Masked-face recognition using deep metric learning and FaceMaskNet-21

Rucha Golwalkar[1] · Ninad Mehendale[1]

Accepted: 24 December 2021 / Published online: 25 February 2022

Abstract

The coronavirus disease 2019 (COVID-19) has made it mandatory for people all over the world to wear facial masks to prevent the spread of the virus. The conventional face recognition systems used for security purposes have become ineffective in the current situation since the face mask covers most of the important facial features such as nose, mouth, etc. making it very difficult to recognize the person. We have proposed a system that uses the deep metric learning technique and our own FaceMaskNet-21 deep learning network to produce 128-d encodings that help in the face recognition process from static images, live video streams, as well as, static video files. We achieved a testing accuracy of 88.92% with an execution time of fewer than 10 ms. The ability of the system to perform masked face recognition in real-time makes it suitable to recognize people in CCTV footage in places like malls, banks, ATMs, etc. Due to its fast performance, our system can be used in schools and colleges for attendance, as well as in banks and other high-security zones to grant access to only the authorized ones without asking them to remove the mask.

Keywords Masked-face recognition · FaceMaskNet-21 · COVID-19 · Deep metric learning · CNN

1 Introduction

World Health Organization (WHO) recommends everyone to continuously wear masks in public, as it has been proven that wearing a facial mask plays a major role in preventing the spread of the coronavirus [1]. Removing the mask can risk the health of people as it can lead to the transmission of the virus. Wearing a face mask may lead to a significant increase in thefts and robberies as the mask hides most of the face and even humans find it difficult to identify a known face. Facial recognition is also carried out to take attendance in schools and colleges, at security checks at the airports and railway stations [2], which provides rich information for many downstream applications, including human-computer interaction [3], artificial intelligence [4, 5], robot vision [6], and intelligent control [7], and so on. The traditional face recognition systems are unable to accurately recognize the human face with a mask, as most of the meaningful facial features are hidden by the mask.

Droplets and aerial transmission of the virus caused a tremendous impact on human health during the COVID-19 pandemic [8]. Hence, taking the mask off for security checks at banks, airports, and other restricted areas can compromise the health of the people. Touching one's nose, mouth, or eyes after touching a surface contaminated with coronavirus can be a way for transmission of the virus [9]. Thus, using a fingerprint scanner to identify a person is also risky, as the virus can spread if the surface of the scanner is contaminated. Implementation of a face recognition system that recognizes masked faces in security checks at railway stations, airports, and other restricted areas will enable contactless checking, restricting the transmission of the virus. Thus, masked face recognition has become an important task for researchers because of its diverse applications.

Most of the existing systems perform face recognition with high accuracy. However, the difference in appearances, head poses, skin color, lighting conditions, etc. can hamper the system's performance [10]. Image denoising can be used to further increase the accuracy of the system [11]. But the efficacy of these systems is affected when these are applied to recognize masked faces. Face masks cover many of the important features of the human face, such as the mouth, nose, cheeks, that contribute largely to the process

✉ Ninad Mehendale
ninad@somaiya.edu

[1] K. J. Somaiya College of Engineering, Vidyavihar,
Mumbai, 400077, India

of face recognition. This makes masked face recognition very tough. Due to COVID-19, face masks have become very popular in use and are available in a variety of colors, shapes and may contain designs. This makes it even more difficult for the system to separate the partially visible part of the human face from a mask. Thus, face recognition of people wearing masks is even more challenging to perform due to the variety of colors and patterns of the face masks. Many methods for face recognition have been developed over the years. However, the problem of masked face detection has been rarely addressed until now, because the need to recognize people with the mask was not much until the outbreak of COVID-19. We have implemented a face recognition system using deep metric learning and FaceMaskNet-21 network to recognize masked faces in real-time videos and from static images. Our system achieved accuracy up to 88.92%. Along with the self-developed FaceMaskNet-21, we have also used a few pre-existing state-of-the-art deep learning models to carry out the process of masked face recognition. The Inception-v4 model gave the second-best accuracy of 82.12% and only showed a marginal difference when compared with the accuracy achieved using our FaceMaskNet-21. Hence, it is best suited to recognize masked faces among the existing state-of-the-art models.

The FaceMaskNet-21 developed by us used a deep metric learning technique to give a 128-d output feature vector, which was achieved by a FaceMaskNet-21 network trained using quadruplets. Many of the existing systems, such as [12] use triplets in order to recognize faces. However, the use of triplets shows weaker performance and hence, we have proposed the use of quadruplets to produce 128-d encodings for the process of masked face recognition. Even though quadruplets have been used previously in other projects [13], quadruplets have not been employed for the recognition of masked faces from images and live video streams until now.

As shown in Fig. 1, we convert all the masked images from the dataset into 128-d face encodings using FaceMaskNet-21. 128-d output feature vectors are encoded from the faces in the input image or live video stream with the help of FaceMaskNet-21. Comparison between face encodings of input with the known face encodings from the dataset is done by calculating the Euclidian distance. Finally, the face image along with the associated name is displayed on the screen based on minimum error. FaceMaskNet-21 network architecture was initially trained using the Labeled Faces in the Wild (LFW) dataset [14]. We then optimized the network using our own dataset comprising images of masked faces. Thus, it makes the proposed system capable of recognizing masked faces by comparing the face encodings of the masked face from the input image or live video stream with that of the masked face in the dataset. Hence, our system is very useful in the light of the COVID-19 pandemic where wearing a face mask has become mandatory in order to prevent the spread of the virus. The FaceMaskNet-21 network developed by us is fast and has only a few layers and thus, it can be used in portable embedded systems.

2 Related work

In order to overcome the problems faced by conventional face recognition systems, various systems have been developed that can be employed to recognize masked

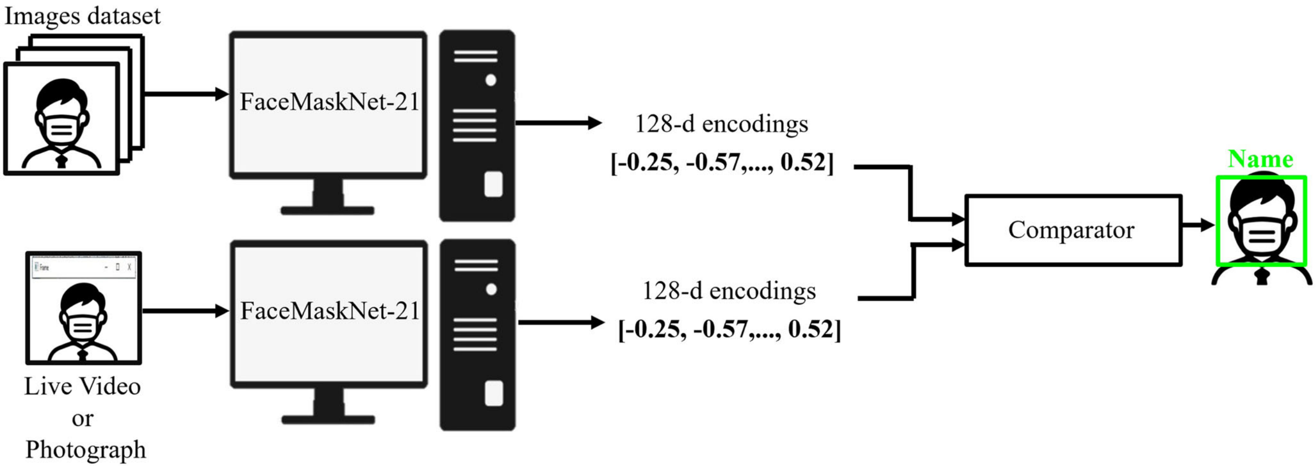


Fig. 1 Concept diagram for masked face recognition using deep metric learning and FaceMaskNet-21. FaceMaskNet-21 converts images from the dataset (during training) and input images or live video (during testing) into 128-d encodings. The encoding achieved (from the test image) is then compared to all encodings achieved from the dataset and the recognized person's name from the dataset is shown based on minimum error decoding

faces. The existing networks such as the ResNet-50 model have been retrained to increase the accuracy of masked face recognition [15, 16]. However, the efficiency can be increased further.

Li et al. [17] have developed a masked face recognition system that focuses on the area around the eyes of a person by cropping the input image. A Convolutional Block Attention Module (CBAM) has been employed to make the system capable of paying more attention to the areas around the eyes that are not hidden by the face mask. The authors have reported that the optimal cropping performed by the proposed system is more efficient when compared to the state-of-the-art methods. Another system has been developed by Hariri et al. [18] that focuses on the unmasked regions of the face such as the forehead and the eyes. The masked area of the face is first discarded and a pre-trained Convolutional Neural Network (CNN) is applied to extract the visible features of the face. The developed system recognizes the masked faces with an accuracy of 91.3%.

Anwar et al. [19] have focused on the development of a tool to generate masked faces by identifying the key points of the face that gets covered by the face mask. This tool has been tested on the existing Facenet system for masked as well as unmasked faces. The conventional face recognition system that failed while recognizing the masked faces, showed an increase in efficiency by approximately 38%. Vu et al. [20] have employed a combination of CNN and Local Binary Pattern (LBP) to extract the features such as eyes and eyebrows from the masked input images. The efficiency of the system is better than the state-of-the-art models. However, the developed system cannot be employed on portable embedded systems due to relatively high energy consumption. The robustness of masked face recognition has been improved by Sha et al. [21] by proposing a face alignment network that uses a data augmentation module.

3 Literature review

Traditional face recognition (non-masked faces) has gained more and more importance over the years due to its extensive applications. Many methods have been developed which mainly involve the extraction of facial features to recognize non-masked faces from images. Face recognition can be a tedious process due to the variation in illumination and misalignment of faces. Thus, Wagner and his team [22] used a projector to illuminate the subject's face in the image to accurately recognize the face. Another method developed by Weyrauch et al. [23] involves the use of morphable models and needs fewer images per subject in the dataset. Emami et al. [24] have implemented face recognition using OpenCV, which requires a comparison of the faces in the input image with the known faces in the database. This technique is becoming increasingly popular. Face recognition can be challenging due to different illumination conditions and head positions of the people in the images.

Goudai et al. [25] proposed a face recognition method based on the computation of local autocorrelation coefficients by extracting features. A database consisting of 11,600 face images of 116 people has been used to test the performance of the system. The feature vectors are classified with the help of linear discriminant analysis. The input feature vectors are rejected if the measure falls below a certain threshold. Gao et al. [26] has presented a technique to align faces accurately by overcoming the occlusions and variations in illuminations in the images. Local Binary Pattern (LBP) is used to subside the shape parameters of the facial input images. A Histogram of Oriented Gradient (HOG) is employed to note the landmark positions of the LBP images. Further weighted integration was performed to successfully align faces. Another method to overcome the problem of pose variations in face recognition has been proposed by He et al. [27]. The proposed Deformable Face Net (DFN) involves deformable convolution models to extract features and learn the alignment of faces to minimize the difference in the features caused due to the variations in poses. Lu et al. [28] in their manuscript, have employed Direct Linear Discriminant Analysis (D-LDA) in combination with the Fractional step LDA (F-LDA) approach called DF-LDA. The face recognition method utilizes D-LDA to remove the null space of the between-class scatter matrix. This system is less likely to overfit. Another method proposed by Nazeer et al. [29] employs feature extraction algorithms. Before extracting the features, the face recognition system pre-processes and normalizes the input image by applying histogram methods to reduce the variations in face illuminations. Three classification techniques, such as Principle Component Analysis (PCA), LDA, and Euclidean distance, have been employed, out of which the Euclidean distance classifier outperforms the other two. Huang et al. [30] has developed a two-level system that computes a 3D face model. The first level consists of component classifiers. The output of the component classifiers was used as inputs to a combination classifier to perform the detection of the face. Moon et al. [31] have developed a face recognition system based on a convolution neural network that can be used for surveillance to detect multiple faces at different distances. Their work focuses on improving the accuracy of the system when the face image is obtained from far. Deep Learning (DL) has successful applications in the fields of image classification [32] and natural language processing [33, 34]. It has also achieved considerable progress in DL-based masked face recognition tasks. Another technique to overcome the problem faced during face recognition due to the large distance between the human and camera has been proposed by Gao et al. [35]. The Multi-Scale

Patch-based Representation Feature Learning (MSPRFL) was trained using features of each patch that were robust with respect to the resolution. The results of all patches were fused based on the votes to recognize facial images of low resolutions.

All the systems reported in the literature work very well for face detection without the mask. On the other hand, when we tried to apply the same systems to masked faces, these algorithms failed. To overcome the issue of detecting masked faces, we decided to implement an algorithm that focuses on masked faces only. Our proposed method creates 128-d face encodings using FaceMaskNet-21 for face recognition followed by Euclidean distance to match the input encodings with the known encodings. Our face recognition system is capable of recognizing masked faces that can be very helpful for security and surveillance purposes during the COVID-19 pandemic.

4 Methodology

We propose a system that can implement face recognition on masked faces for both static images and real-time videos. For this purpose, we used our deep learning network called FaceMaskNet-21. The network was tuned to produce 128-d long encodings. The execution of the proposed method used OpenCV, python, and deep learning libraries. To produce accurate and speedy results on a face with a mask, we used deep metric learning. It is a different technique compared to traditional deep learning because instead of taking one input and classifying it; we output real floating-point numbers of length 128 (128-d encodings). The training of the network was done using a set of images called quadruplet. In this quadruplet (shown in Fig. 2), there are two images of two persons each wearing a face mask. FaceMaskNet-21 quantifies the face input image and constructs 128-d long feature encodings. The FaceMaskNet-21 neural network keeps trying to tweak the weights until both people's images are distinctly away in terms of their 128-d encoding. The proposed method performed better than the systems that use triplets to train the network, as the two images of each person wearing masks in the quadruplet helped train the network to improve its performance. Furthermore, the use of masked faces as quadruplets enabled the network to recognize masked faces successfully. Masked face recognition is very difficult since most of the facial features are hidden by the mask. We can see distinctly only the features of eyes and eyebrows when people are wearing facial masks. Thus, we have designed FaceMaskNet-21 in such a way that it mainly takes into consideration the unmasked features of the face, such as eyes and eyebrows, for the process of masked face recognition. The FaceMaskNet-21 was initially trained with approximately 50,000 images of the Labeled Faces in the Wild (LFW) dataset [14]. Subsequently, the trained network was further optimized for masked faces with 13 people and 204 images. We generated an auxiliary dataset of 2000 images for each class (13 people) using operations, such as crop, rotate, and magnify. For testing, we used images of the same 13 people, but this time 25 images of each person were utilized. Each person was wearing a face mask that was different in type, shape, color, and was of different texture in those 25 images. It was important because this ensured that the network was not considering the mask as a feature. The

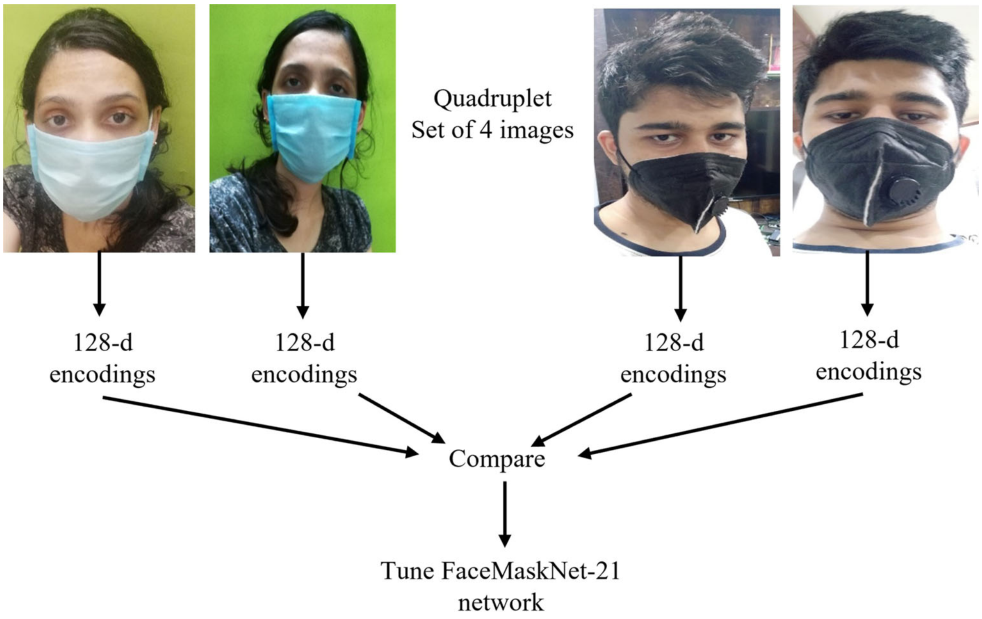


Fig. 2 Tuning process of the FaceMaskNet-21 network. A set of four images called quadruplets is used to tune the FaceMaskNet-21 network. Each image in the quadruplet is converted into 128-d encodings and we compare these face encodings with each other for tuning the network

regular facial tuning acted as a coarse tuning and the final local tuning with a masked face acted as fine-tuning for the network. The testing was done with people from the local dataset.

In order to further test the accuracy of the proposed network, we created an auxiliary dataset that comprised 1,992 images. We chose 13 adults and 7 children as subjects while creating this dataset. The FaceMaskNet-21 network was trained using 1192 images of 13 adults and 800 images of 7 children. The masked face images of children were taken as subjects in our dataset to check the accuracy since children may have fewer features of the face that are exposed while wearing a mask.

For contrast testing, all the images of those 13 people were modified to obtain different contrast ratios. In all, there were five different contrast ratios tested to see the effect of contrast variation on our proposed solution. The different contrast ranges selected were 0 to 255, 50 to 200, 100 to 200, 100 to 150, and original contrast. To test the performance of the proposed system, we also compared the proposed algorithm with 6 other algorithms, AlexNet, GoogleNet, ResNet-18, VGG-16, VGG-19, and Inception-v4 (Inception net V4).

4.1 FaceMaskNet-21 network architecture

As shown in Fig. 3, an RGB input image of dimension 227 X 227 was fed to our self-developed FaceMaskNet-21 network. The input image was passed through a convolution layer. All the smaller and larger input images were first normalized to the specified dimensions by the image input layer before passing it to the convolution layer. The convolution layer has dimensions of 55 X 55 and comprises 96 filters. The convolution layer applied sliding filters in order to extract features from the input images and detect specific patterns. The stride and dilation factor of the convolution layer was (1, 1) and the rate factor, weight learn rate factor, bias learn rate factor, and weight L2 factor were all set to 1. The bias L2 factor and bias initializer of this layer was set equal to 0 while ‘glorot’ was used as the weight initializer. After the convolution layer, the Rectified Linear Unit (ReLU) layer was used, which was followed by the cross channel normalization layer with window channel size set to 5 and alpha = 0.0001, beta = 0.75, and k = 2. The ReLU layer passed the positive values as it is and converted the negative values to zero before passing them to the next layer. The cross-channel normalization performed channel-wise normalization.

It was then passed through the convolution layer along with the ReLU layer and max-pooling layer of dimensions 27 X 27, also consisting of 96 filters. The second convolution layer had the same parameters as the first convolution layer, except for the dimensions. This layer was followed by ReLU and max-pooling layer. The max-pooling layer computed the largest value of the region that was set by the pool size of (5, 5). The max-pooling layer had a stride of (1,1).

The third convolution layer was of 13 X 13 dimensions, with 384 filters. All the other parameters of the convolution layer were the same as the previous one. This layer was followed by ReLU and max-pooling layers. The fully connected layer had an output size of 128 and flattened the input values to map it into output classes. This was followed by a Softmax layer and then the 128-d output encodings were generated.

4.2 Construction of face encodings

For the construction of 128-d face encodings, we started by importing various necessary packages. The path to the

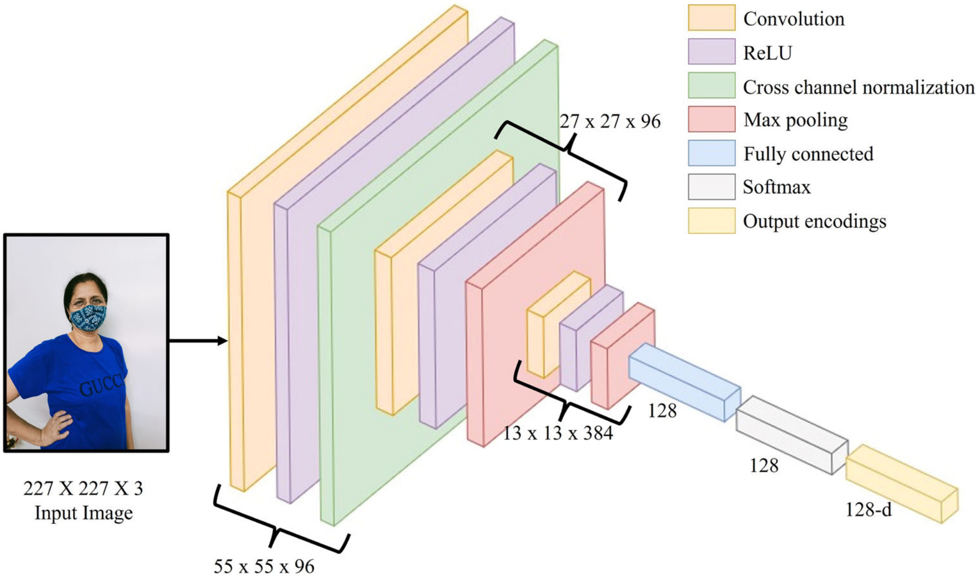


Fig. 3 FaceMaskNet-21 network architecture. An RGB input image of size 227 X 227 X 3 is passed through the Convolution + ReLU + Cross channel Normalization layer of dimensions 55 X 55 and consisting of 96 filters. After this, it is passed through the Convolution + ReLU + Max pooling layer of dimensions 27 X 27 and consisting of 96 filters. Then it is passed through the third set of Convolution + ReLU + Max pooling layer with dimensions 13 X 13 and consisting of 384 filters. This is further passed through 128 fully connected layers followed by a Softmax layer. Finally, 128-d output encodings are produced

dataset and the disk on which the face encodings were written on, were included. We generated the face encodings for all the images using CNN. Lists were initialized to hold the face encodings and the corresponding names for every individual. We iterated the number of images in the dataset and extracted the name of each person from the images. The color spaces of the images in the dataset were swapped from BGR to RGB. The x-y coordinates of the boxes bounding each face were detected, and each face was encoded into a 128-d feature vector using FaceMaskNet-21. A proper list was made consisting of the names and their corresponding encodings before dumping it to the disk in order to make it available during the comparison in real-time testing.

4.3 Face Recognition in static images

Our proposed system successfully recognized masked faces from static images. The necessary libraries, along with the file consisting of the pre-computed face encodings and names, were included to start with the face recognition process. The input image was loaded and immediately converted to RGB from BGR. After detecting the x-y coordinates of the boxes bounding the faces in the input image, facial encodings were computed for every face in the image, and a list was initialized to store the names of the detected faces. We iterated the encodings that were computed for the input image and these encodings were compared with the previously computed encodings of the dataset. The Euclidean distance between the encoding of the input face and that of all the faces in the dataset was calculated. If the distance after comparing with all possible 128-d encodings was minimum, then it indicated a face match. The names of the matched faces were stored in the list previously initialized. A count of the matched faces was maintained. We used the match with the maximum number of votes to recognize the face and the list of names was updated. This was followed by the extraction of coordinates of the bounding box around the face along with the name of the recognized face. Finally, the output image consisting of the recognized face bounded by a box along with the name of the person was displayed.

4.4 Face Recognition in live video stream and video files

Masked face recognition in live video streams and video files was similar to the masked face recognition in static images. Along with the other libraries, we used the VideoStream class from “imutils” to access the camera and take a live video stream as an input. We iterated a loop to read each frame, which was followed by a conversion from BGR to RGB and resizing of the input frame. The face encodings of the people in a single input frame were computed and then compared with the face encodings of the dataset. We found a match between the input frame and the known encodings counted and the name with the highest votes was extracted with the corresponding vote. We iterated over the recognized faces while rescaling the face coordinates and the predicted name of the recognized face was drawn on the image. We then wrote the frames on the disk along with their dimensions. Finally, the video stream with the name of the recognized face was displayed, and the desired results were achieved. Face recognition in video files was carried out similarly. However, instead of the live video stream as an input, a video file was given as an input and an output video file was generated.

4.5 Hardware used

The entire implementation was performed on intel i7 CPU with 24GB DDR4 RAM and 6GB of DDR5 dedicated graphic memory installed.

5 Results and discussions

Figure 4 (a) and (b) show the results that were obtained when the input images were tested on our masked face recognition system based on the deep metric learning technique that used FaceMaskNet-21. All the images in the dataset were first used to create 128-d face encodings and then the face encoding of the input image was compared with these known encodings. The resultant images consisted of a green bounding box that represented the detected faces, along with the name of the recognized person. The person was named ‘Unknown’ if the Euclidean distance is not within the acceptable range, which meant the face in the input image did not belong to one of the subjects from the dataset. Figure 4 (c) shows the results obtained when a live video stream was given as an input to the FaceMaskNet-21. The model could successfully recognize masked faces in the live video stream. These recognized faces were represented using a green-colored bounding box, and the associated name was displayed along with this box. Figure 4 (d) shows the system flow diagram of our proposed system. The process of masked face recognition is carried out in two phases, i.e. encoding and face recognition. The dataset is loaded and the FaceMaskNet-21 network performs 128-d face encodings on each image in the dataset. These encodings are stored on the disk and are called during the face recognition process. An image, a live video stream, or a video file were given as an input and then encoded into 128-d feature vectors. The face encodings of input are compared with the face encodings of the dataset and a recognized face image along with the associated name is displayed on the screen. This process enabled the system to

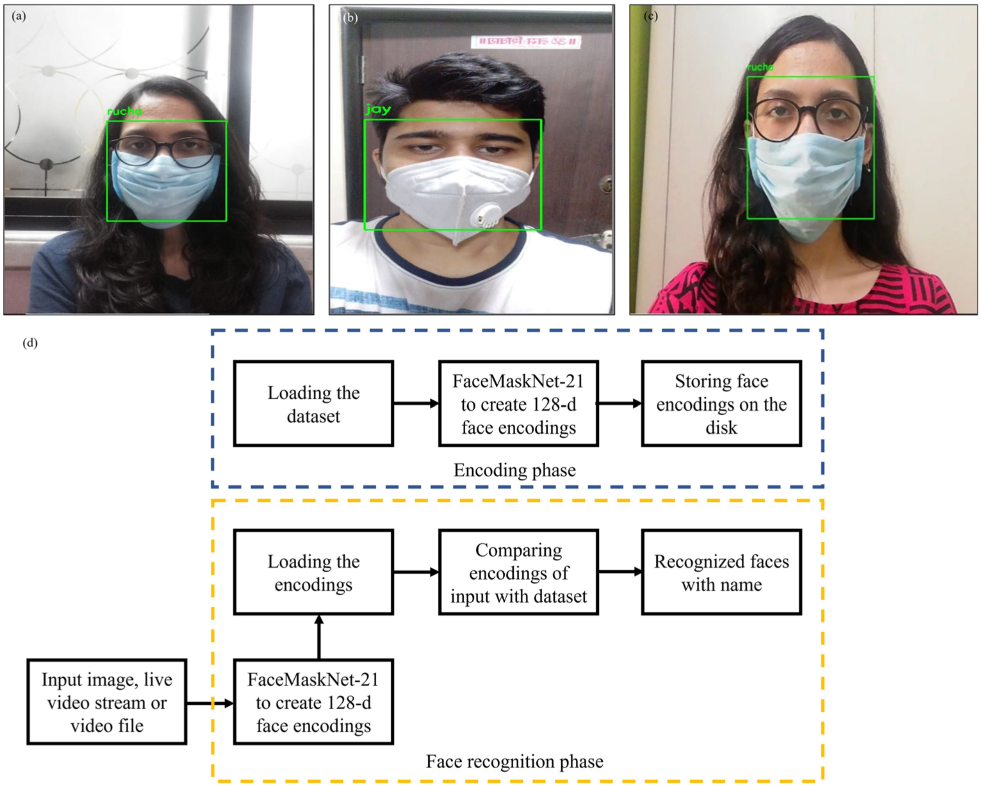

(a)
rucha
(b)
jay
(c)
rucha
(d)
Loading the dataset
FaceMaskNet-21 to create 128-d face encodings
Storing face encodings on the disk
Encoding phase
Loading the encodings
Comparing encodings of input with dataset
Recognized faces with name
Input image, live video stream or video file
FaceMaskNet-21 to create 128-d face encodings
Face recognition phase

Fig. 4 (a), (b) Results obtained when input image was tested on our masked face recognition system. The bounding box shows the person's detected face, and the associated name of the person is displayed near the box in green. (c) Result obtained when a live video stream was presented as an input to the system. The green bounding box was formed around the masked face in the live video stream with the name of the recognized person. (d) System flow diagram for masked face recognition using deep metric learning and FaceMaskNet-21. The dataset is loaded and converted into 128-d encodings using FaceMaskNet-21. The face encodings are stored on the disk. The input image or live video stream is also encoded using FaceMaskNet-21. The input and dataset encodings are compared to display the recognized name on the screen

recognize masked faces from static input images and live video streams successfully.

During the training phase, the accuracy reached with the LFW dataset was 99.76%, whereas the accuracy of FaceMaskNet-21 during local tuning with 204 images was 98.3%. Out of 325 instances, we got the correct output in 289 cases, thereby achieving an overall accuracy of 88.92%. Due to the deep metric learning method, the face recognition with the mask has become extremely accurate and also has an execution speed, such that it can be used in real-time with CCTV video recordings.

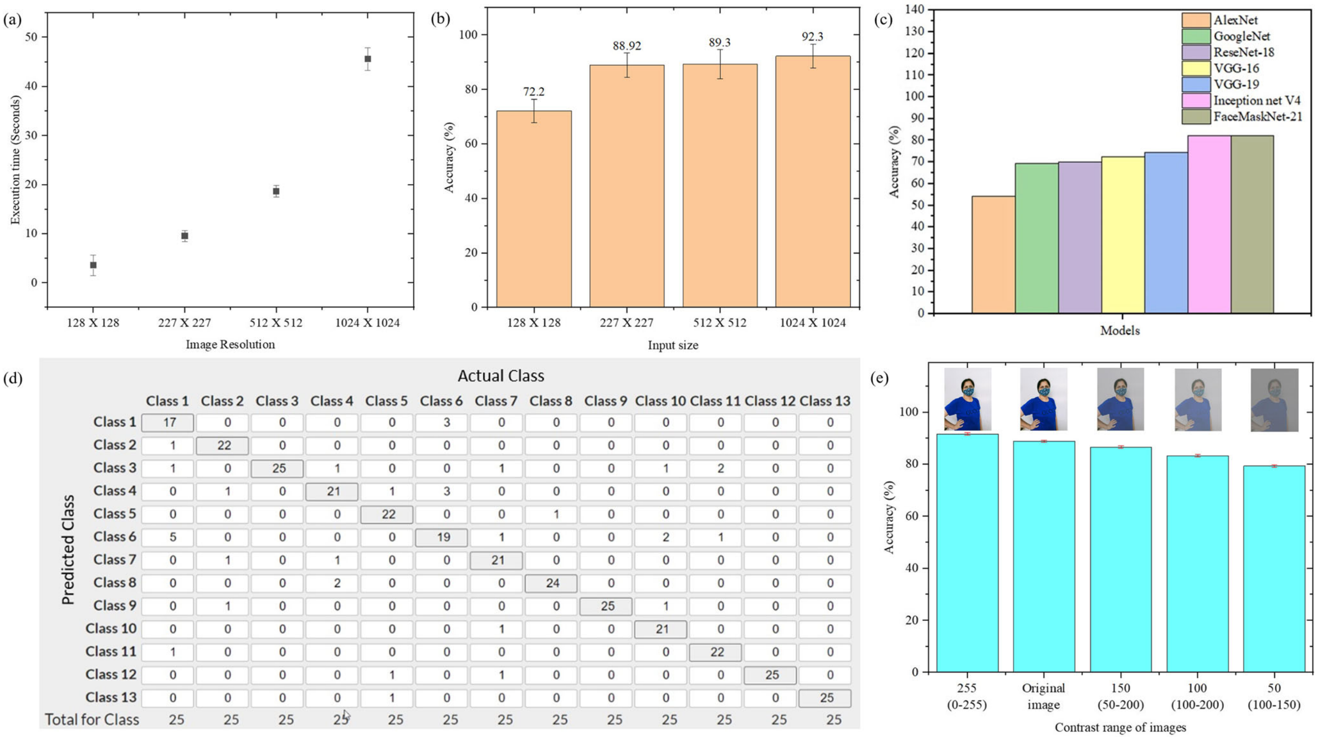
Figure 5 (a) represents a graph showing the comparison between different image resolutions with respect to execution time required. We used our own dataset consisting of 1,192 images of 13 subjects to train the FaceMaskNet-21 model. We have used 3 sets of test images, each set with 75 images. When the input images with 128 X 128 resolution were used to train the network, the lowest average execution time of 3.58 seconds was shown. The average execution time for a total of 75 images is 9.54 seconds when the input images with 227 X 227 resolution are used. The input images with 512 X 512 resolution took an execution time of 18.66 seconds. The input images of 1024 X 1024 resolution took the highest execution time of 45.6 seconds. Figure 5 (b) shows a graph presenting the accuracy achieved by FaceMaskNet-21 when it was trained using input images of different sizes. The lowest accuracy 72.2% was achieved by FaceMaskNet-21 when the input images of the size 128 X 128 were fed to it. A comparable accuracy of 88.92% and 89.3% was achieved by feeding the model with input images of sizes 227 X 227 and 512 X 512, respectively. The highest accuracy of 92.3% was achieved when the network was trained using input images of the size 1024 X 1024. Thus, we have set the input size as 227 X 227 for the input layer of the FaceMaskNet-21 to achieve high accuracy in less execution time.

Figure 5 (c) represents a graph showing the comparison between our FaceMaskNet-21 and 6 different existing deep learning networks based on the accuracy achieved by each network. All the models were trained using a Real World Masked Face Recognition Dataset (RMFRD) [2]. We obtained the samples from the website. After cleaning and labeling, it contains 5,000 masked faces of 525 people and 90,000 normal faces. We have not considered the 90,000 normal faces and only the 5,000 masked faces of 525 people were considered for masked face recognition. Thus, the accuracy achieved by each deep learning model was derived by classifying the images into 525 different classes. The proposed FaceMaskNet-21 achieved the highest accuracy of 82.22% when it was trained and tested on RMFRD. Inception net V4 model achieved the second-highest accuracy of 82.12% which was comparable with the accuracy of our network. Accuracy of 74.35% and 72.33% was seen in the VGG-19 and VGG-16 models, respectively. GoogleNet and ResNet-18 had comparable accuracies with a slight difference in them. GoogleNet achieved 69.44% accuracy, while ReseNet-18 showed an accuracy of 70.13%. The lowest accuracy of 54.23% was achieved by AlexNet. Hence, the proposed FaceMaskNet-21 achieved higher accuracy as compared to the other state-of-the-art deep learning models.

Figure 5 (d) shows the confusion matrix generated for 13 different subjects. 25 images of each person with different type, color, and shape of face masks were given as input to our proposed model. A total of 325 images were used. 4 people got 100% accurate identification whereas only 2 people were identified with less than 80% accuracy. The average precision achieved was 89.46%, while the average recall and F1 score achieved was 88.92%. It was observed that people wearing glasses with different kinds of frames were also recognized, indicating that the system is robust. Only a few cases were such that the system couldn't recognize the person. The reason for this is that in most of the cases, angled face along with other variations were present simultaneously. In all, we tested 7 videos on our system, which consisted of a single person at a time in each frame. 100% accuracy was achieved while recognizing a single face from all the video files. However, the algorithm happened to fail when multiple people with masks were present in the same frame.

When the FaceMaskNet-21 was trained using the dataset consisting of both adults and children wearing masks, the overall accuracy of 88.186% was achieved. The accuracy of the proposed system happened to reduce when the masked faces of children were included in the dataset since the children have fewer exposed facial features when they wear a face mask. This made it difficult for the network to successfully detect the masked faces.

Figure 5 (e) shows a graph comparing the accuracy versus the input images with different ranges of contrasts. The system was tested on 100 images of 13 subjects each. Before testing the network, the FaceMaskNet-21 model was trained using a dataset consisting of 1,192 images of the size 227 X 227. The contrast of the images was



	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10	Class 11	Class 12	Class 13
Class 1	17	0	0	0	0	3	0	0	0	0	0	0	0
Class 2	1	22	0	0	0	0	0	0	0	0	0	0	0
Class 3	1	0	25	1	0	0	1	0	0	1	2	0	0
Class 4	0	1	0	21	1	3	0	0	0	0	0	0	0
Class 5	0	0	0	0	22	0	0	1	0	0	0	0	0
Class 6	5	0	0	0	0	19	1	0	0	2	1	0	0
Class 7	0	1	0	1	0	0	21	0	0	0	0	0	0
Class 8	0	0	0	2	0	0	0	24	0	0	0	0	0
Class 9	0	1	0	0	0	0	0	0	25	1	0	0	0
Class 10	0	0	0	0	0	0	1	0	0	21	0	0	0
Class 11	1	0	0	0	0	0	0	0	0	0	22	0	0
Class 12	0	0	0	0	1	0	1	0	0	0	0	25	0
Class 13	0	0	0	0	1	0	0	0	0	0	0	0	25
Total for Class	25	25	25	25	25	25	25	25	25	25	25	25	25

Fig. 5 (a) Comparison graph of execution time in seconds versus the input images of different resolutions used for training the FaceMaskNet-21 model. The input image with 128 X 128 resolution shows an execution time of 3.58 seconds. (b) Comparison graph of accuracy in percentage versus the size of input images used for training the FaceMaskNet-21 model. The input image of size 128 X128 showed the lowest accuracy of 72.2%. (c) Comparison graph of accuracy in percentage versus the different existing deep learning models and the proposed FaceMaskNet-21. The FaceMaskNet-21 achieved the highest accuracy of 82.22% while Inception net V4 shows the second-highest accuracy of 82.12%. (d) Confusion matrix with the actual class on top and predicted class on left. 88.92% is the overall accuracy achieved by the system while recognizing masked faces from images or live video streams. Average precision of 89.46% was achieved along with 88.92% of average recall and F1 score. (e) Comparison graph representing the accuracy in percentage versus different contrast ranges of the input images. The highest accuracy of 91.67% was achieved when the model was tested with contrast-enhanced images with a contrast range of 0-255

initially enhanced by setting the contrast range to 0–255. The accuracy achieved while testing these images on our FaceMaskNet-21 was 91.67%. We achieved an accuracy of 86.67% and 83.33% when the contrast range of the images was set to 50–200 and 100–200, respectively. The lowest accuracy of 79.33% was achieved when the contrast range of the input images was set to 100-150 The accuracy of the system achieved while testing the images with original contrast was 88.92%. Thus, the accuracy increased when the network was presented with test images having the contrast range set to 100% i.e. 0–255.

As most of the existing methodologies are not suitable for recognizing faces covered with masks, we have presented the comparison results with systems that can perform face recognition without face masks. Table 1 shows a comparison between existing systems based on the accuracy achieved during the face recognition process and whether the systems can recognize masked faces. The comparison is also made based on the dataset used and the number of images in each dataset, along with the technique used for face recognition. The existing systems use different methods for recognizing the faces from images accurately. The system proposed by Goudail et al. [25] gave 95% accuracy while performing face recognition whereas Bah et al. [36] gave an accuracy of 99%. Although the accuracy of most of the systems is very high, these systems will not be able to perform face recognition on masked faces. We have used our own dataset consisting of 204 images of subjects wearing facial masks which was converted into an auxiliary dataset of 2000 images of 13 people each. Training the FaceMaskNet-21 using this dataset enabled the model to recognize masked faces with an accuracy that is comparable with the accuracy of other existing systems. Our system achieved an overall accuracy of 88.92% with our auxiliary dataset of 2000 images and can recognize masked faces as well.

Table 1 A comparison table with the existing systems, based on the accuracy obtained during face recognition and whether or not the systems can recognize masked faces

Method	Accuracy (%)	Masked face recognition	Dataset	Number of images in dataset	Technique
Goudail et al. [25]	95	No	User dataset	11600	Local Autocorrelations Multiscale Integration
Huang et al. [30]	90	No	User dataset	1200	SVM 3D morphable model
Moon et al. [31]	88.90	No	IPES - 1280 face DB	–	CNN with multiple distance face
Bah et al. [36]	99	No	User dataset	–	LBP
Hariri et al. [18]	91.30	Yes	RMFRD	5000 masked face images 90000 non-masked face images	VGG-16 pre-trained model
Lu et al. [28]	–	No	ORL UMIST	400 - ORL 575 - UMIST	DF-LDA
Hwang et al. [37]	81.49	No	FRGC	12776	Fourier-based LDA
Proposed method	88.92 82.22	Yes	User dataset RMFRD	2000 - User dataset 5000 - RMRD masked face images	FaceMaskNet-21
	88.186		User dataset including images of children	1992 - User dataset including images of children	

The table also shows a comparison based on the technique and dataset used, along with the number of images in each dataset. All the methods perform face recognition with high accuracy. However, not all the existing methods are able to recognize masked faces. The proposed system recognizes masked faces successfully with high accuracy

Only one of the existing systems can recognize masked faces accurately (Table 1). The accuracy reported by Hariri et al. [18] achieved is 91.30%, which is similar to our work. In their project, they have applied transfer learning with a pre-trained VGG-16 model. On the other hand, our proposed system uses a self-developed FaceMaskNet-21 network. The use of VGG-16 makes it difficult for the system to be employed on mobile phones, ARM and other portable embedded systems. On the other hand, FaceMaskNet-21 is smaller and faster and thus, our proposed system can be employed easily in portable embedded systems.

In the light of the COVID-19 pandemic, it has become important to develop face recognition systems that can recognize masked faces in the images. The FaceMaskNet-21 successfully overcomes the problem faced by the existing face recognition systems about not being able to recognize masked faces. The network recognizes masked faces in static images as well as live video streams with high accuracy. This makes the system suitable to be deployed in security areas at airports, railway stations, etc., where masked faces can be recognized instantaneously from live CCTV footage. The self-developed network has very few layers and hence, can be employed in portable embedded systems such as mobile phones. The ability to recognize masked faces accurately makes the proposed system fit to be used in the COVID-19 pandemic where it has become compulsory to continuously wear face masks in public in order to reduce the spread of coronavirus.

6 Conclusions

Our work recognizes the face of a person in an input image, live video stream, or a video file and achieved an overall accuracy of 88.92% when it was tested on our dataset. We converted the faces in the input images or videos into 128-d encodings using our FaceMaskNet-21 model. These encodings were then compared with the known encodings of the faces from the dataset. Euclidean distance was employed to decide whether the input face encoding matched with the known face encoding. Finally, a green bounding box was created around the person's face and the recognized name was displayed on the screen. The FaceMaskNet-21 when tested on RMFRD, achieved an accuracy of 82.22% which was better as compared to the other existing state-of-the-art techniques. The COVID-19 pandemic has forced people to wear facial masks to avoid the spread of the virus in society. This has added to the challenges faced by face recognition systems because along with the variations in imaging conditions, masked faces make the face recognition process tougher as most of the facial attributes such as nose, mouth, that contribute significantly to the face recognition process are hidden by the facial mask. Our deep metric learning-based face recognition system can recognize people wearing masks. This makes it fit for various applications in the COVID-19 pandemic. However, it was observed that the system failed to recognize masked faces when the frame consisted of angled faces or other variations. The algorithm happened to fail when multiple people were present in a single frame of the live video stream or video file. The self-developed FaceMaskNet-21 model gives instantaneous results as the execution time is about 9.52 ms to predict results. Thus, the proposed system can be effectively used to recognize faces from CCTV video recordings in malls, markets, airports, and high-security areas. It can also be employed to take attendance in schools and colleges. Security authentication can be carried out using this system instead of using fingerprint scanners to identify a person, as it will avoid the transmission of the corona virus since it spreads through contact. Masked face recognition is a new and unique idea and hence, it leaves scope for improvement. Occlusion handling can be implemented in the proposed system, which will further increase the accuracy of the system and the system will be able to recognize multiple masked faces present in a single input frame.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10489-021-03150-3.

Acknowledgements Authors would like to thank all the volunteers who gave their photos with mask for training of FaceMaskNet-21.

Author Contributions Conceptualization was done by R. Golwalkar (RG) and N. Mehendale (NM). All the literature reading and data gathering were performed by RG. All the experiments and coding was performed by RG and NM. The formal analysis was performed by NM. Manuscript writing- original draft preparation was done by RG. Review and editing was done by NM. Visualization work was carried out by RG and NM.

Funding No funding was involved in the present work.

Code Availability Code will be made available upon reasonable request to the authors.

Declarations

Ethics approval All authors consciously assure that the manuscript fulfills the following statements: 1) This material is the authors' own original work, which has not been previously published elsewhere. 2) The paper is not currently being considered for publication elsewhere. 3) The paper reflects the authors'own research and analysis in a truthful and complete manner. 4) The paper properly credits the meaningful contributions of co-authors and co-researchers. 5) The results are appropriately placed in the context of prior and existing research.

Consent to participate This article does not contain any studies with animals or humans performed by any of the authors. The informed consent was taken from all the participants whose photos were used in the manuscript. All the necessary permissions were obtained from Institute Ethical committee and concerned authorities.

Consent for Publication Authors have taken all the necessary consents for publication from participants wherever required.

Conflict of Interests Authors R. Golwalkar and N. Mehendale declare that there has been no conflict of interest.

References

1. Organization WH et al (2020) Advice on the use of masks in the context of covid-19: interim guidance, 5 June 2020. Tech. rep., World Health Organization
2. Wang Z, Wang G, Huang B, Xiong Z, Hong Q, Wu H, Yi P, Jiang K, Wang N, Pei Y et al (2020) Masked face recognition dataset and application. arXiv:2003.09093
3. Liu H, Fang S, Zhang Z, Li D, Lin K, Wang J (2021) MFDNet: Collaborative Poses Perception and Matrix Fisher Distribution for Head Pose Estimation. IEEE Transactions on Multimedia
4. Li D, Liu H, Zhang Z, Lin K, Fang S, Li Z, Xiong NN (2021) CARM: Confidence-aware recommender model via review representation learning and historical rating behavior in the online platforms. Neurocomputing 455:283–296
5. Shen X, Yi B, Liu H, Zhang W, Zhang Z, Liu S, Xiong N (2019) Deep variational matrix factorization with knowledge embedding for recommendation system. IEEE Transactions on Knowledge and Data Engineering
6. Liu T, Liu H, Li Y, Zhang Z, Liu S (2018) Efficient blind signal reconstruction with wavelet transforms regularization for educational robot infrared vision sensing. IEEE/ASME Transactions on Mechatronics 24(1):384–394
7. Liu T, Liu H, Li YF, Chen Z, Zhang Z, Liu S (2019) Flexible FTIR spectral imaging enhancement for industrial robot infrared vision sensing. IEEE Transactions on Industrial Informatics 16(1):544–554
8. Kumar D, Malviya R, Sharma PK (2020) Corona virus: a review of COVID-19. EJMO 4(1):8–25
9. De Campos Tuñas IT, Da Silva ET, Santoro Santiago S, Maia KD, Silva-júnior GO (2020) Coronavirus disease 2019 (COVID-19): A preventive approach to Dentistry. Rev Bras Odontol 77:E1766
10. Deng X, Zhu Z, Chang J, Ding X (2021) Algorithm Research of Face Recognition System Based on Haar. In: Advances in Computer Science and Ubiquitous Computing, pp. 317–323. Springer
11. Zeng J, Qiu X, Shi S (2021) Image processing effects on the deep face recognition system. Math Biosci Eng 18(2):1187–1200
12. Kang BN, Kim Y, Kim D (2017) Deep convolutional neural network using triplets of faces, deep ensemble, and score-level fusion for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 109–116
13. Karaman K, Akkaya İB, Solmaz B, Alatan AA (2020) A Face Recognition Technique by Representation Learning with Quadruplets. In: 2020 28th Signal Processing and Communications Applications Conference (SIU) (IEEE), pp. 1–4
14. Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech Rep, 07-49, University of Massachusetts, Amherst
15. Mandal B, Okeukwu A, Theis Y (2021) Masked Face Recognition using ResNet-50. arXiv:2104.08997
16. Montero D, Nieto M, Leskovsky P, Aginako N (2021) Boosting Masked Face Recognition with Multi-Task ArcFace. arXiv:2104.09874
17. Li Y, Guo K, Lu Y, Liu L (2021) Cropping and attention based approach for masked face recognition. Appl Intell 51(5):3012–3025
18. Hariri W (2021). Efficient masked face recognition method during the covid-19 pandemic. arXiv preprint arXiv:2105.03026
19. Anwar A, Raychowdhury A (2020) Masked face recognition for secure authentication. arXiv:2008.11104
20. Vu HN, Nguyen MH, Pham C (2021) Masked face recognition with convolutional neural networks and local binary patterns. Applied Intelligence, pp 1–16
21. Sha Y, Zhang J, Liu X, Wu Z, Shan S (2021) Efficient face alignment network for masked face. In: IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (IEEE), vol 2021, pp 1–6
22. Wagner A, Wright J, Ganesh A, Zhou Z, Ma Y (2009) Towards a practical face recognition system: Robust registration and illumination by sparse representation. In: IEEE Conference on Computer Vision and Pattern Recognition (IEEE), vol 2009, pp 597–604
23. Weyrauch B, Heisele B, Huang J, Blanz V (2004) Component-based face recognition with 3D morphable models. In: Conference on Computer Vision and Pattern Recognition Workshop (IEEE, vol 2004, pp 85–85
24. Emami S, Suciu VP (2012) Facial recognition using OpenCV. Journal of Mobile, Embedded and Distributed Systems 4(1):38–43
25. Goudail F, Lange E, Iwamoto T, Kyuma K, Otsu N (1996) Face recognition system using local autocorrelations and multiscale integration. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(10):1024–1028
26. Gao N, Wang X, Wang X (2019) Multi-layer progressive face alignment by integrating global match and local refinement. Appl Sci 9(5):977
27. He M, Zhang J, Shan S, Kan M, Chen X (2020) Deformable face net for pose invariant face recognition. Pattern Recogn 100:107113
28. Lu J, Plataniotis KN, Venetsanopoulos AN (2003) Face recognition using LDA-based algorithms. IEEE Transactions on Neural networks 14(1):195–200
29. Nazeer SA, Omar N, Khalid M (2007) Face recognition system using artificial neural networks approach. In: International Conference on Signal Processing, Communications and Networking (IEEE), vol 2007, pp 420–425
30. Huang J, Heisele B, Blanz V (2003) Component-based face recognition with 3D morphable models. In: International Conference on Audio-and Video-Based Biometric Person Authentication (Springer), pp. 27–34
31. Moon HM, Seo CH, Pan SB (2017) A face recognition system based on convolution neural network using multiple distance face. Soft Comput 21(17):4995–5002
32. Liu H, Nie H, Zhang Z, Li YF (2021) Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction. Neurocomputing 433:310–322
33. Li Z, Liu H, Zhang Z, Liu T, Xiong NN (2021) Learning knowledge graph embedding with heterogeneous relation attention networks. IEEE Transactions on Neural Networks and Learning Systems
34. Zhang Z, Li Z, Liu H, Xiong NN (2020) Multi-scale dynamic convolutional network for knowledge graph embedding. IEEE Transactions on Knowledge and Data Engineering

35. Gao G, Yu Y, Yang M, Huang P, Ge Q, Yue D (2020) Multi-scale patch based representation feature learning for low-resolution face recognition. Appl Soft Comput 90:106183
36. Bah SM, Ming F (2020) An improved face recognition algorithm and its application in attendance management system. Array 5:100014
37. Hwang W, Wang H, Kim H, Kee SC, Kim J (2010) Face recognition system using multiple face model of hybrid Fourier feature under uncontrolled illumination variation. IEEE transactions on image processing 20(4):1152–1165

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

Rucha Golwalkar is currently a student at the University of Stuttgart, Germany in M.Sc. in Information Technology (INFOTECH) specializing in Embedded Systems. She completed her B.Tech. in Electronics Engineering from K. J. Somaiya College of Engineering, Vidyavihar where she was a silver medalist. She did her Diploma in the same field from Veermata Jijabai Technological Institute, Matunga. She has worked on multiple projects in the Embedded Systems domain during her Diploma and Bachelors. Her work as a research intern at Biospark Inc. (incubated at IIT Mandi, India) involved projects like Autonomous Bioreactor, Digital microscope, and so on. She also worked as a research intern at Ninad's Research Lab, Thane, India where she developed multiple systems in the healthcare domain. She also holds a degree in Indian Classical Music.

Dr. Ninad Mehendale is currently working as an Associate professor at K.J. Somaiya College of Engineering, Vidyavihar. He has worked as an Associate professor at Vidyalankar Institute of Technology and as a lecturer as D. J. Sanghavi College of Engineering, Mumbai. He has worked as a Scientist in Karlsruhe Institute of Technology (KIT), Germany. He has completed his Ph.D. from the Indian Institute of Technology Bombay in the field of microfabrication and signal processing. He holds an M. Tech. in embedded system (gold medalist) and Bachelors in Electronics and Telecommunication. He was working as a Principal Investigator at Ninad's Research Lab, Thane, India.