# Studies Advanced in Mask Face Recognition based on Deep Learning

Muheng He *

School of Xiamen University Malaysia, Selangor, Malaysia

* Corresponding author email: CST2109187@xmu.edu.my

**Abstract.** In the era of epidemic, wearing a mask when going out has become a more common phenomenon. And how to complete face recognition and guarantee certain accuracy without letting users take off their masks has been a challenge in recently years' face recognition field. In this paper, the traditional masked face recognition method and the new MTCNN deep learning masked face recognition method are compared and analyzed, and other similar masked experiments are introduced as data references. It is synthetically found that the accuracy of MTCNN masked face recognition using CASIA-FaceV5 as dataset is more than 90%, which is far better than other traditional recognition methods, but its requirement for dataset image quality is high, and the training time is long and costly. We hope that the process analysis and result comparison of different methods in this paper can shed light on the future development of masked face recognition technology.

**Keywords:** Face Recognition; Mask; Deep Learning.

## 1. Introduction

Face recognition is in full swing in the field of machine vision, which aims to build a model to obtain identity information for a given image. Thanks to the rapid development of pattern recognition technology, both the speed and accuracy of face recognition have made breakthroughs, which has been widely used in many fields. However, occluded face recognition is still an open problem, especially in the context of the global COVID-19 epidemic since 2019. Occluded face recognition needs to accurately identify incomplete facial features or fuzzy facial feature information in application scenarios, which can greatly improve work efficiency and user experience.

The difficulties of occluded face recognition technology are mainly reflected in the lack of facial feature points caused by occlusion, incorrect alignment of feature points, and local fusion. According to different causes, occlusion mainly includes: (1) light occlusion: facial occlusion caused by too bright or too dark external light; (2) self-occlusion: lack of feature points caused by face angle and human body posture, such as side face, head down, etc.; (3) Physical occlusion: objects that cover the face, such as scarves, eyes, masks, etc. For the classical face recognition technologies, their core idea is to complete the judgment and recognition through the extraction, analysis, and comparison of facial feature points. The clarity and completeness of features is a very critical factor. If there are too many external interference factors in the recognition sample, the traditional face recognition method that relies too much on feature point information may not be able to match the sample with the database, resulting in algorithm failure.

For the challenging occluded face recognition task, previous efforts can mainly be divided into the conventional method and deep-learning recognition models. The mainstream conventional method for occluded face recognition has three main categories: subspace regression, Robust classifier (Robust mistake coding) and Robust features extracting [1]. (1) The method based on subspace regression divides the face into many different regional sub-spaces while the occlusion is treated as a separate subspace. Then, the non-occluded face image is superimposed with the occluded one to synthesize the occluded face image. Thus, it is possible to transform the recognition problem into one in which the unobstructed face image and the occluded are each returned to the subspace they belong to and then recognized independently. (2) The method based on robust error coding mainly contains two modules " Addition Model" and "multiplicative model". The addition Model converts the masked face image into an initial additive combination of the error value k caused by the unmasked image

and the masked object (simplified), which can be formulated as y=y0+k. The algorithm focuses on separating the error value k from the formula, so as to continuously fit to obtain an infinite approximation of the occluded face image to the initial unobscured image. On the other hand, the multiplicative model treats the occluded face image as a collocation of the unobscured initial image and the occlusion. By continuously increasing the weight of the original image y0 and reducing the weight of the occlusion, such methods can continuously and accurately reconstruct the original unobscured image y0 to achieve effective separation of occluded and unobscured regions. The superposition of the two models will minimize error coding and thus improve the robustness of the recognition system. (3) The method based on robust feature extraction is most in line with the deep learning recognition method. It is important to know that a face image usually contains a wealth of features, including not only low-order features such as color, brightness, texture, orientation, but also high-order features such as pose, expression, age, ethnicity, etc. Robust feature extraction methods are needed to decompose these features. For example, Gabor filter decomposes features at multiple scales and multiple orientations when processing image textures, while attribute features decompose images into multiple describable attributes. The decomposition reduces the mutual interference between features and thus provides sufficiently fine features for machine learning. Such a combination of algorithm and machine learning ideas is simultaneous the thinking pattern of the new deep learning recognition method [2].

Focusing on the above points, in this paper, we focus on the mask face-recognition models based on the robust feature extraction, and introduce the most representative method of MTCNN in Section 2. MTCNN, as the name implies, is a multi-task convoluted neural network including three cascade networks for face framing, such as P/R/O-Net [3]. All the images of the MTCNN model training input are positive samples (photos of wearing masks), and no negative samples are used as model input. In the subsequent recognition task module, the MTCNN model detection results are imported and the feature points are encoded and compared for recognition. In Section 3, we report the results of MTCNN and explore its application range. In Section 4, we discuss the existing issues and their corresponding possible solutions in mask face recognition task.

## 2. Method

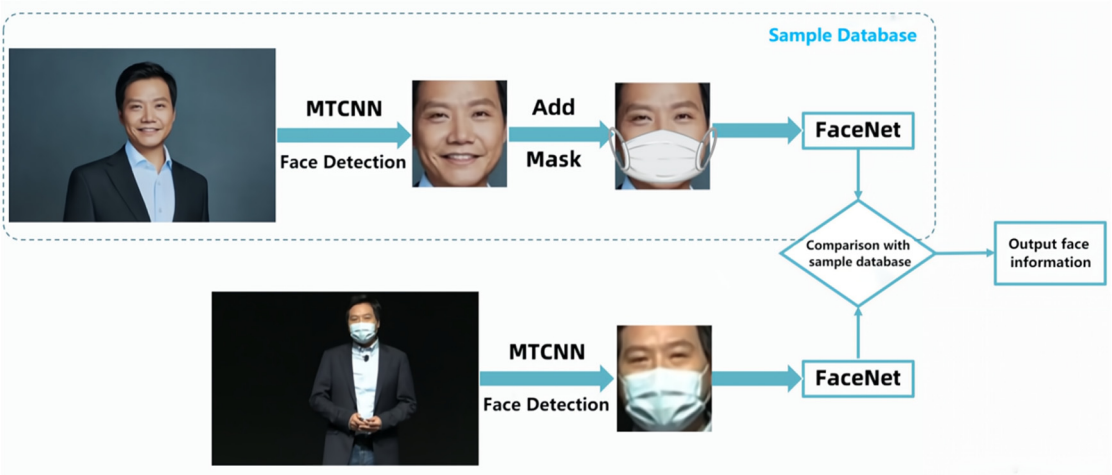



**Fig 1.** The framework of MTCNN

The framework of mask face recognition based on MTCNN can be separated into four stages.: face detection with mask; face alignment with mask; face encoding with mask; identity recognition corresponding to face with mask. As the Figure 1 shown, two sets of input images (unmasked face image and masked face image) are firstly needed for the same target during training. The machine automatically adds mask to the unmasked image and then puts it into the sample library for feature comparison with the masked image. This process is divided into two chains. The first chain is to first

use the three sub-networks of MTCNN technology and coarse to fine extraction of the face part, and then use MobileNet component for mask detection [4]. If the input is found to be the original image (no masks), the mask will be added in the mouth and nose feature point region and the processed "masked face images" will be fed into the recognition sample database. The second link is the cascade correction of MTCNN. The collected masked face image is cropped to the same pixel size as the initial sample library.Then two links are added to the sample database in parallel, where the sample database is compared with FaceNet (the automatically processed masked faces are compared with the original masked faces) to predict the final face information [5].

Here, we assume a situation that the unmasked face image is collected when creating the recognition sample database, and the masked face image is collected when performing recognition (here, we assume that MTCNN is robust to both cases, and we find that it works in practice). After that, the mask is added to the unmasked face image, and then the masked face is compared with the manually processed masked face and the collected masked face. Finally, the identity information of the person is output. For the MTCNN network, the thresholds of P/R/O-Net are simply adjusted. The three thresholds control the cropping frame to output high-precision facial information. The face-plus-mask program mainly applies the 68 key points of the face. As the Figure 2 shown, the main part of adding a mask to a face is as follows [6].

(1) Searching the 68 Critical Points of the face.

(2) Determining the nose and face outline of a person.

(3) Determining the left-side point of the face, the bottom-side point of the face and the right-side point of the face by the facial contour.

(4) Determining the height, center line of the mask size from the nose and the bottom point of the face.

(5) Divide the mask evenly between the left and right parts; size the left side of the mask using the distance between the extreme facial left-side point and the center axis as the width. Adjust the right-mask, the width is the distance between the facial right-side point and the center axis. Combine the left-mask and right-mask into a new one.

(6) Adjust and rotate the new one with the rotating angle of the center axis with respect to the Y-axis, and eventually place the mask in the proper position on the previous image.

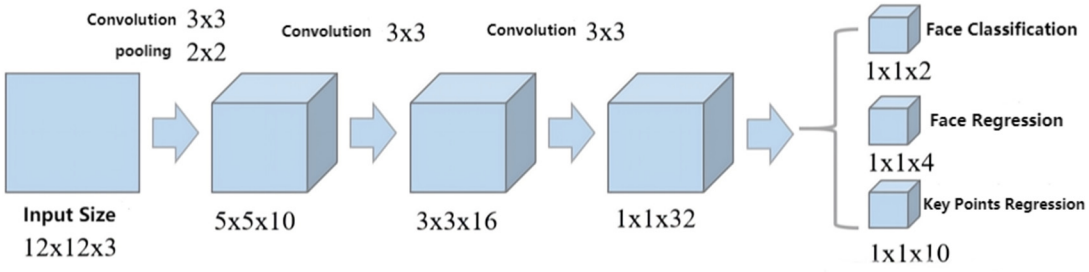



**Fig 2.** The visualization of key points detection and mask adding

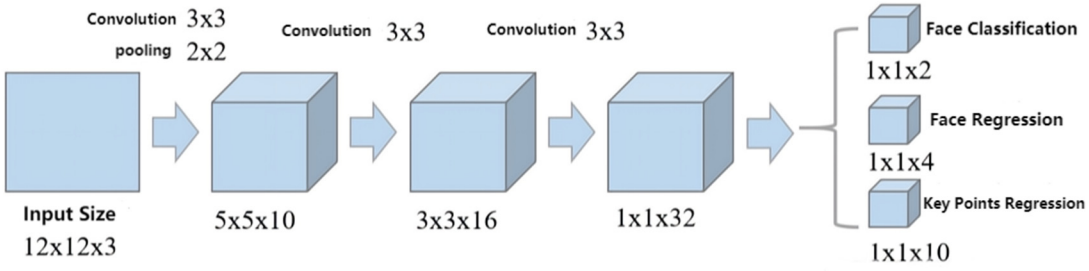
Finally, the two sets of data are compared and recognized based on the FaceNet open-source module. The first method is more intuitive and simpler, while the second method is based on MTCNN to build the recognition area and then use the "dlib" open-source library to automatically supplement the feature points.

### 2.1 Masked Face Detection Part

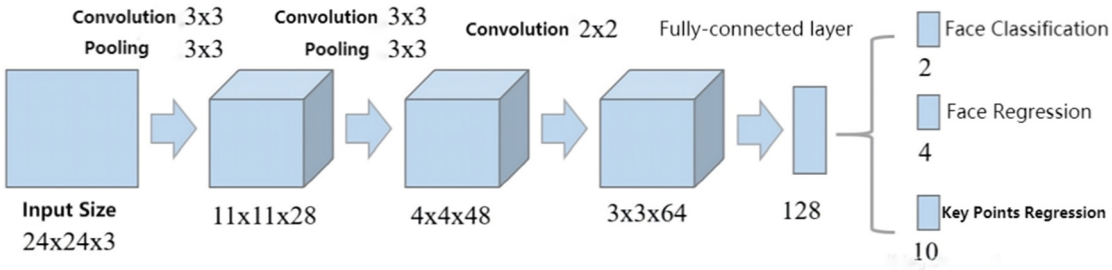
The input sample database of positive sample pictures (face pictures of wearing masks) use uniform size pictures, because there may be body parts such as arms and shoulders in the obtained pictures, which may generate more noise for training, so the MTCNN method should be used to crop the face area of the mask picture; and MTCNN, as the name implies, is a multi-task convoluted neural network, the network model, mainly through three Cascaded networks for face framing, P/R/O-Net. P-net is used for rapidly generate facial intercept windows. The execution process is as follows: after convolving the input features in three layers, a face classifier, border regression and facial key point localization are used for primary selection of face regions. Then, the P-Net primary selection is fed to R-Net for the next step. R - Net serves to filter the face regions intercepted from the previous step in a more refined way. The process is: all the candidate windows obtained from P - Net are fed into R-Net, and the fewer effective candidates are eliminated and further prediction windows are obtained by edge regression and non-maximal suppression. The role of O-Net is to generate the final recognition boundaries and the critical points of the face. The running process is similar to R-Net, but with the addition of regression prediction of the location of human facial feature points. Finally, five face-feature points of human face are output. The network structures of three models are shown in Figure 3.




(a) The structure of P-Net




(b) The structure of R-Net




**(c)** The structure of O-Net

**Fig 3.** The structure of three sub-networks

### 2.2 Face Alignment Part for Mask Wearing

This task module, mainly calls the "Dlib" open source library to extract 128 feature points, the input face picture wearing a mask, for the two parts of the nose and mouth, under the mask cover, the model automatically complements the feature points for face feature point extraction [7]; and the deep learning part uses the Face-net model. This model is not applied to the classical soft-max used for the classification learning, but learns the encoding method from image to Euclidean space by extracting one of the layers as features. The algorithm mainly applies the established CNN model directly (e.g. GoogleNet, etc.) and change the loss function on top of that, to facilitate the mapping of face images to a high-level spatial hierarchy[8]. The loss function is used to optimize the Euclidean distance between faces so that the error distance from the face picture about the same person is minimized and the error distance from the face images about different people is maximized. Based on the acquired feature vectors, the "Euclidean distance" is calculated for face recognition. The network structure is as follows in Figure 4. After this step the missing feature points of the face will be filled and involved in the recognition.

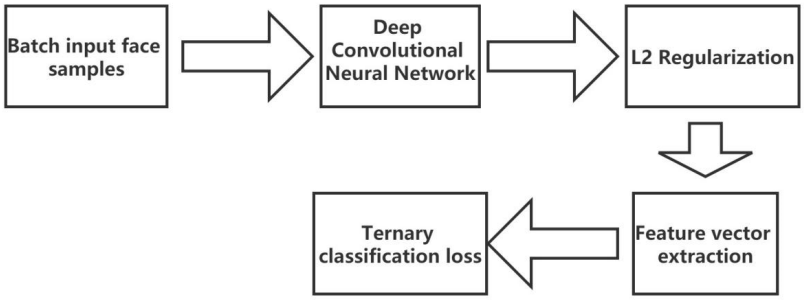



**Fig 4.** The structure of FaceNet

### 2.3 Face Coding Part for Mask Wearing

Obtaining the feature encoding of 128 points of the face and calculating the distance between different faces based on the face encoding information matrix.

### 2.4 Face Recognition Part for Mask Wearing

After all the methods for calculating the distance have been arranged, the final step of face recognition is performed. Calculate the face data in the database, and store the pictures after information encoding into the list of known face information.

## 3. Experiment

### 3.1 Dataset

This experimental module contains two parts, which are the selection of database and the accuracy of the results of MTCNN compared with other methods. In the MTCNN method, the Casia-FaceV5 dataset is used as the training data. Five hundred people, five images for one person, and two thousand five hundred pictures in total, 640*480 images are used in the CASIA-FaceV5 Asian face dataset, which is characterized by a small number of samples and is easier to train on individual user computers. The effect of the sample set is as follows in Figure 5 [9].

After that, a database for testing was created. The database contains photos of 8 people, and a folder is set up for each person to place 5 to 10 images, which are processed into 160*160-pixel images after steps such as face detection and adding masks. When using the entire CASIA-FaceV5 dataset as a test set, it is necessary to consider not only identifying the same person, but also distinguishing different people, which requires generating "pairs" files corresponding to the same person and different people. Because the number of pairs varies greatly with the same person and

different people, the pairs files were generated separately and then tested on the dataset. Finally, the experimental results and recognition accuracy are collected with this training set of the database.



(a) Original samples



(b) Samples with mask

**Fig 5.** Visualization of the training samples before/after masking

### 3.2 Performance Analysis

We report the results of different methods on various datasets, whose results can be seen in Table 1. The AR database contains more than four thousand facial pictures from One hundred and twenty-six subjects. For every subject, twenty-six pictures of the participants were taken in two periods of time separated by two weeks. In each session, images were taken with different facial expressions (smiling, angry, screaming), different illumination, and two different types of masking. In the experiment, a subset of the database was utilized which contained the complete images of each subject and consisted of one hundred and twenty participants. The grayscale image is trimmed and adjusted to a resolving power of $128 \times 128$. The training images have different neutral representations of illumination. The test image is differently illuminated while wearing sunglasses or a scarf. Figure 6 shows that masking has the highest performance in scarves and sunglasses, with a recognition rate of 90.0% for scarves (similar to mask type) and 78.2% for sunglasses [10].



**Fig 6.** Visualization of samples with different masks

**Table 1.** Results comparison of different methods on different datasets

| Method | Experimental | Recognition accuracy(masked) |
|---|---|---|
| MTCNN | CASIA-FaceV5 | 92.3% |
| CFs | AR dataset | 90.0% |
| ZACF | AR dataset | 87.2% |
| mCFs | CMU-PIE | 60.3% |

We also visualize the recognized results in Figure 7, whose recognition accurate all reached over 90% (with repeated training of the original image and automatic supplementation of the mask image).

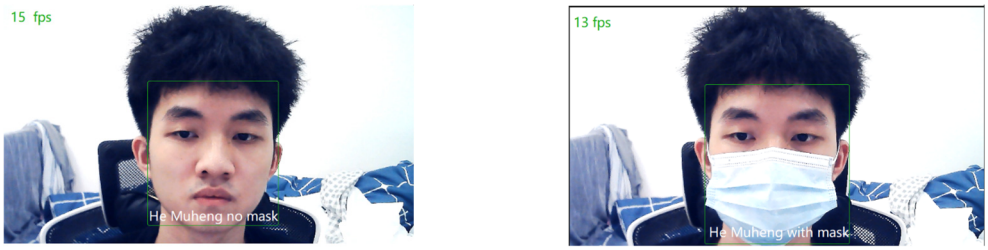



**Fig 7.** Visualization of recognition results

## 4. Conclusion

Combining the experimental data, we find that the new deep learning model MTCNN mask-obscured face recognition method improves the recognition accuracy to a great extent, although it requires higher accuracy and longer training time for the dataset. In this paper, we only introduce two simple recognition frameworks based on the MTCNN detection module, but there is still much potential to improve the accuracy and recognition code block under more severe image environments (e.g., unstable light factors, overly biased face orientation, etc.). Therefore, the combination of MTCNN and masked face recognition is very promising, and it may make another qualitative leap in recognition environment diversity and recognition accuracy when combined with a larger database more efficient training equipment and a more perfect recognition module.

## References

[1] Ahmadlou M, Adeli H. Enhanced probabilistic neural network with local decision circles: A robust classifier[J]. Integrated Computer Aided Engineering, 2010, 17(3):197-210.

[2] Jürgen Schmidhuber. Deep learning in neural networks[J]. Neural Netw, 2015.

[3] Zhang K, Zhang Z, Li Z, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.

[4] Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[5] Schroff F, Kalenichenko D, Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering: IEEE, 10.1109/CVPR.2015.7298682[P]. 2015.

[6] Chai Y, Liu J, Li Y. Facial Target Detection and Keypoints Location Study Using MTCNN Model[J]. Journal of Physics: Conference Series, 2021, 2010 (1):012097.

[7] King D E. Dlib-ml: A Machine Learning Toolkit[J]. Journal of Machine Learning Research, 2009, 10(3):1755-1758.

[8] Tang P, Wang H, Kwong S. G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition[J]. Neurocomputing, 2016, 225(feb.15):188-197.

[9] Umer S, Dhara B C, Chanda B. Face recognition using fusion of feature learning techniques[J]. Measurement, 2019, 146:43-54.

[10] He E J. Masked correlation filters for partially occluded face recognition[C]//IEEE International Conference on Acoustics. IEEE, 2016.