AUTOMATED BOLUS DETECTION IN VIDEOFLUOROSCOPIC IMAGES OF SWALLOWING USING MASK-RCNN

Handenur Caliskan[1]*, Amanda S. Mahoney*[2]*, James L. Coyle*[2]*, Ervin Sejdić*[1]

[1]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA
[2]Department of Communication Science and Disorders, University of Pittsburgh, Pittsburgh, PA, USA
E-mail: esejdic@ieee.org

ABSTRACT

Tracking a liquid or food bolus in videofluoroscopic images during X-ray based diagnostic swallowing examinations is a dominant clinical approach to assess human swallowing function during oral, pharyngeal and esophageal stages of swallowing. This tracking represents a highly challenging problem for clinicians as swallowing is a rapid action. Therefore, we developed a computer-aided method to automate bolus detection and tracking in order to alleviate issues associated with human factors. Specifically, we applied a state-of-the-art deep learning model called Mask-RCNN to detect and segment the bolus in videofluoroscopic image sequences. We trained the algorithm with 450 swallow videos and evaluated with an independent dataset of 50 videos. The algorithm was able to detect and segment the bolus with a mean average precision of 0.49 and an intersection of union of 0.71. The proposed method indicated robust detection results that can help to improve the speed and accuracy of a clinical decision-making process.

1. INTRODUCTION

Oropharyngeal dysphagia (OPD) is defined as difficulty or discomfort in the oral or pharyngeal cavities or esophagus during swallowing. A variety of factors can lead to OPD including structural abnormalities, head-neck cancer, though mostly neurological causes, such as a stroke, neurodegenerative diseases or brain trauma [1][2]. In people with OPD, several of the kinematic events that propel a bolus may be produced in errant sequential patterns that lead to inefficient clearance into the digestive system which results in accumulations of pharyngeal residue that can enter the respiratory system (aspiration) after the swallow. Aspiration can lead to an airway obstruction or to pneumonia, which is associated with mortality rates up to 50% [3]. OPD can also result in malnutrition, dehydration, immune system compromise and decreased quality of life [4]. In the US, 1 in 25 adults is affected by dysphagia annually [5]. Dysphagia is even more prevalent in older adults: it is estimated that the prevalence of dysphagia in community-dwelling older adults over 50 years is between 15% and 22%. The cases in assisted health care facilities are more prevalent: it is reported that 40% to 60% of the population experience feeding difficulties [6]. Adverse medical consequences of OPD add up to $6 billion to US health care expenditures [7].

Clinicians diagnose and develop management plans for patients with OPD through imaging-based assessments, including fiberoptic endoscopic evaluation of swallowing (FEES) or videofluoroscopic swallow studies (VFSS). VFSS is a real-time X-ray video, which allows clinicians to view and evaluate the biomechanical characteristics of all stages of swallowing [8]. It is considered the gold standard to assess the oral and pharyngeal dynamics of swallowing [9]. Among measures performed by OPD diagnosticians using VFSS images to predict risk are quantification of pharyngeal residue and penetration of the upper airway (larynx) and aspiration into the trachea. Human judgment has been the gold standard for making these measurements. However, technological developments and increases in computational horsepower over the past decade are enabling the addition of algorithm-based electronic data analysis as an adjunct to human judgment which, in turn, reduces human measurement errors that can diminish the consistency of measurement and lower inter- and intra-judge reliability [10] and is more prone to fluctuations due to evaluator's level of expertise. Likewise, proper kinematic analysis is time-consuming: clinicians need to analyze swallowing videos frame by frame to capture key features, which takes time and effort. Regarding these challenges, an automated and more reliable measurement of the bolus passage have the potential to decrease the subjectivity and increase the robustness of the evaluation.

Computer aided solutions for medical imaging area provide compelling results to detect, evaluate and diagnose dis-

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute Of Child Health and Human Development of the National Institutes of Health under Award Number R01HD092239, and by the National Science Foundation under Award Number 1652203.

eases. Specifically, there has been a growing interest in deep learning applications in recent years due to its power to perform certain tasks that have been performed traditionally by humans. The state-of-the-art object detection algorithms such as Single Shot Multi-box detectors (SSD), You-Only-Look-Once (YOLO) are powerful to create real-time results with a desired accuracy [11][12]. Few researchers have attempted to apply deep learning on VFSS to address swallowing related problems. Zhang et.al. analyzed hyoid bone movement via the application of an object tracking algorithm called Single Shot Detector [1]. They reached the mean average presicison of 0.89. Mao et.al. proposed a different approach applying deep learning methods for hyoid bone movement tracking with the support of neck sensors. They explored the relationship between sensor signals and hyoid bone movement by utilizing Recurrent Neural Networks [13]. They successfully obtained a feasible way to track hyoid bone movement solely based on sensor information. To the best of our knowledge, applying machine learning methods for bolus movement in VFSS has not been investigated yet. Therefore, in this work, we proposed an automated method for bolus detection with Mask-RCNN to trace the bolus accurately in all frames of a VFSS video which would potentially help to automate and improve the accuracy of the evaluation process.

2. METHODS

2.1. Data Collection

Data collection occurred at the University of Pittsburgh Medical Center Presbyterian Hospital following Institutional Review Board approval and participant informed consent. Videofluoroscopic images were collected from 30 patients with suspected dysphagia referred for a swallowing evaluation with VFSS. Data was collected within the course of standard clinical care, raising the external validity of our data collection methods to correspond to ordinary clinical management, and to test the robustness of the analysis methods with typically-obtained data rather than imposing laboratory-controlled conditions. The number of swallows performed by participants, their head-neck positions during swallowing, and bolus consistencies were controlled by the evaluating clinician based on each patient's OPD patterns. The consistencies of swallowed material used were Varibar nectar (300 cPs viscosity), Varibar pudding (5000 cPs viscosity), E-Z-EM Canada, Inc., Varibar thin (Bracco Diagnostics, Inc.) (<5cPs viscosity) and Keebler Sandies Mini Simply Shortbread Cookies (Kellogg Sales Company) coated with Varibar pudding. Clinicians administered some boluses by spoon (3-5mL) and in other conditions patients self-administered a comfortable sized sip by cup to replicate typical drinking patterns. The videofluoroscopy unit was set to a pulse rate of 30 PPS and digitized at a sampling rate of 60 frames per second (FPS) via a video card (AccuStream Express HD, Foresight Imaging, Chelmsford, MA). The videos were recorded as 2D movie clips of 792 x 1080 pixel resolution via LabVIEW software. To eliminate duplicate frames, the videos were then down-sampled to 30 FPS. 450 swallows constituted the dataset analyzed in this study.

2.2. Network Architecture

Mask-R-CNN was first proposed by He et. al. from Facebook AI research [14]. Mask-R-CNN is an extended version of Faster-R-CNN with an additional branch for the semantic segmentation task, which works in parallel with the existing structure. As shown in Figure 1, Mask-R-CNN has two main modules: The first module involves feature extraction from a given image through a base network and the region proposal network, which are both convolutional structures. For the first feature extraction, ResNet50 and ResNet101 architectures are commonly used as the base network structure [15]. After this step, a Region Proposal Network (RPN) with a fully convolutional structure takes the output feature map of the base network and proposes a set of rectangular box predictions for an object based on its objectness score. ZFNet and VGGNet are commonly used CNN structures for the region proposal networks [16][17][18]. To generate region proposals, a kernel window is slid on the last output feature map of the backbone network. The kernel creates k number of anchors, which are prototypical object boxes for each location. RPN makes 2k predictions, as to whether the window has the object or not, and 4k predictions for the bounding box regression.

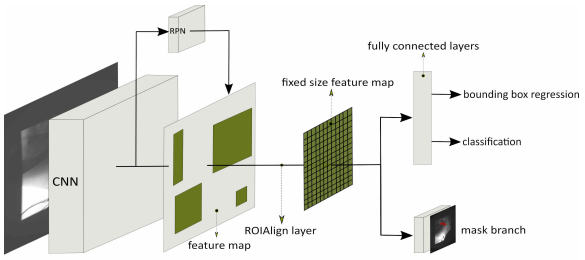

Fig. 1: Mask-R-CNN Architecture: Module 1: Initial features are extracted from images through a CNN backbone. RPN makes predictions about the regions of interest, giving the remaining part of the module the knowledge to find a potential object in the image. The ROIAlign layer normalizes the misalignment between the ROI and the extracted features. Module 2: After the fixed size feature map is created, a fully connected layer performs the bounding box regression and the object classification. The additional convolutional mask branch creates a mask for every region of interest.

The feature map is sent to a box-classification layer and a box-regression layer. After the proposed regions layers, there is an additional region of interest alignment (ROIAlign) layer that fixes the quantization misalignment between the ROI and extracted features that stem from ROI Pooling. Without the ROIAlign layer, the pixel-wise accuracy of the mask predic-

tion is negatively affected. The second module of the Mask-R-CNN structure is called the network head, which includes fully connected layers for the bounding box recognition and an instance mask prediction, which is applied separately for each ROI [14]. The loss function is defined as the multi-task loss on every sampled ROI as $L = L_{cls} + L_{box} + L_{mask}$. L_{mask} is defined as the average binary cross-entropy loss [14].

2.3. Training and Testing

The videos were manually labeled by experts trained in frame by frame labeling. The videos were first segmented for the swallowing period from the bolus head entering the pharynx to the clearance of the entire bolus from pharynx area [13]. By delineating the boundaries of the bolus with freehand drawing in ImageJ software, the labels were created as binary mask images for every individual VFSS frame. The raters maintained inter-rater and intra-rater reliability on a separate set of 10 swallows. For this study, we measured the reliability as an average percentage of overlap between two labels throughout a complete swallow. The bolus has a lot of inconsistent shapes due to its fluid structure. For instance, in some frames its shape is extremely thin in a bifurcated form. Another type of a complicated case depends on the rapid movement of the bolus that exceeds the sampling rate of the image intensifier and therefore creates blurry areas. Considering these complicated scenarios, the reliability measures are held as 60% and above. This challenge will be eliminated further by creating a bounding box for the segmented boluses. The masks need to be resized to reduce the computational cost, however, this reduction could affect the accuracy of the algorithm badly after a certain scale. Therefore, we reduced the resolution from 720 x 1080 pixels to 512 x 512 to keep the details of the image while reducing the computational cost. Nearly 10K masks were created and frames were extracted for the training task. To eliminate the computational burden, the images were cropped due to lack of information from boundaries of the frames. The same rationale applied to every individual corresponding mask.

We utilized Mask-RCNN model on Github which is provided by Matterport [19]. A ResNet101 backbone structure pre-trained with a Common Objects in Context (COCO) image dataset was used to extract the key features from the images. The COCO dataset was primarily designed for object detection and segmentation and is frequently used in instance segmentation tasks [20]. The bounding boxes are determined from a set of proposals according to their intersection over union (IoU) ratios. IoU is a metric used in object detection algorithms to express the ratio of the overlap between predicted bounding boxes and its union with the ground truth bounding box: $IoU = Area\,of\,intersection / Area\,of\,union$. We used mean average precision (mAP) to measure the accuracy of the detection. The area under the precision-recall curve represents the mAP value that is between 0 and 1. Precision is defined as the ratio of the true positive predictions to the all positive predictions. Recall is defined as ratio of the true positive predictions to the total of ground truth positives.

If the proposed bounding box has an IoU greater than 0.5, it is assigned as the final detected bounding box proposal of the object. After a bunch of predicted boxes, non-maximum suppression was applied to these boxes and the lower scored predictions were eliminated. The learning rate was chosen to be 0.001, the learning momentum was chosen as 0.9. We used SGD optimizer for our model. The model was trained with 120 epochs with 500 iterations each. 15% percent of the patient-specific training data was used for validation purposes. The test data was created independent from the training data. All the experiments were conducted on NVIDIA Tesla 24 GB GPU using Keras library over a Tensorflow backend.

3. RESULTS

A step-by-step detection process is demonstrated on a randomly selected frame from a randomly selected swallow video in Figure 2. Figure 2(a) illustrates the dotted boxes that are proposed by the region proposal network and the solid line boxes which are going to be refined by ROIAlign layer in the next step.

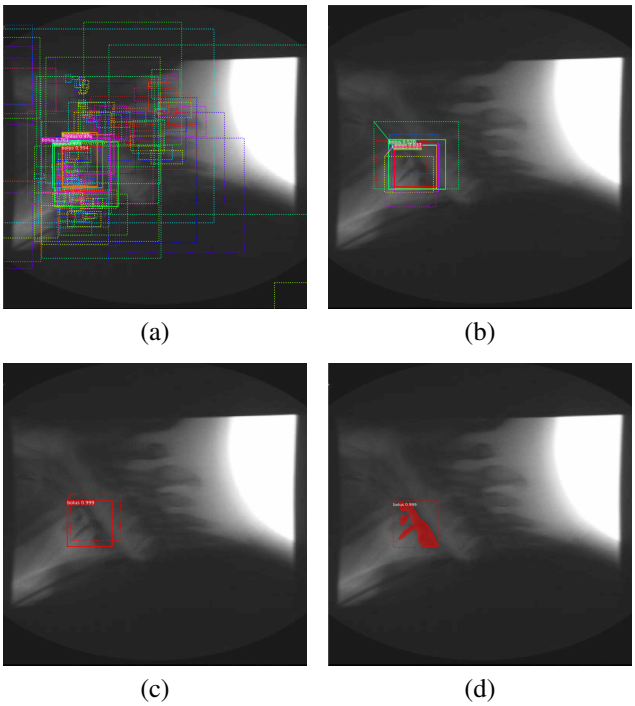

Fig. 2: Bounding box detection steps: (a)ROIs before refinement, (b) ROIs after refinement, (c) Detection after non-maximum suppression, (d) Final detection with mask.

After ROI refinement, the refined bounding boxes are shown in Figure 2(b). Non-maximum suppression was applied to obtain the final bounding box as illustrated in Figure 2(c). The mask with randomly assigned color in Figure 2(d) represents the segmented bolus with its final bounding box. Final detection and segmentation examples with a series of consecutive frames from another randomly selected swallow video are shown in Figure 3.

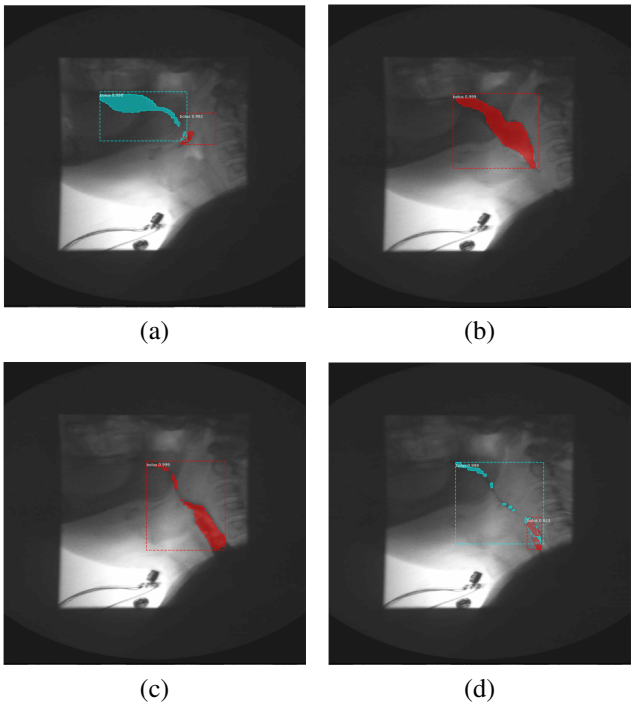

Fig. 3: Exemplary results of detected bolus on different phases of a single swallow: (a) oral phase, (b) early pharyngeal phase, (c) late pharyngeal phase, (d) bolus residue at the end of the swallow.

The colored regions show the bolus segmented from its background. The color palette that the algorithm uses is random coloring. Therefore, both blue and red regions represent multiple instances of the same bolus class. A mAP of 0.49 and IoU of 0.71 for the bounding boxes were achieved on the training data. A mAP of 0.42 was achieved for the independent test data.

4. DISCUSSION

The aim of this study is to automatically identify all parts of a bolus in VFSS images to generate an analysis method designed to reduce human errors. In VFSS, observing bolus shape and structure through a swallow plays an important role in the assessment and treatment of swallowing disorders. It is advantageous for clinicians to have a computer-aided and automated system to improve their efficiency while at the same time preserving accuracy of measurement. We achieved this by applying Mask-RCNN deep learning algorithm that detects and segments objects in a given scene. This work also provides a computational basis for bolus features, which opens the door for further research to be conducted on VFSS. Certain limitations and challenging problems for our study are presented further.

The algorithm performs well when the swallow is clear of any shadows of other structures within the image field (e.g., bones such as the mandible, laryngeal cartilages). The confidence scores are obviously higher when the bolus is more distinguishable among the other parts. In some cases that have no bolus, the algorithm correctly rejects the presence of bolus components while in situation in which a very small amount of bolus is present, the algorithm fails to identify as the bolus is almost invisible. This can often occur due to the high velocity of bolus flow and the limited pulse rate of fluoroscopy units (i.e. 30PPS) which exposes the regions of interest to X-rays once every 1/30 of a second (0.0333 sec.). Furthermore, in case of the presence of materials or structures that contain similar grayscale densities as a bolus (e.g., bones, surgically implanted hardware), the algorithm falsely identified these parts as boluses as well.

We evaluated the performance of Mask-RCNN with a different base network to detect each bolus in VFSS images. We experimented with ResNet50, which is a shallower version of ResNet101. However, the trained network presented a considerably lower accuracy level than ResNet101. Our experimentation showed that bolus detection with Mask-RCNN needs a more complex base than ResNet50.

The findings from this study provide exciting prospects for the future evaluation of swallowing during VFSS by providing a way to computationally evaluate bolus displacement and quantify post-swallow residue patterns more objectively. The performance of the developed algorithm for bolus tracking has some limitations: low quality images and inconsistent images between various X-ray machines can be given as an example of challenging factors in some clinical settings. Moreover, due to low contrast, the bolus was occasionally localized at different ROIs. Therefore, future work should consider improving accuracy despite varied image qualities.

5. CONCLUSION

The aim of this study is to automatically identify bolus in VFSS to eliminate human rater effects. By using a well-known object detection method called Mask-R-CNN, bolus detection and segmentation was accomplished with an mAP of 0.49 and overlap of 0.71. Although we have clear detection results, the accuracy has a room for improvement. This investigation provides a computational base, a faster evaluation process, and a more objective method for clinicians to track bolus flow during VFSS. Future studies intend to analyze bolus further such as detecting presence of bolus residue after a swallow with deep learning applications.

6. References

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

[2] J. M. Dudik, I. Jestrovic, B. Luan, J.L. Coyle, and E. Sejdić, "A comparative analysis of swallowing accelerometry and sounds during saliva swallows," *Biomedical Engineering Online*, vol. 14, no. 1, pp. 3–1–15, 01 2015.

[3] P. Clavé, V. Arreola, M. Romea, L. Medina, E. Palomera, and M. Serra-Prat, "Accuracy of the volume-viscosity swallow test for clinical screening of oropharyngeal dysphagia and aspiration," *Clinical Nutrition*, vol. 27, no. 6, pp. 806 – 815, 2008.

[4] D. L. Doggett, K. A. Tappe, M. D. Mitchell, R. Chapell, V. Coates, and C. M. Turkelson, "Prevention of pneumonia in elderly stroke patients by systematic diagnosis and treatment of dysphagia: An evidence-based comprehensive analysis of the literature," *Dysphagia*, vol. 16, no. 4, pp. 279–295, Oct. 2001.

[5] N. Bhattacharyya, "The prevalence of dysphagia among adults in the united states," *Otolaryngology Head and Neck Surgery*, vol. 151, no. 5, pp. 765–769, 2014.

[6] M. Aslam and M.F. Vaezi, "Dysphagia in the elderly," *Gastroenterology and Hepatology*, vol. 9(12), pp. 784–95, Dec. 2013.

[7] C.-P. Wu, Y.-w. Chen, M.-J. Wang, and E. Pinelis, "National trends in admission for aspiration pneumonia in the united states, 2002-2012," *Annals of the American Thoracic Society*, vol. 14, no. 6, pp. 874–879, 2017.

[8] L. East, K. Nettles, A. Vansant, and S. K. Daniels, "Evaluation of oropharyngeal dysphagia with the videofluoroscopic swallowing study," *Journal of Radiology Nursing*, vol. 33, no. 1, pp. 9 – 13, 2014.

[9] M. Costa and B. Melciades, "Videofluoroscopy: the gold standard exam for studying swallowing and its dysfunction," *Arquivos de Gastroenterologia*, vol. 47, no. 4, pp. 327 – 328, Dec. 2010.

[10] L. Baijens, A. Barikroo, and W. Pilz, "Intrarater and interrater reliability for measurements in videofluoroscopy of swallowing," *European Journal of Radiology*, vol. 82, no. 10, pp. 1683–1695, June 2013.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, June 27-30 2016, pp. 779–788.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, Scott R. E., Cheng-Yang F., and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015.

[13] S. Mao, Z. Zhang, Y. Khalifa, C. Donohue, J. L. Coyle, and E. Sejdić, "Neck sensor-supported hyoid bone movement tracking during swallowing," *Royal Society Open Science*, vol. 6, pp. 181982–1–11, July 2019.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceeding of the 2017 IEEE International Conference on Computer Vision (ICCV 2017)*, Venice, Italy, Oct 22-29 2017, pp. 2961–2969.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, Boston, MA, June 7-12 2015, pp. 770–778.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, June 2017.

[17] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *Proceedings of the Third IAPR Asian Conference on Pattern Recognition (ACPR 2015)*, Piscataway, NJ, Nov. 3-6 2015, pp. 730–734.

[18] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of European Conference Computer Vision (ECCV 2014)*, Zurich, Switzerland, Sep. 6-12 2014, pp. 818–833.

[19] Waleed Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow," `https://github.com/matterport/Mask_RCNN`, 2017.

[20] T. Yi Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of European Conference Computer Vision (ECCV 2014)*, Zurich, Switzerland, Sep. 6-12 2014, pp. 740–755.