Face mask detection and counting using you only look once algorithm with Jetson Nano and NVDIA giga texel shader extreme

Article *in* International Journal of Artificial Intelligence · September 2023

DOI: 10.11591/ijai.v12.i3.pp1169-1177

Nawaid Hasan
Multimedia University

Nur Asyiqin Amir Hamzah
Multimedia University

Hatem Fahd Al-Selwi
Multimedia University

Face mask detection and counting using you only look once algorithm with Jetson Nano and NVDIA giga texel shader extreme

Hatem Fahd Al-Selwi[1], Nawaid Hassan[1], Hadhrami Bin Ab Ghani[2], Nur Asyiqin binti Amir Hamzah[1], Azlan Bin Abd. Aziz[1]

[1]Center for Engineering Computational Intelligence, Faculty of Engineering and Technology, Multimedia University Melaka, Malaysia
[2]Department of Data Science, Universiti Malaysia Kelantan, Kelantan, Malaysia

Article Info

Article history:

Received Mar 12, 2022
Revised Jul 14, 2022
Accepted Aug 25, 2022

Keywords:

Deep learning
Internet of things
Jetson Nano
Nvidia
You only look once

ABSTRACT

Deep learning and machine learning are becoming more extensively adopted artificial intelligence techniques for machine vision problems in everyday life, giving rise to new capabilities in every sector of technology. It has a wide range of applications, ranging from autonomous driving to medical and health monitoring. For image detection, the best reported approach is the you only look once (YOLO) algorithm, which is the faster and more accurate version of the convolutional neural network (CNN) algorithm. In the healthcare domain, YOLO can be applied for checking the face mask wearing of the people, especially in a public area or before entering any closed space such as a building to avoid the spread of the air-borne disease such as COVID-19. The main challenges are the image datasets, which are unstructured and may grow large, affecting the accuracy and speed of the detection. Secondly is the portability of the detection devices, which are generally dependent on the more portable like NVDIA Jetson Nano or from the existing computer/laptop. Using the low-power NVDIA Jetson Nano system as well as NVDIA giga texel shader extreme (GTX), this paper aims to design and implement real-time face mask wearing detection using the pretrained dataset as well as the real-time data.

Corresponding Author:

Hadhrami Ab Ghani
Department of Data Science, Universiti Malaysia Kelantan
Karung Berkunci 36, Pengkalan Chepa, 16100 Kota Bharu, Kelantan, Malaysia
Email: hadhrami.ag@umk.edu.my

1. INTRODUCTION

Deep learning has sparked significant attention in a spectrum of uses, such as machine vision [1]–[4]. It attempts to discover the target visual features from randomly generated image sources. Some examples of applicable deployments include facial recognition [5]–[7] motion detection [8]–[11], image classification [12], [13], and vehicle detection [14]–[17]. Deep learning creates new opportunities for the development of intelligent interactions between people and their devices or technology, paving the path for these new possibilities to emerge. As a result of the current global epidemic situation, individuals are mandated to wear facial masks, which has resulted in the challenge of inspecting individuals when they are wearing facial masks in settings such as public or open spaces. Particularly in the wake of the present worldwide pandemic scenario, when certain healthcare protocols must be followed, including the wearing of facial masks and social separation, face recognition has emerged as one of the topics that is garnering a lot of attention as a hot button issue [18].

Currently, different approaches have been devised for detecting facial masks based on deep learning [19] such as the region proposal network (RPN) [20], [21] and the faster region-based convolutional neural networks (R-CNN) network methods [22]–[24]. On the other hand, the detection speed of these methods is relatively slow, and this is especially the case when they are implemented on low-power processing units like the NVDIA Jetson Nano. The you only look once (YOLO) algorithm [23], [25]–[27] is a novel method that is based on deep learning. It is an enhanced and faster alternative to the traditional approach. It has been established that the efficacy of YOLO is ten times higher than that of faster R-CNN. As a consequence of this, it is of the utmost importance to make certain that the categorization software is installed in a system that has a limited amount of central processing unit (CPU) power and a computational capacity that is relatively low. It is also of the utmost importance to install the software in a system that has a relatively low computational capacity.

NVIDIA's Jetson is a feasible enabler for the introductory phase of machine and computer vision due to its low-power processing capability that it offers as one of forefronts in artificial-intelligence hardware development for computer vision [28]–[30]. The CPU- graphics processing unit (GPU) architecture of Jetson Nano [31], [32], enables the CPU to load faster while the GPU seamlessly runs the machine-learning techniques. It has a sleek design, is portable, and uses little power, making it perfect for application domains which require constraints in weight and power. Because of the improved processing time and accuracy, YOLOv5 is anticipated to operate feasible image detection functions like facial mask wearing detection identification and counting using Jetson Nano. Therefore, this paper presents a YOLOv5-based algorithm for face mask wearing detection and counting using both the NVDIA's GTX and Jetson Nano platforms to address these healthcare and monitoring issues.

2. METHOD

This section describes the step-by-step methodology employed in the study. There are four main steps employed in this paper, which are the development of the deep learning model, the data creation, the model training and the model inferencing. In the next subsection, the employed deep learning model will be presented.

2.1. Deep learning model

A one-stage algorithm based on YOLOv5, which is considerably fast in processing detection and prediction, is used. The algorithm has a unique characteristic in which it is capable of redefining the detected object as a regression problem so that it can be computed at a high computation rate [1], [2]. This is essential to ensure fast detection performance on a standalone platform like Jetson Nano which has a limited processing capacity.

The one-stage architecture of YOLOv5 consists of three main components: the backbone, neck and head. The backbone (CSPDarknet) is a convolutional neural network (CNN) which is responsible for performing feature extraction by collecting and shaping the image features at various granularities. It utilizes the center and scale prediction (CSP) technique to produce the image features. Next, the features will be forwarded to the neck (PANet) stage for feature fusion, where image features are combined. Finally, the combined features will be fed to the head (YOLO layer) for prediction and classification. Figure 1 illustrates the YOLOv5 architecture. In this paper, a recent deep model based on YOLO v5 [23] is proposed and implemented onto a Jetson TX1 [1] for facial mask wearing detection and counting. In the next section, the YOLO v5 model implemented with Jetson Nano TX1 is presented before the results and conclusion are presented in section III and IV respectively.

2.2. Data creation

To train the model, a dataset of images was created using public images of face masks. The images were divided into three categories; i) with a face mask, ii) without a face mask, and iii) masks worn incorrectly. The dataset included 848 images. A data augmentation technique is then applied to generate more images and add some noise to the images to ensure the proposed model is robust against noise. As a result, a total of 2034 images with and without noise were generated. All the images were annotated using the YOLOv5 format for training – PyTorch version of YOLOv5. For training, 87% of the dataset was allocated, with 8% for validation and 4% for testing. The framework of the detection is shown in Figure 2.

2.3. Model training

Two YOLOv5 models have been trained with and without using a pretrained model. Both models were trained using the Nvidia GTX 1660 6 GB GPU for 100 epochs. From the results obtained, it has been observed that the models achieved acceptable and satisfactory accuracy values in the training.

2.4. Model inferencing

The trained models were tested in two different testing modes with distinct computational power. First, the mode is tested using an Nvidia GTX 1660 6 GB. For the second test, an embedded system using a Jetson Nano Board 4 GB has been used.

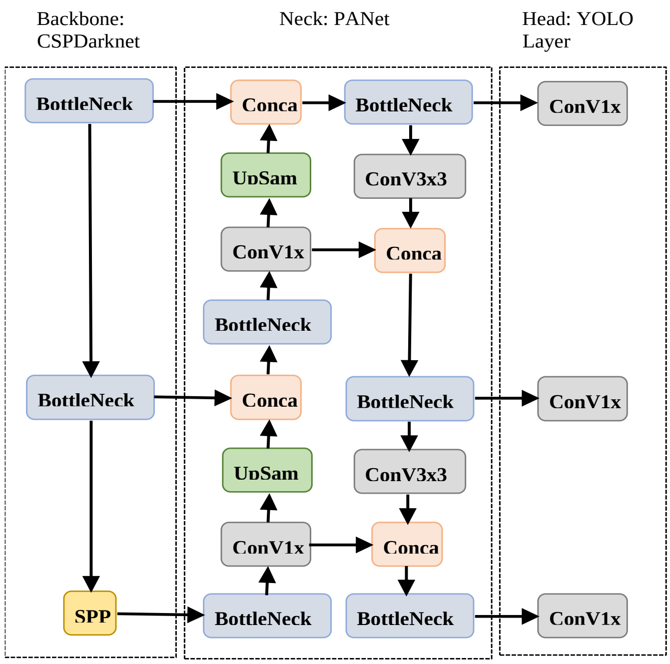

Figure 1. YOLOv5 architecture. CSP (cross stage partial network), SPP (spatial pyramind pooling), conv (convolutional layer) and concat (concatenate function) sub-components

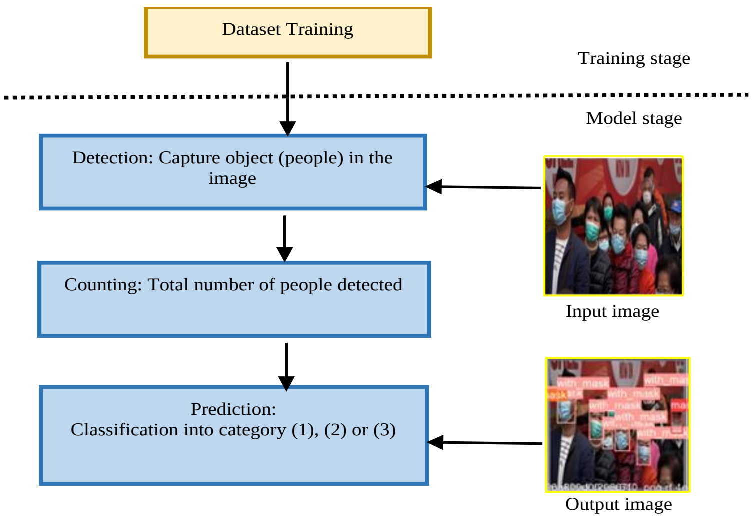

Figure 2. Framework of the proposed model

3. RESULTS AND DISCUSSION

The proposed YOLOv5 model has been trained in two configurations, either with or without the pretrained model (from scratch). The dataset used in this training is the Kaggle facemask detection dataset. The training and testing of the proposed YOLOv5 model are carried out on this dataset. Before delving deeper into YOLOv5, a comparison study was conducted between the proposed YOLOv5 model and other current deep learning-based approaches for face mask recognition, the results of which are displayed in Table 1. It is clear that the YOLOv5 model has attained the highest mAP when compared to the other approaches. Furthermore, YOLOv5 has been recorded to be capable of performing the detection operation at very high frame rates frames per second (FPS), which are 120 FPS using the NVDIA GTX 660 6 GB and 20 FPS with the Jetson Nano.

Table 1. Comparison between different deep learning-based methods for face mask detection

Face mask detection models	mAP Measurement	NVDIA GTX 660 6GB	Jetson Nano
Centernet resnet50 v2	0.57	20FPS	-
Faster R-CNN resnet50 v1	0.59	8.3FPS	-
Ssd mobilenet v1 fpn	0.61	15FPS	-
Ssd resnet50 v1 fpn	0.57	12FPS	-
YOLOv5	0.70	120FPS	20FPS

In the training phase, various input images containing people with/without face masks from the public Kaggle dataset have been fed as the input dataset to the YOLOv5 algorithm. Some of the images which have been detected with the face mask wearing have been recorded and shown in Figure 3. It can be observed in this figure that the algorithm is capable of detecting the people wearing the facial masks and counting them, regardless of the number of people, in most of the images found in the dataset. The numbers of people detected wearing the masks correctly in each of the images, starting from the top left image in figure are: 4, 7, 0, 1, 1, 0, 4, 10, 3, 1, 1, 1, 1, 1, 2.

Figure 3 shows the confusion matrix of the model. It can be noticed that the model confuses mostly between with mask class with mask weared incorrect, which is by a rate of 0.43. In other words, almost 43% of the people who wear the maks incorrectly have been mistakenly predicted as the people who wear the masks correctly. This is expected due to the similarity between the two classes. Another reason is the small number of samples for the mask weared incorrect class. However, the model faces no issue differentiating between the two with mask and without mask classes as it shows a 0.02 confusion rate. In other words, only 2% of the people who do not actually wear masks that are mistakenly predicted as the people who wear masks.

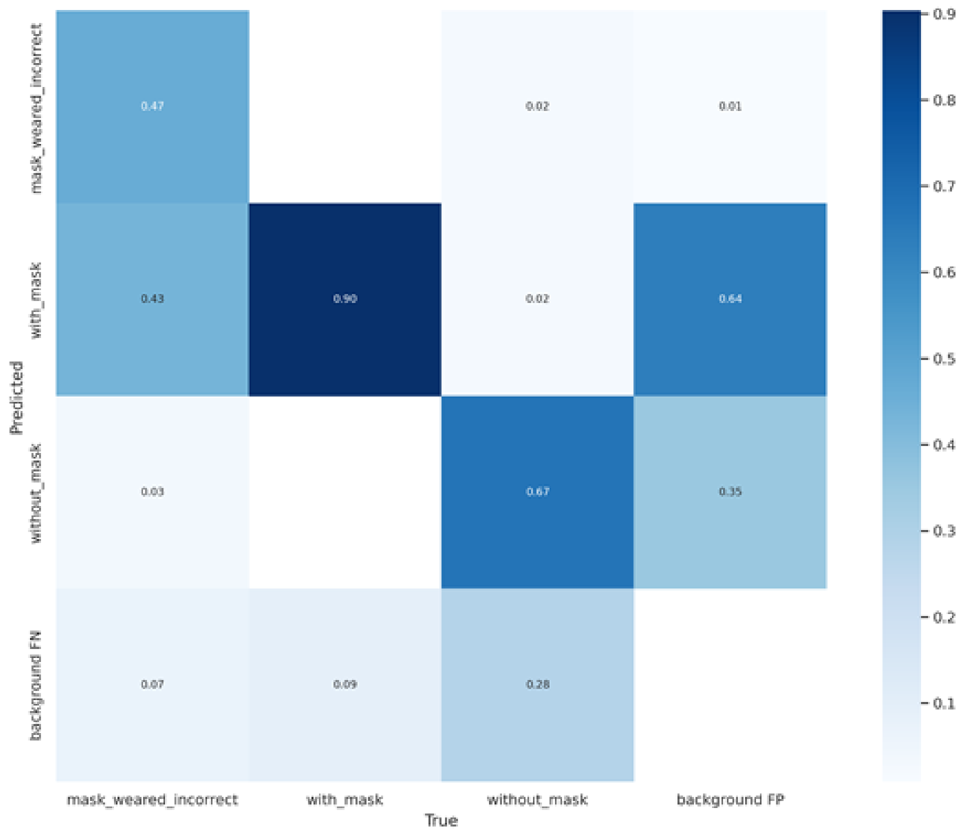

Figure 3. The confusion matrix

Figure 4 shows the F1 score for the model over different confidence levels. As can be seen in the figure, the model shows the highest F1 score and confidence level for the with mask class compared to the other classes. The mask weared incorrect has the lowest F1 score over confidence, which agrees with the confusion matrix result presented before.

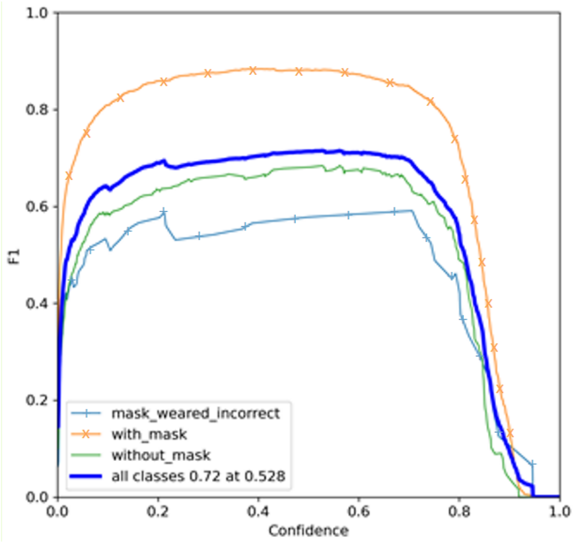

Figure 4. F1 score

Figure 5 shows the overall analysis result for the model, including the training loss, the box loss, the recall and the precision (mAP) measurement. It can be seen that the model progresses well over the 100 epochs. The mAP value keeps increasing as the number of epochs increases and the training loss reduces as the number of epochs increases.

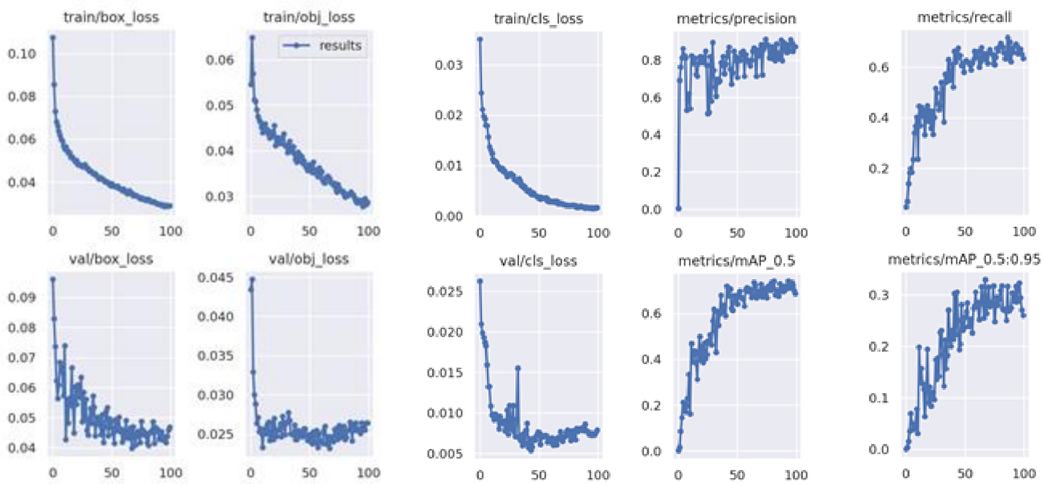

Figure 5. Various metrics measured over 100 epochs, where the horizontal axis represents the number of epochs (The horizontal axis represents the number of epochs and the vertical axis represents the corresponding loss, precision or recall values accordingly)

Figure 6 shows the detection results obtained when the model is tested without using the pretrained model. Although most of the images shown in Figure 6 have been successfully detected as either with a mask, without a mask, or a mask worn incorrectly, the overall performance is slightly lower than the result obtained when the algorithm is tested using the pre-trained model. The reason behind the slight decrease in the performance of the algorithm when tested without the pre-trained model is straight forward.

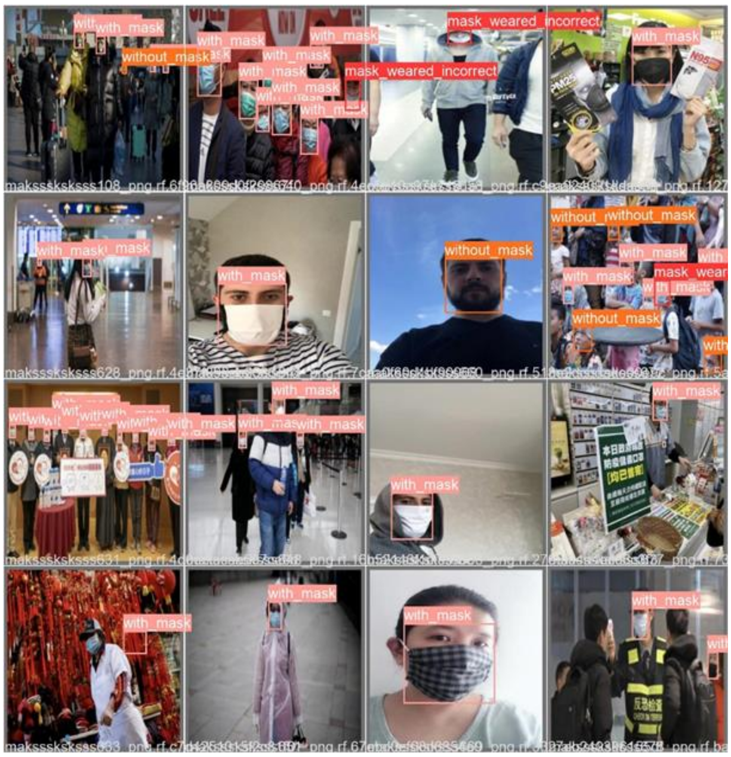

Figure 6. Detection results using the pretrained model

Figure 7 shows the F1 score obtained without the pre-trained model over different confidence levels. As seen in the figure, the highest F1 score and confidence level are obtained when detecting the with mask class as compared to the other classes. The mask weared incorrect class has the lowest F1 score over confidence, which the confusion matrix explains. It can be noticed that the F1 score for the mask weared incorrect class decreases radically as compared to the result obtained when using the pre-trained model.

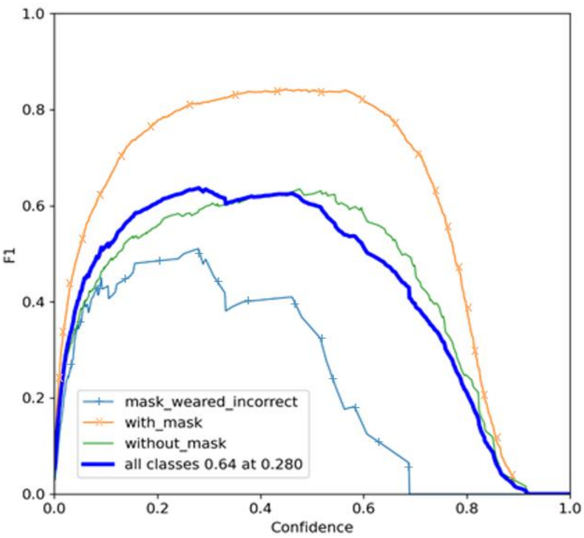

Figure 7. F1 score without the pre-trained model

Finally, the overall metrics are measured and plotted in Figure 8 for the case without the pre-trained model. Although the performance is slightly lower than the pre-trained model, the performance is still considerably acceptable, especially when the number of people being detected in the image is not so large. Therefore, YOLOv5 is suitable for facial mask wearing detection and counting in closed areas or public buildings that have some limitations on the number of people allowed to be in them, such as offices, schools, and religious places such as prayer halls or mosques, and so forth. So, the novelty highlighted in this paper is the ability of the proposed YOLOv5 model to detect face masks worn by people while counting the number of those wearing masks correctly or otherwise. With the relatively high detection accuracy achieved, this model is able to better calculate or estimate the number of people wearing masks in the image under consideration.

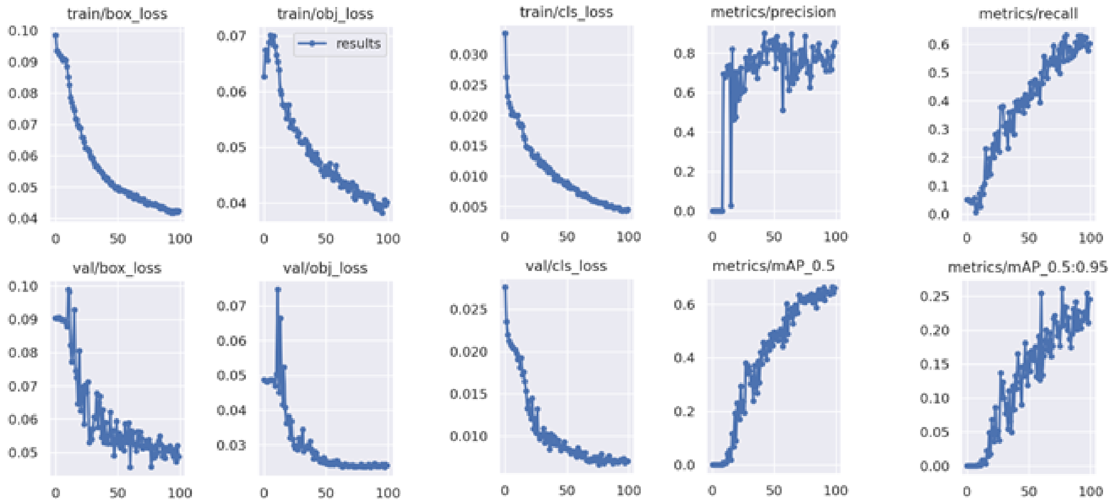

Figure 8. Metrics measurement over 100 epochs without using the pre-trained model, where the horizontal axis represents the number of epochs (The horizontal axis represents the number of epochs and the vertical axis represents the corresponding loss, precision or recall values accordingly)

4. CONCLUSION

Based on the results obtained and presented in this paper, it can be concluded that YOLO v5 is useful for detecting and counting people wearing facial masks. This is essential to control the spread of viruses, especially when the building or area to be entered by the people is closed and limited in space. By counting the number of people wearing the facial masks correctly, necessary and further actions can be taken to stop the people who are not wearing the masks from entering the building, besides ensuring that the number of people wearing the masks is within the maximum number of people allowed to be in the building.

ACKNOWLEDGEMENTS

The authors acknowledge and thank all the support from the Tabung Amanah Zakat Multimedia University for funding this research under the FRDGS fund, with the grant code of No. 11 (MMUE/210032). Not to forget the owner of Ultralytics Github (available at: https://github.com/ultralytics/yolov5) for YOLOv5 coding guidelines. Finally, we appreciate the support from our respective faculties, as well as all other individuals who are directly or indirectly involved in preparing this paper.

REFERENCES

[1] Q. Cheng, D. Gan, P. Fu, H. Huang, and Y. Zhou, “A novel ensemble architecture of residual attention-based deep metric learning for remote sensing image retrieval,” *Remote Sens.*, vol. 13, no. 17, 2021, doi: 10.3390/rs13173445.
[2] H. Bolhasani, M. Mohseni, and A. M. Rahmani, “Deep learning applications for IoT in health care: A systematic review,” *Informatics Med. Unlocked*, vol. 23, 2021, doi: 10.1016/j.imu.2021.100550.
[3] D. E. M. Nisar, R. Amin, N. U. H. Shah, M. A. A. Ghamdi, S. H. Almotiri, and M. Alruily, “Healthcare Techniques through Deep Learning: Issues, Challenges and Opportunities,” *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3095312.
[4] C. Bisogni, A. Castiglione, S. Hossain, F. Narducci, and S. Umer, “Impact of Deep Learning Approaches on Facial Expression Recognition in Healthcare Industries,” *IEEE Trans. Ind. Informatics*, vol. 18, no. 8, 2022, doi: 10.1109/TII.2022.3141400.
[5] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, “MagFace: A universal representation for face recognition and quality assessment,” 2021, doi: 10.1109/CVPR46437.2021.01400.

[6] P. Mishra and P. V. V. S. Srinivas, “Facial emotion recognition using deep convolutional neural network and smoothing, mixture filters applied during preprocessing stage,” *IAES Int. J. Artif. Intell.*, vol. 10, no. 4, 2021, doi: 10.11591/ijai.v10.i4.pp889-900.
[7] S. Yallamandaiah and N. Purnachand, “Convolutional neural network-based face recognition using non-subsampled shearlet transform and histogram of local feature descriptors,” *IAES Int. J. Artif. Intell.*, vol. 10, no. 4, 2021, doi: 10.11591/IJAI.V10.I4.PP1079-1090.
[8] M. A. Khan, M. Mittal, L. M. Goyal, and S. Roy, “A deep survey on supervised learning based human detection and activity classification methods,” *Multimed. Tools Appl.*, vol. 80, no. 18, 2021, doi: 10.1007/s11042-021-10811-5.
[9] S. Mekruksavanich and A. Jitpattanakul, “Biometric user identification based on human activity recognition using wearable sensors: An experiment using deep learning models,” *Electron.*, vol. 10, no. 3, 2021, doi: 10.3390/electronics10030308.
[10] M. Zahid, M. A. Khan, F. Azam, M. Sharif, S. Kadry, and J. R. Mohanty, “Pedestrian identification using motion-controlled deep neural network in real-time visual surveillance,” *Soft Comput.*, 2021, doi: 10.1007/s00500-021-05701-9.
[11] N. V. Kousik, Y. Natarajan, R. Arshath Raja, S. Kallam, R. Patan, and A. H. Gandomi, “Improved salient object detection using hybrid Convolution Recurrent Neural Network,” *Expert Syst. Appl.*, vol. 166, 2021, doi: 10.1016/j.eswa.2020.114064.
[12] S. Niu, Y. Liu, J. Wang, and H. Song, “A Decade Survey of Transfer Learning (2010–2020),” *IEEE Trans. Artif. Intell.*, vol. 1, no. 2, 2021, doi: 10.1109/tai.2021.3054609.
[13] H. A. Ghani, M. R. A. Malek, M. F. K. Azmi, M. J. Muril, and A. Azizan, “A review on sparse Fast Fourier Transform applications in image processing,” *International Journal of Electrical and Computer Engineering*, vol. 10, no. 2. 2020, doi: 10.11591/ijece.v10i2.pp1346-1351.
[14] D. Wang, J. G. Wang, and K. Xu, “Deep learning for object detection, classification and tracking in industry applications,” *Sensors*, vol. 21, no. 21. 2021, doi: 10.3390/s21217349.
[15] S. J. S and E. R. P, “LittleYOLO-SPP: A delicate real-time vehicle detection algorithm,” *Optik (Stuttg).*, vol. 225, 2021, doi: 10.1016/j.ijleo.2020.165818.
[16] Y. Chen, R. Qin, G. Zhang, and H. Albanwan, “Spatial temporal analysis of traffic patterns during the covid-19 epidemic by vehicle detection using planet remote-sensing satellite images,” *Remote Sens.*, vol. 13, no. 2, 2021, doi: 10.3390/rs13020208.
[17] J. Li, Z. Xu, L. Fu, X. Zhou, and H. Yu, “Domain adaptation from daytime to nighttime: A situation-sensitive vehicle detection and traffic flow parameter estimation framework,” *Transp. Res. Part C Emerg. Technol.*, vol. 124, 2021, doi: 10.1016/j.trc.2020.102946.
[18] J. Chelliah, M. Alagarsamy, K. Anbalagan, D. Thangaraju, E. S. Wesley, and K. Suriyan, “Automatic wireless health instructor for schools and colleges,” *Bull. Electr. Eng. Informatics*, vol. 11, no. 1, 2022, doi: 10.11591/eei.v11i1.3330.
[19] A. M. Alkababji and O. H. Mohammed, “Real time ear recognition using deep learning,” *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 19, no. 2, 2021, doi: 10.12928/TELKOMNIKA.v19i2.18322.
[20] J. Zhu, G. Zhang, S. Zhou, and K. Li, “Relation-aware Siamese region proposal network for visual object tracking,” *Multimed. Tools Appl.*, 2021, doi: 10.1007/s11042-021-10574-z.
[21] Y. Nagaoka, T. Miyazaki, Y. Sugaya, and S. Omachi, “Text detection using multi-stage region proposal network sensitive to text scale†,” *Sensors (Switzerland)*, vol. 21, no. 4, 2021, doi: 10.3390/s21041232.
[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-December, doi: 10.1109/CVPR.2016.91.
[23] D. H C, “An Overview of You Only Look Once: Unified, Real-Time Object Detection,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 8, no. 6, 2020, doi: 10.22214/ijraset.2020.6098.
[24] M. Karthikeyan and T. S. Subashini, “Automated object detection of mechanical fasteners using faster region based convolutional neural networks,” *Int. J. Electr. Comput. Eng.*, vol. 11, no. 6, 2021, doi: 10.11591/ijece.v11i6.pp5430-5437.
[25] A. Wong, M. Famuori, M. J. Shafiee, F. Li, B. Chwyl, and J. Chung, “YOLO Nano: A Highly Compact You only Look Once Convolutional Neural Network for Object Detection,” 2019, doi: 10.1109/EMC2-NIPS53020.2019.00013.
[26] D. T. Phan *et al.*, “A smart LED therapy device with an automatic facial acne vulgaris diagnosis based on deep learning and internet of things application,” *Comput. Biol. Med.*, vol. 136, 2021, doi: 10.1016/j.compbiomed.2021.104610.
[27] N. Rachburee and W. Punlumjeak, “An assistive model of obstacle detection based on deep learning: YOLOv3 for visually impaired people,” *Int. J. Electr. Comput. Eng.*, vol. 11, no. 4, 2021, doi: 10.11591/ijece.v11i4.pp3434-3442.
[28] S. Cass, *Hands on, NVIDIA MAKES IT EASY TO EMBED AI*. 2020.
[29] Z. M. Sani, H. A. Ghani, R. Besar, A. Azizan, and H. Abas, “Real-time video processing using contour numbers and angles for non-urban road marker classification,” *Int. J. Electr. Comput. Eng.*, vol. 8, no. 4, pp. 2540–2548, 2018, doi: 10.11591/ijece.v8i4.pp2540-2548.
[30] H. A. Ghani *et al.*, “Advances in lane marking detection algorithms for all-weather conditions,” *Int. J. Electr. Comput. Eng.*, vol. 11, no. 4, 2021, doi: 10.11591/ijece.v11i4.pp3365-3373.
[31] X. Tang and Z. Fu, “CPU-GPU Utilization Aware Energy-Efficient Scheduling Algorithm on Heterogeneous Computing Systems,” *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2982956.
[32] C.-K. Lai, C.-W. Yeh, C.-H. Tu, and S.-H. Hung, “Fast profiling framework and race detection for heterogeneous system,” *J. Syst. Archit.*, vol. 81, pp. 83–91, Nov. 2017, doi: 10.1016/j.sysarc.2017.10.010.

BIOGRAPHIES OF AUTHORS

Hatem Fahd Al-Selwi holds a degree from Universiti Teknikal Malaysia Melaka (UTeM). He then pursued his postgraduate study at Multimedia University (MMU) doing research in the fields of computer and machine vision. Apart from his good research track, he has various experiences in coding and apps development. His research interests and study also revolve, but not limited to, aroung wireless communications, cellular networks, 5G and beyond and so forth. Currently he works as a research assistant at MMU. He can be reached at 1201402635@student.mmu.edu.my.

Nawaid Hasan received the B.S. degree in Computer engineering from Sir Syed University of Engineering & Technology, Pakistan in 2001 and the M.S. degree in Telecommunication engineering from Hamdard University, Pakistan in 2010. Currently doing Ph.D. degree in engineering from Multimedia university, Melaka Malaysia. He has more than two-decade IT & telecom academic and industrial experience with various universities and multinationals companies. His research interests cover from wireless communication, design and analysis of algorithm, signal processing and deep learning for 5G and vehicular communication applications. His current research includes V2V asynchronous NLOS vehicle sensing in vehicular networks. He can be contacted at email: nawaidhasan@hotmail.com.

Hadhrami Bin Ab Ghani received a PhD degree in 2011 at Imperial College London (ICL). He also holds an MSc. Degree from ICL and another Meng Degree in Telecommunications Engineering, which was awarded by The University of Melbourne in 2004. His first degree was awarded by Multimedia University Malaysia in 2002. He has been serving as a lecturer since 2002 and involved in various research projects. His research interests are machine vision, computational intelligence, advanced communications and Internet of Things. He is currently attached to Universiti Malaysia Kelantan as a senior lecturer and can be contacted at email1: hadhrami.ag.@umk.edu.my.

Nur Asyiqin binti Amir Hamzah is a lecturer working at the Faculty of Engineering and Technology at Multimedia University, Melaka Campus, Malaysia following an electronic engineering degree (major in computer) in 1999-2003 at the same university. She then obtained a master degree in science engineering (telemedicine) in 2011, again from the same university. Her research focuses telecommunication engineering, signal processing, telemedicine, biomedical engineering and wavelet. Also includes cross-disciplinary research in science/engineering field related to Quran studies. She can be contacted at email: asyiqin.hamzah@mmu.edu.my.

Azlan Abd. Aziz is a TM staff and currently attached to the Faculty of Engineering and Technology, Multimedia University, Melaka. He has been in telecommunication industry for more than 15 years with a couple years in TM R&D working on next generation wireless networks. He can be contacted at email: n.abdaziz@mmu.edu.my.