Passenger Surveillance Using Deep Learning in Post-COVID-19 Intelligent Transportation System

Srimanta Kundu1,2 · Ujjwal Maulik1

Received: 13 January 2022 / Accepted: 25 April 2022 / Published online: 26 May 2022
© Indian National Academy of Engineering 2022

Abstract
Intelligent Transport System should be renovated in many aspects in post-pandemic situation like COVID-19. The passenger-count inside a car will be restricted based on the vehicle capacity and the COVID-19 hot-spot zone. Traffic rules will be impacted to align with a similar contagious outbreak. The on-road ‘Yellow-Vulture’ cameras need to incorporate such surveillance rules to monitor related anomalies for preventing contamination. To maintain safe-distance, an automatic surveillance system will be preferred by the Government very soon. Moreover, facial mask usage during the journey has become an essential habit to stop the spread of the infection. In this article, we have proposed a deep-Learning based framework that employs an augmented image data set to provide proper surveillance in the transport system to maintain the health protocols. Fast and accurate detection of the number of passengers inside a car and their face masks from the traffic inspection camera feed has been demonstrated. We have exploited the advantages of the popular Transfer Learning approach with novel variations of images while performing the training. To the best of our knowledge, this is the first attempt to watch over in-vehicle social-distancing in post-pandemic circumstances through deep-Learning based image analysis. The superiority of the proposed framework has been established over several state-of-the-art techniques using different numerical metrics and visual comparisons along with a support of statistical hypothesis test. Our technique has achieved 98.5% testing accuracy in various adverse conditions. Zero-shot evaluation has been explored for the Real-Time-Medical-Mask-Detection data set Wang et al. (Real-Time-Medical-Mask-Detection, 2020a https://github.com/TheSSJ2612/Real-Time-Medical-Mask-Detection/, Accessed 14 Nov 2020), where we have attained 96.4% accuracy that manifests the generalization of the network.

Keywords COVID-19 · Vehicle surveillance · Mask detection · Passenger counting · Transfer learning · Inception-V3

Introduction

The current pandemic situation will impact every industry in the world. The automotive industry will not be an exception. All the domains in this industry such as car design, manufacturing, marketing strategy, maintenance, ride safety, etc. will need to address the upcoming restrictions. The vehicle monitoring system will have a domain shift too from its existing version. In general, Government and the Traffic Management Authority install surveillance systems to increase on-road safety and security. Those systems inspect pollution, speed of a vehicle, anomalies in driving behavior and other traffic rule violations. ‘Yellow-Vulture’ cameras have become very popular in UK’s road which captures several offenses, such as smoking, drinking, etc. inside the car. However, in the near future, some additional constraints related to COVID-19, such as social-distancing, mask-wearing, etc., need to be compulsorily considered. Therefore, it is essential to address issues related to contagious diseases in Intelligent Transportation System (ITS) as well.

In the last few years, several research works have been reported to address issues related to the surveillance system in the vehicular field. Baran et al. (2016) showed a smart camera-based system to detect anomalies in ITS. In the paper, Mehboob et al. (2019) have presented one intelligent traffic management system. Zhang et al. (2019)
proposed an efficient vehicle detection technique for traffic surveillance data in real-time. In the paper, Manzoor et al. (2019), recognized vehicles on the basis of generation, make and model which will be used for surveillance. In Gu et al. (2018), the authors proposed a unique network model for multi-UAV surveillance. Drone-based traffic management is also becoming popular in the last few years (Ali 2019; Ong and Kochenderfer 2017). However, in most cases, video processing is commonly used in traffic management systems (Xia et al. 2016; Hsieh et al. 2006). Pramanik et al. (2021) have developed a video surveillance system for road safety with pre-event detection feature. Different vehicle detection and counting techniques have been applied by the researchers for the last decades (Tang et al. 2017; Balid et al. 2017). Detection of several driving anomalies has been deployed for traffic management in many articles (Yuan et al. 2016; Hua et al. 2018). Seat-belt violation has been captured efficiently by Elihos et al. (2018). In recent pandemic state, researchers also have put their effort to identify and enforce social distancing norms (Punn et al. 2020; Yang et al. 2020; Saponara et al. 2021). In (Hörcher et al. 2021), authors have shown different possibilities of implementing social distancing in public transport. The spread of COVID through transport is becoming a huge challenge to the world now. ITS should impose such relevant norms to break the chain of COVID-transmission.

In this paper, we have mainly focused to prevent the spread of such contagious diseases through enforcing rules in ITS. The key intention here is to identify the passengers’ mask-wearing status inside a vehicle. We have deployed a Deep Learning (DL) based approach to deal with the mask status of any passenger seating inside the vehicle. At the same time, the passenger count is another important criterion to maintain during the pandemic time; this indirectly takes care of the social distancing aspect. Though face detection is a well-known problem in computer vision, detection of tiny degraded faces inside a vehicle for passenger counting is not a widely explored problem. A near real-time integrated process flow for the passenger counting and mask detection from image frames has never been explored to the best of our knowledge. In this work, we not only design a robust framework for counting the number of passengers inside a vehicle but also designed a face mask detection algorithm using transfer learning with very high accuracy.

Contributions:

- A robust data set with huge variations and real-life use cases have been prepared. Special attention has been given to the side-face, hand-covered and low-resolution images which will be essential for the traffic surveillance. The data set is publicly available at https://github.com/srimantacse/MaskSurveillance.
- An efficient deep learning model to identify face mask over the human face with the adoption of very tiny face detection has been implemented. We have shown that the proposed model outperforms the other state-of-the-art techniques with the extensive comparison as well as statistical hypothesis testing.
- As per the authors’ knowledge, this is the first work to find out the mask-wearing status in inside-vehicle surveillance. This sort of traffic monitoring will be very essential in the current COVID-19 related pandemic context.

The rest of this paper is organized as follows. Literature survey for the mask detection has been presented in Sect. “Related Works of Mask Detection”. In Sect. “Some Basics of Deep Learning”, an overview of different deep learning concepts, used in this article, have been briefed. The data set preparation and model surveillance camera setup have been presented in Sect. “Modeling Image Data Set”. Thereafter, the experimental setup and the methodology are discussed in Sect. “Proposed Framework”. We have summarized the results under different scenarios in Sect. “Results and Evaluations”. Finally, we discuss the overall achievement in short and conclude the paper in Sect. “Conclusions and Discussion”.

Related Works of Mask Detection

With the prevalence of the COVID-19 pandemic and the implementation of social-distancing, road transportation rules have been modified and imposed with new criteria. In this article, we have outlined a different aspect of surveillance considering the post-COVID-19 circumstances. Here, we have focused on the two important ideas in the vehicular inspection process. First, we have put our attention into the safe distance norm inside the car. The government of different countries has enforced regulations to control the COVID-19 spread through transport. In several countries, passenger count is getting restricted, where four-wheelers can only have three occupants, including the driver, to maintain social-distancing (Online 2000a, b). For the next few years, this kind of rule should be there to control the contagious spread. In this article, we have incorporated a robust face detection technology for counting the number of passengers within a car. In this regard, from the road-side installed surveillance camera, video frames or snippets will be taken. Second, the use of face-mask is another mandatory criterion to prevent the spread via the transport system. Therefore, the identification of facial mask has become an important topic of research in very recent times. Researchers have started putting efforts in this particular area to detect...
mask over the human face which is very critical preventive step (Rahmani and Mirmahaleh 2020). A feature-based technique that uses the images from mobile camera, has been developed by Chen et al. (2021). They have put focus on the feature extraction for the specific types of masks. K-Nearest Neighbor algorithm has been utilized for the detection purpose. Ejaz et al. (2019) have utilized the well known PCA technique for the recognition of mask over human face. In Nieto-Rodríguez et al. (2015), the authors have experimented to identify the presence of a mandatory medical-mask in the operation theater. Though, their application areas were quite novel, the performance of the proposed algorithms were not satisfactory due to the use of traditional machine learning based methods, and recently developed DL approaches provide better performance in the mentioned tasks in terms of accuracy. Loey et al. (2021) have recently presented a Support Vector Machine (SVM) based approach over ResNet architecture for face mask detection. Though, the accuracy of the model is high, it has not been validated for the noisy image set. Moreover, the ResNet feature computation is bit complex due to the higher number of parameters. In Wang et al. (2020b), authors have achieved 95% recognition accuracy using the face–eye-based multi-granularity model. Single-shot detector (SSD) has been experimented in Nagrath et al. (2020) for identifying the faces from the images and afterward, the MobileNet-V2 baseline architecture has been used for the classification task. Though, specifically the tiny face use cases, side-face, hand-covered use cases were not considered in this paper. Thus, the robustness of the proposed framework is low and it is difficult to use in real-world conditions. At the same time, several other CNN frameworks with different backbone architectures can be deployed for serving the same purpose. For example, AlexNet (Krizhevsky et al. 2012) can be used as a backbone of a low complexity process. As VGG (Simonyan and Zisserman 2014) is known to be a robust feature extractor, VGG based models such as VGGFace-ResNet (Parkhi et al. 2015) have also been developed for this purpose. Specialized models for face classification such as FaceNet (Schroff et al. 2015) can be modified to detect face mask in images. VGGNet (Simonyan and Zisserman 2014), AlexNet (Krizhevsky et al. 2012) have similar drawback of having huge numbers of parameters 138M and 62M, respectively. In this respect, Inception-V3 is comparatively lighter network to train to adapt in any Transfer Learning (TL) framework. Table 1 outlines a comparative sketch of the state-of-the-art approaches in this domain.

Table 1 Overview of different related DL based state-of-the-art algorithms: date set used, the face positions and conditions, such as front or side, hand-covered, image quality on which they operate, face detection strategy, zero-shot evaluation status

| Algorithm               | Data set used                                      | Special attention |
|-------------------------|---------------------------------------------------|--------------------|
|                         |                                                    | Front face | Side face | Hand covered | Low resolution | Face detection | Zero-shot evaluation |
| VGGFace (Parkhi et al. 2015) | LFW (Kawulok et al. 2016), Youtube Face DB (Wolf et al. 2011) | ✓          | ✓         | ×            | ×               | ×               | ×                   |
| FaceNet (Schroff et al. 2015) | LFW (Kawulok et al. 2016), Youtube Face DB (Wolf et al. 2011) | ✓          | ✓         | ×            | ×               | ×               | Picasa type         |
| SSDMNV2 (Nagrath et al. 2020) | Mikolaj (Kaggle), Prajna (github) (Bhandary 2020), RTMDD (Wang et al. 2020a) | ✓          | ×         | ×            | ×               | Single shot     | ×                   |
| ResNet-SVM (Loey et al. 2021) | RMFD (Wang et al. 2020b), Prajna (github) (Bhandary 2020), LFW (Kawulok et al. 2016) | ✓          | ×         | ×            | ×               | ×               | ×                   |
| **Proposed Approach**   | RTMDD (Wang et al. 2020a), Prajna (github) (Bhandary 2020), **Proposed data set** | ✓          | ✓         | ✓            | ✓               | Tiny face       | ✓                   |

Bold signifies the better index values obtained by the proposed approach
outcomes. The Transfer Learning (TL) process has been exploited in our framework. We have explored the popular Deep Neural Network (DNN) Inception-V3 in the experiment. Using TL mechanism, the model has been re-trained through the proposed network with our data set. The result has demonstrated the efficiency of the framework. We have compared our approach with the recent state-of-the-art DL techniques, such as SSDMN2V2, ResNet-SVM, AlexNet, VGGFace-ResNet, FaceNet using several numerical metrics viz. Accuracy, Precision, Recall, F1-Score, Specificity and Jaccard score (JS). Visual comparison and other related results also prove the superiority of the proposed approach. Along with these, we have showcased the efficiency of the proposed method through zero-shot evaluation steps with the Real-Time-Medical-Mask-Detection (RTMDD) data set experimented in Nagrath et al. (2020); this supports that the method is not bottleneck by the over-fitting problem. The research benefits of the proposed framework mainly include real-time traffic surveillance. Preventing the spread of COVID-19-like contagious diseases through the transport system is one important motive of this research.

Some Basics of Deep Learning

Deep Learning (DL) is an enhanced and flexible extension of Machine Learning (ML) to improve learning algorithms’ structure and make them easy to use. Being a subset of ML, DL is immensely used with very large data with a simpler framework to deploy any complex model (Schmidhuber 2015; LeCun et al. 2015; Ghosh et al. 2019). DL along with Convolutional Neural Networks (CNNs) has been tremendously applied in several computer vision researches. The major advantage of DL is having the potential to support huge amounts of data during the learning. As an obvious impact, the accuracy will increase automatically by a considerable margin.

Transfer Learning

Transfer Learning (TL) (Pan and Yang 2009) is a popular concept in ML which has been successfully applied in different application areas (Bird et al. 2020; Pan et al. 2010). It utilizes the knowledge acquired while resolving one problem and applies it to a different but related domain. In CNN based classification problems, usually, the output layer (i.e., the final layer used in the network) needs to be replaced with a new layer as per the problem domain. The main idea behind this approach to achieve higher accuracy with fewer data. It allows any DL-based framework to generate a robust feature set from the original input and with a small connected network we can achieve better accuracy. The baseline performance might improve due to this knowledge transfer along with a faster model development time. Better generalization can be attained with this approach.

The same notions have been utilized in our setup to enrich our proposed baseline network. We have achieved higher accuracy with comparatively less data (compare to ImageNet Deng et al. 2009) in our experiment.

ImageNet Database

ImageNet (Deng et al. 2009) is an image database categorized as per the WordNet hierarchy. Typically, the training data set was comprised of 1,000,000 images, with 50,000 for a validation set and 150,000 for a test set. This data set has been used by researchers in several articles and machine vision projects (Russakovsky et al. 2015; Deng et al. 2014).

Inception-V3

Google has provided this popular Deep Learning Convolutional architecture. The model includes a combination of symmetric and asymmetric logical blocks (Szegedy et al. 2016a, 2015). Inception-V3 has been trained using a set of 1,000 classes from the ImageNet data set. There is a total of 42 deep layers (Tsang 2018) in the architecture and it is significantly efficient than VGGNet (Simonyan and Zisserman 2014). In addition to the Inception-V2 architecture, it used RMSProp Optimizer, Factorized 7x7 convolutions and BatchNorm in the Auxiliary Classifiers. The overall architecture is shown in Fig. 1.

Modeling Image Data Set

In post-COVID-19 road transportation, the mask should be a mandatory item that has to be used all the time while going outside mainly. But some people are deliberately guarding their faces by hand that should be properly identified and penalized. We have focused mainly on two aspects while preparing the data set. First, we planned to prepare a substantial data set for the generic mask detection in the human face. Second, we have focused on the on-road inside car image set which will be very much important for the surveillance purpose.

Mask Data Set Preparation

Several perspectives are required for accurate mask detection. The result of the human face and mask identification strongly depends on the image resolution and dots per inch (dpi). These two qualities will be affected if the distance from the camera increases. Even the position and size of the face will be another vital point in this context. Therefore, while designing the data set, we have intentionally
introduced the blurring effect for increasing data variation. The side face images with masked and non-masked variations have been added as well to predict them correctly irrespective of face direction. Different angles of the face need to be considered here. We have contemplated all these in the data set for mask detection.

To construct a substantial database, we took an available data set of Bhandary (2020) as a baseline which has 690 masked and 686 non-masked faces. In the mentioned data set, all the face images were taken from the front side and the mask has been superimposed on the non-masked images to prepare the same. The data set has been enriched by adding another 1906 images with the existing one. The major contribution is that different side face images have been added and used during the training. At the same time, some images of the hand located on top of the mouth have been additionally included which is obviously a restricted use-case in the post-COVID-19 situation. A total of 650 such images are there in the data set. Again there is another important aspect, in the testing phase, if the image quality of the individual face is not good then it may hamper the testing accuracy. The chance of getting noisy images is higher in case of gantry camera input while capturing the vehicle in motion. Keeping this limitation in mind, we have introduced blurred images in the data set. Total 400 masked and 130 non-masked images are converted to blurred with the kernel size $(20 \times 20)$. Table 2 shows the categorywise distribution for the blurred image set. Considering the feasible angle of head position in the real scenario, we have augmented those images with two rotation angles $10^\circ$ and $20^\circ$ to increase the variation. The total number of images in the data set is now 3282. Figure 2 shows the comparative statistics of images in the overall database. Figure 3 shows some sample images of the enriched data set in different conditions.

**Model of Surveillance Camera Setup**

In our experiment, we have mostly considered low-height surveillance cameras, such as gantry-setup. The tentative height of this camera should be around 6.5 ft above ground level. It will not be elevated so high which may prevent getting the proper view of the inside vehicle. As shown in Fig. 4, the region of interest should be focused on the side windows and the windshield of a vehicle from a certain height. We have collected such images that will actually replicate the real gantry camera scenario for the test simulation. Figure 5 shows such images which have been used in our experiment. This kind of mask surveillance will be
the part of traffic ‘Yellow-Vulture’ shortly. The height of the surveillance camera has to be set experimentally based on the dimension of several vehicles. Mostly for the Light Motor Vehicle (LMV) the optimum height range is 4 ft to 7 ft. In our experiment, we have mostly considered LMV and the height from the ground level has been considered to be 6.5 ft. For the windshield images (a maximum distance of 25 to a minimum distance of 6.5 from the camera), the angle range is +75° to +45°. For the side window images, a short distance range (a maximum distance of 10 ft to a minimum distance of 4.5 ft from the camera) will be considered, where the angle range lies approximately in between +55° to +35°.

Figure 4 shows this sample calculation for the front window screen only. On the other hand, for SUVs or big buses, this height should be higher. To examine all the use cases we need to consider a multi-camera setup installed in different heights. Inside vehicle camera is another solution for this surveillance. However, this will increase the cost and create an extra burden for individual installation and necessary changes.

**Proposed Framework**

We have deployed our framework in google colab system, where the training of the proposed DNN network and testing of several use cases have been performed separately. We have tuned different DL parameters of the proposed network and used them after proper optimization.

**Experimental Setup**

The google colab environment which has been chosen for our experiment has a baseline operating system (OS) Linux – 4.19.104 (Ubuntu-18.04) with 12.7 GB RAM. We have deployed our experiment using tensorflow and keras environment along with the required APIs.

We have utilized the pre-trained Inception-V3 weights at the beginning and appended fully connected network components with average pooling followed by flatten. Subsequently, 3 sets of consecutive dense and different activation layers have been used with an intermediate dropout. The pool size in the average pooling step was kept as (5 x 5). The three activation layers used in the network are relu, sigmoid and softmax consecutively. The initial dropout percentage has been kept as 50%. Subsequently, it is reduced to 40%. Figure 6 shows the overall training architecture of the Deep Neural Network (DNN) used in the experiment. 70% of the images from the newly built data set have been used for training and the remaining images have been kept for testing.

---

2 https://www.pendlelease.co.uk/news/new-uk-speed-cameras.
3 https://colab.research.google.com/.
Again, 10% of the training data has been used as the validation set during the training. Adam optimizer has been used during the training with a learning rate $\eta$. We have tested the range of $\eta$ from 0.000001 to 0.01 at 9 discrete levels [0.000001, 0.000005, 0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01] to select the optimal value. The experimentally found optimum value of $\eta$, where the proposed model has achieved the best result is 0.0001. The learning rate $\eta$ has been uniformly distributed in the decay parameter of the optimizer over the epochs which has been set to 30.

To note, for all the other algorithms too, we have run for the same number of iterations and compared the results. "binary-cross-entropy" has been utilized for this binary classification task as the loss function. The two dropout layers have been used for better generalization in the network which will reduce the chance of over-fitting.

**Methodology**

In the proposed approach, we have used the enhanced data set as mentioned in Sect. “Modeling Image Data Set” to train the deep model. The tiny face detector model (Hu and Ramanan 2017) has been adopted to find out the face images from the given input. In our experiment, the TL concept has been deployed to build the deep network for mask detection. The baseline Inception-V3 model has been built on imagenet data set. This data set has few variations of mask (normal-mask, gas-mask, ski-mask and oxygen-mask) images as well. Therefore, our mask identifying target problem domain, $D_1$, is a subset of the Inception-v3’s problem domain, $D_s$ ($D_1 \subset D_s$). Subsequently, the learning task $T_1$ is also a subset of the corresponding source domain’s learning task $T_s$ ($T_1 \subset T_s$). From a thousand class problem, we have boiled down to two class problem using this TL framework. After removing the last output layer of the base model, we have introduced unique two-class layers (masked and non-masked) at the end in Mask Prediction node, as shown in Fig. 6. Algorithm 1 has shown the detailed structure of the process. Two input parameters, the vehicle image $I$ with size $M \times N \times 3$ and the threshold parameter $\delta$ for the passenger limit will be taken by the algorithm for providing the proper on-road surveillance. The $\delta$ parameter will be set based upon the vehicle type and capacity. The social-distancing norm will be strongly dependent on this parameter. For example, for four-wheeler LMV, the value of $\delta$ should be set to less than or equal to ($\leq$) 3 including the driver. Let’s assume, there are $N_{\text{Faces}}$ faces present in the given input $IM$. The function $R$ has performed the resizing operation over every face image to make them a fixed size of $(m \times m \times 3)$. Here, in the experiment, we have kept $m=n=224$. For the $i^{th}$ face, the corresponding feature set $\text{feature}_i^{(f_1, f_2, ..., f_k)}$ with cardinality $k$ has been computed using $E$. The generated Inception-V3 based TL model ($G$) has considered this feature set to provide the confidence level of prediction which has been noted as $\text{Conf}_{\text{Test}}^i$. Subsequently, the mask-wearing status of $i^{th}$ face has been processed from $\text{Conf}_{\text{Test}}^i$ using the function $\text{Stat}$. Based on the status and face count, the rule violation decision will be taken. During the training time, following the similar process, $\text{Conf}_{\text{Train}}^i$ has been generated for the $i^{th}$
For computing the loss (binary-cross-entropy) during the training phase, we have used the below Eq. 1. $N_{\text{Train \ total}}$ denotes the total number of images used in the training phase of the experiment:

\[
\text{Loss}_{\text{BCE}} = -\frac{1}{N_{\text{Train \ total}}} \sum_{i=1}^{N_{\text{Train \ total}}} \text{Target}_i \cdot \log \text{Conf}_{\text{Train}}^i + (1 - \text{Target}_i) \cdot \log (1 - \text{Conf}_{\text{Train}}^i)
\]

In the testing phase, we have used the tiny face detector model to extract the individual face images with the proper bounding box. As per the flow shown in Fig. 7, the frame extractor module will extract images from the video feed taken from the surveillance camera. Based on the face count, we can make a decision about the rule violation. The threshold value would be taken as prior information to the system. For a special case, only the driver without a mask inside a car would have been warned for ignoring the safety measure. At the final step of the testing phase,

![Diagram](image-url)

**Fig. 7** Testing flow of the proposed framework for passenger surveillance to find mask-wearing status inside vehicle

**Algorithm 1:** TL based passenger surveillance for mask detection inside vehicle

**Input:** Input RGB Image $I_{M \times N \times 3}$, Passenger limit threshold $\delta$

**Output:** Face set with status & confidence value: $F_{\text{out}}(\text{Status, Confidenti} = \{F_{\text{out}1}, F_{\text{out}2}, \ldots, F_{\text{out}m}\}$.

\[
\sigma:: \text{TinyFace detector API};
\]

1. $i \leftarrow 1$;

2. $\text{Face}_{\text{out}} = []; /* Initialize to empty list */$;

3. $F_{\text{out}} = []; /* Initialize to empty list */$;

4. $G \leftarrow \text{Inception} - \text{V3 based TL model as shown in Fig. 6};$

5. while At least one face remains in the image $I$ do

6.   $\text{Face}_{\text{out}}^{m \times n \times 3} = \sigma(I_{M \times N \times 3}, \text{Face}_{\text{out}});$;

7.   $\text{Face}_{\text{out}}^{m \times n \times 3} = R(\text{Face}_{\text{out}}^{m \times n \times 3}); /* Resize the image to (m x n x 3) */;$

8.   $F_{\text{out}}[i] = \text{Face}_{\text{out}}^{i};$

9.   $i++ = 1;$

10. $N_{\text{Faces}} \leftarrow \text{count}($Face$_{\text{out}}$);

11. for $i \leftarrow 1$ to $N_{\text{Faces}}$ do

12.   $\text{feature}_i(f_1, f_2, \ldots, f_k) = \mathcal{E}(\text{Face}_{\text{out}}[i]); /* Extract k features from the $i^{th}$ face */;$

13.   $\text{Conf}_{\text{Test}}^i = P(G, \text{feature}_i(f_1, f_2, \ldots, f_k)) /* Predict the confidence level */;$

14.   $\text{Stat}_{\text{Test}}^i = \text{Stat}(\text{Conf}_{\text{Test}}^i);$;

15.   $F_{\text{out}}[i] = (\text{Stat}_{\text{Test}}^i, \text{Conf}_{\text{Test}}^i);$;

16. if $\text{count}(F_{\text{out}}) = 1 \& F_{\text{out}}.\text{Status} = \text{Non - Mask}$ then

17.   Vehicle needs to be warned;

18. else

19.   if $\text{count}(F_{\text{out}}) \geq \delta$ or $\text{count}(F_{\text{out}}.\text{Status} = \text{Mask}) \neq \text{count}(F_{\text{out}}$) then

20.     Rule has been violated !!!;


using the proposed mask detection model we can detect the mask from the face images.

The learning has been demonstrated through the visualization of the saliency map for individual training images. Figure 8 demonstrates six sample saliency images with the respective original images. These sample images indicate that the model has learned the overall facial part in case of normal images and the face part without the covered zone in case of masked images. As shown in Fig. 9 the training accuracy has been gradually increased and the loss has been decreased with the epoch while training. Moreover, for statistically demonstrating the efficiency of the generated model we have run the repeated K-Fold validation process separately. The repetition parameters we have set to 5 and the K has been set to 5 as well in K-Fold. For displaying the outcome, we have drawn the boxplot of the corresponding validation accuracy for different folds in different run.

Figure 10 shows that our model exhibits the minimal span of the box in the boxplot, while the other models have shown varying accuracy in different runs. To hypothetically demonstrate the supremacy, we have computed the pairwise (state-of-the-art technique vs proposed model) p value of the Wilcoxon rank Wilcoxon (1945) which has been shown in Table 3. Here we can see, there is not such similarities in between the proposed models and other models used for the comparison in the experiment. As per the hypothesis, the obtained p values, less than 0.05, are supposed to be statistically significant. The mean and standard deviation of the different models have been noted in the same Table 3. This clearly shows the primacy of the proposed model.

![Sample saliency-map images generated for our proposed model](image)

**Fig. 8** Sample saliency-map images generated for our proposed model

| Method     | Wilcoxon p-Value (vs Proposed) | Mean  | Standard deviation |
|------------|-------------------------------|-------|--------------------|
| AlexNet    | 0.000012                      | 0.907 | 0.018              |
| VGGFace    | 0.000012                      | 0.827 | 0.019              |
| FaceNet    | 0.000012                      | 0.874 | 0.038              |
| SSDMNv2    | 0.000012                      | 0.934 | 0.011              |
| Res-SVM    | 0.016599                      | 0.964 | 0.017              |
| Proposed   | NA                            | **0.974** | **0.006**         |

**Table 3** Wilcoxon hypothesis comparison for the different models on the K-Fold validation with mean and standard deviation of the achieved by the different algorithms

![Accuracy and loss curve during the training and validation (10% of the training data) process](image)

**Fig. 9** Accuracy and loss curve during the training and validation (10% of the training data) process
Results and Evaluations

To demonstrate the superiority of the proposed approach we have tested on multiple variant images as well as on-road images of the vehicle. The primary objective of the method is to count the number of passengers inside the car, and at the same time, we need to check whether they wear a mask or not during the journey. The testing car database contains images from different angles and heights taken by the surveillance camera, night mode pictures, mirror-reflected face images, glass-covered faces etc. These images are mainly used for testing purposes. For the evaluation process, we have utilized several numerical indices viz. Accuracy, Precision, Recall, F1-Score, Specificity and Jaccard Score (JS) using Eqs. 2, 3, 4, 5, 6 and 7, respectively, along with the visual comparisons:

\[
\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}
\] (2)

\[
\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}
\] (3)

\[
\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}
\] (4)

\[
F1 - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\] (5)

\[
\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}
\] (6)

Table 4 Numerical Index (Average Weighted) comparison with the state-of-the-art techniques and other variations over 1066 [Masked (634) & Non-Masked (432)] testing images

| Model Used       | Accuracy | Precision | Recall | F1-Score | Specificity | JS     |
|------------------|----------|-----------|--------|----------|-------------|--------|
| SSDMNV2          | 0.935    | 0.910     | 0.854  | 0.891    | 0.940       | 0.880  |
| ResNet-SVM       | 0.958    | 0.964     | 0.964  | 0.965    | 0.951       | 0.939  |
| AlexNet          | 0.905    | 0.913     | 0.904  | 0.893    | 0.924       | 0.805  |
| VGGFace(ResNet)  | 0.905    | 0.904     | 0.906  | 0.897    | 0.941       | 0.765  |
| FaceNet          | 0.825    | 0.857     | 0.827  | 0.817    | 0.845       | 0.784  |
| Proposed Net     | **0.985**| **0.990** | **0.983** | **0.990** | **0.986** | **0.970** |

Bold signifies the better index values obtained by the proposed approach

Table 5 Special test set comparison with state-of-the-art techniques over different unusual conditions of mask images

| Image Type       | Model Used   | Correct Prediction |
|------------------|--------------|--------------------|
| Blurred (400)    | SSDMNV2      | 174 (43.50%)       |
|                  | ResNet-SVM   | 351 (87.75%)       |
|                  | Proposed Net | **389 (97.25%)**   |
| Side Face (50)   | SSDMNV2      | 31 (62.00%)        |
|                  | ResNet-SVM   | 46 (92.00%)        |
|                  | Proposed Net | **49 (98.00%)**    |
| Hand-Covered (200)| SSDMNV2   | 180 (90.00%)       |
|                  | ResNet-SVM   | 196 (98.00%)       |
|                  | Proposed Net | **198 (99.00%)**   |

Bold signifies the better index values obtained by the proposed approach
Table 4 shows the different indexes used for the evaluation. We have computed those indexes for all the mentioned algorithms. From Table 4, it is clear that in terms of accuracy over the testing data set, our method has outperformed other techniques. While the proposed network has achieved 98.5% accuracy, the best accuracy among other techniques is 95.8% obtained by ResNet-SVM. A similar observation is applicable for the other indexes as well. The proposed model distinctly outperforms the existing algorithms in various aspects.

\[ JS = \frac{TruePositive}{TruePositive + FalsePositive + FalseNegative} \]  

(7)

For comparing the confidence level of the prediction, we have chosen the best two models, i.e., SSDMNV2 and ResNet-SVM as per Table 4. Here we have used several unusual and critical environment conditions, such as low-quality images, side-face images and hand-covered images for the comparison. Table 5 clearly demonstrates the superiority of the proposed approach in all those conditions. 400 blurred, 50 side-face and 200 hand-covered images have been experimented. SSDMNV2 and ResNet-SVM have correctly identified 174 and 351 blurred images, respectively, while our proposed model has predicted 389 successfully. Similar types of results have been observed, while testing side-face and hand-covered category too is shown in Table 5. Figures 11 and 12 show the confidence level of prediction.
where the state-of-the-art techniques predicted correctly. For example, as shown in Fig. 11, among SSDMNV2’s correct prediction of 174 images, the confidence value of the prediction of our proposed model is higher for 71% of such images. For side-face and hand-covered images, our model has beaten SSDMNV2 by 100% and 97% margin in terms of confidence level. The same observation is applicable for ResNet-SVM model as well for the same scenarios that are depicted in Fig. 12. Figure 13 shows few sample test images, where the recommended approach correctly predicted but other existing models failed.

Figure 14 shows one sample testing flow with four passengers inside a car and all of them have used masks while traveling. The confidence level of prediction has appeared with the bounding box. The back seat passengers’ tiny faces also have been captured and predicted the mask-wearing status successfully. Figure 15 shows the stepwise results for these best-performed algorithms in various road conditions with the confidence level of successful predictions. Those conditions have been mentioned explicitly in the first column of the same Fig. 15.

Zero-Shot Evaluation

In this zero-shot evaluation process, we have used the data set that has been experimented by Nagrath et. al. in the paper (Nagrath et al. 2020). We have tested on 10,212 images in this data set with 5520 non-masked and 4692 masked images, respectively. Our model has successfully predicted 5511 non-masked and 4336 masked images. Collectively 9847 (96.4%) successful prediction is a remarkable figure without deploying training, while SSDMNV2 has achieved 92.6% accuracy after the training phase. Figure 16 shows some sample evaluation output over the mentioned data set. Figure 16a and b show few masked and non-masked outputs, respectively.

Testing Time Analysis

The face detection module has taken almost uniform time (around 3 s) for any number of human faces. However, the mask detection process has been conducted sequentially for every face image. It has linearly grown with the number of faces present in a single image. The average time taken for this process is around 0.5 s for one human face. The time graph is shown in Fig. 17. So, generally, the surveillance framework will take around 4 s to complete the process for a standard vehicle with 3 passengers (maintaining social-distancing norm). The timing measurement has been performed on the same google colab setup as mentioned in Sect. “Experimental Setup”. With Graphics Processing Unit (GPU) or Tensor Processing Unit (TPU) system, the timing graph can be improved further.

Conclusions and Discussion

The advantages of Transfer Learning have been exploited to develop new traffic surveillance which is extremely useful during COVID-19-like pandemic situation. Our proposed network has outperformed all the other state-of-the-art techniques in terms of several metrics. The zero-shot evaluation has reached 96.4% accuracy as well which supports the potency of the technique. This framework will also be applicable for similar contagious diseases, where the mask will be compulsory to use. Considering the post-COVID-19 urgency, different traffic rules need to be imposed in the transport system. The surveillance strategy should also be...
| Condition Type                      | Original Image | Detection of Faces | Detection of Mask by Proposed Model | Proposed Model | SSDMN V2 | RESNET-SVM |
|------------------------------------|----------------|--------------------|-------------------------------------|----------------|----------|------------|
| 1. Blurred Image With Mask         | ![Image](image1) | ![Image](image2) | ![Image](image3) | Pass 99.80%  | Fail     | Pass 99.20% |
| 2. Side View Image With Mask       | ![Image](image4) | ![Image](image5) | ![Image](image6) | Pass 90.46%  | Pass 85.92| Pass 96.44% |
| 3. Night Mode Image With Mask      | ![Image](image7) | ![Image](image8) | ![Image](image9) | Pass 99.60%  | Pass 74.45| Pass 83.79% |
| 4. Very Bright Image With Mask     | ![Image](image10) | ![Image](image11) | ![Image](image12) | Pass 98.50%  | Fail     | Pass 92.79% |
|                                   |                |                    |                                    | 99.30%         | Fail     | 79.86%     |
| 5. Mirrored Image With Mask        | ![Image](image13) | ![Image](image14) | ![Image](image15) | Pass 99.73%  | Fail     | Pass 99.08% |
| 6. Reflected Tiny Image With Mask  | ![Image](image16) | ![Image](image17) | ![Image](image18) | Pass 99.5%   | Fail     | Pass 85.17% |
|                                   |                |                    |                                    | 99.665%        | Fail     | 82.26%     |
| 7. Image With No Mask              | ![Image](image19) | ![Image](image20) | ![Image](image21) | Pass 99.65%  | Pass 98.50|% Pass 99.10 |
| 8. Side Face With No Mask          | ![Image](image22) | ![Image](image23) | ![Image](image24) | Pass 98.90%  | Pass 99.91%| Pass 97.90% |
| 9. Sunlight reflection on windscreen | ![Image](image25) | ![Image](image26) | ![Image](image27) | Pass 99.26%  | Fail     | Pass 78.02% |
| 10. Night mode with Poor illumination | ![Image](image28) | ![Image](image29) | ![Image](image30) | Pass 98.43%  | Fail     | Fail       |
| 11. Passengers at the Back seat    | ![Image](image31) | ![Image](image32) | ![Image](image33) | Pass 99.18%  | Pass 90.10%| Pass 83.64% |
| 12. Very Bright light over window  | ![Image](image34) | ![Image](image35) | ![Image](image36) | Pass 99.66%  | Pass 81.88%| Fail        |

**Fig. 15** Results of two best-performed state-of-the-art techniques and proposed model over unusual and challenging on-road conditions with inside-vehicle passengers for mask surveillance along with confidence levels of individual prediction.
modified accordingly. We have demonstrated an efficient way to identify the on-road rule-breakers. The developed deep model has achieved high accuracy in different conditions. As a scope of future research, one can work to enhance the training data set from a different perspective related to mask-wearing. Designing proper network to deal with the hazy and rainy images is another future scope. Researchers can apply better deep learning based tracking mechanism for the surveillance. Hyper-parameter optimization is another area to work on. For handling the night-mode on-road images more accurately, we can generate a large number of such samples using Generative Adversarial Networks. For managing varying heights of several types of vehicles, one can use a multi-camera setup. In addition, the multi-view of a specific vehicle will help to provide higher confidence in the outcome. Real field testing with this type of multi-camera setup will be another challenging task. Currently, the authors are working in these directions.

References

Ali BS (2019) Traffic management for drones flying in the city. Int J Crit Infrastruct Protect 26(100):310
Balid W, Tafish H, Refai HH (2017) Intelligent vehicle counting and classification sensor for real-time traffic surveillance. IEEE Trans Intell Trans Syst 19(6):1784–1794
Baran R, Rusc T, Fornalski P (2016) A smart camera for the surveillance of vehicles in intelligent transportation systems. Multimedia Tools and Applications 75(17):10,471–10,493
Bhandary P (2020) https://github.com/prajnasb/observations/tree/master/experiments/data
Bird JJ, Kobylarz J, Faria DR, Ekárt A, Ribeiro EP (2020) Cross-domain mlp and cnn transfer learning for biological signal processing: eeg and emg. IEEE Access 8:54,789–54,801
Chen Y, Hu M, Hua C, Zhai G, Zhang J, Li Q, Yang SX (2021) Face mask assistant: Detection of face mask service stage based on mobile phone. IEEE Sensors Journal 21(9):11,084–11,093
Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 248–255
Deng J, Russakovsky O, Krause J, Bernstein MS, Berg A, Fei-Fei L (2014) Scalable multi-label annotation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp 3099–3102
Ejaz MS, Islam MR, Sifatullah M, Sarker A (2019) Implementation of principal component analysis on masked and non-masked face recognition. In: 2019 1st international conference on advances in science, engineering and robotics technology (ICASERT), IEEE, pp 1–5
Ellios A, Alkan B, Balci B, Artan Y (2018) Comparison of image classification and object detection for passenger seat belt violation detection using nir & rgb surveillance camera images. In: 2018 15th IEEE International conference on advanced video and signal based surveillance (AVSS), IEEE, pp 1–6
Ghosh S, Das N, Das I, Maulik U (2019) Understanding deep learning techniques for image segmentation. ACM Comput Surv (CSUR) 52(4):1–35
Gu J, Su T, Wang Q, Du X, Guizani M (2018) Multiple moving targets surveillance based on a cooperative network for multi-UAV. IEEE Commun Mag 56(4):82–89
Hörcher D, Singh R, Graham DJ (2021) Social distancing in public transport: mobilising new technologies for demand management under the covid-19 crisis. Transportation pp 1–30
Hsieh JW, Yu SH, Chen YS, Hu WF (2006) Automatic traffic surveillance system for vehicle tracking and classification. IEEE Trans Intell Trans Syst 7(2):175–187
Hu P, Ramanan D (2017) Finding tiny faces. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 951–959
Hua S, Kapoor M, Anastasius DC (2018) Vehicle tracking and speed estimation from traffic videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 153–160
Kawulok M, Celebi E, Smolka B (2016) Advances in face detection and facial image analysis. Springer
Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Adv Neural Inform Process Syst 25:1097–1105

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521

Loey M, Manogaran G, Taha MHN, Khalifa NEM (2021) A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic. Measurement 167(108):288

Manzoor MA, Morgan Y, Bais A (2019) Real-time vehicle make and model recognition system. Mach Learn Knowl Extraction 1(2):611–629

Mehboob F, Abbas M, Rauf A, Khan SA, Jiang R (2019) Video surveillance-based intelligent traffic management in smart cities. In: Intelligent Video Surveillance, IntechOpen, p 19

Nagrath P, Jain R, Madan A, Arora R, Kataria P, Hemant J (2020) SSDMNv2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2. Sustainable cities and society p 102692

Nieto-Rodríguez A, Macientes M, Brea VM (2015) System for medical mask detection in the operating room through facial attributes. In: Iberian Conference on Pattern Recognition and Image Analysis, Springer, pp 138–145

Ong HY, Kochenderfer MJ (2017) Markov decision process-based distributed conflict resolution for drone air traffic management. Journal of Guidance, Control, and Dynamics pp 69–80

Online (2020) https://auto.hindustantimes.com/auto/news/driving-during-lockdown-3-0-here-is-faqs-answered-41589167300775.html, [Online;]

Online (2020) https://gulfnews.com/world/gulf/qatar/covid-19-qatar-issues-restrictions-on-number-of-passengers-in-one-car-1.1590517292042, [Online;]

Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans knowl Data Eng 22(10):1345–1359

Pan W, Xiang E, Liu N, Yang Q (2010) Transfer learning in collaborative filtering for sparsity reduction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 24

Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: British Machine Vision Conference

Pramanik A, Sarkar S, Maiti J (2021) A real-time video surveillance system for traffic pre-events detection. Accident Anal Prevent 154(106):019

Punn NS, Sonbhadrak SK, Agarwal S (2020) Monitoring covid-19 social distancing with person detection and tracking via fine-tuned yolo v3 and deepsort techniques. arXiv preprint arXiv:2005.01385

Ramani AM, Mirmahaleh SYH (2020) Coronavirus disease (covid-19) prevention and treatment methods and effective parameters: A systematic literature review. Sustainable cities and society p 102568

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis (IJCV) 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y

Saponara S, Elhanashi A, Gagliardi A (2021) Implementing a real-time, ai-based, people detection and social distancing measuring system for covid-19. Journal of Real-Time Image Processing pp 1–11

Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117

Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016a) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016b) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826

Tang Y, Zhang C, Gu R, Li P, Yang B (2017) Vehicle detection and recognition for intelligent traffic surveillance system. Multimedia Tools Appl 76(4):5817–5832

Tsang SH (2018) https://medium.com/@sh.tsang/review-inception-v3-1st-runner-up-image-classification-in-ilsvrc-2015-1791542177c

Wang Z, Wang G, Huang B, Xiong Z, Hong Q, Wu H, Yi P, Jiang K, Wang N, Pei Y, et al. (2020b) Masked face recognition dataset and application. arXiv preprint arXiv:2003.09093

Wilcoxon F (1945) Individual comparisons by ranking methods. biom. bull., 1, 80–83

Wolf L, Hassner T, Maoz I (2011) Face recognition in unconstrained videos with matched background similarity. In: CVPR 2011, IEEE, pp 529–534

Xia Y, Shi X, Song G, Geng Q, Liu Y (2016) Towards improving quality of video-based vehicle counting method for traffic flow estimation. Signal Process 120:672–81

Yang D, Yurtsever E, Renganathan V, Redmill KA, Özgüner Ü (2020) A vision-based social distancing and critical density detection system for covid-19. arXiv preprint arXiv:2007.03578 pp 24–25

Yuan Y, Wang D, Wang Q (2016) Anomaly detection in traffic scenes via spatial-aware motion reconstruction. IEEE Trans Intell Trans Syst 18(5):1198–1209

Zhang F, Li C, Yang F (2019) Vehicle detection in urban traffic surveillance images based on convolutional neural networks with feature concatenation. Sensors 19(3):594

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.