Face Mask Detection using MediaPipe Facemesh

B. Thaman*, T.Cao* and N.Caporusso*
* Department of Computer Science, Northern Kentucky University, Highland Heights, United States
thamanb1@nku.edu, caot1@nku.edu, caporusson1@nku.edu

***Abstract* – Recently, face masks have received increasing attention due to the COVID-19 pandemic, as their correct use can reduce and prevent the spread of outbreaks. Thus, several research studies focused on developing new strategies for identifying if individuals are wearing a face mask before they can be admitted into public spaces, buildings, and transportation systems. In this paper, we present an alternative approach to face mask detection pipeline for automatically detecting whether an individual is equipped with a face mask. Our proposed solution utilizes MediaPipe, a popular image segmentation and object detection machine learning model designed especially for cross-platform operation, with specific regard to mobile devices. We present the architecture of our pipeline, detail its operation, and report the results of an evaluation study in which we analyzed the performance of our model in real-world scenarios.**

Keywords – machine learning; Mediapipe; Facemesh; health safety.

I. INTRODUCTION

Since the beginning of the COVID-19 pandemic, several health-safety recommendations have been introduced to limit contagion and prevent the spread of the SARS-CoV-2 virus and its variants. Although most restrictive measures (e.g., lockdown and shelter in place orders) have been lifted, social distancing, the use of face masks, and temperature monitoring are still mandated in several public spaces, and they often are a requirement to enter airports, schools, museums, and shopping centers. Although face masks have been among the most effective solutions to stop the spread of COVID-19 [1], studies investigated the causes of individuals' low compliance with them [2]. Consequently, several municipalities in multiple cities all over the world have started using surveillance cameras in public spaces and on transportation systems to evaluate if the population adheres to the prescribed health safety recommendations in terms of face mask use. In the last months, several research groups focused on developing systems for measuring individuals' adoption of health-safety requirements, with specific regard to personal protective equipment and social distancing [3]. To this end, machine learning (ML) and, specifically, neural networks, represent an invaluable resource. Most works describe the use and performance of transfer learning and convolutional neural networks in detecting whether individuals are wearing a face mask [3] [4] [5].

In this paper, we present a study in which we explored an alternative solution to existing systems. Specifically, our work leverages MediaPipe Facemesh (FM), a landmark detection model specifically designed for detecting multiple facial geometries in images and video. However, we utilize images indirectly: our processing pipeline integrates a lightweight algorithm that detects whether the user wears a face mask based on the dynamics of the facial topology estimated by FM. Specifically, the proposed system uses FM to accurately predict the facial landmarks in an image or live video feed and subsequently uses specific landmarks located around the mouth as an input for the detection algorithm, which is based on the fast Fourier transform (FFT). By doing this, our system does not require any additional training, which makes it particularly suitable for deployment in contexts that require computational efficiency.

II. RELATED WORK

In the last decades, research in image processing has made significant progress thanks to the use of machine learning frameworks, which have enabled researchers to optimize the performance of algorithms focusing on image segmentation and object identification tasks. As a result, nowadays, face detection is a common use of machine learning models, and it has applications in several different scenarios such as biometric identification [6]. Nevertheless, the COVID-19 pandemic has given new impulse to realize new research in the context of machine learning with the objective of detecting whether users are wearing a mask.

Most published works focused on identifying unmasked subjects in video surveillance contexts. To this end, the authors of [7] created a dataset of masked faces and then trained a convolutional neural network model that achieved 76% accuracy in detecting faces covered by a mask. In [8], the Viola and Jones face detector is utilized to determine if medical professionals are wearing masks in an operating room: the system achieved above 95% accuracy and supported real-time image processing up to 20 frames per second within a distance of the subject of up to five meters from the camera. In a different study, researchers used ResNet50 for feature extraction and then tested different types of classifiers against four different image datasets. The accuracy varied based upon the specific classifier and dataset used, but almost all achieved above 90% [9]. Furthermore, multi-stage face mask detectors were utilized for use in surveillance systems. The authors of [10] compared three different face detection models, as well as three different mask classifiers. They found that RetinaFace and NasNetMobile held the best results for face and mask classification, respectively. Despite a high testing accuracy, the system was only able to run at five frames per second, which is

not suitable for many applications that require faster speed. Some of the more promising methods proposed make use of OpenCV for face detection and an object detection model to detect the masks. In [11], the authors utilized OpenCV's single shot multi-box detector for face detection and the pre-trained MobileNetV2 model for mask detection. They obtained an average accuracy of 93% and were able to achieve a processing speed of 15.71 frames per second. The work of [12] also uses OpenCV for face detection and MobileNetV2 for mask detection, achieving an even better accuracy of 99.87%.

Other studies evaluated OpenFace, a face landmark detection system based on OpenCV, and tested it against three different object detection models: ResNet50, AlexNet, and MobileNet [13]. The highest accuracy (98.92%) was obtained with ResNet50, which produced excellent results. In addition, the model proposed by [13] works with live video, though the system was tested using mobile phones only. Indeed, although mask detection systems have been increasingly studied in the last months, the optimal accuracy trade-off has not been achieved yet.

III. System Design

The objective of our work was to evaluate a different approach to mask detection that could potentially increase efficiency. While other systems directly analyze image data, our aim is to indirectly detect whether a user is wearing a face mask based on the quantitative characteristics of their facial topology. To this end, we designed a mask detection pipeline organized into three components, as described in Fig. 1:

- the video acquisition and processing system, which captures and analyzes video streams
- a landmark detection system that identifies the topology of the user's face
- the mask detection system evaluates whether the user is wearing a face mask.

The video acquisition system consists of any type of camera compatible with a PC (e.g., webcam). Additionally, it supports external video feeds (including streaming) having media formats compatible with the web (i.e., JPEG, PNG, or MP4).

As per the landmark detection system, our solution leverages Tensorflow's MediaPipe FaceMesh model [14], which is an open-source machine learning tool developed by Google for face landmark estimation. MediaPipe is a cross-platform library consisting of customizable ML solutions for processing images and real-time video. MediaPipe includes several ML models, each addressing a specific image segmentation or object detection problem (i.e., motion tracking, hair segmentation, posture detection, hand and finger recognition).

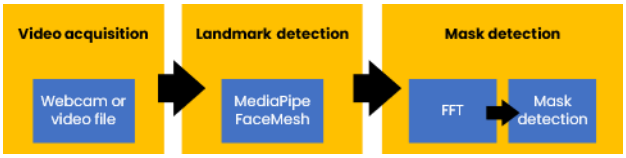

Figure 1. The ML and processing pipeline employed in our face mask detection solution.

FaceMesh [14] is a face landmark ML model developed via transfer learning and specifically designed to recognize the user's facial topology in three dimensions. The machine learning architecture of FM is built on top of another model, that is, BlazeFace [15], which has the purpose of finding a face in the image or video frame and estimating its bounding box. BlazeFace is an improvement on the MobileNetV2 model [16] in both speed and accuracy, and it supports multiple faces. After the bounding box surrounding a face has been identified by BF, FM uses it to estimate the three-dimensional coordinates of 468 landmarks that describe the facial topology (see Fig. 2). The landmarks identified by MP are organized into 32 groups, each describing a facial component (e.g., silhouette, left cheek, and nose tip).

The mask detection component of our system is designed to evaluate whether users are wearing a face mask and if they are positioning it correctly over their mouth and nose. To this end, our system does not analyze the image directly. On the contrary, it leverages the interference of face masks with the landmarks detected by MP, which results in a noisier signal. Specifically, the mask detection component of our solution has the purpose of evaluating the changes in the landmarks of interest estimated by MP when the user wears a mask and covers key facial features in contrast to when the user is unmasked. In addition, the mask detection component of our system takes into consideration a number of landmarks in other regions that are not interested by face masks (e.g., silhouette and eyebrows), which are utilized as a control. Therefore, we utilize the landmarks of eight facial regions of interest, that is, lipsUpperOuter (i.e., 61, 185, 40, 39, 37, 0, 267, 269, 270, 409, and 291), lipsLowerOuter (i.e., 146, 91, 181, 84, 17, 314, 405, 321, and 375), lipsUpperInner (i.e., 78, 191, 80, 81, 82, 13, 312, 311, 310, 415, and 308), lipsUpperInner1 (i.e., 76, 184, 74, 73, 72, 11, 302, 303, 304, 408, and 306), lipsUpperInner2 (i.e., 183, 42, 41, 38, 12, 268, 271, 272, and 407), lipsLowerInner (i.e., 95, 88, 178, 87, 14, 317, 402, 318, and 324), lipsLowerInner1 (i.e., 62, 77, 90, 180, 85, 16, 315, 404, 320, 307, and 292), and lipsLowerInner2 (i.e., 96, 89, 179, 86, 15, 316, 403, 319, and 325), noseBottom (i.e., landmark 2), noseTip (i.e., landmark 1),

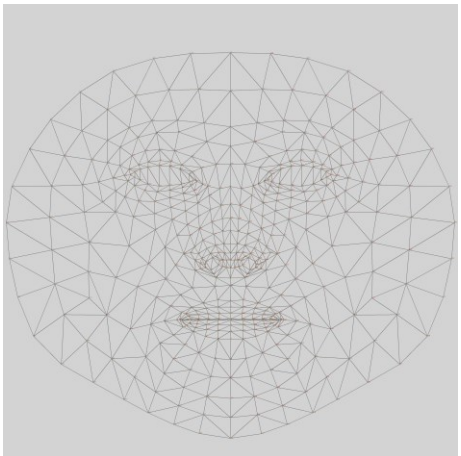

Figure 2. The model employed by MediaPipe Facemesh, which represents the facial topology with 468 landmarks.

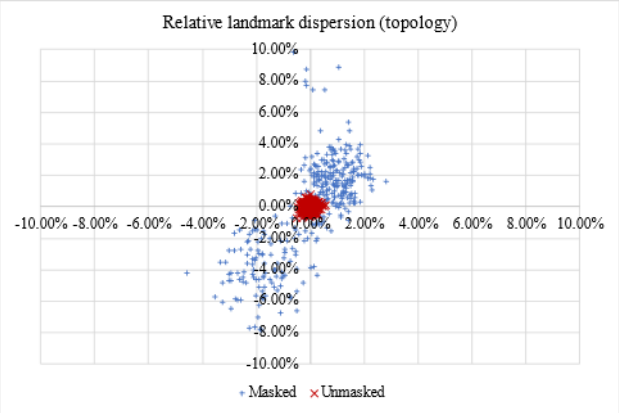

Figure 3. Dispersion of key landmark locations predicted by MediaPipe FaceMesh when the user is unmasked (red) and masked (blue). The chart represents a total of 720 samples (i.e., 360 for each subject condition).

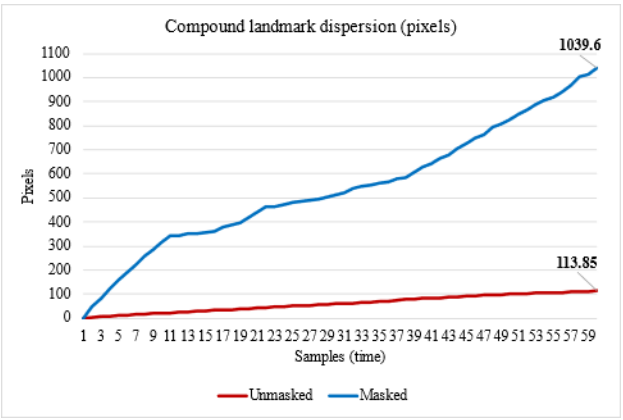

Figure 4. Compound dispersion (60 samples, i.e., approximately two seconds) of key landmark locations predicted by MediaPipe FaceMesh when the user is unmasked and masked.

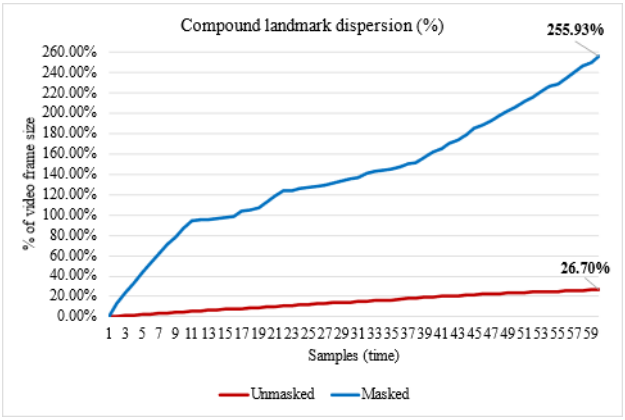

Figure 5. Compound dispersion (60 samples, i.e., approximately two seconds) calculated as a percentage of the size of the video frame.

noseLeftCorner (i.e., landmark 327), and noseRightCorner (i.e., landmark 98).

Specifically, we calculate the pixel dispersion (i.e., variance) between the estimated coordinates of a landmark over a period of time and their average. However, calculating the dispersion on key landmarks during a defined time window could be suitable in circumstances in which the subject does not move and hold still in front of the camera for the duration of the acquisition. However, it

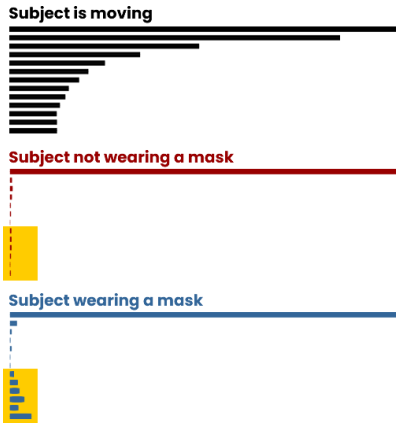

Figure 6. Frequency profile of the subject obtained by the face mask detection component of our system using FFT.

is not a reliable method for determining whether the user is wearing a mask or not in the context of video feed and situations in which the subject can move. Indeed, spontaneous movements or posture changes can cause the predicted landmarks to shift widely, resulting in higher dispersion. Therefore, our mask detection algorithm utilizes FFT to identify the main frequency components of the landmarks estimated by FM and use them to estimate whether the dispersion in the signal is caused by spontaneous movements or by the presence of a mask covering key facial features. By encoding the position of the landmarks into the frequency domain, we obtain a frequency profile of the subject that describes whether the dispersion of the landmarks is caused by spontaneous movement or by visual obstacles that prevent FM from correctly estimating key facial landmarks. The three profiles are described in Fig. 6, which highlights the frequency bands that differentiate users who are wearing a face mask from users who are not masked. Subsequently, our face mask detection subsystem implements a high-pass filter that classifies low-frequency signals (usually produced changes of the posture of the head and facial expressions) as spontaneous movements and isolates high-frequency signals changes, which, as detailed below, indicate that FM is not able to correctly estimate the location of the landmarks because the user is wearing a face mask. The filter can be customized to adjust the threshold depending on specific circumstances such as the frame rate, the distance of the subject from the camera, and the signal-to-noise ratio.

In designing our model, we analyzed the individual landmarks estimated by FM in circumstances in which the facial features are visible and in situations in which they are covered in order to evaluate their differences and identify a viable mask detection strategy. In our preliminary work, we collected data from three subjects utilizing a webcam with an HD resolution (i.e., 1280×720 pixels). The acquisition protocol consisted of 30 sessions consisting of two trials, each lasting 30 seconds. During the first trial, the subject was unmasked, whereas in the second, they had a face mask on. Subjects were located in a standard position that enabled the webcam to capture their entire face.

At this stage, our acquisition protocol did not take into consideration different mask types. Fig. 3 represents a

comparison of the dispersion in the estimated coordinates of a landmark of interest located on the subject's lips when the facial features are visible and when they are hidden. The actual coordinates have been repositioned with respect to the target position of the landmark, and the dispersion has been normalized with respect to the width and height of the video frame. As shown in the image, in conditions in which the user is not wearing a face mask, the predicted landmarks are consistent with the position of the actual landmark and show very little dispersion (i.e., 0.27%±0.15%, on average) over a total of 360 measurements (i.e., approximately 12 seconds, though the actual acquisition time depends on the frame rate). On the contrary, when the facial features are covered (i.e., the user is wearing a face mask), the estimated landmarks show higher dispersion (i.e., 2.96%±2.02%, on average), and the estimated prediction in the subregion of the facial topology is less accurate, resulting in a more dispersed set of data points. As shown in Fig. 4, the normalized compound dispersion over a 60-sample interval (i.e., approximately two seconds of recording, though the actual acquisition time depends on the frame rate) clearly marks the difference when the subject wears a face mask, which results in a higher dispersion in the predicted landmarks. The figure shows that the dispersion is consistent over the entire duration of the sampling interval. Similarly, the normalized compound dispersion (see Fig. 5) enables appreciating the difference of one order of magnitude over a two-second measurement.

IV. Performance Analysis

First, we evaluated the performance of our system in terms of speed. To this end, in our study, we utilized a webcam to record video at 30 frames per second (FPS). Then, we evaluated the performance of the next stage, that is, the landmark detection component. All the models in the MediaPipe framework are optimized for use in mobile applications. As a result, they are extremely fast. However, lighting impacts their performance considerably in terms of FPS. Therefore, we tested the landmark detection component of our system in five different lighting conditions (i.e., very good, good, medium, poor, and very poor). As shown in Fig. 7, the FM performs very well when there is enough environmental light. Specifically, the average FPS is 28.40±1.16 in optimal lighting conditions, 26.20±1.76 in good lighting conditions, and 23.58±2.19 in medium lighting conditions. Unfortunately, performances degrade significantly in situations of poor lighting (12.18±2.65 FPS) or darkness (6.14±1.43 FPS), which, in turn, affects the accuracy of the mask detection component and makes our system inapplicable when environmental lighting is insufficient. Nevertheless, this factor is strictly related to the performance of the landmark detection model and can be addressed by making sure that there is appropriate lighting in the area in which the system is utilized.

In addition to environmental lighting, several other factors can influence the speed of FM and result in lower FPS. They include the hardware characteristics of the computer processing the data, the available CPU, the presence of GPUs, and the resolution of the camera. Therefore, we analyzed the performance of the mask

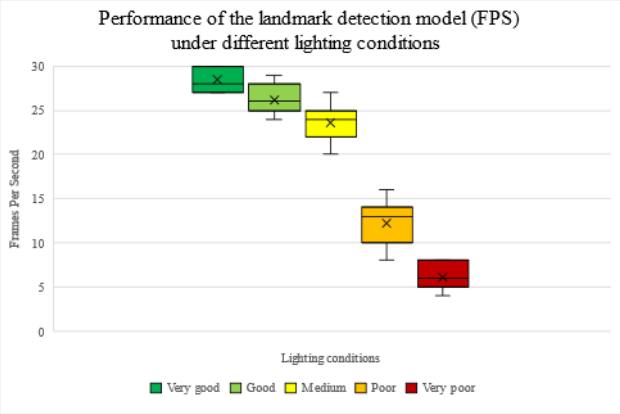

Figure 7. Performance of the landmark detection model (i.e., FM) in environments with optimal, good, medium, poor, and very poor lighting.

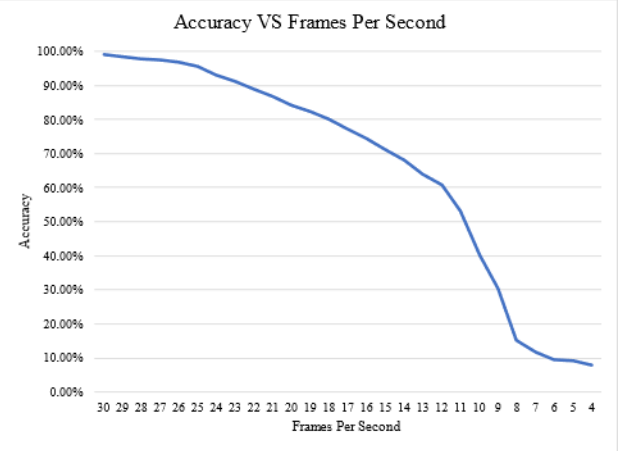

Figure 8. Accuracy of the face mask detection system in different FPS rates.

detection component in different circumstances as it relates to FPS. To this end, we created a post-processing routine that reduced the frame rate after FM estimated the facial topology. Fig. 8 reports the accuracy of the system with respect to different FPS configurations: as represented in the chart, the accuracy of the face mask detection component is heavily influenced by the frame rate. In particular, it is above 90% when FM produces an output having at least 23 FPS. Also, it is greater than 70% when the landmark detection system has an output of at least 15 FPS. Unfortunately, it drops significantly when the frame rate is less than 11 FPS. Also, this performance aspect is related to the characteristics of MediaPipe and can be addressed by ensuring that the computer executing the system has enough resources or by outsourcing the landmark prediction task to an external system with sufficient computational resources.

Subsequently, we focused on the accuracy of the mask detection subcomponent. To this end, we realized a total of 100 trials (i.e., 50 masked and 50 unmasked) with a single subject in optimal lighting conditions. We utilized one participant only because our algorithm is independent of the specific facial topology of the individual. The subject was positioned in front of the camera (an HD webcam) at a distance of approximately 40 centimeters from the lens. The subject was asked to act naturally in front of the camera and was allowed to move their head

and change posture as needed. The results are described by the confusion matrix shown in Table 1.

When tested in optimal circumstances, the mask detection component of our system was able to correctly identify the presence of a mask with a precision of 1, a recall of 0.96, and an F1 score of 0.98. The average time required by the system to detect whether a subject is wearing a face mask is approximately one second. This is mostly due to the time required by MediaPipe to process enough frames to accurately represent the facial topology. During this time, the subject should remain still in front of the camera. Although the system can detect if the subject is moving, avoiding any changes in face posture enables to prevent further delay. The processing time of the face detection component is less than ten milliseconds, on average.

V. CONCLUSION

Face mask detection has become an increasingly popular research topic since the beginning of the COVID-19 pandemic, which has originated the need for solutions that are able to detect whether individuals are complying with health safety measures. In this paper, we have presented a face mask detection solution that aims at achieving optimal efficiency in the context of real-time video. In contrast to other systems based on image processing, our solution leverages the output of a facial geometry pipeline, which consists in three-dimensional coordinates. The novelty of our approach consists in leveraging face landmark detection combined with an FFT algorithm, which dramatically reduces processing times. Furthermore, our approach does not rely on the specific facial topology of the subject, which makes it particularly flexible for multiple application scenarios.

Indeed, the present work results in a more efficient detection algorithm, though our approach has several limitations related to the indirect measurement used to detect if the user is wearing a mask. One main drawback of our solution is that it works with video feed only due to the nature of FFT. As a result, our method is not suitable for processing still images and frames. In addition, our approach is slower than models working on images because our system must acquire a number of frames before it can produce any result. Furthermore, our algorithm is not suitable for detecting the type of mask worn by the user, though further image processing layers can could be added to our pipeline to evaluate the quality and appropriateness of the face mask. A more effective approach incorporating MP could consist in training an image segmentation and object detection model designed to recognize face masks. This approach, which we will explore in a follow-up paper, might identify the type of mask worn by individuals. Additionally, it would remove two steps from the pipeline, which might result in significant performance improvements and in the possibility of using the mask detection system on low-power devices (e.g., Raspberry Pi).

TABLE I. CONFUSION MATRIX OF THE FACE MASK DETECTION COMPONENT

		Predicted	
		Masked	**Unmasked**
Actual	**Masked**	50	0
	Unmasked	2	48

REFERENCES

[1] J. Howard κ.ά., 'An evidence review of face masks against COVID-19', Proceedings of the National Academy of Sciences, τ. 118, τχ. 4, 2021.

[2] R. V. Tso και B. J. Cowling, 'Importance of face masks for COVID-19: A call for effective public education', Clinical Infectious Diseases, τ. 71, τχ. 16, σσ. 2195–2198, 2020.

[3] S. Yadav, 'Deep learning based safe social distancing and face mask detection in public areas for covid-19 safety guidelines adherence', International Journal for Research in Applied Science and Engineering Technology, τ. 8, τχ. 7, σσ. 1368–1375, 2020.

[4] G. Jignesh Chowdary, N. S. Punn, S. K. Sonbhadra, και S. Agarwal, 'Face mask detection using transfer learning of inceptionv3', στο International Conference on Big Data Analytics, 2020, σσ. 81–90.

[5] K. Suresh, M. B. Palangappa, και S. Bhuvan, 'Face Mask Detection by using Optimistic Convolutional Neural Network', στο 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, σσ. 1084–1089.

[6] R. Raghavendra και C. Busch, 'Novel presentation attack detection algorithm for face recognition system: Application to 3d face mask attack', στο 2014 IEEE International Conference on Image Processing (ICIP), 2014, σσ. 323–327.

[7] S. Ge, J. Li, Q. Ye, και Z. Luo, 'Detecting masked faces in the wild with lle-cnns', στο Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, σσ. 2682–2690.

[8] A. Nieto-Rodriguez, M. Mucientes, και V. M. Brea, 'System for medical mask detection in the operating room through facial attributes', στο Iberian Conference on Pattern Recognition and Image Analysis, 2015, σσ. 138–145.

[9] M. Loey, G. Manogaran, M. H. N. Taha, και 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, τ. 167, σ. 108288, 2021.

[10] S. K. Addagarla, G. K. Chakravarthi, και P. Anitha, 'Real time multi-scale facial mask detection and classification using deep transfer learning techniques', International Journal, τ. 9, τχ. 4, σσ. 4402–4408, 2020.

[11] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, και J. Hemanth, 'SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2', Sustainable cities and society, τ. 66, σ. 102692, 2021.

[12] S. Asif, Y. Wenhui, Y. Tao, S. Jinhai, και K. Amjad, 'Real time face mask detection system using transfer learning with machine learning method in the era of COVID-19 pandemic', στο 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), 2021, σσ. 70–75.

[13] S. Sethi, M. Kathuria, και T. Kaushik, 'Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread', Journal of biomedical informatics, τ. 120, σ. 103848, 2021.

[14] Y. Kartynnik, A. Ablavatski, I. Grishchenko, και M. Grundmann, 'Real-time facial surface geometry from monocular video on mobile GPUs', arXiv preprint arXiv:1907. 06724, 2019.

[15] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, και M. Grundmann, 'Blazeface: Sub-millisecond neural face detection on mobile gpus', arXiv preprint arXiv:1907. 05047, 2019.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, και L.-C. Chen, 'Mobilenetv2: Inverted residuals and linear bottlenecks', στο Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, σσ. 4510–4520.