Proceedings of the ASME 2020
International Design Engineering Technical Conferences
and Computers and Information in Engineering Conference
IDETC/CIE2020
August 17-19, 2020, Virtual, Online

DETC2020-22190

AUTONOMOUS CRICOTHYROID MEMBRANE DETECTION USING NEURAL NETWORKS FOR FIRST-AID SURGICAL AIRWAY MANAGEMENT

Xiaoxue Han *
Robotics and Mechatronics Lab
Department of Mechanical Engineering
Virginia Tech,
Blacksburg, Virginia 24060

Hailin Ren *
Robotics and Mechatronics Lab
Department of Mechanical Engineering
Virginia Tech,
Blacksburg, Virginia 24060

Pinhas Ben-Tzvi †
Robotics and Mechatronics Lab
Department of Mechanical Engineering
Virginia Tech,
Blacksburg, Virginia 24060

ABSTRACT

Airway management is one of the most important priorities when dealing with patients with severe injuries, but knowledge of the important anatomy and physiology is needed for providers to perform a successful surgery . This paper provides a solution for the precise cricothyroid membrane detection problem for real-time surgical airway management applications. With a commercial compact and portable cricothyrotomy kit, the proposed method will enable providers with general knowledge to perform successful first-aid airway management. In this paper, we propose a Hybrid Neural Network (HNNet), consisting of two parallel computing ensembles. The first ensemble takes as an input a low-resolution global image and outputs the Region-of-Interest (ROI) from the predefined grids. The high-resolution image is then cropped according to the ROI, and fed into the second ensemble to achieve precise keypoint detection. Global features and their spatial information from the first ensemble are also fed into the second ensemble to improve the precision. A dataset that consists of over 16,000 images from 13 subjects is built, and the location of the cricothyroid membrane in each image is precisely labeled by medical experts. The training results are presented to show both the efficiency and improved performance of our proposed method compared to existing ones.

*Authors contributed equally
†Corresponding author – bentzvi@vt.edu

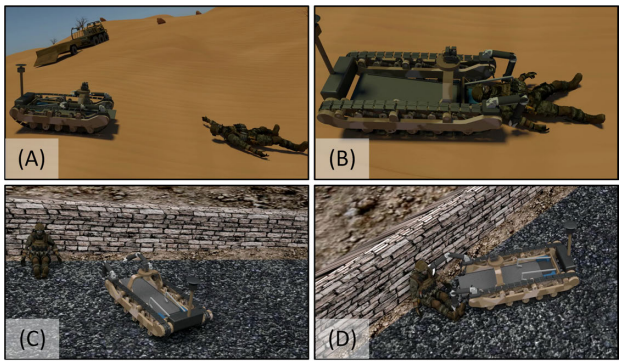

Figure 1. A Semi-Autonomous Victim Extraction Robot (SAVER) serves for victim extraction

1 INTRODUCTION

Keeping the airway open and clean is crucial when dealing with severely injured patients. Maintaining adequate oxygenation to the patient's lungs remains a high priority after natural or human-made disasters [1, 2]. Compared to common non-invasive airway management, surgical cricothyrotomy requires specialized medical equipment or advanced training [3, 4]. Using a cricothyrotomy kit to establish the airway through the skin and cricothyroid membrane is a much easier and quicker method compared to other methods [5].

The motivating example this work serves for is victim extraction using a Semi-Autonomous Victim Extraction Robot (SAVER), as shown in Fig.1. A semi-autonomous approach

to first-Aid airway management can be performed by high-level remote command from an operator. Robotic manipulators equipped with cricothyrotomy kit will perform autonomous cricothyrotomy on the victim with visual feedback. Compared to teleoperated operation, the proposed autonomous cricothyroid membrane (CTM) detection and manipulation will save the communication cost and achieve a fast response. This paper focuses on solving precise cricothyroid membrane keypoint detection problem using RGB images to help medical providers locate the invasion point for the cricothyrotomy kit.

Keypoints detection is an important research topic in computer vision with a wide range of applications in human-computer interaction [6], Augmented Reality (AR) and Virtual Reality (VR) [7], gaming, and filmmaking [8], and security and surveillance [9]. Various machine learning techniques with carefully engineered feature extractors have been developed to make precise predictions of different keypoints of interest [10, 11]. In the last decade, as large datasets and powerful computation become available, deep learning, associated with deep neural networks, has been gaining popularity in solving highly nonlinear problems with delicate performances that match or exceed humans [12,13]. This generates more research in an effort to design and apply different neural networks in computer vision. Deep neural networks that consist of multiple layers are developed to make more precise predictions but increase both the training and running time [14–16]. Compressed images are always used to improve the overall processing speed [14–16], while high-resolution input images provide more features and thus improve the final prediction precision [17, 18]. However, the motivating cricothyrotomy application requires not only accurate prediction generated from the learning-based approach, but also computation efficiency for real-time processing.

This paper presents a Hybrid Neural Network (HNNet), which consists of two parallel computing ensembles. The first ensemble is deployed to generate the Region of Interest (ROI) using compressed images in low-resolution. The second ensemble takes in the ROI selected from the first ensemble in high-resolution and the intermediate global features extracted in the first ensemble to generate precise predictions in the original high-resolution images. For validating the performance of the proposed neural network, a dataset consisting of 16,415 images was built with the locations of the cricothyroid membrane labeled by medical experts. Comparison with other state-of-the-art methods in prediction performance is also presented.

2 RELATED WORKS

Very deep neural networks are developed for applications with strict requirements in accuracy [18–22]. Low-level features are extracted by the first couple layers while later couple layers are in charge of generating high-level features from the low-level ones. Various neural network architectures have been developed

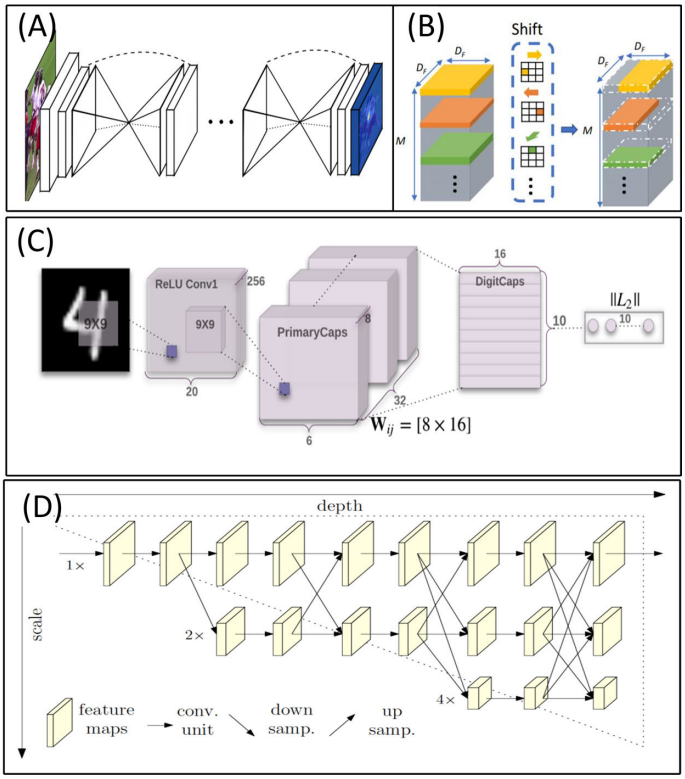

Figure 2. Gallery of some related works. (A) Hourglass network, (B) Shift layers network, (C) Capsule network, (D) High resolution network

to improve the estimation, as shown in Fig. 2. Features can be up/down-sampled to different resolutions and then combined to generate high-level features and thus improve the final prediction [19, 23]. Convolutional layers and pooling layers were developed to analyze image data efficiently. However, the spatial information in the images will get lost and lead to the "Picasso" problem [24, 25]. For maintaining the spatial information of the features through the neural network processing, the capsule neural network and its variants were proposed [22, 25, 26]. Recent studies also found that sparse shift layers can also preserve the spatial information efficiently [20, 27]. Another way to improve the prediction accuracy is to maintain the high resolution representation of the input image through the whole process [18]. However, keeping the high-resolution features through the neural network processing poses high requirements for the computing device and lowers the overall prediction speed. Using a region proposal as the preprocessor and implementing a keypoint detection neural network on the high-resolution proposals is another alternative, but global features outside of the proposal get lost [21]. Each of the above-mentioned approaches has its own specific strengths and weaknesses. As such, the specific application requirements need to be taken into consideration.

3 PROPOSED APPROACH

The proposed cricothyroid membrane keypoint detection neural network serves as the fundamental process for the cricothyrotomy application. In this application, the proposed neural network needs to balance both the running time and the prediction accuracy of the processing of high-resolution images. To solve the dilemma of speed and accuracy, a hybrid architecture, including one low-resolution region proposal ensemble and one high-resolution keypoint detection ensemble, is proposed, as shown in Fig. 3. The tensor size is presented in the form of $height \times width \times channel$ in the rest of the paper, expressing the height, width, and channel numbers of the tensor. The default channel size is 1 and it will not be shown. These numbers are determined by hyperparameters tuning during the training process.

Region Proposal Ensemble

The regional proposal ensemble takes in the low-resolution image with the size of 48×72, I_l, compressed from the original image of size $1024 \times 1536 \times 3$, I, to propose the ROI, ROI, out of a predefined 12×12 grid. I_l first pass through an hourglass model to efficiently generate the global features of size $48 \times 72 \times 128$, F. F is further reduced by following convolutional layers to the size of $12 \times 12 \times 8$. A sequence of sparse shift layer operations is performed to roll the resulting features in either the horizontal or vertical direction, as shown in Fig. 4. The shifted features are then concatenated to a combined feature map, F_s. F_s, embedded with spatial information, which is then passed through convolutional layers to generate a 12×12 heatmap, H_{ROI}, presenting the predicted region-of-interest.

Centered by the predicted ROI, a larger-region high-resolution image, I_R, is cropped from the original images. To preserve the global cues for the final prediction, the low-resolution features surrounding the ROI with a larger span, F_R, is also fed into the second ensemble along with I_R.

Keypoint Detection

The resulting cropped image from the region proposal ensemble is of size $256 \times 384 \times 3$ while the cropped features is of size $128 \times 192 \times 128$. The Keypoint Detection Ensemble takes in both the I_R and F_R, but at different depth levels of the neural network to make a final prediction. I_R passes one bottleneck inside the first hourglass module [19] and then concatenated with F_R. These combined features then pass through the rest of the two stacked hourglass module to generate the heatmap, H_R, of size 64×96. The peak value point in H_R represents the predicted location of the cricothyroid membrane. The regional heatmap is then padded to the global coordinate and generates the global heatmap, H, of size 256×384.

The training process is divided into two separate procedures, the training procedure of the region proposal ensemble and the training procedure of the keypoint detection ensemble. In the training procedure of the region proposal ensemble, the model takes in I_l as the input and generates H_{ROI} as the output. The best weights of the region proposal ensemble, $\tilde{\theta}_{rp}$, are stored and used for the training procedure of the keypoint detection ensemble. In the training procedure of the keypoint detection ensemble, two ensembles are connected as a single end-to-end network. This neural network takes in I as the input and produces H as the output. In this procedure, only the parameters inside the keypoint detection ensemble, θ_{kd}, are trainable while the parameters of the region proposal ensemble are fixed as $\tilde{\theta}_{rp}$. Both the parameters of the region proposal ensemble and the keypoint detection ensemble are updated by minimizing the objective functions during the training process, which are introduced in the following section.

In the final deployment of the proposed neural network, the region proposal takes in the compressed images as the input and generates H_{ROI} and F as output. I_R and F_R, cropped from I and F separately according to the H_{ROI}, are fed into the keypoint prediction model as the input to predict H_R. H_R is then padded back into the global heatmap, H, according to the spatial information provided by H_{ROI}.

Table 1. The Diversity Statistics of the Subjects in the Dataset

Race		Age		Gender	
Mongoloid	6	18-21	2	Male	10
Caucasian	2	22-25	7		
Negroid	5	26-29	4	Female	3

4 EXPERIMENTS

Cricothyroid Membrane Dataset

To train and validate the proposed cricothyroid membrane keypoint detection neural network, a dateset containing 16,415 images is created with the pixel location information of cricothyroid membrane on each image. The dataset contains images from 13 subjects among different genders, races, and ages to guarantee the diversity of the data, as shown in Tab. 1

A Kinect V2 [28] is used to collect the RGB image data from the subjects, as shown in Fig. 5. (A). Each subject is asked to rotate their neck about three different directions: 1) rotate the neck from side to side, 2) extend the neck to lift the chin upward, 3) bend the neck laterally to bring the ear to the shoulder, as shown in Fig. 5. (B-G).

For each of the above three motions, images are captured from different points of view. The points of view are decided by the combinations of the different relative height of the camera to the subject, h, the relative horizontal distance, w, and the angle between the neutral axis of the camera and the one of the subject, ψ.A summary of these combinations is provided in Tab. 2. A scene from the image collection process is shown in Fig. 5.

In the process of the image collection, the camera captures

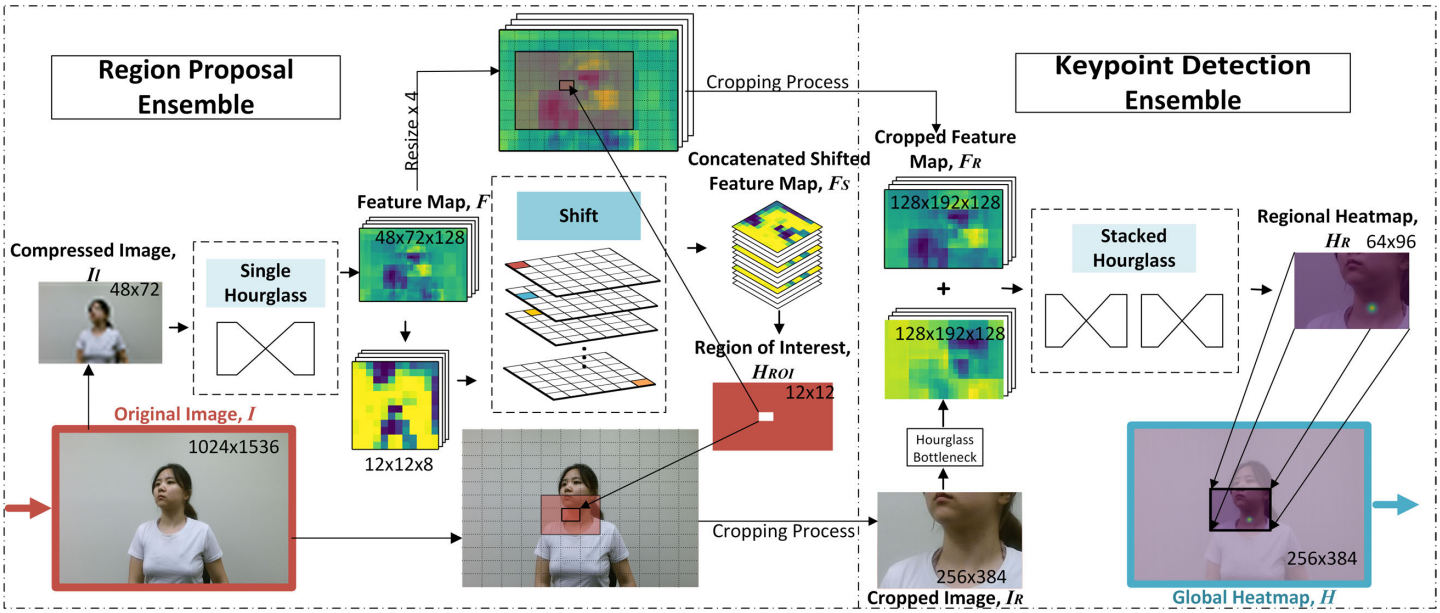

Figure 3. Proposed Hybrid Neural Network (HNNet)

Table 2. Summary of shooting angles, distance and height

Angle, ψ	$[-90°, -45°, 0°, 45°, 90°]$
Distance, w	$[0.75m, 1.25m]$
Height, h	$[-0.5m, 0m, 0.5m]$

50 images for each combination. A total 58,500 images are captured in the image collection and 16,145 images that provide unobstructed views of the human face and neck area are selected to build the dataset.

To make the annotation of the dataset efficiently and accurately, a user-friendly MATLAB based Graphical User Interface (GUI) labeling program is built, as shown in Fig. 6 (A C). In this program, the position of the keypoint in pixel is labeled (if visible), and three different levels of visibility can be selected in this labeling process: '0' (invisible in the image), '1' (visible, with no distinct feature), and '2' (visible, with distinct feature). The detailed mouse and keyboard operations required during the labeling process are explained in Fig. 6 (D).

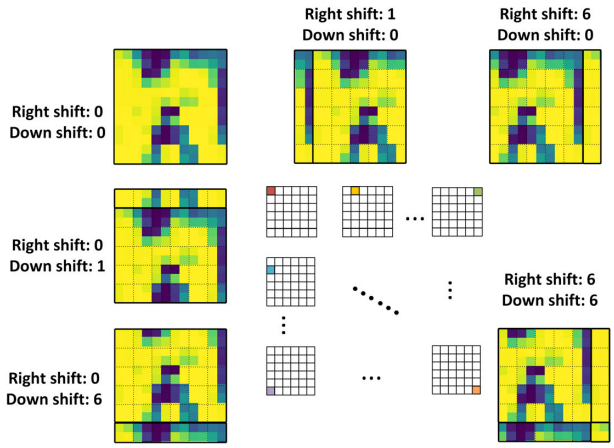

Figure 4. Shift layer operation

Training Process and Results

Among all of the images in the dataset, 3283 of them (20%) are randomly selected for the validation dataset, and 13,123 are selected for the training dataset. The original RGB images collected from Kinect v2 have a resolution of $1080 \times 1920 \times 3$. To fit the size of the neural networks, they are cropped around the

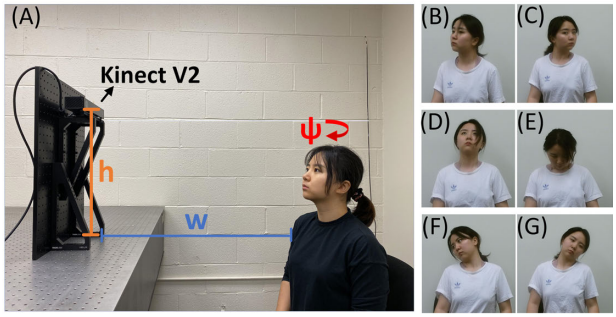

Figure 5. (A) A scene of image collection process, (B-G) Demo of Degree-of-freedom of neck movements on (B-C) Yaw axis, (D-E) Pitch axis and (F-G) Roll axis

center to the size of $1024 \times 1536 \times 3$. To augment the dataset, processes of rotation($-30 \sim 30$ degrees), scaling($0.8 \sim 1.2$), transportation($0 \sim 1/2$ of the distance from the keypoint to each edge) are performed to the original images, followed by normalization to $0 \sim 1.0$ on each RGB channel.

For the region proposal models, the groundtruth is a one-hot map with a single high bit as labeled *ROI*. Let $x_i \in \mathbb{R}^2$ be the location of the *ROI*. The value, S_i, of each pixel, $x \in \mathbb{R}^2$, of the one-hot map is expressed as follows.

$$S_i = \begin{cases} 1 & \text{if } x = x_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

For the keypoint detection models, the groudtruth are heatmaps with 2-D Gaussian centered on the keypoint location. Let $x_j \in \mathbb{R}^2$ be the position of keypoint. The value, S_j, of each pixel, $x \in \mathbb{R}^2$, of the of heatmap is expressed as follows:

$$S_j = exp\left(-\frac{|| x - x_j ||_2^2}{\sigma^2}\right) \tag{2}$$

Where σ is a constant that controls the spread of the high bits.

The neural networks are trained using Keras [29] with Tensorflow [30] as the backend on a NVIDIA Xp GPU. All models are trained for 20 epochs.

To evaluate the proposed region proposal model, a single hourglass model [19] and a stacked hourglass model [19] are also trained on the same dataset with the same input and output size, 48×72 and 12×12, respectively. To fit the GPU memory and optimize the performance of the training process, the batch size

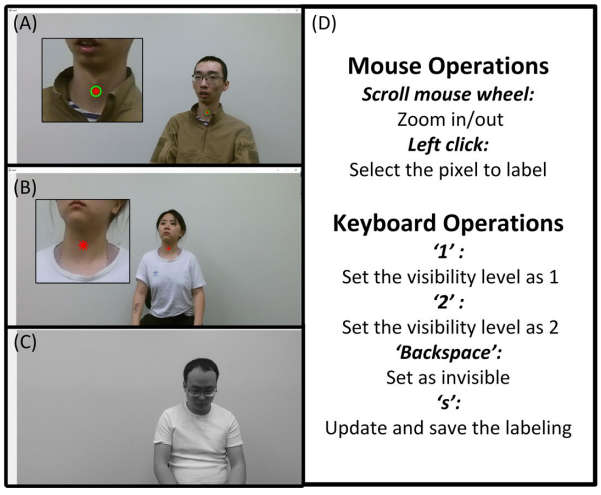

Figure 6. (A-C) Screenshots of MATLAB GUI for labeling process for cases when the visibility level of the keypoint labeled as (A) '2', (B) '1', (C) '0', and (D) The instructions to label the image with GUI

Table 3. Summaries of Training Processes and Validation Result of Region Proposal Model

Models	PRP [a]	Single Hourglass	Stacked Hourglass
Training process (%)			
Batch size	32	32	24
Training parameter	3,505,827	3,426,163	6,562,470
Prediction Accuracy (%)[b]			
$P^{PD=0}$ [c]	73.1	69.5	70.0
$P^{PD=1}$	99.7	99.4	99.6
Running Time (ms)[d]	23.4	12.9	21.0

[a] PRP stands for proposed region proposal model
[b] Based on results of prediction of 1946 images from validation dataset with CTM labeled as visible.
[c] $P^{PD=n}$ stands for the percentage of the images that the euclidean distance between predicted position of CTM and the ground truth is less than n pixels.
[d] The average time taken for single prediction on one image. (The models run on 1000 images in total).

is set as 32 for the proposed region proposal model and the single hourglass model, and 24 for the single stacked-hourglass-model. An optimizer, RMSprop with learning rate of 5e−4, is chosen for all of the models. A margin loss function is deployed for calculating the training loss for the region proposal model, as follows:

$$MarginLoss = err_{true} + err_{other} \tag{3}$$

where

$$err_{true} = (\frac{1}{n}) \sum_{i=1}^{n} y_{true} \cdot (1 - y_{predict}{}^2)) \tag{4}$$

$$err_{others} = (\frac{1}{n}) \sum_{i=1}^{n} (1 - y_{true} \cdot (y_{predict}{}^2)) \tag{5}$$

The number of trainable parameters of the proposed region proposal model is 3,505,827. Summary of the training details of the models are provided in Tab. 3.

The Ladder Capsule Network [22] was also tested as another region proposal model. The network was originally created for classification of the image in MNIST dataset, with the input size of 28×28. The dynamic routing algorithm implemented inside the network requires a larger space compared to other models. As a result, it could not be implemented within a limited computation unit for applications with large-resolution inputs and rich intermediate features.

To evaluate the proposed keypoint detection model, High-Resolution Network(HRNet) [18] and stacked hourglass network [19] are trained with the input size and the output size 256×358 and 64×96, respectively. The batch size of all these models is 8. Mean square error (MSE) was selected as the loss function, as follows:

$$MSE = (\frac{1}{n}) \sum_{i=1}^{n} (y_{true} - y_{predict})^2 \tag{6}$$

We chose RMSprop with the learning rate of 5e−4 to optimize the models. The number of the trainable parameter of

Table 4. Summaries of Training Processes and Validation Results of Keypoint Detection Models

Models	HNNet	Hourglass	HRNet	HNNet*[a]
Training process (%)				
Training parameter	6,626,162	6,562,370	1,453,376	—
Prediction Accuracy (%)[b]				
$P^{PD=5}$ [c]	29.1	7.9	1.1	30.0
$P^{PD=10}$	61.7	26.3	5.7	64.0
$P^{PD=20}$	78.3	64.9	19.7	80.9
$P^{PD=30}$	84.3	82.5	37.7	87.1
Running Time (ms)[d]	23.4+28.8	26.4	20.0	—

[a] The performance of HNNet when the correct Region of Interest is fed into the keypoint detection model.
[b] Based on results of prediction of 1946 images from validation dataset with CTM labeled as visible.
[c] PD (pixel deviations) stands for euclidean distances in pixels, and $P^{PD=n}$ stands for the percentage of the images with PD between predicted position of CTM and the ground truth is less than n pixels.
[d] The average time taken for single prediction on one image. (The models run on 1000 images in total).

HNNet is 6,626,162. Training details of the models are provided in Tab. 4.

The precision of the region proposal models is evaluated using two different measures: $P^{PD=0}$, which represents the percentage of the prediction that the prediction locations of ROI are exactly the same as the groundtruth; and $P^{PD=1}$, which represents the percentage of the prediction that the prediction locations has a deviation of less than or equal to 1 pixel in both x and y directions compared with the groundtruth. As long as the deviation of less than or equal to 1 pixel, the correct ROI would be included in the cropped image. We consider the situation as acceptable.

The evaluation results of the models are presented in Tab. 3. All the prediction accuracies are based on the prediction of 1946 images from the validation dataset with CTM labeled as visible. It shows that $P^{PD=0}$ and $P^{PD=1}$ of the proposed region proposal network model are 73.1% and 99.7%, which are the highest among the three models. With 1000 frames predicted in total, the average time to predict on each frame of the proposed region proposal network is 23.8 ms.

To evaluate the performance of the keypoint detection models, Euclidean distance between the predicted position of the cricothyroid membrane and the labeled position is used to calculated the pixel deviation:

$$d = \sqrt{(x - k \cdot (x_p - \frac{1}{2}))^2 + (y - k \cdot (y_p - \frac{1}{2}))^2} \tag{7}$$

Where x, y is the location of the groundtruth keypoint on the original high-resolution image, and x_p, y_p is the predicted location transformed back to the original high-resolution image. k is the scale factor from the neural network output to the original image. The evaluation result is shown in Tab. 4. PD_n stands for the percentage of the frames that the Euclidean distance between the predicted position of CTM and the ground truth is less than n pixels. PD_5 to PD_{30} provides measures of precision among different stringencies. The maximum Euclidean distance is set to be 30 pixels since a deviation of the estimated position of cricothyroid membrane in real-world coordinates should be no more than 0.5 cm.

The results show that HNNet achieves a PD_{30} of 84.3%. The performance of HNNet is highly dependent on the prediction of the region proposal model, although a PD_1 as high as 99.7% guarantees that almost every image can be cropped with the region around cricothyroid membrane included. We also tested the performance of the keypoint detection procedure by feeding the correct Region of Interest into the model. The last column in Tab. 4 shows that PD_{30} can reach 87.1% using a correct cropped region.

The proposed neural network is deployed in GPU and the running time is invariant to the number of cricothyroid membrane shown on the image, with runtime complexity of $O(1)$. It takes 52.2 ms for HNNet to predict on a single frame, with 1000 frames tested in total. The prediction time consists of two portions: 23.4 ms for the proposed region proposal ensemble, and 28.8 ms for the proposed keypoint prediction. It would allow the two ensembles to run on two computers synchronously and thus shorten the running time.

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed a Hybrid Neural Network model(HNNet) that achieved high accuracy and high efficiency on the task of real-time CTM keypoint detection. Existing models that depend on either very deep neural networks or maintaining high-resolution representations of images haven't achieved a good trade-off between prediction precision and running speed. Our proposed network further explored this dilemma and improved precision by taking in the high-resolution image without requiring large computations. 84.3% of the predictions had deviations of less than 30 pixels in the validation dataset, and 61.7% of them had deviations of less than 10 pixels. The model can also be adapted to a wide range of applications that require high-precision keypoint detection.

HNNet serves as the perception process of the proposed autonomous first-aid airway management system. Robotic manipulation that is capable of basic behaviors and using a cricothyrotomy kit tool will be learned using reinforcement learning. The integrated system, including both perception and control, will be set up. Experimental validation with a full-size medical manikin will be performed to validate the research on autonomous first-aid airway management. The setup of the system is shown in Fig. 7.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of medical expert, Zhiping Liu, in Respiratory Department of XuZhou

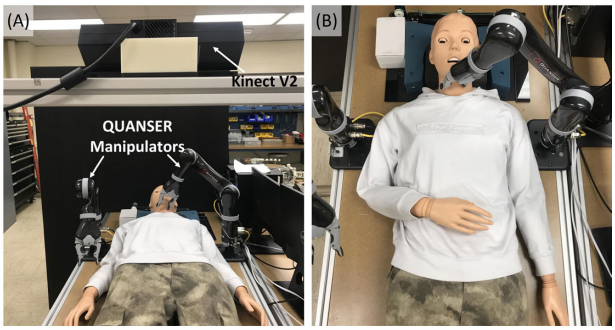

Figure 7. Setup of the Proposed Autonomous First-aid Airway Management System in A) Lateral view B) Top view

Central Hospital for labeling the position of the cricothyroid membrane in our dataset. The authors would also like to gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] Khan, R. M., Sharma, P. K., and Kaul, N., 2011. "Airway management in trauma". *Indian Journal of Anaesthesia,* ***55***(5), 9, pp. 463–469.

[2] Kennedy, C. C., Cannon, E. K., Warner, D. O., and Cook, D. A., 2014. "Advanced airway management simulation training in medical education: A systematic review and meta-analysis". *Critical Care Medicine,* ***42***(1), 1, pp. 169–178.

[3] Lockey, D. J., Healey, B., Crewdson, K., Chalk, G., Weaver, A. E., and Davies, G. E., 2015. "Advanced airway management is necessary in prehospital trauma patients". *British Journal of Anaesthesia,* ***114***(4), 4, pp. 657–662.

[4] Crewdson, K., Rehn, M., and Lockey, D., 2018. "Airway management in pre-hospital critical care: A review of the evidence for a 'top five' research priority". *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine,* ***26***(1), 12, p. 89.

[5] Katos, M. G., and Goldenberg, D., 2007. "Emergency cricothyrotomy". *Operative Techniques in Otolaryngology - Head and Neck Surgery,* ***18***(2), 6, pp. 110–114.

[6] Ren, H., Kumar, A., Wang, X., and Ben-Tzvi, P., 2018. "Parallel deep learning ensembles for human pose estimation". In ASME 2018 Dynamic Systems and Control Conference, DSCC 2018, Vol. 1, American Society of Mechanical Engineers, p. V001T07A005.

[7] Fangbemi, A. S., Liu, B., Yu, N. H., and Zhang, Y., 2018. "Efficient human action recognition interface for augmented and virtual reality applications based on binary descriptor". In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10850 LNCS. Springer Verlag, pp. 252–260.

[8] Huang, C., Gao, F., Pan, J., Yang, Z., Qiu, W., Chen, P., Yang, X., Shen, S., and Cheng, K. T. T., 2018. "ACT: An Autonomous Drone Cinematography System for Action Scenes". In Proceedings - IEEE International Conference on Robotics and Automation, IEEE, pp. 7039–7046.

[9] Kaur, J., and Bathla, A. K., 2017. "Video stabilization for an aerial surveillance system using sift and surf". In Proceedings on 2016 2nd International Conference on Next Generation Computing Technologies, NGCT 2016, IEEE, pp. 742–747.

[10] Rosenthaler, L., Heitger, F., Kübler, O., and von der Heydt, R., 1992. "Detection of general edges and keypoints". In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 588 LNCS. Springer Verlag, pp. 78–86.

[11] Sirmacek, B., and Unsalan, C., 2009. "Urban-area and building detection using SIFT keypoints and graph theory". *IEEE Transactions on Geoscience and Remote Sensing,* ***47***(4), 4, pp. 1156–1167.

[12] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., and Hassabis, D., 2017. "Mastering the game of Go without human knowledge". *Nature,* ***550***(7676), 10, pp. 354–359.

[13] OpenAI, 2019. OpenAI Five.

[14] Krizhevsky, A., Sutskever, I., and Hinton, G. E., 2017. "ImageNet classification with deep convolutional neural networks". *Communications of the ACM,* ***60***(6), 5, pp. 84–90.

[15] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., and Schiele, B., 2016. "DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model". *Arxiv,* ***9910 LNCS***(4), 5, pp. 34–50.

[16] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y., 2017. "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields". In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2017-Janua, IEEE, pp. 1302–1310.

[17] Scott, G. J., England, M. R., Starms, W. A., Marcum, R. A., and Davis, C. H., 2017. "Training Deep Convolutional Neural Networks for Land–Cover Classification of High-Resolution Imagery". *IEEE Geoscience and Remote Sensing Letters,* ***14***(4), 4, pp. 549–553.

[18] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., and Xiao, B., 2019. "Deep High-Resolution Representation Learning for Visual Recognition". *Computer Vision and Pattern Recognition*, 8.

[19] Newell, A., Yang, K., and Deng, J., 2016. "Stacked Hourglass Networks for Human Pose Estimation". In *European*

Conference on Computer Vision, Vol. 9912 LNCS. Springer Verlag, 3, pp. 483–499.

[20] Wu, B., Wan, A., Yue, X., Jin, P., Zhao, S., Golmant, N., Gholaminejad, A., Gonzalez, J., and Keutzer, K., 2018. "Shift: A Zero FLOP, Zero Parameter Alternative to Spatial Convolutions". In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, pp. 9127–9135.

[21] Ren, S., He, K., Girshick, R., and Sun, J., 2017. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.". *IEEE transactions on pattern analysis and machine intelligence,* ***39***(6), 6, pp. 1137–1149.

[22] Jeong, T., Lee, Y., and Kim, H., 2019. "Ladder Capsule Network". In Proceedings of the 36th International Conference on Machine Learning, pp. 3071–3079.

[23] Ronneberger, O., Fischer, P., and Brox, T., 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 9351. Springer Verlag, 5, pp. 234–241.

[24] Bengio, Y., Bengio, Y., Lecun, Y., and Lecun, Y., 1995. *Convolutional Networks for Images, Speech, and Time-Series*. The MIT press.

[25] Sabour, S., Frosst, N., and Hinton, G. E., 2017. "Dynamic Routing Between Capsules". *Advances in Neural Information Processing Systems,* ***2017-Decem***, 10, pp. 3857–3867.

[26] Paoletti, M. E., Haut, J. M., Fernandez-Beltran, R., Plaza, J., Plaza, A., Li, J., and Pla, F., 2019. "Capsule Networks for Hyperspectral Image Classification". *IEEE Transactions on Geoscience and Remote Sensing,* ***57***(4), 4, pp. 2145–2160.

[27] Chen, W., Xie, D., Zhang, Y., and Pu, S., 2019. "All You Need Is a Few Shifts: Designing Efficient Convolutional Neural Networks for Image Classification". In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 7234–7243.

[28] Zhang, Z., 2012. "Microsoft Kinect Sensor and Its Effect". *IEEE Multimedia,* ***19***(2), 2, pp. 4–10.

[29] Chollet, F., et al., 2015. Keras.

[30] Abadi;, M., Agarwal;, A., Barham;, P., Brevdo;, E., Chen;, Z., Citro;, C., Corrado;, G. S., Davis;, A., Dean;, J., Devin;, M., Ghemawat;, S., Goodfellow;, I., Harp;, A., Irving;, G., Isard;, M., Jozefowicz;, R., Jia;, Y., Kaiser;, L., Kudlur;, M., Levenberg;, J., Mané;, D., Schuster;, M., Monga;, R., Moore;, S., Murray;, D., Olah;, C., Shlens;, J., Steiner;, B., Sutskever;, I., Talwar;, K., Tucker;, P., Vanhoucke;, V., Vasudevan;, V., Viégas;, F., Vinyals;, O., Warden;, P., Wattenberg;, M., Wicke;, M., Yu;, Y., and Zheng;, X., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.