A hybrid tiny YOLO v4-SPP module based improved face mask detection vision system

Akhil Kumar¹ · Arvind Kalia¹ · Akashdeep Sharma² · Manisha Kaushal³

Received: 20 December 2020 / Accepted: 8 October 2021

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Law offenders take advantage of face masks to conceal their identities and in the present time of the COVID-19 pandemic wearing face masks is a new norm which makes it a daunting task for the investigation agencies to identify the offenders. To address the issue of detection of people wearing face masks using surveillance cameras, we propose a novel face mask vision system that is based on an improved tiny YOLO v4 object detector. The face masks detection network of the proposed vision system is developed by integrating tiny YOLO v4 with spatial pyramid pooling (SPP) module and additional YOLO detection layer and tested and validated on a self-created face masks detection dataset consisting of more than 50,000 images. The proposed tiny YOLO v4-SPP network achieved a mAP (mean average precision) value of 64.31% on the employed dataset which was 6.6% higher than tiny YOLO v4. Specifically, for detection of the presence of a small object like a face mask on the face region, the proposed tiny YOLO v4-SPP based vision system achieved an AP (average precision) of 84.42% which was 14.05% higher than the original tiny YOLO v4 thus, ensuring that the proposed network is capable of accurate detection of a mask on the face region in real-time surveillance applications where visibility of complete face area is a guideline.

Keywords Tiny YOLO v4 · YOLO v4 · Face mask detection · SPP module

1 Introduction

The development of effective vision systems for automatic surveillance technology is a challenging and cornered problem in the area of object detection. New proposals of reliable vision systems will entrench the progress of automatic surveillance in uncontrolled environments without manual monitoring. The uncontrolled environments are surveillance

areas where people are monitored round the clock with CCTV cameras without human intervention. In the field of security, law enforcement, military applications, commercial and public setups such automatic vision systems are need of time and a security asset. For real-world applications, automatic surveillance can be achieved by integrating surveillance cameras such as CCTV cameras and UAVs with evolutionary deep neural network-based algorithms to localize, classify, and detect a wide array of objects. Amid COVID-19 wearing face masks has become a new normal so it has emerged as a new problem of detection of identities of people behind the face mask. The present-day face recognition system works by recognizing essential features of the face such as jaw area, nose area, and shape of eyebrows, etc. which tends to hide if a person wears a face mask, therefore, make the present facial recognition systems difficult to determine the identity of a person at places like ATMs, banks, biometric attendance systems having a guideline of complete vision of the face area. Furthermore, in this pandemic time government has issued a regulation of mandatory wearing of face masks in the public area as a prevention to spread the infection of COVID-19 thus, it has become a dire need of time to come up with automatic

✉ Akashdeep Sharma
akashdeep@pu.ac.in

Akhil Kumar
akhil.hpucs@gmail.com

Arvind Kalia
arvkalia@gmail.com

Manisha Kaushal
manisha.kaushal@thapar.edu

¹ Department of Computer Science, Himachal Pradesh University, Shimla, India

² CSE, UIET, Panjab University, Chandigarh, India

³ CS&ED, Thapar Institute of Engineering and Technology, Dera Bassi Campus, Patiala, India

surveillance systems that can monitor whether the people are wearing face masks in the public areas or not. Moreover, the literature (Hamdan and Mokhtar 2018; Chen and Sang 2018) indicates face masks are used by criminals to breach the security in surveillance areas by concealing their identities. Considering these research issues, in this work we effort to propose a robust face mask vision system that is capable of detecting the faces with masks, faces without masks, and masks on the face region with high detection accuracy. The proposed vision system is developed by combining conventional computational hardware with an evolutionary convolutional neural network (Lecun et al. 1998) based object detection algorithm. To carry out this work we have selected a state-of-the-art object detection algorithm and proposed improvements by enhancing its feature extraction network to perform detection of a small object like a face mask on the face area, faces with masks, and faces without masks in image sequences. The proposed face mask detection network can be integrated with present surveillance systems and facial recognition systems to detect the presence of face masks on the face region and look and recognize the residual features of the face to determine the identity of a person.

With the advent of GPUs and high-end computation resources, the commonly used methods for object detection are largely based on spatial pyramid pooling networks (Kaiming et al. 2014), R-CNN (Girshick et al. 2014), SSD (Liu et al. 2016), and YOLO series (Redmon et al. 2016; Redmon and Farhadi 2017, 2018; Bochkovskiy et al. 2020) that are based on deep neural network inspired convolutional neural networks. The present state-of-the-art YOLO v4 is a single-stage object detector that tackles object detection as a regression problem. The YOLO v4 object detector localizes and classifies an object using a back-to-back architecture by combining feature extraction and detection layers. In comparison to other algorithms, the YOLO v4 network has the fastest detection speed and accurately addresses the problem of regression. The disadvantage of YOLO series algorithms is their inability to detect small objects, and detecting face mask targets on the face of a person is a challenging task. Over the years several improvements are proposed in YOLO series algorithms and the latest is YOLO v4, which uses CSPDarknet-53 as the backbone feature extraction network, mish as the activation function, provides self-augmentation of data, and uses C-IoU as the loss function. On similar lines, authors of YOLO v4 released a small-scale version of YOLO v4 with the name tiny YOLO v4, which is compact and small and uses fewer computation resources for training on custom data. Due to the ability of generalized object detection YOLO series has shown fascinating results in various real-world applications such as general category object detection (Liu et al. 2019; Yin et al. 2020), car license plate recognition (Hendry and Chen 2019), pedestrian detection (Yi et al. 2019), road marking detection (Ye et al. 2020) and

vehicle detection (Mahto et al. 2020). Amid the COVID-19 pandemic outbreak, few researchers (Roy et al. 2020; Loey et al. 2021b) have exploited YOLO v2, YOLO v3, and tiny YOLO v3 for face mask detection tasks. Following a similar strategy, in this work, we have employed variants of YOLO and proposed a new variant of tiny YOLO v4 that is capable of detecting small objects such as a face mask for an automatic vision system with high detection accuracy. The proposed detection network is developed by improving the feature extraction network of the tiny YOLO v4 that makes it capable of accurate detection of a face mask on the face region, faces with masks, and faces without masks.

The following are the primary contributions of this work:

1. We propose a robust face masks vision system which is developed by employing an improved variant of tiny YOLO v4 network and further combining it with a spatial pyramid pooling module and an additional YOLO detection layer.
2. We evaluate and compare the performance of YOLO v1, YOLO v2, tiny YOLO v1, tiny YOLO v2, tiny YOLO v3, and tiny YOLO v4 for face mask detection tasks with the proposed network.
3. The proposed tiny YOLO v4-SPP network achieved a state-of-the-art mAP (mean average precision) of 64.31% which was 6.6% higher than tiny YOLO v4, and higher than other evaluated object detection networks in this work.
4. Specifically for detection of a face mask of the face region, the proposed tiny YOLO v4 achieved an AP (average precision) of 84.42% which was 14.05% higher than the tiny YOLO v4 and significantly higher than other employed networks in this work thus, signifying accurate detection of a small object.

This work contributes novel proposals for the development of an effective face mask vision system for automatic surveillance in uncontrolled environments that can detect faces with masks with high accuracy. The proposed tiny YOLO v4-SPP face mask detector network developed in this work have high implications in this time of COVID-19 pandemic to monitor whether the people are wearing face masks or not in the public areas with an automatic surveillance system. Furthermore, the proposed network can be combined with existing face recognition systems to develop an effective identity recognition system that can detect face masks on the face regions and recognize the identity of a person by looking at the residual features of the face.

The sections of this work are as follows: Sect. 2 presents the related work in literature; Sect. 3 describes the proposed approach; Sect. 4 set-out the experiments, result analysis, and comparison with related work; and Sect. 5 presents the conclusion and future scope.

2 Related work

The recent published literature on face recognition, face mask classification, and detection largely addresses the issue of security in surveillance environments and the classification and detection of people wearing face masks. Novel proposals by improving the existing face detectors and face masks detectors can achieve enthralling results and escalate the research in this domain of object detection. In Hamdan and Mokhtar (2018) authors proposed a 2D identity recognition system for the detection of spoofing by wearing 3D face masks to recognize the identities of the hackers. The proposed work addresses the security attack of identity theft by hackers used for breaching the face recognition systems. The proposed method was developed by extracting pertinent features of the face using Angular Radial Transformation (ART) and classified using the SVM classifier to recognize the presence of a face in the image. On the 3DMAD dataset, the proposed method achieved a classification accuracy of 63.2%. In Zhang et al. (2018) researchers proposed a face occlusion detection technique to detect arbitrarily occluded faces in ATM surveillance environments. The proposed technique use shape features of shoulder and head movement to determine the face occlusion by constructing an energy function for head contour detection. Furthermore, to track the head a Bayesian framework that uses the gradient and shape cues was utilized. Finally, to detect the occlusion of head features of skin and head were passed to the Ada-Boost classifier. On a real-world dataset, the developed technique achieved an accuracy of 98.6% for face detection and accuracy of 98.5% for face occlusion detection. The authors (Chen and Sang 2018) proposed a fraud prevention method for biometric security environments. The proposed system combined OpenCV, Dlib, and Gaussian Mixture Model (GMM) to construct the model of human faces and used the features of faces and compared with face masks to determine whether it is a human face or a face mask. The features in this model were extracted using a CNN model and classification was performed using a Gaussian Mixture Model. The proposed OpenCV + Dlib + GMM technique achieved a classification accuracy of 94.8% on the CAS-PEAL dataset. In Niu and Chen (2018) authors proposed a face detection technique for the detection of normal faces and abnormal faces i.e. faces covered by masks, occluded faces, and faces with sunglasses in security fields like residential and bank detection access. The proposed technique combines Dlib, the residual network of convolutional neural networks, and Gaussian Mixture Model (GMM) to classify normal faces and abnormal faces like faces covered by masks. On the CAS-PEAL face dataset, the proposed technique achieved an accuracy of 96%.

The authors (Ejaz et al. 2019) proposed a method based on PCA that can recognize the persons with face masks and without face masks in varied conditions. A hybrid approach to classify persons with masks and without masks was proposed in Loey et al. (2021a) by integrating the ResNet-50 feature extractor with classical machine learning algorithms. In the proposed work, the combination of ResNet-50 with SVM classifier achieved a classification accuracy of 99.6% on the RMFD dataset consisting of samples of persons with masks and without masks. The authors in Khandelwal et al. (2020) proposed a deep learning model based on MobileNetv2 image classifier capable of binarizing an image into two classes namely, mask present and mask not present. The authors trained the proposed classifier on a custom dataset consisting of 380 images for class mask present and 460 images for mask not present and achieved an AUROC value of 97.6% for the face masks classification task. The authors (Li et al. 2020) used a CNN feature extractor to learn HGL features from the images to classify the head pose with a mask. To develop this classification model authors used color texture and line portraits to distinguish between frontal view and side view having presence and absence of a mask on the face. For frontal view, this method achieved accuracy of 93.64% and for side view, the achieved accuracy was 87.17% for the presence of a mask on the face area. A face mask-wearing classification model was proposed in Qin and Li (2020) capable of quantifying the mask-wearing conditions in 2-D facial pictures. The proposed model was developed using a classification network capable of integrating image super-resolution to quantify conditions such as mask present, mask not present, and mask worn incorrectly. The authors (Nagrath et al. 2021) proposed a face masks classification model by combining Single Shot Detector (SSD) with MobileNet v2 classifier and performed classification of people wearing or not wearing face masks on a custom dataset and achieved a classification accuracy of 92.64%.

A benchmark work (Ge et al. 2017) proposed a face mask detection model by integrating locally linear embedding (LLE) with CNNs. The authors also introduced the Masked Faces (MAFA) dataset with 35,806 images. In the proposed work LLE-CNNs based model achieved an average precision of 76.1% on the MAFA dataset. The authors (Inamdar and Mehendale 2020) proposed a real-time deep learning-inspired face mask detection model achieving an accuracy of 98.6% on a custom dataset. The authors used a very small dataset consisting of 35 images where the class with mask and without mask had 10 images and class mask worn improperly had 15 images. The authors (Roy et al. 2020) proposed the MOXA dataset consisting of 3,000 images for class labels persons with masks and without masks. The authors tested for detection with tiny YOLO v3 attaining mAP of 41.06%, YOLO v3 detecting with mAP value of 63.99%, SSD attaining mAP of 46.52%, and Faster-RCNN

detecting with mAP value of 60.05%. In the proposed work authors straightforwardly implemented the object detectors on the MOXA dataset and evaluated the detection accuracy without improving the state-of-the-art to achieve better results. In Loey et al. (2021b) authors proposed a hybrid face masks detection model by integrating ResNet-50 classifier with YOLO v2 to detect medical masks on the face regions. The authors used ResNet-50 as the feature extractor network and employed the strategy of transfer learning. The feature vector produced by ResNet-50 was passed to the YOLO v2 detection network to perform the detection task. On the employed dataset the proposed hybrid model achieved an AP of 81%.

The literature indicates a scarcity of exploration of state-of-the-art object detection methods that are enhanced and improved to achieve reliable results in real-time surveillance applications. We, therefore, propose a state-of-the-art face mask vision system based on tiny YOLO v4 that is improved and enhanced to achieve accurate and better results.

3 The proposed approach

To propose an effective face mask detection system we have integrated a vision system with the proposed improved tiny YOLO v4 network. In this section, firstly in detail, the face mask vision system is introduced. Then, we present the improved backbone feature extraction network of tiny YOLO v4 that specifies our evaluation approach.

3.1 The face mask vision system

The face mask vision system developed for this work consists of three major units, i.e. the vision system, face mask detector, and alarm system. Specifically, the vision system based on a camera captures the information in form of images and videos in unsupervised surveillance environments. The information is then passed to the face mask detector which detects face masks on face regions and faces

with masks and without masks in sequences passed by the vision system. Finally, the alarm system that produces a siren sound if a face mask is detected on the face region by the face mask detector system. The face mask detector system is the most important part of the proposed face mask vision system since it is responsible for detecting the presence of a face mask on the person's face therefore, efforts are required to develop an effective face mask detector that can detect the presence of masks in varied conditions with high detection accuracy. The detailed flow of working of the proposed face mask vision system is shown in Fig. 1.

In our case, a CCTV camera is used to capture the video sequence in real-time that is connected to a computer system running the proposed tiny YOLO v4-SPP network as a pre-trained face mask detector. The pre-trained tiny YOLO v4-SPP based face mask detector draws bounding boxes for face masks in frames containing faces with masks. The system produces an alarm sound if the pre-trained detector detects a bounding box containing a face mask. In the next subsections, we focus on the tiny YOLO v4-SPP based face mask detector system, which is the most important part of the face mask vision system.

3.2 Proposed tiny YOLO v4-SPP network

The tiny YOLO v4 network is a small-scale variant of the YOLO v4 object detection algorithm. The network tiny YOLO 4 works as a single-stage object detector. The feature extraction network of any object detection method is very important; the detection accuracy of the method largely depends on the ability of the feature extraction network to extract rich features from the set of input data. If a network is capable of extracting more features, then it does detection with high accuracy. YOLO v4 uses CSPDarknet-53 as the feature extraction network and tiny YOLO v4 uses a smaller version of the same for feature extraction. The network of tiny YOLO v4 performs fast detection but struggles with detecting small objects thus, leading to false and inaccurate detections. Hence, some improvements are necessary

Fig. 1 Proposed face mask vision system

to build an effective face mask detection algorithm that is capable of detecting small objects like the presence of a face mask on the face. For our specified task of face masks detection, the proposed work enhances the feature extraction network of tiny YOLO v4 detector by combining a spatial pyramid pooling module and an additional YOLO detection layer to detect face masks with high accuracy and precision. This work combines Complete-IoU (C-IoU) loss function proposed in (Zheng et al. 2020) to achieve faster convergence and accurate regression. The anchors boxes for the proposed network were determined using a k-means++ clustering mechanism specific for the employed face masks detection dataset.

In Table 1, the bold highlighted blocks represents the spatial pyramid pooling (SPP) module, convolutional layers and the third YOLO detection layer applied to tiny YOLO v4 to develop the proposed tiny YOLO v4-SPP network.

The feature extraction network of the original tiny YOLO v4 is composed of twenty-one convolutional layers and three max-pooling layers. The tiny YOLO v4 performs detection using two YOLO detection layers which are connected after the feature extraction layers. To upgrade the tiny YOLO v4 network and improve its overall detection accuracy for face mask detection we incorporated few enhancements in its feature extraction network. We added a spatial pyramid pooling module after the feature extraction layers to extract rich features and added one more YOLO detection layer that makes the network capable of detecting small objects. Using a spatial pyramid pooling module reduces computational complexity by preventing the processing of features extracted by the convolutional layers over and over, therefore, increasing the speed of the network. The feature extraction network of the proposed tiny YOLO v4-SPP is composed of twenty-five convolutional layers and three max-pooling layers. Furthermore, a spatial pyramid pooling module consisted of three max-pooling layers of size 1×1 , 3×3 , and 5×5 is applied after the feature extraction layers that changed the dimension of feature map from $26 \times 26 \times 512$ to $13 \times 13 \times 2048$. For detection three YOLO detection layers are applied after the feature extraction network. The difference between tiny YOLO v4 and proposed tiny YOLO v4-SPP network is presented in Table 1.

The proposed network uses YOLO detection layers to predict the bounding boxes. Furthermore, anchor boxes are used to predict the bounding boxes for each object class present in the face masks detection dataset. The three YOLO detection layers of the network process the bounding box, objectness score, anchors, and class predictions. The filter size before each YOLO layer is calculated by $(classes + 5) * 3$ thus, in this work as the number of classes was 4 hence; the filter size was set to 27. As compared to the original tiny YOLO v4, the proposed network has four additional convolutional layers and an SPP module composed of three max-pooling

layers that aid in extracting rich information from the input data. Furthermore, the third YOLO detection layer enhances the accuracy of localization and prediction of objects by using a high dimension feature map of size $52 \times 52 \times 27$. The detailed architecture of the proposed tiny YOLO v4-SPP network is illustrated in Fig. 2.

3.3 Loss function

For accurate object detection, two questions always surface while designing and using a loss function: (1) To achieve faster convergence, how to straightforwardly curtail the normalized distance between the predicted box and target box; (2) How to speed up and make the regression accurate? To address these questions, researchers in (Zheng et al. 2020) proposed loss functions based on C-IoU loss. To address the task of object detection as a regression problem, the YOLO v4 object detector uses C-IoU loss function which can be represented as:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(B, B^{gt})}{c^2} + \alpha v \quad (1)$$

where IoU signifies intersection over union; $\rho(\cdot)$ represents the Euclidean distance; c signifies the diagonal length of the smallest enclosing box covering the two boxes; B represents the predicted box; B^{gt} represents the target box; α signifies a positive trade-off parameter; v is used to measure the consistency of aspect ratio.

The face mask detection task performed in this work was a bounding box regression problem thus, following the above-specified approach of C-IoU loss function we used it with our proposed network to achieve better convergence speed and detection accuracy.

3.4 Anchor boxes using k-means++ clustering

Inspired by the anchor box technique of Faster-RCNN, the series of YOLO algorithms in YOLO v2, YOLO v3, and YOLO v4 adopted the mechanism of k-means and k-means++ clustering to extract the best priors for anchor boxes on the employed dataset. The clustering mechanism helps in extracting the maximum number and dimension of the anchor boxes and prevents convergence at training time and lead to local optimization.

The k-means++ clustering mechanism to extract anchor boxes used in YOLO v4 and tiny YOLO v4 is based on the IoU score proposed in (Nowozin 2014; Ahmed et al. 2016) to determine the clustering result thus, prevents the error caused by the varying size of the bounding boxes. To compute the distance between the box and centroid, the k-means clustering method uses the Euclidean Distance Function as shown in the following equation:

Table 1 Difference between tiny YOLO v4 and tiny YOLO v4-SPP network

Type	Filters	Size/Stride	Output
Convolutional	32	3x3/2	208x208x32
Convolutional	64	3x3/2	104x104x64
Convolutional	64	3x3	104x104x64
Route			104x104x32
Convolutional	32	3x3	104x104x32
Convolutional	32	3x3	104x104x32
Route			104x104x64
Convolutional	64	1x1	104x104x64
Route			104x104x128
Maxpool		2x2/2	52x52x128
Convolutional	128	3x3	52x52x128
Route			52x52x64
Convolutional	64	3x3	52x52x64
Convolutional	64	3x3	52x52x64
Route			52x52x128
Convolutional	128	1x1	52x52x128
Route			52x52x256
Maxpool		2x2/2	26x26x256
Convolutional	256	3x3	26x26x256
Route			26x26x128
Convolutional	128	3x3	26x26x128
Convolutional	128	3x3	26x26x128
Route			26x26x256
Convolutional	256	1x1	26x26x256
Route			26x26x512
Maxpool		2x2/2	13x13x512
Convolutional	512	3x3	13x13x512
Maxpool			
Route			
Maxpool			
Route			
Maxpool			
Route			
Convolutional	256	1x1	13x13x256
Convolutional	512	3x3	13x13x512
Convolutional	27	1x1	13x13x27
YOLO			
Route			13x13x256
Convolutional	128	1x1	13x13x128
Upsample		2	26x26x128
Route			26x26x384
Convolutional	256	3x3	26x26x256
Convolutional	27	1x1	26x26x27
YOLO			
Route			26x26x256
Convolutional	128	1x1	26x26x128
Upsample		2	52x52x128
Route			52x52x256
Convolutional	256	3x3	52x52x256
Convolutional	512	3x3	52x52x512
Convolutional	27	1x1	52x52x27
YOLO

Fig. 2 Proposed tiny YOLO v4-SPP network

$$d(box, centroid) = 1 - IoU(box, centroid) \quad (2)$$

where *box* signifies the sample; *centroid* signifies the center point of the cluster; *IoU(box, centroid)* represents the overlap ratio of the cluster box and the center box.

In k-means++ based clustering firstly the centroid is picked randomly then the distance is computed for all the points in the dataset from the picked centroid. From the

farthest centroid, the distance of x_i point can be computed by following equation:

$$d = \max_{(j: 1 \rightarrow m)} \|x_i - centroid_j\|^2 \quad (3)$$

where d denotes the distance of x_i point from the farthest centroid and m is the number of centroids already picked.

Following the above strategy, we determined anchor boxes using the k-means++ clustering mechanism on the employed dataset. Considering the complexity of the network proposed in this work we choose the k value of 9 and compared IoU scores for different k values. On the employed dataset the anchor box values obtained after applying k-means++ clustering mechanism were: $[(9 \times 13), (17 \times 27), (28 \times 43), (37 \times 69), (58 \times 91), (75 \times 148), (133 \times 121), (131 \times 229), (231 \times 301)]$.

4 Experiments

The experiments were carried out using the Darknet repository (Bochkovskiy 2020) written in C language. For computation resources we used Google Colab providing us the 12 GB of RAM and Tesla K 80 GPU with CUDA 10.1 GPU acceleration library. In total, we carried out seven experiments exploiting YOLO v1, YOLO v2, tiny YOLO v1, tiny YOLO v2, tiny YOLO v3, tiny YOLO v4, and the proposed tiny YOLO v4-SPP for performance evaluation on face mask detection task. The total number of iterations to train the proposed tiny YOLO v4-SPP network and compared networks was set to 8000, the learning rate was set to 0.001 with the momentum of 0.9, the decay of 0.0005, and a batch size of 64 with a sub-division of 16. The input size was initialized to 416×416 and set to change randomly during training to obtain better detection accuracy for input images of varying sizes. To achieve accurate detection results for each class of the dataset, the confidence threshold was set to 0.25 and the IoU threshold to 50% for the proposed network and compared YOLO networks. The proposed network utilized 2 GB of RAM and 1.8 GB of GPU memory and entire training took around 3.5 h.

4.1 Dataset

To carry out this work, we used a self-created custom face masks detection dataset. The dataset was created by crawling images from the internet using Google and Bing APIs. In total, we collected approximately 11,000 images using Google and Bing APIs. The images collected from the internet were of varying size and dimensions and to meet the input size

requirement of the YOLO network, we resized the collected images to a size of 416×416 . The dataset consists of images of people wearing face masks, medical masks and people not wearing face masks. To specify our evaluation approach we annotated the collected 11,000 images with the LabelImg annotation tool (Tzutalin 2015) for four classes with labels with mask, without mask, mask worn incorrectly, and mask area. We applied approximately 50,000 bounding boxes over 11,000 images to gather rich and precise information for each class of the dataset. Furthermore, to enhance and increase the number of images of the dataset we applied data augmentation techniques such as rotation, shearing, flipping, and HSV shift. Applying data augmentation increased our dataset from 11,000 images to 52,635 images. The dataset is available on request at link. The image samples corresponding to different class labels are shown in Fig. 3.

4.2 Evaluation criteria

To specify our evaluation approach and to determine the efficacy of the proposed tiny YOLO v4-SPP network, we evaluated standard performance metrics such as AP (average precision) for each dataset class, precision, recall, F-1 score, and mAP (mean average precision) on the entire dataset. The formulae for performance metrics are shown in Eqs. 4, 5, and 6, where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negative instances.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F-1Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (6)$$

Fig. 3 Image samples from the dataset for different class labels

Table 2 Performance of proposed tiny YOLO v4-SPP network

Network	Class	AP (%)	Precision (%)	Recall (%)	F-1 Score (%)	mAP (%)
Original tiny YOLO v4	With mask	83.94	79	65	72	57.71
	Without mask	53.16				
	Mask Incorrectly	23.99				
	Mask area	70.37				
Proposed tiny YOLO v4-SPP	With mask	86.31	84	70	77	64.31
	Without mask	58.86				
	Mask Incorrectly	27.64				
	Mask area	84.42				

5 Results and analysis

In order to determine and evaluate the effectuality of the proposed tiny YOLO v4-SPP network we combined the above work, employed the above-specified custom face masks detection dataset, and divided the dataset into 80% for the train set, 10% for the test set, and 10% for the

validation set and trained and tested the network to determine the performance metrics values as shown in Table 2.

To compare the performance of the proposed network with its YOLO counterparts, we trained and tested YOLO v1, YOLO v2, tiny YOLO v1, tiny YOLO v2, tiny YOLO v3, and tiny YOLO v4 on the employed face masks detection dataset and evaluated for performance metrics. The comparative results for performance metrics achieved by the

Fig. 4 Comparative values for average precision

Table 3 Comparative values for performance metrics

YOLO variant	Precision (%)	Recall (%)	F-1 Score (%)	mAP (%)
YOLO v1	63	54	61	52.4
YOLO v2	68	59	63	55.34
Tiny YOLO v1	29	42	34	30.75
Tiny YOLO v2	33	49	39	33.78
Tiny YOLO v3	74	56	63	49.03
Tiny YOLO v4	79	65	72	57.71
Proposed	84	70	77	64.31

proposed tiny YOLO v4-SPP network and its counterparts are shown in Fig. 4 and Table 3.

Comparing the performance metrics of the proposed tiny YOLO v4-SPP with other YOLO networks employed for the face mask detection task, the method proposed has achieved good performance, the mAP (mean average precision) value reached 64.31% which was 6.6% higher than the original tiny YOLO v4, which was more than that of YOLO v1, YOLO v2, tiny YOLO v1, tiny YOLO v2, and tiny YOLO v3. The tiny YOLO v4-SPP network proposed in this research work is an incremental improvement to the feature extraction network of the original tiny YOLO v4 and its overall performance by showing fascinating results in detecting the presence of masks in the images where the face mask was existent in any area of the face. Specifically for the class mask area, the proposed network achieved an AP (average precision) of 84.42% which was 14.05% higher than tiny YOLO v4, which was more than that of YOLO v1, YOLO v2, tiny YOLO v1, tiny YOLO v2, and tiny YOLO v3 thus, enabling the proposed network capable of detecting small object such as a face mask in almost all the images of the dataset under varied conditions. Furthermore, the proposed tiny YOLO v4-SPP network achieved an intersection of union (IoU) value of 57.22% on the employed dataset indicating good detection accuracy of the object detection network. The detection results of the proposed tiny YOLO v4-SPP network on the employed dataset in varied real-world face masks wearing conditions are illustrated in Fig. 5. The advantage of the proposed network is its ability to detect people wearing or not wearing a face mask and specifically detecting the mask on any region of the face as illustrated in the results of Fig. 5a, b, d, and e that makes the network a useful tool in this time of COVID-19 pandemic. Figure 5a illustrates the results in public sparsely crowded area where all the faces were successfully identified. Furthermore, the proposed network has high implications in automatic and uncontrolled surveillance environments where there is a need to check whether the people are trying to hide their

identities by wearing a face mask as shown in the detection results of Fig. 5e and f. Both Fig. 5e and f gave an almost correct estimate about the amount of area covered by face masks. The results of the proposed network in densely crowded scenes are also presented in Fig. 5c illustrating that the proposed detector is capable of correct detection for almost all of the face instances present in image. There are few missed instances in Fig. 5c but this is primarily due to occlusions and overlapping of face masks with other parts of the body but still the performance of the detector seems convincing. The missing results in densely crowded scene also provide an opportunity to explore possible solutions in this regard. The detection accuracy and precision for such instances can be further improved by calculating small anchor boxes and training the network with image samples of occluded faces and face masks classes.

5.1 Comparison with related work

To evaluate the efficacy of the proposed tiny YOLO v4-SPP network more intuitively, we tested it on MOXA (Roy et al. 2020) face mask detection dataset. MOXA dataset is a collection of images of people wearing face masks, medical masks and people without masks. For validation of the proposed network, the test set of the MOXA dataset consisting of 200 images was selected and re-annotated and labeled for dataset classes employed in this work. For testing, the weights trained on our dataset were used and further mAP was evaluated. On the MOXA dataset, the proposed network achieved mAP of 63.41% which was higher than the methods employed by (Roy et al. 2020). The detailed validation comparison is presented in Table 4.

The comparison analysis presented in Table 4 shows that the proposed tiny YOLO v4-SPP outperforms mAP of state-of-the-art deep learning-based object detectors. Furthermore, as shown in Table 3 proposed tiny YOLO v4-SPP network performance based on the mAP advantage over its YOLO predecessors and tiny YOLO v4. To get the more intuitive performance comparison we compared the proposed tiny YOLO v4-SPP network with recent related work. The detailed comparison with recent related work is presented in Table 5.

The work carried out in this work is a detection task therefore from the comparison drawn with other related work this can be concluded that the proposed tiny YOLO v-SPP network outperforms methods employed by (Ge et al. 2017; Roy et al. 2020; Loey et al. 2021b) as shown in Tables 4 and 5. The recent related work is largely based on the employment of CNN based object detectors like F-RCNN, SSD, YOLO v2, and YOLO v3 for face masks detection task. The employed object detectors in related work have a deeper feature extraction network which requires more computation resources such as high end

Fig. 5 Detection results of tiny YOLO v4-SPP detector under varied face masks wearing conditions and varied environments

(a) Face masks detection results in sparsely crowded public area

(b) Face masks detection results on individual faces without face mask

(c) Face masks detection result in densely crowded public area

(d) Face masks detection results for medical face masks in COVID-19 pandemic

(e) Face masks detection results for presence of face mask on any region of the face

(f) Face masks detection results for criminals wearing face mask captured using a CCTV camera

Table 4 Performance comparison on MOXA dataset

Work	Network	Dataset	mAP (%)
Roy et al. (2020)	tiny YOLO v3	MOXA	56.27
	SSD 300 MobileNet v2		46.52
	F-RCNN 300 Inception v2		60.50
Proposed	tiny YOLO v4-SPP		63.41

processors and GPUs and high training time like SSD, YOLO v2, and YOLO v3 takes 8–10 h for training on a dataset with approximately 50,000 images. The tiny YOLO v4-SPP network proposed in this work has a comparatively smaller feature extraction network compared to F-RCNN, SSD, YOLO v2, and YOLO v3 therefore can be trained on a system with an Intel i5 processor, 8 GB of RAM, and 4 GB of GPU in 3–4 h on a dataset with approximately 50,000 images. This advantage of the proposed network over F-RCNN, SSD, YOLO v2, and YOLO v3 makes it capable of integrated with low computation power devices. The depth of the feature extraction network, computation power, training time, and high detection accuracy are the four most important factors for the development of an effective face mask detector. Moreover, the dataset is also equally important in development of an effective face mask detector. In supervised learning conditions, correctly labeled and richly annotated data with minimum class imbalance also aid in achieving high detection accuracy. The results as shown in Tables 3, 4 and 5 indicates that the proposed tiny YOLO v4-SPP with a twenty-five convolution layer-based feature extractor, trained in 3.5 h outperforms state-of-the-art object detectors like F-RCNN, SSD, YOLO v1, YOLO v2, and YOLO v3 by achieving high detection accuracy and average precision, therefore, proposed as a powerful and efficacious face mask detector.

6 Conclusion

In this work to propose a face mask vision system, the feature extraction mechanism of tiny YOLO v4 was improved by adding a spatial pyramid pooling module and an extra YOLO detection layer for the extraction of rich features from the input images. The k-means++ clustering mechanism was used to determine the best prior values of the anchor boxes on the employed dataset. To achieve faster and accurate regression C-IoU loss function was used. The proposed tiny YOLO v4-SPP achieved a 6.6% higher mAP value comparing to tiny YOLO v4 and higher than other tested networks on the employed dataset. Specifically for face masks, the proposed network achieved 14.05% higher AP than tiny YOLO v4 thus, enabling the proposed network suitable for the detection of the small objects. The proposed network performed better as compared to its YOLO predecessors and other state-of-the-art detectors exploited in the literature. The proposed tiny YOLO v4-SPP based vision system has high implications in automatic surveillance in uncontrolled environments to detect people wearing or not wearing face masks and using face masks as an aid to hiding identities. The proposed vision system detects face masks on the face area with much high accuracy thus; it is suitable for deployment in real-world applications such as facial-biometric attendance checks, banks, and military applications. The proposed tiny YOLO v4-SPP based face mask detector struggles to detect the faces with masks in crowded scenes where a large number of faces and masks are present at distance with partial visibility. Improvements can be incorporated in future work by developing a robust and more rich feature extraction network that is capable of detecting overlapping objects like faces and masks in images with a crowd of people wearing face masks and for other classes like where the face masks are worn incorrectly. The proposed work has the future scope of integration with Generative Adversarial Networks (GANs) to generate the covered

Table 5 Comparison with related work

Work	Method	Dataset	Task	Result
Ge et al. (2017)	LLE-CNNs	MAFA	Detection	AP = 76.1%
Ejaz et al. (2019)	PCA	ORL face dataset simulated with face masks	Classification	Accuracy = 70%
Roy et al. (2020)	YOLO v3	MOXA	Detection	mAP = 63.99%
Loey et al. (2021a)	ResNet-50 + SVM	RMFD	Classification	Accuracy = 99.64%
Loey et al. (2021b)	YOLO v2 + ResNet-50	Medical Masks Dataset (https://www.kaggle.com/vtech6/medical-masks-dataset)	Detection	AP = 81%
Nagrath et al. (2021)	SSD MobileNet v2	Custom (Images collected from various sources like Kaggle etc.)	Classification	Accuracy = 92.64%
Proposed	tiny YOLO v4-SPP	Custom self-created	Detection	AP = 86.31% mAP = 64.31%

face mask area that will help in generating the complete image of the face of the human behind the mask.

Acknowledgements Not applicable.

Author contributions AkK and AS planned and executed the experimental work. AkK had additional responsibility of drafting the manuscript with result analysis and plotting of graphs. MK helped in drafting and proof-reading the manuscript. ArK supervised the entire work with additional responsibility of giving this manuscript a final shape.

Funding The authors are thankful to All India Council of Technical Education, India for funding this work. This research work is funded under Research Promotion Scheme of AICTE, India vide file No. 8-108/FDC/RPS(POLICY-1/2019-20).

Availability of data and material The custom face masks detection dataset employed in this work is available on request at link.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Consent for publication Not applicable.

References

- Ahmed F, Tarlow D, Batra D (2016) Optimizing expected intersection-over-union with candidate-constrained CRFs. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp 1850–1858. <https://doi.org/10.1109/ICCV.2015.215>
- Bochkovskiy A (2020) Darknet: open source neural networks in C. <https://github.com/AlexeyAB/darknet>
- Bochkovskiy A, Wang CY, Lio HY (2020) YOLOv4: optimal speed and accuracy of object detection. arXiv. <https://arxiv.org/abs/2004.10934>
- Chen Q, Sang L (2018) Face-mask recognition for fraud prevention using Gaussian mixture model. *J vis Commun Image Represent* 55:795–801. <https://doi.org/10.1016/j.jvcir.2018.08.016>
- 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), pp 1–5. <https://doi.org/10.1109/ICASERT.2019.8934543>
- Ge S, Li J, Ye Q, Luo Z (2017) Detecting masked faces in the wild with LLECNNs. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp 426–434. <https://doi.org/10.1109/CVPR.2017.53>
- Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv. <https://arxiv.org/abs/1311.2524>
- Hamdam B, Mokhtar K (2018) The detection of spoofing by 3D mask in a 2D identity recognition system. *Egypt Inform J* 19(2):75–82. <https://doi.org/10.1016/j.eij.2017.10.001>
- Hendry CRC (2019) Automatic license plate recognition via sliding-window darknet-YOLO deep learning. *Image vis Comput* 87:47–56. <https://doi.org/10.1016/j.imavis.2019.04.007>
- Inamdar M, Mehendale N (2020) Real-Time face mask identification using face masknet deep learning network. SSRN. <https://doi.org/10.2139/ssrn.3663305>
- Kaiming H, Xiangyu Z, Shaoqing R, Jian S (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. arXiv. <https://arxiv.org/abs/1406.4729>
- Khandelwal P, Khandelwal A, Agarwal S (2020) Using computer vision to enhance safety of workforce in manufacturing in a post COVID world. arXiv. <https://arxiv.org/abs/2005.05287>
- Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324. <https://doi.org/10.1109/5.726791>
- Li S, Ning X, Yu L, Zhang L, Dong X, Shi Y, He W (2020) Multi-angle head pose classification when wearing the mask for face recognition under the COVID-19 coronavirus epidemic. In: 2020 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS), Shenzhen, China, 2020 pp 1–5. <https://doi.org/10.1109/HPBDIS49115.2020.9130585>
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg A (2016) SSD: single shot MultiBox detector. arXiv. <https://arxiv.org/abs/1512.02325>
- Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Xinwang L, Pietikäinen M (2019) Deep learning for generic object detection: a survey. *Int J Comput vis* 128:261–318. <https://doi.org/10.1007/s11263-019-01247-4>
- Loey M, Manogaran G, Taha MHN, Khalifa NEM (2021a) A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. *Measurement* 167:108288. <https://doi.org/10.1016/j.measurement.2020.108288>
- Loey M, Manogaran G, Taha MHN, Khalifa NEM (2021b) Fighting against COVID-19: a novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. *Sustain Cities Soc* 65:102600. <https://doi.org/10.1016/j.scs.2020.102600>
- Mahto P, Garg P, Seth P, Panda J (2020) Refining Yolov4 for vehicle detection. *Int J Adv Res Sci Eng Technol* 11(5):409–419. <https://doi.org/10.34218/IJARET.11.5.2020.043>
- Nagrath P, Jain R, Madan A, Arora R, Kataria P, Hemanth J (2021) SSDMNv2: a real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2. *Sustain Cities Soc* 66:102692. <https://doi.org/10.1016/j.scs.2020.102692>
- Niu G, Chen Q (2018) Learning an video frame-based face detection system for security fields. *J vis Commun Image Represent* 55:457–463. <https://doi.org/10.1016/j.jvcir.2018.07.001>
- Nowozin S (2014) Optimal decisions from probabilistic models: the intersection-over-union case. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 2014, pp 548–555. <https://doi.org/10.1109/CVPR.2014.7>
- Qin B, Li D (2020) Identifying facemask-wearing condition using image super-resolution with classification network to prevent COVID-19. *Sensors* 20(18):5236. <https://doi.org/10.3390/s20185236>
- Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp 6517–6525. <https://doi.org/10.1109/CVPR.2017.690>
- Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement. arXiv. <https://arxiv.org/abs/1804.02767>
- Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA 2016, pp 779–788. <https://doi.org/10.1109/CVPR.2016.91>
- Roy B, Nandy S, Ghosh D (2020) MOXA: A deep learning based unmanned approach for real-time monitoring of people wearing medical masks. *Trans Indian Natl Acad Eng* 5:509–518. <https://doi.org/10.1007/s41403-020-00157-z>
- Tzutalin (2015) LabelImg. Git code. <https://github.com/tzutalin/labelImg>

- Ye XY, Hong DS, Chen HH, Hsiao PY, Fu LC (2020) A two-stage real-time YOLOv2-based road marking detector with lightweight spatial transformation invariant classification. *Image vis Comput* 102:103978. <https://doi.org/10.1016/j.imavis.2020.103978>
- Yi Z, Yongliang S, Jun Z (2019) An improved tiny-yolov3 pedestrian detection algorithm. *Optik* 183:17–23. <https://doi.org/10.1016/j.ijleo.2019.02.038>
- Yin Y, Li H, Fu W (2020) Faster-YOLO: an accurate and faster object detection method. *Digit Signal Process* 102:102756. <https://doi.org/10.1016/j.dsp.2020.102756>
- Zhang T, Li J, Jia W, Sun J, Yang H (2018) Fast and robust occluded face detection in ATM surveillance. *Pattern Recognit Lett* 107:33–40. <https://doi.org/10.1016/j.patrec.2017.09.011>
- Zheng Z, Wan P, Li W, Li J, Ye R, Ren D (2020) Distance-IoU Loss: faster and better learning for bounding box regression. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp 12993–13000. <https://doi.org/10.1609/aaai.v34i07.6999>

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