2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871978

Creating Computer Vision Models for Respiratory Status Detection*

Quan T. Do[a], and Jamil Chaudri[b], [a]*Mayo Clinic, Rochester, MN, USA* [b]*Marshall University, Huntington, WV, USA.*

***Abstract*—This study aims to use computers to detect and recognize ventilation objects (masks and tubes) and their positions on the patient's face. We created two models: the You Only Look Once (YOLO) and the Transfer Learning (TL) models, to perform this computer vision task. The development processes and comparison of performance will be described in this paper. The TL model had a better performance (98%) compared to the YOLO model (93%).**

***Clinical Relevance*— Healthcare providers and researchers interested in the field of computer vision applied in medicine, specifically automatic object detection using video streams or real-time video streaming may benefit from findings reported.**

I. Introduction

Digital Video Recordings (DVRs) have played an important role in patient care, from preventive to intensive [1-3]. For example, analyzing recordings of gait helped identify unsafe habits and reduce incidence of falls [4]. Furthermore, DVRs have been implemented in healthcare to prevent errors, lower cost, and improve quality by optimizing the workload of healthcare providers [2-5]. In critical care, real-time video monitoring has been successfully deployed to aid diagnosis [6], prevent hospital-acquired infections [7], detect patient mobilization activities [8] and perform security, safety, education, and quality assessment [1,7,9].

In order to facilitate early detection of adverse events, deterioration in clinical status, and medical error-prevention round the clock monitoring is required. This task is time and labor intensive, therefore real-time monitoring using computer vision technology is a potential solution while also reducing the monitoring cost. Besides, the limited quantitative data on patient's respiratory status available through Electronic Medical Records (EMR) falls short of bedside assessment in evaluating patient-device interaction, detecting deteriorations or adverse events. Shifting of ventilation devices (both adjuncts and facial interfaces and supported ventilation machines) may indicate a change in the patient's clinical status. For example, changing from intubation (invasive mechanical ventilation) to extubating status (either room air, Nasal Cannula (NC), High-Flow Nasal Cannula (HFNC) or non-invasive ventilation (NIV) denotes an improvement in respiratory support. Failure to recognize patient deterioration and timely intervention to prevent poor outcomes continues to be the focus of research.

II. Background

Image classification is the process of assigning a specific class (also called label or class-label) to an image (or a video frame). Object localization is the process of drawing a bounding box around one or multiple objects in an image. The combination of these two tasks is called object detection, where a bounding box is drawn around each object of interest and a specific class label is also assigned to these objects. Each respiratory device interface has a different shape, and each must be positioned in its specific fashion in order to properly function. Thus, the object of the study is to detect the shape and appropriate positioning of the objects (masks and tubes) on the patient's face. Subsequent to the detection, the bounding box around the face is assigned a label, which classifies the type of tube or mask. In short, object detection aims to recognize and locate objects in an image frame using computer vision techniques. The aim of this study is creating computer vision models then deploy them with streaming data from a camera to detect and recognize ventilation objects such as NC, face mask, HFNC, and endotracheal tube and their positions (of respiratory devices) on the patient's face.

Our literature review persuaded us to conduct a comparative study using the models: (1) You Only Look Once (YOLO) and (2) Transfer Learning (TL) model. YOLO algorithm has been known to be much faster compared to other object detection algorithms but also maintains a high accuracy rate. YOLO is most appropriate to be implemented for real-time streaming object detection problems (e.g. the detection of adverse events) [12]. Besides, TL is becoming more and more popular in the field of computer vision for its ability to build highly accurate models in a timesaving and limited resource manner [13]. The TL is pre-trained on somewhat similar tasks, referred to as Source Tasks. This training imbues the model greater efficiency on target or Destination Tasks. The utility of pre-training is that when the model confronts the destination tasks, it has to only to contrive adjustments, even additions, to the retained memory - ab initio learning is reduced or avoided, while dealing with new data. TL is best implemented where the source task's model has been trained on a much larger training set than could be obtained for the destination task. Overall, if appropriately used, transfer learning may bring tremendous benefits such as lower loss, higher accuracy, faster convergence and higher asymptotic accuracy [15].

III. Method

The presence of respiratory devices on the patient's face displays ongoing assisted ventilation. The presence of such a

*Research approved by Mayo Clinic institutional review board.

J. Chaudri is with the Department of Computer Science, Marshall University, Huntington, WV, 55902, USA. (phone: 304-696-2694; e-mail: chaudri@marshall.edu).

Q. Do is with the Department of Radiology, Mayo Clinic, Rochester, MN, 55902, USA. (email: do.quan@mayo.edu).

device in the near vicinity of the face could indicate that the respiratory device becomes detached (perhaps unintentionally). Besides, the availability of different respiratory devices on the patient's face at two different points in time, say T1 and T2, could be indicative of the patients shifting need for assisted respiration.

A. Study Design

Video recordings of participants who were admitted to the Intensive Care Unit (ICU) were collected. Images were extracted from these videos to obtain the data for training computer vision models.

One of the most important assets of this study was the designed logic to detect the deterioration in clinical status of the patients. This logic was based on clinical knowledge and then was converted into a computer algorithm. Table 1 displays the clinical logic of this project. When there is no respiratory device attached on the patient's face, it indicates no respiratory support. Otherwise, there is ongoing respiratory support. Shifting from one respiratory device to a more advanced respiratory device indicates increasing respiratory support and vice versa. NC has the lowest advancement level while oxygen mask (face mask) has the second lowest advancement level and HFNC has the third lowest advancement level and endotracheal tube has the highest advancement level. Furthermore, intubation can be recognized with the presence of an endotracheal tube at the moment, but no endotracheal tube attached in the past. On the other hand, extubation is the event that there is no endotracheal tube attached at the moment but there was an endotracheal tube attached in the past.

TABLE I. CLINICAL LOGIC

Type	**Rules Applied**		
	Name	***Code***	***Rules***
Object	Room Air[a]	1.1.0	Ordered by advancement[b]
	Nasal Canula	1.1.1	
	Face Mask	1.1.2	
	High-flow Nasal Cannula	1.1.3	
	Endotracheal tube	1.1.4	
Action	Intubation	A.1.1	1.1.4 detected at T+1 AND NO 1.1.4 detected at T
	Extubation	A.1.2	1.1.4 detected at T AND NO 1.1.4 detected at T+1
Clinical Phenotype	No respiratory support	C.1.0	Object < 1.1.1
	Respiratory support	C.1.1	Object > 1.1.0
	Increasing respiratory support	C.1.2	Object Code at T+1 > Object Code at T

a. No respiratory support needed. b. Ordered by the advancement in need of respiratory support (advancement levels).

B. Study Subjects

Patients who were admitted to the ICU and volunteered for this study. Patients or their families had to sign the consent form and have a clear understanding about the recording process. When a patient was incubated, the recruiters had to seek permission from the patient's family. A consent form had to be signed by a family member if she/he permitted the recordings and agreed to participate in the study. After the patient was extubated, the recruiters then had to explain the study and the recording process to the patients. If the patient agreed to participate, she/he had to sign a consent form. Patients' participants are voluntary, and patients have the right to withdraw his/her consent or discontinue participation at any time without penalty.

a. Eligibility Criteria

ICU patients who are equal to or more than 18 years old and provided with information about this study.

b. Exclusion Criteria

Patients who are younger than 18 years old. No minority populations will be included.

c. Sample Size

We recruited 25 volunteering patients. Each patient was recorded for a maximum of 24 hours or within the time frame suggested by the participants and/or healthcare providers. We recruited both male and female patients. Since skin color could be a bias factor in AI models, we recruited patients with different races / ethnicities. Also, patients in different age groups were included. Multiple videos were recorded for each patient in an effort to avoid label leakage which may lead to inaccurate estimates of the model's performance. There were a total of 257 recorded videos. The recordings at different time frames and multiple daytime and night hours were encouraged.

C. Data Collection

Data collection/data preparation is crucial for training an effective AI model since it decides the generalization ability of an AI model. For this project, group discussions helped us brainstorming a clear recording plan in advance in order to be able to collect images at different angles, backgrounds, lighting conditions and noises.

a. Camera Types

There were 2 types of cameras which were used for recording the videos. The focus camera was used to collect closed up videos/images of the face. And the wide angle camera to collect the video/image of the entire or a large section of the patient room. We bought cameras and lenses from FLIR (Blackfly S BFS-U3-200S6C)

b. Three Angles Recording

To suit our special needs we fashioned a mobile cart to pull the recording devices and accompanying gadgetry to patients beds. Each monopod carried one camera. Each camera was connected with the computer so we could save the data into the folder inside our LAN network. The cameras were placed at three different angles inside the patient room depending on the space availability. The camera could be placed at the bottom left corner of the room and points to the patient from the left. The camera could also be placed at the bottom right corner of the room and points to the patient from the right. Occasionally, the camera was placed in the middle of the room and pointed straight to the patient. The position of the cameras could be

interchangeable. No camera was just used to record for a specific angle.

Figure 1. Recording Cart

This is a visualization of the recording cart. Since the clinical staff allow only one cart at a time in a room, we actually later attach another pole into the cart so 2 cameras (wide angle and focus) are streaming data at the same time.

c. Multiple Background

Each ICU room was different in shape which helped collect videos with different backgrounds.

d. Lighting Condition

The videos were recorded both during day-time and at night if permitted. The lighting condition might also vary since patients might ask for switching off the light.

e. Noisy Data

Noisy data in this case is the recordings that contain additional meaningless information. We collected noisy data in situations where a large group of people gathered in a frame or there were many surrounding devices/cables inside the room.

f. Camera Settings

We selected the H.264 compression method to save space. This is the latest and most efficient lossless compression method. All other setting options were default. Len focuses adjustment might be needed.

g. Multiple Clinical Scenarios

We focused on recording videos of various clinical scenarios. Therefore, shifting in clinical status was closely monitored so we could collect data of multiple scenarios as much as possible.

D. Data Preprocessing

Video recordings were manually checked for quality and selection since a more robust and representative dataset helped increase the generalization ability and the robustness of a model. Selected videos were converted into images using a written script. The images were then manually labeled. Images of wrong positioning ventilation objects were difficult to find in the real world since it happened in very rare events such as when the process of attaching or taking off the device was not yet finished or when a patient tried to grab devices off unintentionally or intentionally without knowing the risk. Since there was not enough data collected to efficiently build a model that can recognize wrong positioning of devices at this stage, we decided to add another class label named "0.0.0" for the wrong positioning images and trained the model with 6 class labels. This approach to recognize wrong positioning devices is less effective compared to the use of face landmarks. We will come back to this later when more data will be collected.

E. YOLO Model

We trained a YOLO [12] ventilation object detection model. Images were manually labeled with 6 different class labels (room air, NC, face mask, HFNC, endotracheal tube, and wrong position) using labelImg software. YOLO requires only 1 forward propagation pass through the neural network to make predictions. Each input image is divided into a set of grid cells. Each grid cell has an associated vector in the output that indicates (1) whether an object exists in that grid cell (2) the class label of that object (3) the estimated location of the object (bounding box). With YOLO, one Convolutional Neural Network (CNN) simultaneously predicts multiple bounding boxes and prediction probabilities for those boxes.

The whole dataset contained 5,267 images. Except for the "wrong position" class, there were about 1000 images for each class label. 70% of the data was randomly selected for training and 30% was randomly selected for validation.

The network architecture was similar to a CNN model. There were 24 Convolution Layers, 4 MaxPooling Layers of size (2,2), and 2 fully connected layers. The input size was (160*160*3). There were 125 epochs and the batch_size was 30.

F. Transfer Learning Model

In order to train a transfer learning model for real-time face-object detection, we employed a pre-train face detection model and trained a classification model to classify the type of respiratory objects such as masks and tubes on the face. Both models will be called concurrently to recognize different respiratory devices on human faces. After experimenting with multiple open-source face object detection models, we decided to use the Multi-task Cascade Convolutional Neural Network (MTCNN). MTCNN is one of the most popular and accurate open-source libraries for face bounding boxes of face detection together with their 5 point facial landmarks [10,11]. It was trained with more than 400 thousand labeled images [10] and has been known for gaining state-of-art results on standard benchmark face detection datasets. The next step would be training a classification model for classifying images into 6 categorical classes: room air, NC, face mask, HFNC, endotracheal tube, and wrong position.

The images were not needed to be labeled in this case. A total of 5,267 images were selected. Except for the "wrong position" class, there were about 1000 images for each class label. 70% of the data was randomly selected for training and 30% was randomly selected for validation. Data augmentation methods were applied to boost model performance by diminishing overfitting and therefore increased the model's generalization ability. For this model, the input size was (160*160*3). There were two Convolution Layers followed by MaxPooling Layers of size (2,2), one Flatten Layer, one Dense Layer, and one final output Dense Layer with Softmax as activation function. There were 125 epochs and the batch_size was 30.

IV. RESULTS

A. YOLO Model

The current YOLO model for object detection of ventilation devices on patient's faces reached an accuracy rate of 95% while validation accuracy rate is 93%.

B. *Transfer Learning Model*

The classification model reached the training accuracy rate of 99% while validation accuracy rate was 98%.

C. *Comparison*

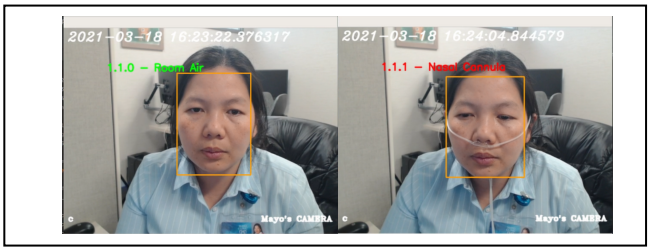
Transfer learning had better performance with fewer resources required. Since the labeling process is labor and time intensive, the absence of this labeling task helped us save time and resources.

TABLE II. PERFORMANCE COMPARISON

Model	*Training Accuracy*	*Validation Accuracy*
YOLO	95%	93%
Tranfer Learning	99%	98%

We tested the models against several videos recorded at the ICU to measure the difference in speed of recognizing objects of the two models. However, the performance was pretty good. The difference in the processing speed of the two models was not significant.

Figure 2. Sample Screens while Testing the model. *(pictures of author)*



V. DISCUSSION

Transfer learning model is a novel approach. The use of transfer learning would make it much easier for maintenance and improvement while saving time and resources. Later, if there is a new device needed to add to the model for detection, the process will be much easier and less time consuming especially data labeling will not be needed. Further, this approach makes the detection of wrong positioning devices on a patient's face more accurate since face landmarks can be detected easier. We can also foresee the need of face recognition for a patient's identity in the future for the monitoring processes. This approach makes it possible to improve the model with such directions.

As mentioned before, we are continuing with the video recordings of volunteer subjects to collect more data, especially images of wrong positioning devices on a patient's face. Current approach to recognizing wrong positioning is not efficient without the use of face landmarks. Further modification will be added later.

It is a good idea to also detect and track ventilation machines inside the room since this additional information may help recognizing changes in ventilation support. For the next step, we aim to (2) detect ventilation machines such as wall oxygen, high flow nasal cannula, non-invasive ventilation devices (CPAP/BiPAP), and invasive mechanical ventilation (mechanical ventilator) inside the patient room by computer vision, and combine the results of aims 1 and 2 in an algorithm to determine adverse events (such as inadvertent dislocation of endotracheal tube) and clinical deterioration or worsening need of respiratory support. The training of a new YOLO model may be needed in this case. However, these ventilation machines are larger in size and would enhance the accuracy rate.

VI. CONCLUSION

The results of algorithms for mask and tube detection on the human face were encouraging. In this study, we were able to achieve 98% accuracy to recognize ventilation devices on human faces in validation images set using transfer learning approach. In the next phase, we will evaluate the implementation performance of both the models to indicate which model works better in real world scenarios. Computer-vision carries many possible uses in healthcare settings, especially in the ICU. Building on object detection, future directions include recognition of events based on object changes and patterns able to predict adverse events.

REFERENCES

[1] Dašić, P., Dašić, J., & Crvenković, B. (2017). Improving patient safety in hospitals through usage of cloud supported video surveillance. Open access Macedonian journal of medical sciences, 5(2), 101–106.

[2] Goran, S. F. (2010). A second set of eyes: an introduction to tele-ICU. Critical Care Nurse, 30(4), 46-55.

[3] Panlaqui, M., Broadfield, E., Champion, R., Edington, P., & Kennedy, S. (2017). Outcomes of telemedicine intervention in a regional intensive care unit: a before and after study. Anesthesia and intensive care, 45(5), 605-610.

[4] Rajagopalan, R., Litvan, I., & Jung, T. P. (2017). Fall prediction and prevention systems: recent trends, challenges, and future research directions. Sensors (Basel, Switzerland), 17(11), 2509.

[5] Su, L., Waller, M., Kaplan, S., Watson, A., Jones, M., & Wessel, D. L. (2015). Cardiac resuscitation events: one eyewitness is not enough. Pediatric Critical Care Medicine, 16(4), 335-342.

[6] Yagi, T., Koizumi, Y., Aoyagi, M., Kimura, M., & Sugizaki, K. (2005). Three-dimensional analysis of eye movements using four times high-speed video camera. Auris Nasus Larynx, 32(2), 107-112.

[7] Haque, A., Guo, M., Alahi, A., Yeung, S., Luo, Z., Rege, A., ... & Platchek, T. (2017). Towards vision-based smart hospitals: A system for tracking and monitoring hand hygiene compliance.

[8] Yeung S, Rinaldo F, Jopling J, Liu BB, Mehra R, Downing NL, et al. A computer vision system for deep learning-based detection of patient mobilization activities in the ICU. Npj Digit Med. 2019;2.

[9] Taylor, K., VanDenberg, S., le Huquet, A., Blanchard, N., & Parshuram, C. S. (2011). Parental attitudes to digital recording: A pediatric hospital survey. Journal of pediatrics and child health, 47(6), 335-339.

[10] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503.

[11] Luka, D. (2020). Face recognition with FaceNet and MTCNN. Retrieved from https://arsfutura.com/magazine/face-recognition-with-facenet-and-mtcnn.

[12] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[13] Rawat, W. and Wang, Z., 2017. Deep convolutional neural networks for image classification: A comprehensive review. Neural computation, 29(9), pp.2352–2449.

[14] Kim, YG., Kim, S., Cho, C.E. et al. "Effectiveness of transfer learning for enhancing tumor classification with a convolutional neural network on frozen sections." *Scientific reports* vol. 10,1 21899. 14 Dec. 2020, doi:10.1038/s41598-020-78129-0.