Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 20, Iss. 1, pp. 62-73, 2022

A Scalable Analytics Pipeline for COVID-19 Face Mask Surveillance

Clayton Kossoski
Federal University of Techonology - Paraná
claytonutf@gmail.com

Gustavo Schaefer
Federal University of Techonology - Paraná
gstvo.schaefer@gmail.com

Gianlucca Fiori Oliveira
Federal University of Techonology - Paraná
gianlucca.oliveira123@gmail.com

Heitor Silvério Lopes
Federal University of Techonology - Paraná
hslopes@utfpr.edu.br

Abstract – The COVID-19 coronavirus pandemic still causes a global health crisis. An effective protection method is using a face mask in public areas, according to the World Health Organization (WHO). Computer vision systems can be allies in monitoring public areas where the face mask is mandatory. However, face mask detection is challenging due to many factors, including diversity of people, facial features, head accessories, mask design, image position, and lighting changes. To tackle these issues, we present the following contributions: a new balanced face mask dataset named UTFPR-FMD1, consisting of 61,430 images splitted into "face" and "mask" classes; a transfer learning classification model for computer vision tasks, trained with our dataset; a new processing pipeline that allows face mask detection in video streams. Unlike available public datasets with imbalanced class distributions, the UTFPR-FMD1 contains images from different people, gender, and ages to minimize the training difficulty of deep learning models. We introduced a new measure to select valid images to perform inferences. Experimental results show the effectiveness of our model, outperforming the state-of-art methods for face mask detection tasks. Additionally, and different from other authors, we also present qualitative results. The system can detect heads with up to 60 degrees of rotation and process up to 10 FPS. In future work, we will deploy the current framework into production, perform tests in a near real-time environment, and extend it to process multiple video streams.
Keywords – facemask dataset, facemask detection model, distributed processing framework, data augmentation, transfer learning.

1 Introduction

Coronaviruses are a family of viruses that can cause respiratory illness. There is currently an outbreak of coronaviruses called COVID-19[1]. Health authorities suggested social distancing, frequent handwashing with alcohol, avoiding touching the face, and using masks in several countries worldwide as part of non-pharmaceutical actions to control the pandemic [1].

The pace of vaccination is very different across the world. Many countries in Africa and Eastern Europe had low or unknown vaccination coverage [2]. Since the beginning of the pandemic, scientists have warned about future risks [3]. In fact, some variants discovered are more infectious and with higher transmissibility [4]. Still, in 2022, many countries in the world have witnessed a rapid increase of new infections of COVID-19, suggesting that the fourth wave of infections is approaching [5]. In the first months of 2022, many cities and countries released citizens from the putative use of face masks. However, the new wave of infections will soon change this landscape from optional to obligatory. In the worst case, besides face masks for all the population, several cities are in lockdown [6]. That is why experts strongly recommend that face masks shall be used by all the people, in addition to social distancing requirements, in areas with active transmission of COVID-19 [7].

Although the efficiency of face masks for preventing the transmission of the coronavirus was controversial at the beginning of the pandemic [8], the World Health Organization (WHO) suggested preventing contamination from droplets of saliva or nasal discharge [9]. Moreover, face masks (especially the medical ones) proved effective for preventing virus transmission [10, 11]. However, during the last two years of the pandemic, despite the significant efforts of the health authorities worldwide, face masks were not mandatory everywhere. Differently from high-income countries (HICs), where the living conditions allow the population to be better protected from COVID-19 infections, in low-and-middle-income countries (LMICs) face masks are still,

[1] https://www.who.int/health-topics/coronavirus

an accessible (and, sometimes, the only) protective way for a large part of the population [12, 13]. This is particularly true for poor countries/regions in Africa, Asia, and Latin America. Furthermore, as mentioned before, a new wave of infections corroborates former previsions about the recurrence of COVID-19 [14], thus stressing the need to extend vaccination coverage and take precautions such as wearing masks.

Since public cameras are everywhere, they could support the surveillance of face masks and help the worldwide effort to avoid the propagation of the COVID-19 virus. From the computational point of view, face mask detection can be very challenging when the data comes from video streams in the wild. This is the case with most surveillance cameras in public streets, boulevards, or mall aisles. Video streams capture people walking from any direction and under non-ideal pose and lighting conditions.

Deep Neural Networks (DNN) are state-of-art for computer vision detection with higher accuracy results for detecting people, face, eyes, mouth, facial expressions, objects, anomalies, and traffic monitoring. However, most popular DNN face detection models were not designed to detect face masks [15–18], but the source code, models, datasets, and weights are available. Thus, the existing models can be adapted by other researchers to improve results.

The most popular state-of-the-art DNN-based approaches used for face detection, such as OpenCV[2], OpenFace [15], ResNet [19], MTCNN [20], FaceNet [21], and Dlib[3] are not suitable for predicting whether a person is wearing a face mask. In addition, they may present noise and error under real-world uncontrolled imaging conditions.

The current literature on computer vision-based face mask detection has several limitations, lack of population diversity in datasets, lack of headwear and diverse lighting conditions, unbalanced datasets, poor quality models, monolithic processing architecture, and lack of tests on video streams. They limit or prevent their applicability in the real world. All of these issues are addressed in our proposal.

To address these relevant issues, the contributions of this article can be summarized as follows:

- A new dataset containing human faces of different people and ages, with/without accessories (hat, glasses, beard), under other lighting conditions and pose angles. This dataset was created specifically for use in countries/regions with high population diversity, such as Brazil. The dataset, built from repositories of free shared images, will be made public to help researchers and practitioners improve results.
- A face mask detection model is robust to different human characteristics and imaging conditions.
- A framework optimized to process still images and video streams in a real-world surveillance context.
- A comparative study with different datasets from related work and a performance test on videos.

The remaining paper contains the following structure. Section 2 presents the related work and comment its limitations. Section 3 proposes and describes the proposed work comprehending a new curated face mask dataset, a new model for detection of face/masks trained with our dataset, and a scalable face mask detection pipeline. Section 4 presents the experimental and discusses the results. Section 5 concludes the paper and presents future work.

2 Related work

Transfer learning is a popular machine learning technique that reuses a trained model for one task as the starting point for a different but related task. As mentioned above, it uses weights from a good quality network and fine-tunes the model by retraining it with another specialized dataset. For face mask detection, many works found in the literature used transfer learning across the development process [22–29].

Due to relevance of this theme, there are several datasets of masked and unmasked faces appeared [26, 30–34, 34–36]. The main features of these datasets are compared in Table 1. Data selection for training detection models is a very critical activity. However, few of the datasets mentioned above have enough quality to train robust models, considering the diversity of people worldwide. Some of these datasets were conceived for specific populations (asian, for instance). There are several ways to improve the quality of an image dataset, such as class balancing and increasing diversity.

In real life, it is not always possible to have new original images from a specific context. The recent literature suggested data augmentation to increase diversity when the original data is scarce in variety [37–39]. In the face mask context, few authors used data augmentation to increase the number of images or balance classes [27, 28]. Many works merged two or more datasets to build one for training and testing phases [23–26]. About the diversity of dataset images, black people are absent in almost all of them. Existing datasets do not take variety into account. Unfortunately, it reduces the robustness and generality of models trained with these datasets.

Another relevant issue is the image quality. A common practice is using unbalanced datasets to train models and measure the results [22, 24, 26, 28]. Few works have tested their models with images in the wild, as in a surveillance context [31, 36]. In addition, many works have merged different datasets or used datasets that are not publicly available for further comparison. [22, 23, 28]. Other published works have not tested their models with datasets from different authors [24, 27, 36].

[2]See OpenCV and deep learning module in `https://docs.opencv.org/4.5.0/d2/d58/tutorial_table_of_content_dnn.html`
[3]See Dlib C++ library in `http://dlib.net`

Table 1: Comparison of datasets for facemask detection

Dataset	# images	Class balance	Data augmented	Fake masks	Comments
AIZOO [30]	7,959	N/A	No	No	Front and side faces are not in the same amount. Many images have watermarks.
CMFD [26]	67,049	Masked only	Yes	Yes	The mask class contains one type of blue mask.
DEUBa [31]	2,253	44% faces, 56% masked faces	No	No	Based on 11 videos from Youtube from different enviroments.
FMD [32]	853	Masked only	No	No	Poor quality, some repeated images.
IMFD [26]	66,734	Fake mask only	Yes	Yes	The mask class contains one type of blue mask.
MAFA [33]	35,811	Masked only	No	No	Detailed description of facial attributes and face positions.
MFDD [34]	24,771	Masked only	No	No	Dataset unavailable.
MMD	3835	N/A	No	No	Dataset unavailable.
SMFD [35]	1,376	Good for class, bad for intraclass	Yes	Yes	Lack of documentation, diversity in facial features, and mask types.
RMFRD [34]	95,000	Very poor	No	No	Lack of diversity of people. Some images are flattened.
SMFRD [34]	500,000	Fake mask only	Yes	Yes	Dataset unavailable.
SSDMNV2 [36]	11,042	Balanced across classes	Yes	Yes	Lack of diversity of people. Some images are flattened.

3 Methods

First, we created a new face mask dataset following best practices such as class balancing and data variety, which make it robust in different contexts. In the second step, we developed a DNN model suitable for still images and surveillance videos. Then, we train the network with the proposed dataset. Figure 1 presents an overview of the entire workflow.

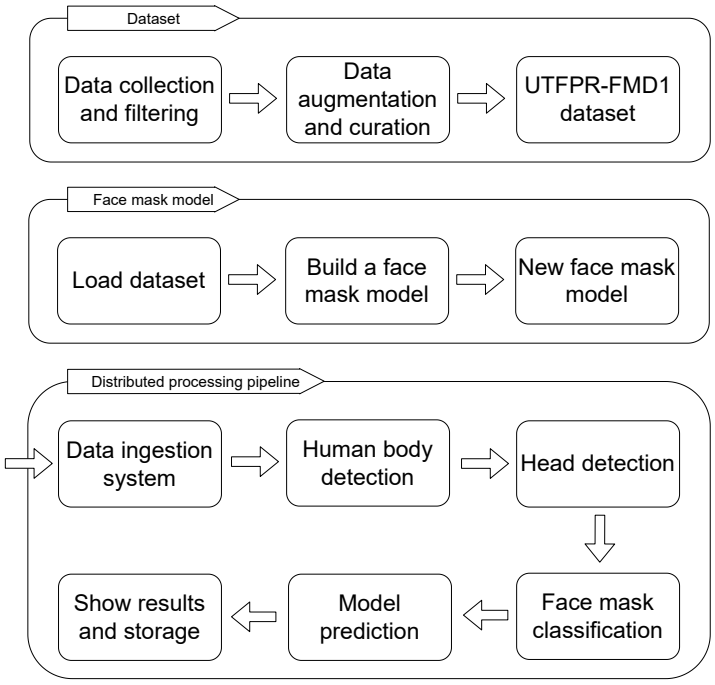

Figure 1: Overview of the entire workflow

3.1 A new dataset for COVID-19 face mask detection

Motivated by the limitations found in the literature, we propose a new dataset, namely UTFPR-Face Mask Dataset 1 (UTFPR-FMD1), designed to achieve, as deeply as possible, high quality in the following characteristics: class balancing, diversity of people, variety of masks (shape, sizes, colors, real/fake), variety of accessories (glasses, beard), different lighting conditions, and poses.

3.1.1 Data collection and filtering

To build this dataset, we used Selenium WebDriver[4] and BeautifulSoap[5] webcrawlers to get images from free shared image sites such as Pexels[6], Unsplash[7], and GettyImages[8] using search queries (asian, black, white, woman, man, kid) and logical operators (and, or) in many combinations. Therefore, the image acquisition process can be reproducible.

After a manual revision, many images were discarded for several reasons: not a person, poor quality, or tiny image. Finally, we used VisiPics[9] to remove similar or duplicated images. After a data augmentation procedure, we organized the remaining files to get the best-as-possible balance of classes. The only criteria for selecting human images as white, black, asian, male, female, and child were the results of the search strings in the image repositories. For example: "white child face", "black man mask", "asian woman", among others. The final dataset contains 61,430 images of human faces with and without a face mask. More detailed information about it is presented in Table 2. To foster further research in this area, the dataset will be publicly available at Github [10].

- All the images were checked for containing a human head/face using the OpenFace detector.
- After the head detection task, the original images were cropped and rescaled to a fixed size (300×300 pixels). Most images were color, but a few images were grayscale.
- The classes and subclasses of the images were standardized by visual inspection. We organized into black, white, and asian; male, female; and two age ranges (adults or kids).
- The final dataset contains 30,715 face images without mask, in which 9,027 are original images and the remaining modified by data augmentation (see Section 3.1.2). Similarly, the dataset contains 30,715 target face images with mask, comprising 3,899 original images and the remaining modified by data augmentation. The criterion of the final class balancing considers random sort with new images produced from data augmentation until it reaches 30,715 images for each class.

Table 2: The distribution classes and sub-classes of the UTFPR-FMD1 dataset for both the original and data augmented images

		Original images		Augmented images	
People	**Gender/Age**	**Face**	**Mask**	**Face**	**Mask**
White	Man	1,758	516	5,033	4,128
	Woman	2,487	641	6,554	4,851
	Kid	1,000	581	4,000	4,448
Black	Man	800	406	3,200	3,248
	Woman	1,044	310	4,176	2,480
	Kid	535	347	2,140	2,776
Asian	Man	457	373	1,828	2,984
	Woman	388	323	1,552	2,584
	Kid	558	402	2,232	3,216
		9,027	3,899	30,715	30,715

3.1.2 Data augmentation

Preliminary experiments showed that the number of high-quality images collected by web scrapping was insufficient to train a deep and robust model for face/mask classification. Therefore, we created new (transformed) images from the original ones using a data augmentation procedure. This process was accomplished by adding different types of noise and applying geometrical

[4] https://www.selenium.dev/
[5] https://www.crummy.com/software/BeautifulSoup/
[6] https://www.pexels.com/
[7] https://unsplash.com/
[8] https://www.gettyimages.com.br/
[9] https://www.fosshub.com/VisiPics.html
[10] https://github.com/bioinfolabic/UTFPR-FMD1

transformations or complex changes (in this case, adding a mask to a face) [40, 41]. The Python image augmentation library (imaug)[11] was used to increase the amount of image per each class. To be effective, the transformations over the original image must be coherent with the context in which the images are used. We considered meaningful transformations reflecting common features of real-world images from surveillance cameras, such as left to right flip, $\pm 45^o$ rotation, random brightness, Gaussian blur, and salt & pepper noise.

Besides improving the robustness of the classifier [42], the data augmentation process also helps to provide a better class/sub-class balance of the dataset, as shown in Table 3. Figure 2 shows images from UTFPR-FMD1 dataset.

Table 3: Final class distribution after data augmentation of the UTFPR-FMD1 dataset.

	Proportion (%)								
	White			**Black**			**Asian**		
	man	**woman**	**kid**	**man**	**woman**	**kid**	**man**	**woman**	**kid**
Face	16.4	21.3	13.0	10.4	13.6	7.0	6.0	5.0	7.3
Mask	13.4	15.8	14.5	10.6	8.1	9.0	9.7	8.4	10.5

Figure 2: Samples of the diversity of UTFPR-FMD1 dataset.

3.2 The Scalable Face Mask Detection Pipeline

The proposed Scalable Face Mask Detection Pipeline (SFMDP) for processing video streams at scale. The SFMDP is depicted in Figure 3, and it is composed of two main modules:

1. **Data ingestion system**: process video streams in real-time or near-real-time, by using Apache Kafka[12] as the stream framework. The incoming videos are producers that publish frames on a Kafka topic which, in turn, serves a group of consumers. A consumer group is a set of processes cooperating to consume data from some topic.

2. **Face mask detection pipeline**: contains three decoupled modules that consume frames from a Kafka topic and perform operations for (a) Human body detection, (b) Head detection, and (c) Face mask classification. The main components of these blocks are DNN architectures as follows: Yolo-v4, ResNet, and our network model (FaceMask-v1), based on transfer learning of MobileNetv2, trained on the UTFPR-FMD1 dataset. The MongoDB database stores the relevant processed frames so the end-user can further access the stored data through structured queries.

The SFMDP is run on two servers. One server for data ingestion with Kafka, in a workstation with Intel Xeon E5-2450 processor, 48 GBytes RAM, and 9 TB of storage. Another server for facemask detection task, in a workstation with an Intel Core i7-5820K processor, 32 GBytes RAM, and Nvidia GeForce RTX-2060 GPU with CUDA 10.1. Tensorflow and Keras libraries for deep learning processing. Both machines run Docker containers under Ubuntu 18.04 operational system and are connected by a gigabit ethernet connection.

[11] https://github.com/aleju/imgaug
[12] https://kafka.apache.org/

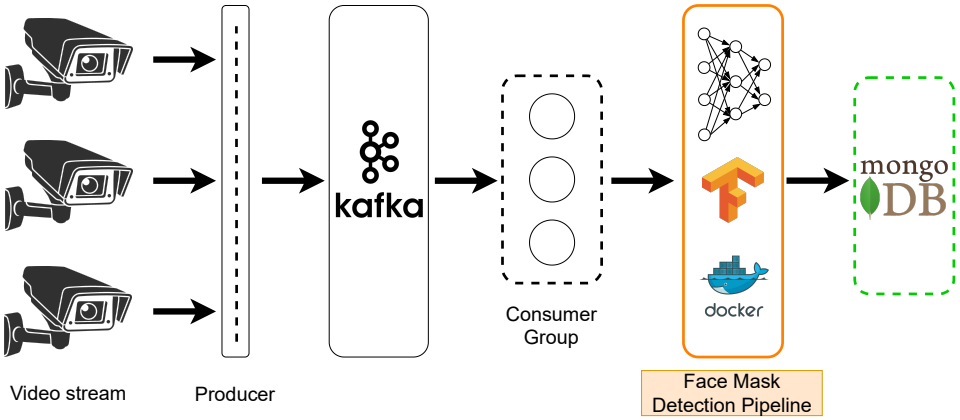

Figure 3: Overview of the Scalable Face Mask Detection Pipeline.

3.2.1 Data ingestion

The producer application reads frames from the incoming videos (or surveillance streams). In the next step, it publishes them to the Kafka topic with the following data: camera or video identification, timestamp, and image frame (in the JPG format). The configuration parameters of the Kafka producer are acknowledge=all, retries=100, compression type=gzip, linger=5 ms, and batch number of messages=32. The remaining configurations used the default values.

In the current version, all input frames are being processed. To increase the performance, it does not wait for confirmation that the message has been delivered or not. However, it could be configured to drop redundant frames using some pre-processing strategy, such as background subtraction, for instance. Also, in some applications where the dynamics of the objects in the video stream are relatively slow (such as people walking in a corridor, for instance), it would be enough to process very few frames per second. The Kafka topic contains one replication factor (runs in one server) and one partition (for one consumer only).

The consumer application is asynchronous, running on another server, different from the one responsible for the data ingestion. The consumer server reads frames in batches, and when a data batch is formed, it is processed by the Face Mask Detection Pipeline (Figure 4), which consists of three main steps: human body detection, head detection, and facemask classification.

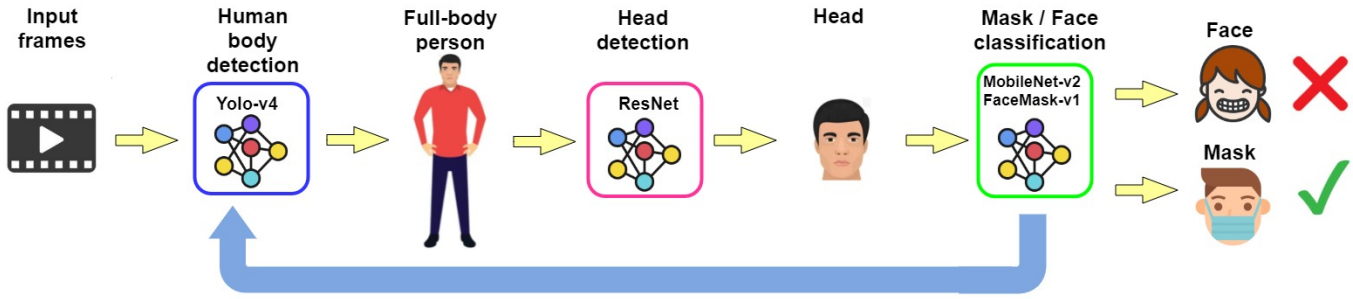

Figure 4: Face Mask Detection Pipeline.

3.2.2 Human body detection

In this step, the inference task uses Yolo-v4 [43], trained on the COCO dataset [44], which is capable of detecting 80 classes of objects. Any bounding box with a label different from "person" is discarded, thus decreasing further useless processing.

However, a common inconvenience of popular models results from multiple bounding boxes for the same object. Again, to avoid redundant or unnecessary processing, we used the Non-Max-Suppression (NMS) algorithm [45]. Good results were obtained with Intersection-over-Union (IoU) $< 30\%$. For each person correctly detected, the bounding box is cropped and forwarded to the next step of the pipeline.

3.2.3 Head detection

Because the input is a video stream, the quality and lighting conditions may not be optimal. In many cases, the bounding box of the person in the previous step is relatively small, related to the size of the image. Thus, preprocessing was needed to improve the classification accuracy in the following steps before trying to detect a human's head.

Figure 5 illustrates how such improvement was done. From a full-body bounding box, the first strategy is to try to detect the head of the person, using a ResNet DNN pre-trained with the Wider Face dataset [46]. If ResNet does not detect the head, the image is resized to 300×300 pixels. If still no head is detected, the image is resized, aiming to keep the original aspect ratio.

At this point, ResNet could detect the head in some cases. Considering that in most images, people will stand up, most parts of the bounding box are not important, and the head will be naturally in the upper part of the image. We pointed this additional step improved confidence in the head detection procedure. In the final step, the image was resized again to keep the new aspect ratio. In short, the procedure adopted here enabled ResNet to achieve high confidence compared to the preliminary steps. Overall, this procedure highlighted the importance of correct resizing before inference.

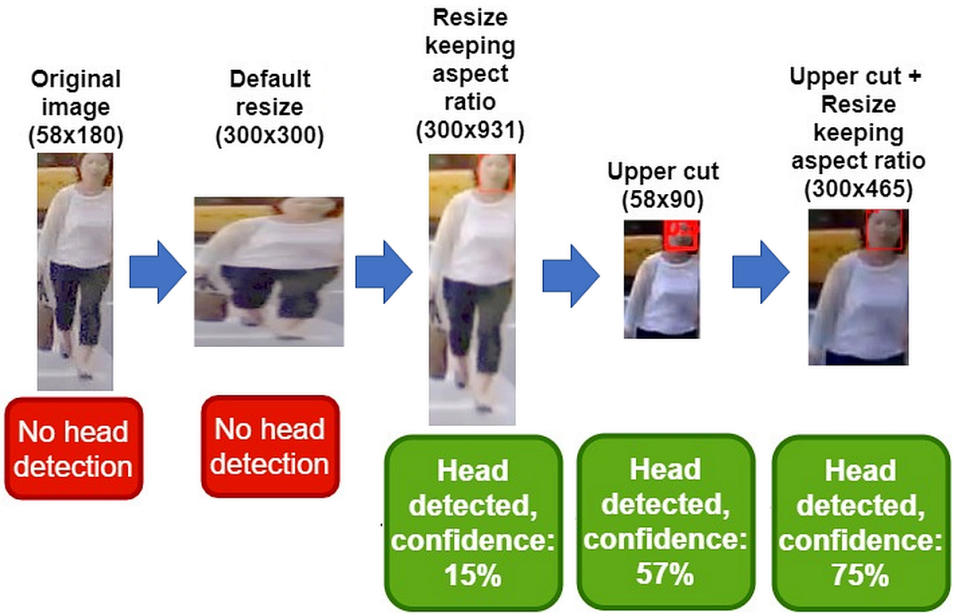

Figure 5: Results of several preprocessing approaches. The images are not in real scale. Uppercut and resizing the width to 300 pixels while keeping the aspect ratio of the image yielded better inferences. This technique also works for partial body image.

After this preprocessing, the ResNet can detect the head and return the corresponding bounding box with high confidence. However, empirical tests in the next step (face mask classification) revealed misclassification errors for low-resolution images. To tackle this issue, the algorithm filter out tiny-sized head bounding boxes according to the video resolution. We established a reference resolution of 854×480, for which the minimum size was empirically set to 15×15. All images below this threshold are discarded. For any other video resolutions, the minimal proportions can be obtained by using: $W_{min} = (\frac{W}{854}).15$ and $H_{min} = (\frac{H}{480}).15$, where W_{min} and H_{min} are the minimum acceptable width and height, and W and H are the width and height of the current resolution of the frame. Additionally, we accepted only square or near-square head bounding boxes. If the difference between height and width was larger than 72, the image was discarded. Otherwise, it is forwarded to the next step.

3.2.4 Face mask classification

Following other works in the literature [23,24,27,47], we also used transfer learning to build a model for face mask classification. The base for our approach is the MobileNet-v2 model [48] previously trained with the ImageNet dataset. The pooling and fully connected layers replaced the last dense layers of the network. This model was called FaceMask-v1, and we trained it with the UTFPR-FMD1 dataset previously described.

To accomplish the transfer learning, we applied an AveragePooling2D layer with a 7×7 kernel size, followed by three convolution layers with 128, 256, and 256 channels, respectively, each one using the ReLU (Rectified Linear Unit) activation function. A single dropout of 0.5 was also applied to improve the model generalization and avoid overfitting. The final convolution layer consists of two channels with the softmax activation function. During training, we used a binary cross-entropy loss.

The images were randomly distributed to compose the train and test sets in the proportion of 70% and 30%, respectively. The input shape of the FaceMask-v1 is a 224×224 RGB image, with pixels normalized in the $[-1..1]$ range. We used the Adam optimizer with default parameters and a learning rate ($alpha = 0.0001$). We defined 20 training epochs and a batch size of 64. Tensorflow and Keras are the DNN frameworks used in this project. With this configuration the results of model fit are: 8.51% training loss, 97.02% training accuracy, 3.16% validation loss, and 98.83% validation accuracy.

The trained model accepts a head image as input and outputs the respective prediction probability for “face” and “mask”. The classification module also draws the bounding boxes in the frame and their probability of belonging to the predicted class. Red bounding boxes correspond to the face class, and blue bounding boxes correspond to the mask class. The result is shown in Figure 6 for a crowded scene.

All frames containing faces, with or without a mask, are stored in a fast-write NoSQL database (in our case, in a MongoDB database). Each stored document contains frame identification, timestamp, labels, and bounding boxes, making it easy to query the data. Other frames was discarded.

[13]Gabriella Clare Marino - unsplash.com

Figure 6: Face mask detection in a crowded real-world scene. Red bounding boxes correspond to "face", and blue bounding boxes correspond to "mask" classes. Images below the described thresholds (see section 3.2.3) are not processed. Original image from Unsplash[13].

4 Experimental results and discussion

In this Section, we present and discuss the processing pipeline results through a qualitative analysis of the FaceMask-v1 model and a comparison of our approach with other related work and different datasets.

4.1 Data ingestion and processing performance

Considering the hardware mentioned in Section 3.2, the average inference time of the FaceMask-v1 model for each head image was $0.03\,\mathrm{s}$. These results make it suitable for working with many pictures, such as in a real-world surveillance context. Additionally, we performed an experiment that evaluates the throughput, in FPS (frames per second) of the entire SFMDP pipeline, from frame ingestion to the final face/mask detection, when using only the CPU and when using CPU and GPU together. It is important to note that the producer only runs on the CPU, while the consumer can run on either CPU or GPU. The processing time includes the I/O operations and frame decoding for a single video stream. All resolution sizes require almost the same processing time. This behavior is due to the resizing procedure performed during the frame ingestion, keeping the aspect ratio.

Table 4 presents the average Kafka ingestion performance for several frame resolutions. It is remarkable that even with Full-HD resolution, the system can ingest many FPS, more than necessary for simple video surveillance tasks. Thus, the system allows ingesting more than one video stream.

Table 4: Throughput of the entire pipeline considering different frame resolutions, Kafka ingestion performance in frames per second (FPS), FaceMask-v1 model FPS without GPU, FaceMask-v1 model FPS with GPU, and ΔGPU is the performance gain with GPU.

Resolution	Ingestion FPS	no-GPU	with-GPU	ΔGPU
432×240	415	3.58	9.41	163%
640×360	153	2.82	5.69	100%
854×480	78	2.80	5.51	98%
1280×720	37	2.77	5.40	95%
1920×1080	16	2.70	5.15	90%

4.2 Qualitative results

Table 5 presents the visual evaluation results of the average detection degree for each head position predicted by the FaceMask-v1 model. Significant results were achieved until 60° of head rotation, even considering complex image/background. These results are remarkable since most published works well only with frontally aligned faces. However, for real-world video surveillance, face alignment cannot control the alignment of the face. The image processing system must be robust for this purpose.

Table 5: Detection angle of the FaceMask-v1 model for each head position. The zero degrees correspond to a frontally aligned view of the head image.

Head position	**Head angle**		
	45^o	**60^o**	**90^o**
Face up	Yes	Yes	No
Face down	Yes	No	No
Face left	Yes	Yes	Yes
Face right	Yes	Yes	Yes
Mask up	Yes	Yes	No
Mask down	Yes	No	No
Mask left	Yes	No	No
Mask right	Yes	Yes	No

4.3 Comparison with other works and datasets

We run several inference experiments to compare the performance of the FaceMask-v1 model, trained on our UTFPR-FMD1 dataset, against other available datasets previously published in the literature. Due to the lack of standardization regarding the way results have been presented in recent papers, we used the four most popular metrics from the literature [49] (accuracy, precision, recall, and F1-score) to present the results, and they are shown in Table 6.

Table 6: Performance measures of the FaceMask-v1 model for different datasets

	Class	**Precision**	**Recall**	**F1-score**	**Acc.**
AIZOOTech [30]	Face	0.950	0.966	0.958	0.966
	Mask	0.939	0.958	0.948	0.940
CMFD [26]	Mask	0.991	0.991	0.991	0.991
FMD [32]	Face	0.813	0.800	0.806	0.777
	Mask	0.962	0.968	0.965	0.940
MAFA [33]	Mask	0.860	0.860	0.860	0.860
RMFRD [34]	Mask	0.843	0.843	0.843	0.843
SMFD [35]	Face	0.864	0.667	0.753	0.767
	Mask	0.622	0.809	0.670	0.735
SSDMNV2 [36]	Face	0.818	0.759	0.787	0.795
	Mask	0.775	0.831	0.802	0.795

In [50] authors presented two detection models and used the AIZOOTech dataset to train, validate, and test the proposed system. The average results for precision and recall were 87% and 93%, respectively. Therefore, our model surpassed these results, obtaining better average results in accuracy (+7%) and recall (+3%). They did not present performance tests on videos or images in the wild as we did.

In [24] authors presented three different detection models trained and tested with four datasets: RMFRD, SMFD, a fusion of RMFRD and SMFD, and LFW with simulated face masks. However, the last two datasets are not available. The authors reported the mean accuracy, recall, and F1 score for the RMFRD dataset as 98%. For the SMFD dataset, the mean accuracy, recall, and F1 score were 83%, 79%, and 81%, respectively. Although the paper showed good results on their dataset, they did not test datasets from different published works. There is no discussion about why some testing results are poor, possibly because the classes were strongly unbalanced. The results do not consider test performance over videos.

Later, the previous work [24] was extended in [25]. They presented results for two datasets: MAFA and a custom merge of two other datasets from Kaggle: Medical Masks Dataset (MMD)[14], and Face Mask Dataset (FMD)[15]. Unfortunately, merged

[14]https://www.kaggle.com/vtech6/me dical-masks-dataset
[15]https://www.kaggle.com/andrewmvd/face-mask-detection

datasets are no longer available for comparison. The tests on the MAFA dataset achieved 81% of precision using YOLOv2 with ResNet50. The authors of MAFA [33] achieved average precision equal to 76.1% using their LLE-CNN. Our model surpassed with 86% in all metrics considered, 5% better than [25] and 9.9% better than [33].

Based on the above-mentioned extensive tests with other datasets, we showed that the FaceMask-v1 model is quite robust for both good and bad image quality and of any size. This suggests its utility for real-world applications, such as in the surveillance context. Considering our dataset, UTFPR-FMD1, the FaceMask-v1 model achieved 99% on all metrics.

5 Conclusions and future work

Although several datasets and deep learning models for processing images of masked and unmasked people have emerged in recent months, they present essential drawbacks. Few datasets contain enough samples to train CNN models with acceptable quality and robustness for real-life images. Almost all existing frameworks and models are useless for real-time or near real-time processing.

The contributions of the present work fill these gaps in the literature and present solutions to many problems. The proposed method was applied to other available datasets and achieved outstanding results, thus highlighting its adequacy for face mask recognition in video streams. In addition, other researchers can use our work as a baseline and comparison. Soon, we will deploy the current framework into production, perform tests in a near real-time environment, and extend it to multiple video streams. Also, the dataset UTFPR-FMD1 will be expanded with more human diversity images, respecting the correct balance and a variety of classes.

Acknowledgements

Author C. Kossoski thanks CAPES for the PhD grant, author H.S. Lopes thanks CNPq for the research grant 311785/2019-0 as well as Fundação Araucária for grant PRONEX 042/2018. All authors thank NVIDIA Corp. for the donation of the Titan-Xp GPUs used in the experiments.

References

[1] European Centre for Disease Prevention and Control. “Using face masks in the community: first update”. Technical Report, Stockholm, Dennmark, 2021.

[2] J. P. Ioannidis. “The end of the COVID-19 pandemic”. *European Journal of Clinical Investigation*, 2022.

[3] A. A. Dawood. “Mutated COVID-19 may foretell a great risk for mankind in the future”. *New Microbes and New Infections*, vol. 35, pp. 100673, 2020.

[4] E. Volz, V. Hill, J. T. McCrone, A. Price, D. Jorgensen *et al.*. “Evaluating the effects of SARS-CoV-2 spike mutation D614G on the transmissibility and pathogenicity”. *Cell*, vol. 184, no. 1, pp. 64–75, 2021.

[5] S. P. Rajeshbhai, S. S. Dhar and Shalabh. “Fourth wave of COVID-19 in India: statistical forecasting”. *MedRxiv*, 2022.

[6] “The way Chinese think about COVID-19 is changing”. *The Economist: China*, Apr 2022. `https://www.economist.com/china/2022/04/16/the-way-chinese-think-about-covid-19-is-changing`. Accessed: 2022-04-30.

[7] C. Wang and J. Han. “Will the COVID-19 pandemic end with the Delta and Omicron variants?”, 2022.

[8] H. J. Schünemann, E. A. Akl, R. Chou, D. K. Chu, M. Loeb, T. Lotfi, R. A. Mustafa, I. Neumann, L. Saxinger, S. Sultan and D. Mirtz. “Use of facemasks during the COVID-19 pandemic”. *The Lancet Respiratory Medicine*, vol. 8, no. 10, pp. 954–955, 2020.

[9] World Health Organization. “Mask use in the context of COVID-19”. Who/2019-ncov/ipc_masks/2020.5, Geneve, Switzerland, 2020.

[10] S. E. Eikenberry, M. Mancuso, E. Iboi, T. Phan, K. Eikenberry, Y. Kuang, E. Kostelich and A. B. Gumel. “To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic”. *Infectious Disease Modelling*, vol. 5, pp. 293–308, 2020.

[11] J. Howard, A. Huang, Z. Li, Z. Tufekci, V. Zdimal *et al.*. “An evidence review of face masks against COVID-19”. *Proceedings of the National Academy of Sciences*, vol. 118, no. 4, pp. e2014564118, 2021.

[12] P. Cotrin, A. C. Bahls, D. O. Silva, V. M. P. Girão, C. R. Maio *et al.*. “The use of facemasks during the COVID-19 pandemic by the Brazilian population”. *Journal of Multidisciplinary Healthcare*, vol. 13, pp. 1169–1178, 2020.

[13] G. T. Tucho and D. M. Kumsa. “Universal use of face masks and related challenges during COVID-19 in Developing Countries”. *Risk Management and Healthcare Policy*, vol. 14, pp. 511—517, 2021.

[14] L. Peeples. "COVID reinfections likely within one or two years, models propose". *Nature*, Out 2021.

[15] B. Amos, B. Ludwiczuk and M. Satyanarayanan. "OpenFace: A general-purpose face recognition library with mobile applications". Technical report cmu-cs-16-118, CMU School of Computer Science, 2016.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu and A. C. Berg. "SSD: Single shot multibox detector". In *Proc. European Conf. on Computer Vision*, volume LNCS 9905, pp. 21–37, 2016.

[17] F. Schroff, D. Kalenichenko and J. Philbin. "FaceNet: a unified embedding for face recognition and clustering". In *Proc. the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

[18] K. Zhang, Z. Zhang, Z. Li and Y. Qiao. "(MTCNN) Multi-task Cascaded Convolutional Networks". *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[19] K. He, X. Zhang, S. Ren and J. Sun. "Deep Residual Learning for Image Recognition". In *Proc. the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 770–778, 12 2016.

[20] 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.

[21] F. Schroff, D. Kalenichenko and J. Philbin. "Facenet: A unified embedding for face recognition and clustering". In *Proc. the IEEE Conf. on computer vision and Pattern Recognition*, pp. 815–823, 2015.

[22] B. Qin and D. Li. "Identifying facemask-wearing condition using image super-resolution with classification network to prevent COVID-19". *Sensors (Switzerland)*, vol. 20, no. 18, pp. 5236, 2020.

[23] S. K. Addagarla, G. Kalyan Chakravarthi and P. Anitha. "Real time multi-scale facial mask detection and classification using deep transfer learning techniques". *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 4, pp. 4402–4408, 2020.

[24] 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, pp. 108288, 2021.

[25] M. Loey, G. Manogaran, M. H. N. Taha and N. E. M. Khalifa. "Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection". *Sustainable Cities and Society*, vol. 65, pp. 102600, 2021.

[26] A. Cabani, K. Hammoudi, H. Benhabiles and M. Melkemi. "MaskedFace-Net -– A dataset of correctly/incorrectly masked face images in the context of COVID-19". *Smart Health*, vol. 19, pp. 100144, 2021.

[27] G. J. Chowdary, N. S. Punn, S. K. Sonbhadra and S. Agarwal. "Face mask detection using transfer learning of InceptionV3". In *Proc. the Int. Conf. on Big Data Analytics*, volume LNCS 12581, pp. 81–90, 2020.

[28] A. Oumina, N. El Makhfi and M. Hamdi. "Control the covid-19 pandemic: Face mask detection using transfer learning". In *Proc. IEEE 2nd Int. Conf. on Electronics, Control, Optimization and Computer Science*, pp. 1–5. IEEE, 2020.

[29] F. Mercaldo and A. Santone. "Transfer learning for mobile real-time face mask detection and localization". *Journal of the American Medical Informatics Association*, 2021.

[30] D. Chiang. "Detect faces and determine whether people are wearing mask". `https://github.com/AIZOOTech/FaceMaskDetection`, 2020.

[31] D. G. Dondo, J. A. Redolfi, D. Garcia and R. G. Araguás. "Application of Deep-Learning Methods to Real Time Face Mask Detection". *IEEE Latin America Transactions*, vol. 19, no. 6, pp. 994–1001, 2021.

[32] M. Andrew. "Face Mask Detection (FMD)". `https://www.kaggle.com/andrewmvd/face-mask-detection`, 2020.

[33] S. Ge, J. Li, Q. Ye and Z. Luo. "Detecting masked faces in the wild with LLE-CNNs". In *Proc. the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2682–2690, 2017.

[34] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong *et al.*. "Masked Face Recognition Dataset and Application". *arXiv*, vol. 2003.09093, 2020.

[35] P. Bhandary. "Simulated Masked Face Dataset (SMFD)". `https://github.com/prajnasb/observations`, 2020.

[36] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria and J. Hemanth. "SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2". *Sustainable Cities and Society*, vol. 66, pp. 102692, 2021.

[37] R. Takahashi, T. Matsubara and K. Uehara. "Data Augmentation Using Random Image Cropping and Patching for Deep CNNs". *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2917–2931, 2020.

[38] K. Wang, B. Fang, J. Qian, S. Yang, X. Zhou and J. Zhou. "Perspective Transformation Data Augmentation for Object Detection". *IEEE Access*, vol. 8, pp. 4935–4943, 2020.

[39] A. Brilhador, M. Gutoski, L. T. Hattori, A. de Souza Inácio, A. E. Lazzaretti and H. S. Lopes. "Classification of Weeds and Crops at the Pixel-Level Using Convolutional Neural Networks and Data Augmentation". In *Proc. IEEE Latin American Conf. on Computational Intelligence*, pp. 1–6, 2019.

[40] C. Shorten and T. M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning". *Journal of Big Data*, vol. 6, no. 60, pp. 1–48, 2019.

[41] A. Mikołajczyk and M. Grochowski. "Data augmentation for improving deep learning in image classification problem". In *Proc. the International Interdisciplinary PhD Workshop*, pp. 117–122, 2018.

[42] M. Romero, M. Gutoski, L. T. Hattori, M. Ribeiro and H. S. Lopes. "A Study of the Influence of Data Complexity and Similarity on Soft Biometrics Classification Performance in a Transfer Learning Scenario". *Learning & Nonlinear Models*, vol. 18, no. 2, pp. 56–65, 2020.

[43] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao. "YOLOv4: Optimal Speed and Accuracy of Object Detection". *arXiv*, vol. 2004.10934, 2020.

[44] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick and P. Dollár. "Microsoft COCO: Common Objects in Context". In *Proc. the European Conf. on Computer Vision*, pp. 740–755, 2014.

[45] N. Bodla, B. Singh, R. Chellappa and L. S. Davis. "Soft-NMS – Improving object detection with one line of code". In *Proc. IEEE Int. Conf. on Computer Vision*, pp. 5562–5570, 2017.

[46] S. Yang, P. Luo, C. C. Loy and X. Tang. "WIDER FACE: a face detection benchmark". In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 5525–5533, 2016.

[47] M. Gutoski, M. Ribeiro, L. T. Hattori, M. Romero, A. E. Lazzaretti and H. S. Lopes. "A Comparative Study of Transfer Learning Approaches for Video Anomaly Detection". *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 35, no. 5, pp. 2152003, 2020.

[48] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. C. Chen. "MobileNetV2: Inverted Residuals and Linear Bottlenecks". In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

[49] D. V. Carvalho, E. M. Pereira and J. S. Cardoso. "Machine Learning Interpretability: A Survey on Methods and Metrics". *Electronics*, vol. 8, no. 8, 2019.

[50] M. Jiang, X. Fan and H. Yan. "Retinamask: A face mask detector". *arXiv preprint arXiv:2005.03950*, 2020.