Advanced Video Laryngoscope and Automatic Data Collection System

Gabrielle Hoyer, Sean Runnels, Samer Merchant, and Kai Kuck.
Departments of Anesthesiology and Biomedical Engineering, University of Utah

***Abstract*— The Advanced Video Laryngoscope is designed to address the high stress situation of an inexperienced healthcare provider performing an intubation on a patient. The technology is superior to current video laryngoscopes in that it utilizes machine learning techniques to guide the healthcare provider in real-time, providing augmented reality cues to anatomical features, feedback to prevent critical levels of deoxygenation of the patient, and an automated system to assess the difficulty of airway and call on the assistance of other physicians if first-pass intubation is not successful. By providing real-time assistance to the operator, this device will increase the success rate of first-pass intubation and decrease the risk of complications for the patient.**

***Index Terms* — intubation, laryngoscopy, object detection, YOLO algorithm**

I. Introduction

Intubation is frustratingly dangerous and difficult to get right; 400,000 procedures require three or more attempts, and 220,000 of these difficult intubation patients die. Complication rates increase dramatically with multiple intubation attempts; it is paramount that first-pass intubations succeed [1]. Laryngoscopy is a technique used to allow a health care provider a view of the throat, specifically the region of the vocal folds. The procedure is often performed to assist in intubation, the delivery of a tube directly to the airway through the vocal folds; this is used to oxygenate the patient. Laryngoscope tools were developed over 75 years ago and were later augmented with image recording hardware and a screen to allow the caregiver a visualization of the airway.

The additions of optics created an leap in laryngoscopy technology. The developments ameliorated the critical weakness of traditional laryngoscopes: the lack of visualization of the vocal cord and esophageal region during intubation [1, 2]. However, when comparing first-pass success rates, the ability of the health care provider to place a tube in the airway on the first try, of video and direct laryngoscopes, the outcomes vary. Some studies indicate an improved first-pass success rate [3] while others show little to no benefit [4, 5]. Furthermore, some studies have even indicated an increased risk of complications with video laryngoscopes [4, 6]. Effectively, studies have demonstrated that the core goal of laryngoscopy or “first pass success rate” was not significantly impacted by these developments. It should be noted that the studies reporting positive results with use of the video laryngoscope allowed users to choose their method, direct or video laryngoscopy; however, studies that assigned the intubation method randomly to users, reported decreased success rates.

Interestingly, studies have shown that video laryngoscopy led to a greater increase in first-pass success rate for inexperienced healthcare providers [3], such as EMTs and first and second year ER Residents. Indeed, these individuals must often perform intubations in high stress situations with limited guidance and must act as first-responders. Mistakes that often occur with inexperienced healthcare providers include insertion of the endotracheal tube into the esophagus or inserting the tube through the vocal folds at the incorrect depth leading to low levels of oxygenation for the patient. Additionally, a stressful situation may cause the healthcare provider to lose track of time when performing a difficult intubation; it is imperative that the intubation process is done in a timely manner or the patient may suffer brain injury or death. Indeed, it is difficult to successfully perform an intubation on a first pass, thus patients are more likely to experience complications if intubation is not done correctly.

The risk of complications increases dramatically with every failed intubation [7], thus it is essential to correctly place the intubation tube into the trachea on the first attempt. Less significant complications, such as tracheal injuries, can cost a hospital $2,000, and a patient approximately $11,000 if readmission to the hospital is necessary [8]. Additional complications include brain damage or death, which may cost hospitals millions of dollars in compensation [9]. This is in addition to the patient's suffering. Therefore, it is in the best interest of patients, hospitals, health care professionals, and insurance providers that intubations succeed.

A. A Smarter Laryngoscope

The Advanced Video Laryngoscope is the next evolution of video laryngoscopes. The Advanced Video Laryngoscope is designed to improve first pass success rates by not only allowing the caregiver to visualize the airway, but to receive real-time guidance and feedback in a stressful procedure with anatomical variation. With the use of artificial intelligence, our laryngoscope records the patient’s unique anatomy, and overlays visual cues on the screen, to guide the caregiver in a time-efficient manner. If needed, this device will be able to call on assistance of another physician.

This device is unique in its use of artificial intelligence and deep learning neural networks. The intubation region may vary immensely from patient to patient as a result of obesity, tumors, trauma, and mucus or saliva buildup. Each of these situations may lead to a difficult intubation that the caregiver has not yet experienced; the Advanced Video Laryngoscope’s use of machine learning can account for all of these degrees of

variation and guide the user through a successful first-pass intubation in real-time. Therefore, the patient will have reduced risks of complications as a result of intubation.

II. Methods

A. Data Collection Device

In order to build an airway management dataset, an automatic data collection device was designed and implemented in the University of Utah hospital. The device was designed to be compatible with a variety of video laryngoscopy systems to collect high-fidelity video data without needing to interact with health care providers [15]. In this way, data is collected more consistently, and the workload of providers is not affected.

Additionally, it was necessary to design the device in such a way as to run a trained neural network for real-time use. The device includes a micro-processor which can store only limited amounts of data at any one time. Overloading the system could potentially affect the performance of any neural network programmed into the device. Hence, a system was developed to automatically upload the procedural data from the device to our server each day. The data from that day would then be cleared from the device and prepared to collect new procedural data for the next day.

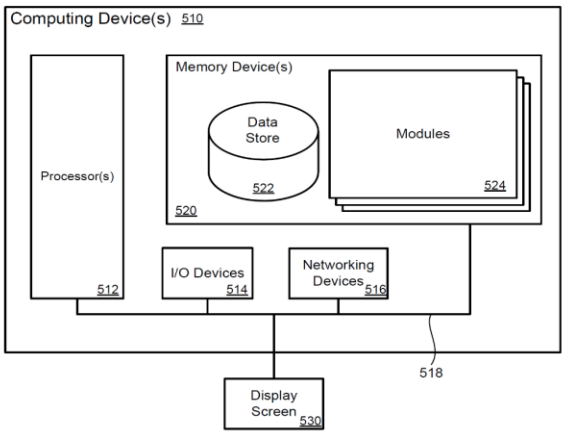

Fig. 1. System to collect, store, and send procedural data to data management system for future processing. System device connects to video laryngoscope tower.

B. Data Processing and Annotation

Once the data was collected by the device, the data was stored in a secure, HIPAA compliant workstation, and ready for processing. Videos were split into frames, and redundant frames were removed. Furthermore, frames with patient identifiers were removed.

An annotation team consisting of senior medical students, anesthesiology residents, and anesthesiologists was formed to annotate images from intubation videos. Each annotator would be assigned a set of images to identify and label features. These features include airway anatomy such as the epiglottis, arytenoids, vocal folds, as well as airway management tools such as an endotracheal tube and introducer. Additionally, features to be labeled included indicators of trauma such as blood and bruising.

Each image was classified by two annotators, one to identify anatomical features and place bounding boxes around them, and another to tighten or correct the location of the label bounding box. In this way a database was created, and two datasets were developed for neural network training. There was a small initial dataset composed of 32 patient cases which contained ~280 images with 4 classes, and a large dataset composed of 114 patient cases which had ~1700 images with 11 classes. The large dataset was composed of 1459 instances of the epiglottis, 1756 instances of the vocal cords, 963 instances of an endotracheal tube, 1689 instances of arytenoids, 190 instances of an introducer, 108 instances of the trachea rings, 142 instances of blood, 100 instances of an NG tube, and 108 instances of the esophagus

Furthermore, a third dataset was formed by performing augmentation techniques on the large dataset. Specifically, the color, hue, and saturation of images within the dataset were randomly altered to introduce additional variability into the dataset, which could not be naturally collected from the intubation procedure. Such augmentation to the data could improve the performance of the object detection models, thereby leading to improved real-time guidance cues and assistance to healthcare providers.

The datasets were split into training and testing sets composed of 90% and 10% of the datasets, respectively. Testing the predictive performance of a trained network was done so on test set images, images not utilized in training. The verification metrics were determined from performance on the training set, during training time.

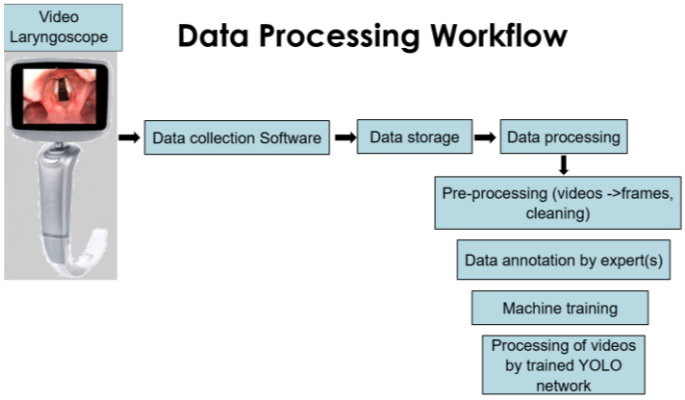

Fig. 2. Workflow of collecting, storing, and processing procedural, high-fidelity data from VLs for machine learning.

C. Anatomical Feature Object Detection

In order to build an anatomical feature recognition system for real-time use, it was necessary to choose a neural network which balanced accuracy and performance with processing time. The YOLOv3 (you only look once) algorithm has such capabilities [14]. The YOLOv3 algorithm reduces the multi-step process of detecting a feature and its location in context to other features in the image, which is common in other object detection algorithms. This consolidation of multiple pipelines

increases the algorithm's efficiency with processing real-time data.

In addition to the standard YOLOv3 model which utilizes the Darknet-53 architecture, we looked at the YOLOv3-tiny model [14] which reduces the number of convolutional layers present in the architecture. This reduction in layers leads to improved processing time, which is useful for real-time applications, but with potentially poorer accuracy and performance for anatomical feature recognition.

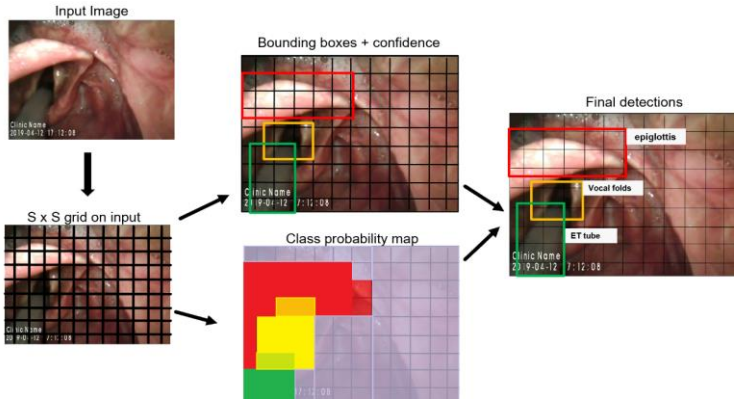

Fig. 3. YOLO algorithm determines location and classification of object in an image. YOLO trains both components within the same network, improving processing time.

D. Neural Network Verification

When determining the performance of our neural networks, several verification metrics were calculated for each network trained on the small dataset, the large dataset, and the augmented large dataset. These metrics include values for Intersection over Union (IoU) (1) which determines how well a trained network places a predictive bounding box over a feature, and then compares the placement to ground truth labels from the expert annotators. This value ranges from 0-1, with one being a perfect overlap of trained network prediction to expert annotation.

Additionally, precision (2), recall (3), and F1-score (4) values were calculated. Each of these values is an indicator for accuracy and performance by considering the number of true positives (TP), false positives (FP), and false negatives (FN) predictions the trained network makes during verification testing. Mean average precision (5) value determines the performance of the trained network for predicting all classes present in a dataset. The mean average precision value is defined as the summation of some threshold, k = 1 to N, of the precision at threshold k, P(k), times change in recall at threshold k, $\Delta r(k)$.

$$IoU = TP/(FP + TP + FN) \quad (1)$$

$$Precision = TP/(TP + FP) \quad (2)$$

$$Recall = TP/(TP + FN) \quad (3)$$

$$F1 = 2 * (precision * recall)/(precision + recall) \quad (4)$$

$$mAP = \sum_{k=1}^{N} P(k)\Delta r(k) \quad (5)$$

Furthermore, we determined total detection time and average inference time for all models when trained on the three datasets. Total detection time is the time it takes a trained network to process predictions for all images in a dataset. The average inference time is the time it takes a trained network to make all predictions for a single image. These time values are indicators for how well an algorithm would perform with real-time tasks.

III. Results

The three tables below describe the six combinations of YOLOv3 models and datasets. Table 1. describes the performance of each model combination for each class in the associated dataset, as well as the mean average precision value for each combination.

When training the standard Yolov3 network model on the large dataset of approximately 1700 images and eleven classes, the model performed well. As seen in Table 2., the mean average precision of the model lies above 85%; precision, recall, and F1-score values all lie at a value of 0.90 and above. The total detection time when performing verification testing of the training set was 44 seconds (Table 3). While the accuracy is immensely high, the detection time is a bit lacking, which is not ideal for real-time processing.

Tables 1 and 2 display the results of training the large dataset on the YOLOv3-tiny model. The mean average precision is slightly below the 85% value. However, the precision, recall, and F1-score of this model are at a value of 0.90 and above. At a value of 14 seconds, the total detection time of the large dataset on the tiny model is substantially lower than that of the YOLOv3 standard model. Indeed, the accuracy of the YOLOv3-tiny model is on par with the standard YOLOv3 model, but with a far improved processing time. This is indicative that the YOLOv3-tiny model would be useful for real-time use.

Feature	yolov3-tiny, large data	yolov3, large data	yolov3-tiny, augmented data	yolov3, augmented data	yolov3-tiny, small data	yolov3, small data
mAP	84.19%	86.29%	83.27%	90.38%	15.53%	69.89%
epiglottis	99.82%	99.46%	99.70%	99.92%	3.41%	91.40%
vocal cords	89.39%	95.82%	87.21%	99.38%	50.99%	99.25%
arytenoid	88.18%	91.06%	84.63%	98.94%	NA	NA
trachea rings	77.61%	89.49%	77.73%	98.14%	NA	NA
esophagus	87.42%	93.95%	80.38%	99.43%	NA	NA
introducer	99.87%	99.80%	98.74%	99.92%	NA	NA
ET tube	99.64%	98.96%	99.50%	99.87%	0.00%	0.00%
blood	84.43%	87.48%	89.62%	98.65%	NA	NA
NG tube	99.75%	93.16%	98.41%	99.94%	NA	NA
unknown	100.00%	100.00%	100.00%	100.00%	7.70%	88.92%

Table 1. Comparison of feature recognition and mean average precision.

Tables 1 and 2 display the results of training the Yolov3 and Yolov3-tiny models on the small dataset, containing less than 300 images and only 4 classes. While the standard Yolov3 model performs decently well, though mean average precision is down to 69.89%. The Yolov3-tiny model, however,

performed quite poorly. Mean average precision dropped to a value below 20%. A comparison of performance for the standard Yolov3 model on small and large datasets can be seen in Table 2.

model	precision	recall	F1-score	average IoU	mean average precision	total detection time
yolov3-tiny, small data	NA	0	NA	0%	15.53%	2 seconds
yolov3, small data	0.87	0.93	0.9	67.48%	69.89%	8 seconds
yolov3-tiny, large data	0.9	0.91	0.9	69.05%	84.19%	14 seconds
yolov3, large data	0.95	0.9	0.92	73.25%	86.29%	44 seconds
yolov3-tiny augmented data	0.88	0.89	0.89	67.66%	83.27%	8 seconds
yolov3, augmented data	0.98	0.98	0.98	79.60%	90.38%	44 seconds

Table. 2. Comparison of verification metric performance, precision and mean average precision.

It was expected that the models trained on the large datasets with augmented data would perform as well if not better than the models trained on the standard large dataset. Augmentation of data provides additional variability to the dataset which may have not been captured traditionally from a video laryngoscope. Indeed, the YOLOv3 standard model trained on the augmented large dataset had the greatest mean average precision value of all combinations and had the greatest value for all verification metrics.

model	mean average precision	total detection time	average inference time
yolov3-tiny, small data	15.53%	2 seconds	7.04 ms
yolov3, small data	69.89%	8 seconds	28.17 ms
yolov3-tiny, large data	84.19%	14 seconds	7.82 ms
yolov3, large data	86.29%	44 seconds	24.57 ms
yolov3-tiny augmented data	83.27%	8 seconds	28.17 ms
yolov3 augmented data	90.38%	44 seconds	24.57 ms

Table. 3. Comparison of processing time.

Interestingly, the YOLOv3-tiny model performed better when trained on the standard large dataset rather than the augmented large dataset. Additionally, this combination had the best inference time.

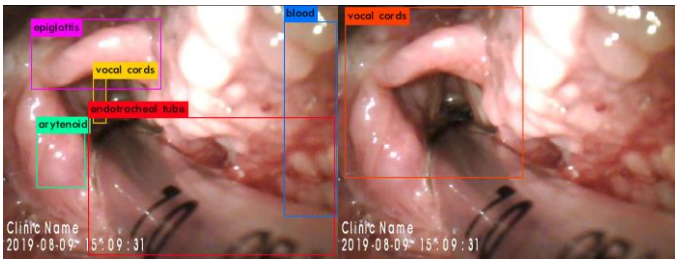

Fig. 4. Left) Image displaying anatomical features with overlaid predictions made by yolov3 network, trained on the large dataset. The network is capable of detecting small detailed features (vocal cords) with notable precision. Right) Image displaying anatomical features with overlaid predictions made by yolov3 network, trained on the small dataset.

IV. Conclusion

A. Performance Outcome

When comparing the various combinations of algorithms and datasets for training, the YOLOv3-tiny algorithm trained on the small dataset performed the worst. This is demonstrative of the necessity for balance between size of dataset and how deep the network is. If a shallower network trains on a limited dataset, the performance will likely be poor, compared to a deep network trained on a limited dataset, or a small and efficient network trained on an extensive dataset. It should be kept in mind, however, that deep networks trained on smaller datasets tend to overfit thereby decreasing how generalizable the model is.

The YOLOv3 algorithm trained on the augmented large dataset performed with the greatest mean average precision, though training and processing time with this algorithm are more extensive. The YOLOv3-tiny algorithm trained on the standard large dataset performed with a mean average precision >.80 and had the shortest inference time. The performance of the tiny algorithm in both accuracy and processing time are indicative of its potential for use in practical applications such as the advanced video laryngoscopy device. The YOLOv3-tiny algorithm will be used in the continued development of this device and system.

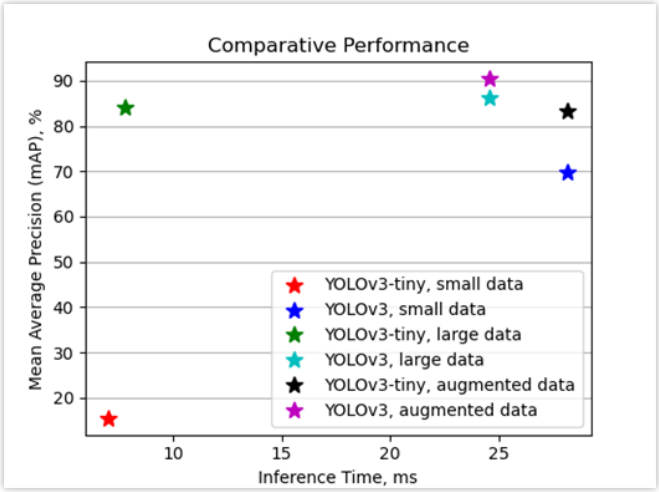

Fig. 5. Mean average precision and inference time, comparison of models.

B. Limitations

When training and testing the models, the datasets were split into two sets for training and testing. This limits the number of images the networks are trained on. Additionally, all verification metrics were calculated from training performance, though visualizations of predictive bounding box placement were made on test images, not used for training of the networks. In the future, cross-validation techniques will be used, which utilize the full extent of the dataset for training and verification, useful for limited data. In this way, we may optimize the training of our network with full use of our dataset and obtain a better representation of performance and accuracy for our network when detecting anatomical features.

References

[1] Paolini, Jean-Baptiste, François Donati, and Pierre Drolet 2013Review Article: VideoLaryngoscopy: Another Tool for Difficult Intubation or a New Paradigm in Airway Management? Canadian Journal of Anesthesia/Journal Canadien d'anesthésie 60(2): 184–191.

[2] Silverberg, Michael J., Nan Li, Samuel O. Acquah, and Pierre D. Kory (2015). Comparison of Video Laryngoscopy Versus Direct Laryngoscopy During Urgent Endotracheal Intubation: A Randomized Controlled Trial. Critical Care Medicine 43(3): 636–641.
[3] Michael F. Aziz, Dawn Dillman, Rongwei Fu, Ansgar M. Brambrink (2012). Comparative Effectiveness of the C-MAC Video Laryngoscope versus Direct Laryngoscopy in the Setting of the Predicted Difficult Airway. Anesthesiology 116(3):629-636. doi: 10.1097/ALN.0b013e318246ea34.
[4] Lascarrou, J. B., Boisrame-Helms, J., Bailly, A., Le Thuaut, A., Kamel, T., Mercier, E., ... & Meziani, F. (2017). Video laryngoscopy vs direct laryngoscopy on successful first-pass orotracheal intubation among ICU patients: a randomized clinical trial. Jama, 317(5), 483-493.
[5] Castillo-Monzón, C. G., Marroquín-Valz, H. A., Fernández-Villacañas-Marín, M., MorenoCascales, M., García-Rojo, B., & Candia-Arana, C. A. (2017). Comparison of the macintosh and airtraq laryngoscopes in morbidly obese patients: a randomized and prospective study. Journal of clinical anesthesia, 36, 136-141.
[6] Kory, Pierre, Keith Guevarra, Joseph P. Mathew, Abhijith Hegde, and Paul H. Mayo (2013). The Impact of Video Laryngoscopy Use during Urgent Endotracheal Intubation in the Critically Ill. Anesthesia and Analgesia 117(1): 144–149.
[7] Divatia, Jigeeshu V, Parvez U Khan, and Sheila N Myatra 2011Tracheal Intubation in the ICU: Life Saving or Life Threatening? Indian Journal of Anaesthesia 55(5): 470–475.
[8] Knox N, Chinwe O, Themba N, Joseph F, Hormoz A. Relationship between intubation rate and continuous positive airway pressure therapy in the prehospital setting. World Journal of Emergency Medicine. 2015;6(1):60-66.
[9] Lubin and Meyer PC (2013). $1.576 Million Verdict in Intubation Death. Retrieved from http://www.lubinandmeyer.com/cases/intubation-malpractice.html
[10] Qingyu Zhao, True Price, Stephen Pizer, Marc Niethammer, Ron Alterovitz, and Julian Rosenman, "The Endoscopogram: A 3D Model Reconstructed from Endoscopic Video Frames," in Medical Image Computing and Computer Assisted Intervention (MICCAI), Oct. 2016, pp. 439-447.
[11] Pfuntner, A. W. L., & Stocks, C. (2010). Most frequent procedures performed in US hospitals. URL http://hcup-us. ahrq. gov/reports/statbriefs/sb149. pdf (Accessed 24/9/15).
[12] Cision PR Newswire (2016, June 14). Global Anesthesia Video Laryngoscope Market 2016- 2020 - Robot-assisted Intubation is on the Rise - Research and Markets. Retrieved from https://www.prnewswire.com/news-releases/global-anesthesia-video-laryngoscope-market-2016- 2020---robot-assisted-intubation-is-on-the-rise---research-and-markets-300284466.html
[13] Technavio (2017, April 27). Global anesthesia video laryngoscope market worth $329.3 million by 2020. Retrieved from https://www.technavio.com/pressrelease/global-anesthesiavideo-laryngoscope-market-worth-3293-million-2020
[14] Redmon, Joseph, and Ali Farhadi. “Yolov3: An Incremental Improvement.” *ArXiv*, 2018.
[15] Runnels, Sean, et al. *TRACHEAL INTUBATION PROCEDURE MONITORING*.
[16] Tzutalin. LabelImg. Git code (2015). https://github.com/tzutalin/labelImg