DEEP LEARNING BASED FPN AND MT-CNN FACE MASK DETECTION SYSTEM

Mr B. Madhava Rao[1], Dr.N.Satheesh[2],Dr.B.Rajalingam[3],Mr Bavankumar[4], Mr E.Lingappa[5],Dr.N.Satheesh[6]

[1,4,5]Assistant Professor,Department of Computer Science and Engineering,
[2,3,6]Associate Professor,Department of Computer Science and Engineering
St.Martin's Engineering College,Dhulapally, Near Kompally,Secunderabad-500 100.Telangana, India

Abstract-

With the fast global spread of Corona virus (COVID-19), wearing face masks in public becomes a need to limit the transmission of this or other pandemics. However, with the absence of on-ground automated preventative techniques, reliance on people to enforce face mask-wearing rules in universities and other organizational facilities is an extremely expensive and time-consuming strategy. Without tackling this difficulty, controlling highly airborne transmittable illnesses would be unfeasible, and the time to respond will continue to expand. Considering the high personnel traffic in buildings and the effectiveness of countermeasures, that is, detecting and offering unmasked personnel with surgical masks, our aim in this paper is to develop automated detection of unmasked personnel in public spaces in order to respond by providing a surgical mask to them to promptly remedy the situation. Our technique comprises of three main components. The first component utilizes a deep learning architecture that mixes deep residual learning (ResNet-50) with Feature Pyramid Network (FPN) to detect the presence of human beings in the videos (or video stream) (or video feed). The second component utilizes Multi-Task Convolutional Neural Networks (MT-CNN) to detect and extract human faces from these videos. For the third component, we develop and train a Convolutional neural network classifier to detect masked and unmasked human participants. Our algorithms were implemented in a mobile robot, Thor, and assessed using a dataset of videos taken by the robot from public locations of an educational university in the U.S. Our assessment findings demonstrate that Thor is quite accurate earning an F1 score of 87.7 percent with a recall of 99.2 percent in a range of scenarios, an acceptable accuracy considering the tough dataset and the issue area.

Keywords--machine learning, Convolutional neural networks, and face detection, deep learning, mask detection, COVID-19.

I. INTRODUCTION

The globe is facing a health catastrophe owing to the fast spread of Corona virus Disease 2019 (COVID-19) (COVID-19). According to the World Health Organization (WHO) COVID-19 dashboard [1], more than more than 109 million persons were infected with COVID-19 across 188 countries. The WHO produced many reports that give advice and mitigating methods to minimize the spread of the illness. According to the abovementioned reports and numerous research studies, wearing a face mask is extremely efficient in avoiding the spread of respiratory viruses like COVID-19 [2]–[4]. For instance, Sim et al. [5] conducted a thorough investigation and showed that the efficiency of wearing N95 mask in avoiding SARS transmission is 91 percent. Since the emergence of COVID-19, several companies have changed their polices to demand wearing face masks in public to safeguard their workers and community

from the illness [6]. Therefore, a significant task of the artificial intelligence and machine learning community is to propose innovative technologies to automatically detect circumstances when individuals fail to use face masks in public settings to assist prevent the spread of COVID19 and other pandemics. For example, France connected an AI-based system to the Paris Metro surveillance cameras [7] to offer data regarding the adherence to the face mask legislation. Recent improvements in deep learning methods and its core component Deep Neural Networks (DNNs), have greatly improved the performance of picture classification and object recognition [8], [9]. Convolutional Neural Networks (CNNs or ConvNets) are a main model of DNNs that have proven greater efficacy in areas such as image recognition and classification. CNNs have been particularly effective at identifying human subjects, faces, and other objects in images and videos due of their tremendous feature extraction capabilities. In this research, we explore the question: can we construct a deep learning-based classifier to detect unmasked faces from low quality images? Our objective is to examine the potential of deep learning to extract significant characteristics from low-quality images captured by a mobile robot (Thor) to construct a classifier that recognizes unmasked personnel with high accuracy. We characterize low-quality images (and videos) not just as low-resolution images, but also other factors that substantially impair feature extraction from images. These factors are as follows:

• The height gap between the camera and the face. Our mobile robot records images using a camera that is 1-foot high from the ground, which produces partial facial images that are more challenging for feature extraction and classification than popular datasets that comprise largely images obtained by cameras at the same height level of the face.

• The angle between the camera and the face. Facial images are not often captured with subjects facing the camera directly, unlike most popular datasets. In actuality, a dataset might comprise images of human subjects that are walking away or at a 90 degree angle from the camera, which results in partial face images that add further hurdles to the image classification and mask detection tasks.

• Quality of light. Unlike most common datasets, utilizing a mobile robot to record videos or images produces in images that are captured in locations with a lighting quality that ranges from low to intense. This variance in the dataset poses a new difficulty to solve. • Distance to human subjects. Capturing images at varying distances between the camera and human subjects makes the task of feature extraction and subsequently image classification more challenging because, at far distances, the areas of interest in the image (i.e., the human subject, face, and mask) are smaller which provides less powerful features to use for face and mask detection in such images.

II. RELATED WORKS

To assist the organizations and the community fight against the fast spread of Corona virus Disease 2019 (COVID-19), there have been considerable efforts to spread awareness and share counter measures with the public to minimize the spread of COVID-19. Wearing a face mask in public is a vital countermeasure to restrict the spread of COVID-19 [4], and hence, many educational and industrial organizations have changed their policy to include having to wear face masks when on campus or within buildings. In general, most research efforts are focused on face creation and recognition for identify-based authentication. Loey et al. [10] demonstrated a machine learning based method for detecting face masks using a collection of

high quality conference-like face photos. Their model attained an accuracy of 99.64 percent -100 percent in detecting face masks. These photographs however, were captured while a face was facing at a computer camera that is a few inches away. This is not ideal to use in genuine circumstances when individuals are strolling about tens of steps away with different angles that may only have partial vision of people's faces and masks. Qin and Li [11] proposed a system that assesses the appropriateness of mask wearing based on its positioning. The approach places each circumstance into one of three categories: accurate placement of the mask, poor placement, and no mask at all. Face masks and mask placements could be detected with 98.7 percent accuracy using the new method. Ejaz et al. [12] examined and compared face recognition accuracy as an identity-based verification utilizing Principal Component Analysis (PCA) to identify a person. They observed that the accuracy of face identification reduced to 73.75 percent while wearing masks. Li et al. [13] proposed an approach for face detection using YOLOv3 algorithm. The approach created a classifier utilizing more than 600,000 photos of human faces given from CelebA and WIDER FACE databases. The approach attained an accuracy of 92.9 percent in detecting faces. Nieto-Rodr´ıguez et al. in [14] and [15] proposed an approach for detecting surgical face masks in operation rooms. The major purpose of this approach is to limit the false positive face recognition in order to only warn workers who are not wearing masks within operating rooms. To do this, the approach takes use of the unique surgical masks color to limit false positives. The approach produced a recall over 95 percent with a false positive rate below 5 percent for the identification of faces and surgical masks. However, unlike operating rooms where the medical team exclusively wears the highly recognized surgical masks, many other individuals use masks with diverse colors and designs. The proposed approach in [14] and [15] will be affected by this fluctuation, and as a result, their stated accuracy may decrease dramatically.

III. PROPOSED METHOD

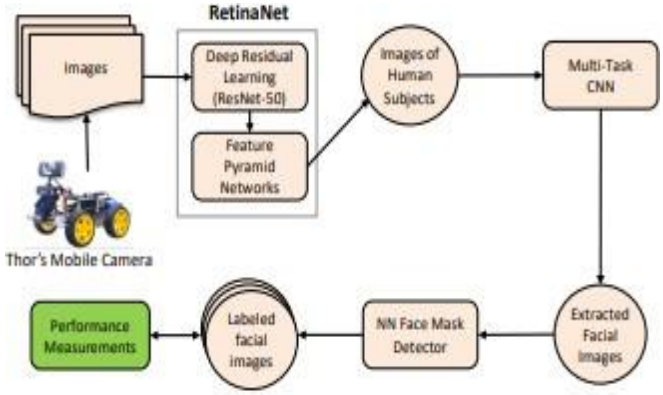

Fig. 1. The Architecture of Thor.

This study is done utilizing a pipeline of three detection models that are developed and evaluated on four publically accessible datasets in addition to our own dataset that we collected to examine our research purpose. Table I summarizes these datasets. In this section, we describe each of these datasets. COCO Dataset. Microsoft Common Objects in Context (COCO) [19] is a large-scale object identification dataset that comprises a total of 330,000 images of 91 item kinds (including human subject) (including human subject). These images

have been tagged in this dataset with 2.5 million item labels. The COCO dataset was utilized to construct a pre-trained model that recognizes human subjects in captured images using our approach. We offer further insights regarding this approach in Section IV. CelebA. The CelebFaces Characteristics Dataset (CelebA) [20] is a large-scale face dataset that comprises 202,599 facial images of celebrities where each of these images has been tagged with 40 binary attributes. This dataset provides a mostly diversified collection of faces with various posture variants of human subjects which makes it a goldmine for training classifiers for face attribute identification, face detection and extraction (of facial portion) from the supplied image. WIDER FACE. This is a face detection benchmark dataset [21] that comprises 393,703 tagged faces with substantial diversity in terms of position and occlusion. The CelebA and WIDER FACE datasets were utilized to construct a pre-trained model that recognizes facial images (the facial region) in captured images using our approach. CMCD. the Custom Mask Community Dataset [22] has 1,376 facial images that are well-balanced in terms of mask wearing. 50.15 percent (or 690) of these images have masked faces and 49.85 percent (or 686) contain unmasked faces. Our approach leverages this data to construct a Convolutional neural network model for mask recognition in the images it collects. Figure 1 demonstrates the architecture of Thor containing an Image Generator (IG), a human subject detector (HSD), a face detector and extractor (FD), a mask detector (MD) (MD). First, the IG regularly captures videos from numerous spaces and hallways in our business. Then, it minimizes the size of the movie by sampling its images by retaining one image every second TABLE I SUMMARY OF DATASETS dataset images content number of images COCO [19] human subjects 330,000 WIDER FACE [21] The CelebA [20] facial images 600,000 CMCD [22] photos of disguised and unmasked faces 1,376 Standard Images Deep Residual Learning (ResNet-50) (ResNet-50) Feature Pyramid Networks Multi-Task CNN Images of Human Subjects Retina Net Extracted Facial Images Identified and categorized images of facial faces Performance Measurements NN Face Mask Detector Thor's Personal Camera on Wheels Fig. 1. The Architecture of Thor. and deleting the other images captured in that second. This sampling is critical because the enormous number of images captured in each second considerably increases the size of the data and stresses the robot’s resources. Therefore, we programmed the robot to save 1 image each second and discarded the other images captured in that second that is unlikely to contribute further information. Then, the HSD recognizes the presence of human subjects in these images and filters out the ones that do not feature human subjects. The FD then recognizes and extracts human faces from these images and delivers them for the MD. The extracted faces are then classified as either "Masked" or "Unmasked" by the physician who performed the procedure. Unmasked people may be alerted by the robot's two speakers, and the robot would then give them a mask.

A. Data Collection and Preprocessing As indicated previously, our robot (Thor) is equipped with a modified Donkey Car for movement and a Raspberry Pi 1080p Camera that collects 20 images/second for data collecting. We have deployed Thor to roam a university campus and it collected over 150 videos from different hallways and spaces. These videos had different durations and most of their material was empty (e.g., no action in images) (e.g., no activity in images). There were 229 human subjects in our dataset after a manual review of all of the video footage. 198 of the human subjects

were facing the camera whilst the other 31 subjects were not facing the camera and hence did not offer any facial film. 133 of the subjects were wearing masks and 65 subjects were not wearing masks. To decrease the size of the data (i.e., number of images), our sampler picked just one image (frame) from the 20 frames captured in each second and discarded the other 19 images captured during that second. This technique decreased the size of our data to 5 percent and enhanced the performance of the following detecting modules by 95 percent. In the following portion, we describe how we recognize human subjects in these images.

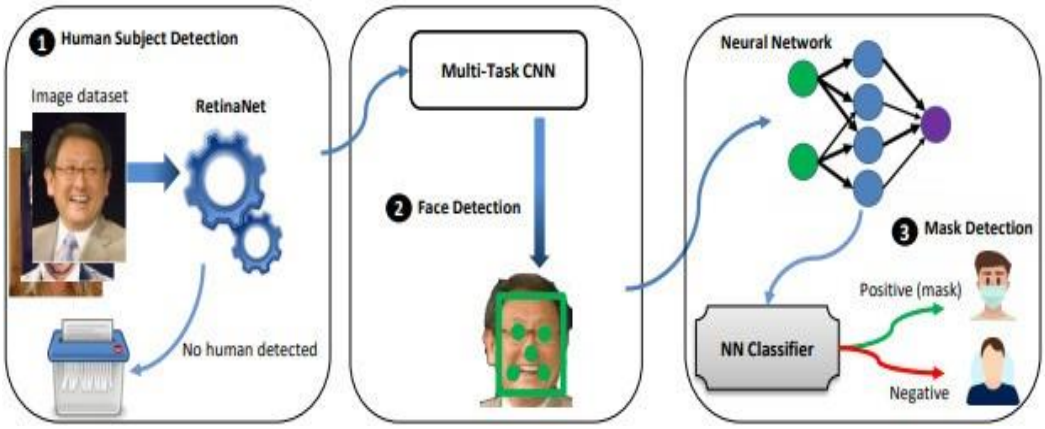

Fig. 2. Workflow of face mask detection using pipeline of deep learning techniques.

IV. RESULTS AND DISCUSSION

According to our findings, Thor goes above and above to collect, identify, and minimize problems involving unmasked people in enclosed areas. With the growing potential of transmitting infections, automatically detecting, notifying, and giving a mask for unmasked individuals may effectively address present and developing air born diseases. However, our present implementation of Thor is very rudimentary and in this part we explain the limitations and potential future research of our approach. Error/misdetection analysis. Our research shows that Thor has a good recall and accuracy considering the nature of our dataset. Thor, on the other hand, is still prone to mistaking real masks for fakes. These challenges generally originate from the limitations of current datasets and consequently the classifiers we employ that are trained on these datasets. Specifically, the images given by these datasets vary from images that are captured by a 1-foot-tall mobile robot in terms of the angle of capture that varies with distance and the available indoor illumination in the image. For example, utilizing the Multi-Task cascaded Convolutional Neural Network (MTCNN) [8] on our dataset to recognize human faces (facial regions) produced an accuracy of 94.4 percent. The data responsible for the loss of 5.6 percent accuracy directly influences the result

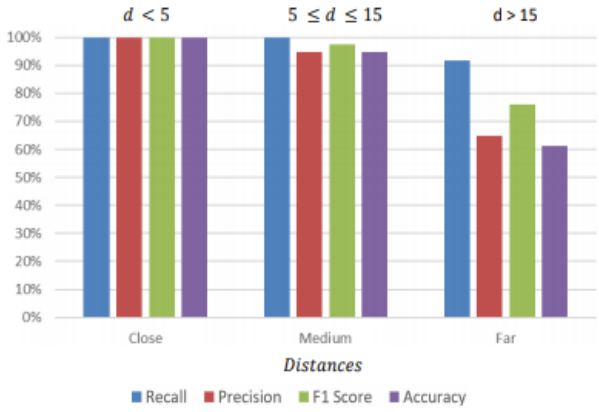

Fig. 3. The impact of the distance (d) between the human subject and the camera on the detection accuracy of face masks

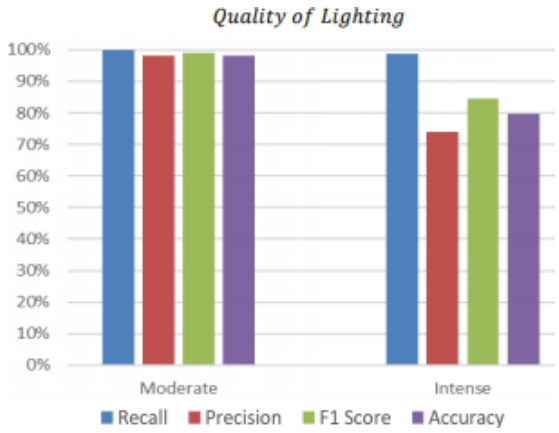

Fig. 4. The impact of the quality of lighting (q) in the space where image is captured on the detection accuracy of face masks.

CONCLUSION

In this research, we describe Thor, a system that employs deep learning-based algorithms for autonomous detection of unmasked personnel in public spaces. Thor devised a novel approach that incorporates multiple forms of deep learning for face mask detection. Our prototype robot is constructed of three components. The first module uses a combination of ResNet-50 and Feature Pyramid Network for feature extraction and human subject detection. The second module uses Multi-Task Convolutional Neural Network (MT-CNN) to recognize and extract faces from images involving human subjects. Then, the third module uses our developed neural network model to categorize the processed images to masked or unmasked. This rating permits spotting unsafe indoor settings of unmasked personnel. To mitigate such situations, Thor offers a surgical mask to the detected unmasked personnel. We assessed our approach using a dataset of 229 human subjects collected by our mobile robot,

Thor. The approach produced a mask detection accuracy of 81.3 percent with a very high recall of 99.2 percent. To the best of our knowledge, this is the first attempt that explores detecting face masks in images that are captured in many demanding situations such as space lighting and distance to camera.

REFERENCES

[1] "WHO Coronavirus Disease (COVID-19) Dashboard," Accessed: January 10, 2021. [Online]. https://covid19.who.int, 2021.

[2] B. J. Cowling, K.-H. Chan, V. J. Fang, C. K. Cheng, R. O. Fung, W. Wai, J. Sin, W. H. Seto, R. Yung, D. W. Chu et al., "Facemasks and hand hygiene to prevent influenza transmission in households: a cluster randomized trial," Annals of internal medicine, vol. 151, no. 7, pp. 437–446, 2009.

[3] S. M. Tracht, S. Y. Del Valle, and J. M. Hyman, "Mathematical modeling of the effectiveness of facemasks in reducing the spread of novel influenza a (h1n1)," PloS one, vol. 5, no. 2, p. e9018, 2010.

[4] S. Feng, C. Shen, N. Xia, W. Song, M. Fan, and B. J. Cowling, "Rational use of face masks in the covid-19 pandemic," The Lancet Respiratory Medicine, vol. 8, no. 5, pp. 434–436, 2020.

[5] S. W. Sim, K. S. P. Moey, and N. C. Tan, "The use of facemasks to prevent respiratory infection: a literature review in the context of the health belief model," Singapore medical journal, vol. 55, no. 3, p. 160, 2014.

[6] H. Elachola, S. H. Ebrahim, and E. Gozzer, "Covid-19: Facemask use prevalence in international airports in asia, europe and the americas, march 2020," Travel Medicine and Infectious Disease, 2020.

[7] "Paris Tests Face-Mask Recognition Software on Metro Riders," Accessed: January 10, 2021. [Online]. https://Bloomberg.com, 2021.

[8] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.

[10] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic," Measurement,vol. 167, p. 108288, 2020.

[11] B. QIN and D. LI, "Identifying facemask-wearing condition using image super-resolution with classification network to prevent covid-19," 2020.

[12] M. S. Ejaz, M. R. Islam, M. Sifatullah, and A. Sarker, "Implementation of principal component analysis on masked and non-masked face recognition," in 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019, pp. 1–5.

[13] C. Li, R. Wang, J. Li, and L. Fei, "Face detection based on yolov3," in Recent Trends in Intelligent Computing, Communication and Devices. Springer, 2020, pp. 277–284.

[14] A. Nieto-Rodr´ıguez, M. Mucientes, and V. M. Brea, "System for medical mask detection in the operating room through facial attributes," in Iberian Conference on Pattern Recognition and Image Analysis. Springer, 2015, pp. 138–145.

[15] A. Nieto-Rodr´ıguez, M. Mucientes, and V. M. Brea, "Mask and maskless face classification system to detect breach protocols in the operating room," in Proceedings of the 9th International Conference on Distributed Smart Cameras, ser. ICDSC '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 207–208. [Online]. Available: https://doi.org/10.1145/2789116.2802655

[16] J.-S. Park, Y. H. Oh, S. C. Ahn, and S.-W. Lee, "Glasses removal from facial image using recursive error compensation," IEEE transactions on pattern analysis and machine intelligence, vol. 27, no. 5, pp. 805–811, 2005.

[17] M. K. J. Khan, N. Ud Din, S. Bae, and J. Yi, "Interactive removal of microphone object in facial images," Electronics, vol. 8, no. 10, p. 1115, 2019.

[18] N. U. Din, K. Javed, S. Bae, and J. Yi, "A novel gan-based network for unmasking of masked face," IEEE Access, vol. 8, pp. 44 276–44 287, 2020.

[19] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollar, "Microsoft coco: ´ Common objects in context," 2015.

[20] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 3730–3738.

[21] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5525–5533

[22] "Custom Mask Community Dataset (DMCD)," Accessed: January 10, 2021. [Online]. Available: https://github.com/prajnasb/observations , 2021.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss ´ for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[24] 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, 2016, pp. 770–778.

[25] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[26] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 2015.

[27] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," UMass Amherst technical report, Tech. Rep., 2010.

[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[29] GIMP, 2020, (Accessed: 12.07.2020). [Online]. Available: https://www.gimp.org/downloads/