boundaries from the background so that it is less affected by the type of CNN models. It was the first attempt to automate cervical vertebrae segmentation using CNN to obtain an adjusted trajectory of the hyoid bone.

When comparing and analyzing the trajectories obtained by the algorithm to the ground truth, the predicted y-axis coordinates tended to fluctuate slightly above and below the ground truth because the maximum value of the y coordinate of the hyoid bone passes through the mandible and is affected by localization errors of the predicted anterosuperior apex of the hyoid bone, c2 and c4 coordinates.

In two-dimensional X-ray images, three-dimensional structures are visualized as overlapped and mixed so cannot be accurately presented [3]. Especially in the analysis of fluoroscopic images, this property may cause abrupt changes of the object contrast during continuous X-ray radiation, which can interfere with tracking the target object properly. In developing algorithms to trace the hyoid bone during swallowing, tracking failures usually occur when it is overlapped by the mandible. The algorithm developed in this study showed robust performance using the online learning method that trains positive samples obtained at a long-term and negative samples at a short-term. The proposed algorithm can be potentially utilized for the analysis of image data from other organ systems obtained by fluoroscopy including pulmonary, cardiovascular, genitourinary, musculoskeletal, and gastrointestinal organs.

The main limitation of this study is that the user has to specify a bounding box of the appropriate size at the hyoid position in the first frame before the tracking system can be used. For an automated system, a deep learning-based detection algorithm such as SSD [27] can be used to detect the hyoid bone in the first frame, however, it is too much to develop a system to implement such a simple function. Also, when comparing the performance of the hyoid tracking systems, deep learning-based detection algorithms often detect the location of the hyoid-like jaw tip, which can be a crucial error when acquiring the trajectory of the hyoid. In addition, there are problems in predicting the results of multiple bounding boxes including the area surrounding the hyoid bone. Furthermore, the position of the hyoid bone must be manually specified in the first frame in SDNet [18], therefore, the hyoid position can be robustly tracked by adaptively changing the bounding box size with respect to the surrounding background during the inference step. Finally, since the proposed model is operated using both tracking and segmentation algorithms in a cascaded manner, the performance of segmentation algorithm can be affected by that of tracking algorithm. Thus, the overall system performance can be mainly dependent on the performance of the tracking algorithm.

VI. CONCLUSION

In this study, we propose a CNN-based online learning algorithm to track and acquire high-accuracy trajectories of the hyoid bone during swallowing with adjustments of the neck movements. We prove that the developed methods are able to track the hyoid bone even when it passes through the mandible. In addition, the hyoid trajectories can be obtained with high-accuracy by semantic segmentation of the cervical vertebrae and hyoid bone, and subsequent transformation of the coordinate axes. The proposed algorithms provide the opportunity to conduct swallowing kinematic analysis for patients with swallowing difficulty in clinical practice. It can potentially enhance physician's decision making in diagnostic and therapeutic modalities based on quantitative swallowing assessments. Future works will focus on the application of the proposed methods on a large-sized dataset of hyoid motions from patients with swallowing difficulty and will develop data-driven diagnostic or prognostic systems based on machine learning to verify clinical significance.

ACKNOWLEDGMENT

(Dongheon Lee and Woo Hyung Lee contributed equally to this work.)

REFERENCES

[1] M. Carr, "Dysphagia in the elderly," *South Afr. Gastroenterol. Rev.*, vol. 8, no. 1, pp. 853–866, May 2010.

[2] M. Crary, L. Sura, A. Madhavan, and G. Carnaby-Mann, "Dysphagia in the elderly: Management and nutritional considerations," *Clin. Interventions Aging*, vol. 7, p. 287, Jul. 2012.

[3] X. Ying, N. J. Barlow, and M. H. Feuston, "Micro-computed tomography and, volumetric imaging in developmental toxicology," in *Reproductive and Developmental Toxicology*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 1183–1205.

[4] G. D. Gramigna, "How to perform video-fluoroscopic swallowing studies," *GI Motility Online*, May 2006. [Online]. Available: https://www.nature.com/gimo/contents/pt1/full/gimo95.html, doi: 10.1038/gimo95.

[5] J. C. Lee, H. G. Seo, W. H. Lee, H. C. Kim, T. R. Han, and B.-M. Oh, "Computer-assisted detection of swallowing difficulty," *Comput. Methods Programs Biomed.*, vol. 134, pp. 79–88, Oct. 2016.

[6] B.-M. Oh, J. H. Lee, H. G. Seo, W. H. Lee, T. R. Han, S. U. Jeong, H. J. Jeong, and Y. J. Sim, "Changes in hyolaryngeal movement during swallowing in the lateral decubitus posture," *Ann. Rehabil. Med.*, vol. 42, no. 3, p. 416, 2018.

[7] W.-S. Kim, P. Zeng, J. Q. Shi, Y. Lee, and N.-J. Paik, "Semi-automatic tracking, smoothing and segmentation of hyoid bone motion from videofluoroscopic swallowing study," *PLoS ONE*, vol. 12, no. 11, Nov. 2017, Art. no. e0188684.

[8] P. M. Kellen, D. L. Becker, J. M. Reinhardt, and D. J. Van Daele, "Computer-assisted assessment of hyoid bone motion from videofluoroscopic swallow studies," *Dysphagia*, vol. 25, no. 4, pp. 298–306, Dec. 2010.

[9] M. S. H. Aung, J. Y. Goulermas, S. Stanschus, S. Hamdy, and M. Power, "Automated anatomical demarcation using an active shape model for videofluoroscopic analysis in swallowing," *Med. Eng. Phys.*, vol. 32, no. 10, pp. 1170–1179, Dec. 2010.

[10] I. Hossain, A. Roberts-South, M. Jog, and M. R. El-Sakka, "Semi-automatic assessment of hyoid bone motion in digital videofluoroscopic images," *Comput. Methods Biomech. Biomed. Eng., Imag. Vis.*, vol. 2, no. 1, pp. 25–37, Jan. 2014.

[11] Z. Zhang, J. L. Coyle, and E. Sejdić, "Automatic hyoid bone detection in fluoroscopic images using deep learning," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 12310.

[12] I. Sia, P. Carvajal, G. D. Carnaby-Mann, and M. A. Crary, "Measurement of hyoid and laryngeal displacement in video fluoroscopic swallowing studies: Variability, reliability, and measurement error," *Dysphagia*, vol. 27, no. 2, pp. 192–197, Jun. 2012.

[13] J. C. Lee, K. W. Nam, D. P. Jang, N. J. Paik, J. S. Ryu, and I. Y. Kim, "A supporting platform for semi-automatic hyoid bone tracking and parameter extraction from videofluoroscopic images for the diagnosis of dysphagia patients," *Dysphagia*, vol. 32, no. 2, pp. 315–326, Apr. 2017.

[14] W. H. Lee, M. H. Lim, H. S. Nam, Y. J. Kim, H. G. Seo, M. S. Bang, M. Y. Seong, B.-M. Oh, and S. Kim, "Differential kinematic features of the hyoid bone during swallowing in patients with Parkinson's disease," *J. Electromyogr. Kinesiol.*, vol. 47, pp. 57–64, Aug. 2019.

[15] J.-H. Leigh, B.-M. Oh, H. G. Seo, G. J. Lee, Y. Min, K. Kim, J. C. Lee, and T. R. Han, "Influence of the chin-down and chin-tuck maneuver on the swallowing kinematics of healthy adults," *Dysphagia*, vol. 30, no. 1, pp. 89–98, Feb. 2015.

[16] M. Y. Seong, B. M. Oh, H. G. Seo, and T. R. Han, "Influence of supraglottic swallow on swallowing kinematics: Comparison between the young and the elderly," *J. Korean Dysphagia Soc.*, vol. 8, no. 1, pp. 23–29, 2018.

[17] W. H. Lee, C. Chun, H. G. Seo, S. H. Lee, and B.-M. Oh, "STAMPS: Development and verification of swallowing kinematic analysis software," *Biomed. Eng. OnLine*, vol. 16, no. 1, p. 120, Dec. 2017.
[18] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302.
[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[20] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 39–51, Jan. 1998.
[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[25] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFS with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.
[26] M.-K. Hu, "Visual pattern recognition by moment invariants," *IEEE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962.
[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[28] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.
[29] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using u-net based fully convolutional networks," in *Proc. Annu. Conf. Med. Image Understand. Anal.* Cham, Switzerland: Springer, 2017, pp. 506–517.
[30] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein, "NnU-net: Self-adapting framework for U-net-based medical image segmentation," 2018, *arXiv:1809.10486*. [Online]. Available: http://arxiv.org/abs/1809.10486
[31] P. F. Jaeger, S. A. A. Kohl, S. Bickelhaupt, F. Isensee, T. A. Kuder, H.-P. Schlemmer, and K. H. Maier-Hein, "Retina U-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection," 2018, *arXiv:1811.08661*. [Online]. Available: http://arxiv.org/abs/1811.08661
[32] A. Sevastopolsky, "Optic disc and cup segmentation methods for glaucoma detection with modification of U-net convolutional neural network," *Pattern Recognit. Image Anal.*, vol. 27, no. 3, pp. 618–624, Jul. 2017.
[33] G. Zeng, X. Yang, J. Li, L. Yu, P.-A. Heng, and G. Zheng, "3D U-net with multi-level deep supervision: Fully automatic segmentation of proximal femur in 3D MR images," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2017, pp. 274–282.
[34] M. A. A. Hegazy, M. H. Cho, M. H. Cho, and S. Y. Lee, "U-net based metal segmentation on projection domain for metal artifact reduction in dental CT," *Biomed. Eng. Lett.*, vol. 9, no. 3, pp. 375–385, Aug. 2019.
[35] S. M. R. Al Arif, K. Knapp, and G. Slabaugh, "Shape-aware deep convolutional neural network for vertebrae segmentation," in *Proc. Int. Workshop Challenge Comput. Methods Clin. Appl. Musculoskeletal Imag.* Cham, Switzerland: Springer, 2017, pp. 12–24.
[36] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support.* Cham, Switzerland: Springer, 2018, pp. 3–11.
[37] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: http://arxiv.org/abs/1804.03999

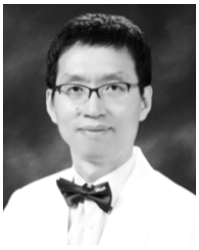
DONGHEON LEE received the B.S. degree in electronic system engineering from Hanyang University, South Korea, in 2008, and the M.S. and Ph.D. degrees in biomedical engineering from Seoul National University, Seoul, South Korea, in 2015 and 2020, respectively. His research interests include application of deep learning and augmented reality to medical images.

WOO HYUNG LEE (Member, IEEE) received the M.D. and Ph.D. degrees in biomedical engineering from the Seoul National University College of Medicine, in 2011 and 2020, respectively. He finished his residency and fellowship with the Rehabilitation Medicine in Seoul National University Hospital, in 2016 and 2017, respectively. As a junior member of the faculty, he has been an Assistant Clinical Professor with the Department of Rehabilitation Medicine, Seoul National University Hospital, since 2020. His research interests include pediatric neurorehabilitation, motor recovery, swallowing rehabilitation, and prognostic modeling.

HAN GIL SEO received the M.D. degree from the Seoul National University College of Medicine, Seoul, South Korea, in 2005, and the M.S. and Ph.D. degrees in medical science from Seoul National University, Seoul, in 2010 and 2018, respectively. From 2014 to 2018, he was an Assistant Professor with the Department of Rehabilitation Medicine, Seoul National University Hospital, Seoul. Since 2018, has been an Associate Professor with the Department of Rehabilitation Medicine, Seoul National University Hospital. His research interests include development and application of rehabilitation technologies for patients with brain disorders, such as stroke and Parkinson's disease. He is a member of the World Federation for NeuroRehabilitation. He was a recipient of the Korean Academy of Rehabilitation Medicine Poster Award, in 2008, and the World Congress of the International Society of Physical and Rehabilitation Medicine Poster Award, in 2013.

BYUNG-MO OH received the M.D. degree from the Seoul National University College of Medicine, Seoul, South Korea, in 2000, and the M.S. and Ph.D. degrees in medical science from Seoul National University, Seoul, in 2003 and 2010, respectively. He has been working as an Assistant Professor, since 2010, and as an Associate Professor, since 2015, with the Department of Rehabilitation Medicine, Seoul National University College of Medicine. His research interest includes the development of effective rehabilitation robotic devices for people with neurological diseases, including stroke, traumatic brain injury, and swallowing difficulty. He is a member of the World Federation for NeuroRehabilitation, the American Heart Association/American Stroke Association, and the Society for Neuroscience. He was a recipient of the Korean Academy of Rehabilitation Medicine Best Scientist Award, in 2009, and the Korea-Japan Neurorehabilitation Conference Best Poster Award, in 2012.

JUNG CHAN LEE (Member, IEEE) received the B.S. degree in mechanical and aerospace engineering and the Ph.D. degree in biomedical engineering from Seoul National University, Seoul, South Korea, in 2001 and 2008, respectively. From 2011 to 2014, he was a Research Professor with the Medical Research Center, Institute of Medical and Biological Engineering, Seoul National University. He joined the Faculty of the Department of Biomedical Engineering, College of Medicine, Seoul National University and Seoul National University Hospital, in 2014, where he is currently a Professor with the Medical Biomechanics and Design Laboratory (MBDL). His current research interests include artificial organs, mechanical pump for medical use, and mechanical design for medical device. He is a member of the Korea Society of Medical and Biological Engineering and the IEEE EMBS.



HEE CHAN KIM (Member, IEEE) received the Ph.D. degree in control and instrumentation engineering (biomedical engineering) from Seoul National University, Seoul, South Korea, in 1989. From 1989 to 1991, he was a Staff Engineer working on National Institute of Health (NIH)-funded Electrohydraulic Total Artificial Heart Project with the Artificial Heart Research Laboratory, The University of Utah, Salt Lake City, UT, USA. He joined the Faculty of the Department of Biomedical Engineering, College of Medicine, Seoul National University and Seoul National University Hospital, in 1991, where he is currently a Professor with the Medical Electronics Laboratory (MELab). He has published over 180 peer-reviewed scientific articles in international journals and holds more than 170 patents. His major research interests include development of intelligent algorithms and electronic instrumentations for medical and biological applications including artificial organs, such as artificial heart and artificial pancreas, biosensors, ubiquitous/mobile healthcare systems, and man–machine interface. He is a member of the Korea Society of Medical and Biological Engineering (KOSOMBE), the IEEE/EMBS, and the American Society of Artificial Internal Organs (ASAIO).