Received: 10 October 2020 | Revised: 9 February 2021 | Accepted: 2 March 2021

DOI: 10.1002/cpe.6331

SPECIAL ISSUE PAPER

WILEY

Multi-angle head pose classification with masks based on color texture analysis and stack generalization

Shuang Li[1,2,3] | Xiaoli Dong[1,2,3] | Yuan Shi[2,4] | Baoli Lu[1,3] | Linjun Sun[1,2,3] | Wenfa Li[5]

[1]Institute of Semiconductors, Chinese Academy of Sciences, Beijing, China

[2]Cognitive Computing Technology Joint Laboratory, Wave Group, Beijing, China

[3]Beijing Key Laboratory of Semiconductor Neural Network Intelligent Sensing and Computing Technology, Beijing, China

[4]Shenzhen Wave Kingdom Co., Ltd., Shenzhen, China

[5]College of Robotics, Beijing Union University, Beijing, China

Correspondence
Linjun Sun, Institute of Semiconductors Chinese Academy of Sciences, Beijing 100083 China.
Email: sunlinjun@semi.ac.cn
Wenfa Li, College of Robotics, Beijing Union University, Beijing 100101, China.
Email: 644085325@qq.com

Funding information
National Natural Science Foundation of China, Grant/Award Number: 61901436 & 61972040; The Premium Funding Project for Academic Human Resources Development in Beijing Union University, Grant/Award Number: BPHR2020AZ03

Summary

Head pose classification is an important part of the preprocessing process of face recognition, which can independently solve application problems related to multi-angle. But, due to the impact of the COVID-19 coronavirus pandemic, more and more people wear masks to protect themselves, which covering most areas of the face. This greatly affects the performance of head pose classification. Therefore, this article proposes a method to classify the head pose with wearing a mask. This method focuses on the information that is helpful for head pose classification. First, the H-channel image of the HSV color space is extracted through the conversion of the color space. Then use the line portrait to extract the contour lines of the face, and train the convolutional neural networks to extract features in combination with the grayscale image. Finally, stacked generalization technology is used to fuse the output of the three classifiers to obtain the final classification result. The results on the MAFA dataset show that compared with the current advanced algorithm, the accuracy of our method is 94.14% on the front, 86.58% on the more side, and 90.93% on the side, which has better performance.

KEYWORDS

color space conversion, head pose classification, line portrait, stacked generalization

1 | INTRODUCTION

Head pose classification, which is to determine the direction of the face in three-dimensional space. And it is a popular research direction in the field of computer vision and pattern recognition. The focus of human attention can be obtained by estimating the pose of the human face, which can be used in the fields of attention monitoring, real-time monitoring of driver fatigue, coordination-based live detection of faces, and human–computer interaction. At the same time, with the rapid development of research on face detection, face landmark detection, face recognition,[1] and face attribute recognition, it has become particularly important to provide high-quality face images for these algorithms. In addition to poor image sharpness and low light will affect the image quality, side face images with larger head deflection angles often also have a negative impact on image quality. Therefore, head pose estimation has always played a very important position in the field of facial modeling and analysis.

With COVID-19 virus[2] global pandemic in 2020, human health and medical treatment are greatly threatened and affected. At the same time, the application of facial biometric recognition technology has also received a huge impact. Since the virus can be spread through droplets, contact, and other means, the authentication method based on password and fingerprint is no longer safe for human health. Therefore, face recognition will replace the original authentication method on a large scale. But during the virus epidemic, most people wear masks for authentication, which challenges the performance of most algorithms related to faces. In order to solve this problem, it is necessary for researchers to improve existing algorithms. This article will focus on the face pose classification algorithm when wearing a mask.

FIGURE 1 Head pose estimation can be regarded a rigid transformation. The rotation of the head can be represented by three Euler angles, (A) yaw; (B) pitch; (C) roll

In the field of facial pattern recognition, head pose classification is to judge the deflection direction and angle of the face in the three-dimensional space through a two-dimensional face image or video. Therefore, it needs to extract the features in the two-dimensional image of the face to restore the pose in the three-dimensional space, which is to find a set of mapping between two-dimensional and three-dimensional space. Usually we regard the head pose estimation as a rigid transformation, which can be represented by a 3D vector of pitch, yaw, and roll,[3] as shown in Figure 1. In practical applications, human face images are affected by light, occlusion, and sharpness, which makes head pose estimation a problem of great concern.

The input for head pose estimation is a two-dimensional face image. Due to the reduction of one-dimensional information, this research is challenging. Some researchers have used various methods to try to solve the problem of head pose estimation. For example, adding the three-dimensional information to the image to simplify the difficulty of research,[4-9] calculation based on face landmark detection,[10-12] use support vector machine to classify images, use convolutional neural network (CNN) for feature extraction,[13-16] use line portraits to highlight important information and feature aggregation methods,[17,18] multi-level classification to regression,[19] and so forth.

The follow-up content of this article is arranged as follows: Section 2 briefly introduces related work on head pose classification and estimation. Section 3 introduces our proposed method of head posture classification when wearing a mask. Section 4 designs related experiments to evaluate and analyze the performance on the MAFA dataset. Section 5 summarizes the work and innovative contributions of this article.

2 | RELATED WORKS

Early methods for head pose classification were solved by realizing a two-dimensional to three-dimensional mapping, such as a method based on face landmark detection. This method divided the problem into two stages. The first stage was to use face landmark detection algorithm[10,11] to obtain the information of some important feature points of the face, generally five points at the corners of eyes, the corners of the mouth, and the tip of the nose. And the obtained feature point position coordinates were passed as input to the second stage. The second stage was usually called perspective-n-point (PnP) problem in computer vision terminology,[12] which was to extract three-dimensional information from two-dimensional data. Researchers used external parameters to reconstruct the three-dimensional information of the input face landmarks, and calculated the values of three Euler angles to obtain head pose information for classification. Another method was to obtain a standard model of a three-dimensional face in advance, and then fitted the model with the input two-dimensional image. This process was usually achieved by using the nonlinear least squares method.[6] At this time, the pose information of the three-dimensional face model was the pose information of the input image.

With the introduction of deep learning technology, it was also used for the first time to solve the head pose classification problem. Deng et al.[20] used a three-layer convolution and two-layer fully connected network to train and classify part of the original training set with noise added. Chang et al.[21] proposed a network called Face-Pose-Net, and align the two-dimensional and three-dimensional faces through this network. They used this process to replace the face landmark detection algorithm, and directly return to the three-dimensional head pose through the two-dimensional image. Feng et al.[5] proposed a method that can reconstruct and align three-dimensional faces, named Position Map Regression Network. The author

designed a UV location map, which can save the three-dimensional information in the face image, and then trained a simple CNN through weighted loss to extract the UV location map from the two-dimensional face image. Hsu et al.[22] proposed using multiple regression loss functions to train CNN. The author used L2 regression loss to predict more accurate angle values and predicts the ordering of tags through ordered regression loss, and then combined the two results to finally return the head posture deflection angle. Ruiz et al.[23] also trained a multiple regression loss CNN to solve the pose problem. This method used ResNet50 as the main network to extract features, and then combined classification and regression with two objective functions to predict Euler angles in three directions of the head.

The latest head pose research is the FSA-Net stage regression method proposed by Yang et al.[19] The author found that the spatial relationship of feature maps was often ignored in previous work. Therefore, this method added the step of grouping features before feature aggregation, and obtained different models by using different features, which could complement each other well. The network used SSR-Net which was used to solve the age estimation problem. It finally obtained the probability of the pose distribution through multi-stage classification, thereby estimated the head pose information of a single RGB image.

Although the above methods have achieved good results in obtaining head posture information. But when the input is a face image wearing a mask, the performance of each algorithm is negatively affected to varying degrees. To solve this problem, we propose a method that combines color texture analysis and face contour. First, the input RGB face image is converted in color space, and then H channel information is extracted from the image converted to HSV color space. This image can better distinguish the mask and the face area. Then use the line portrait algorithm to extract the contours of the face and facial features in the image. These two processes make the network's attention better to focus on the features that are conducive to head pose classification. After that, the two images obtained are fused with grayscale images and input into three CNNs for training. Finally, the stacked generalization[24] technology is used to minimize the generalization error rate of the three classifiers and achieve more accurate classification of head posture.

3 | PROPOSED METHOD

The method proposed in this article involves five parts, including image color space conversion, line portrait generation, CNN-based feature extraction, pixel fusion, and stacked generalization.

3.1 | Face image preprocessing based on RGB to HSV color space

Generally, the color space of the input image used is RGB, when researchers solve the head pose classification problem. And when using CNN for feature extraction, it also operates based on the pixel value of the RGB color space. Because RGB is the most basic and most commonly used color space in image processing.[25]

The RGB color space is represented by three channels, namely the red (R), green (G), and blue (B) channels. The model of the three-dimensional space is shown in Figure 2. Regardless of the color, the colors of these three channels are combined by different components, so the three channels are highly correlated with each other.

Compared with RGB, the HSV color space has a better effect in displaying colors. It can very intuitively reflect the hue, saturation, and value of the image. This makes HSV closer to the human visual perception system than RGB. HSV also represents an image by three channels, the difference is that the three channels are hue (H), saturation (S), and lightness (V), as shown in Figure 3. We found that no matter how the light of the image changes, the H channel value of the HSV color space is not affected by these changes. This feature makes it easier to distinguish objects of different colors through the H channel.[26] We compare the four imaging effects of RGB image, grayscale image, B channel of RGB color space (B-RBG), and H channel of HSV color space (H-HSV), as shown in Figure 4.

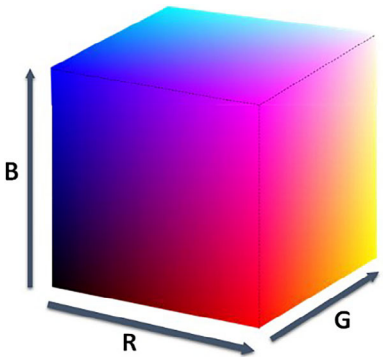

FIGURE 2 RGB three-dimensional space model

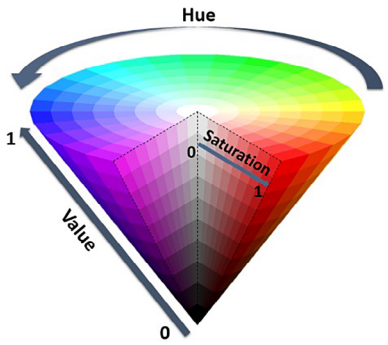

FIGURE 3 HSV inverted cone model

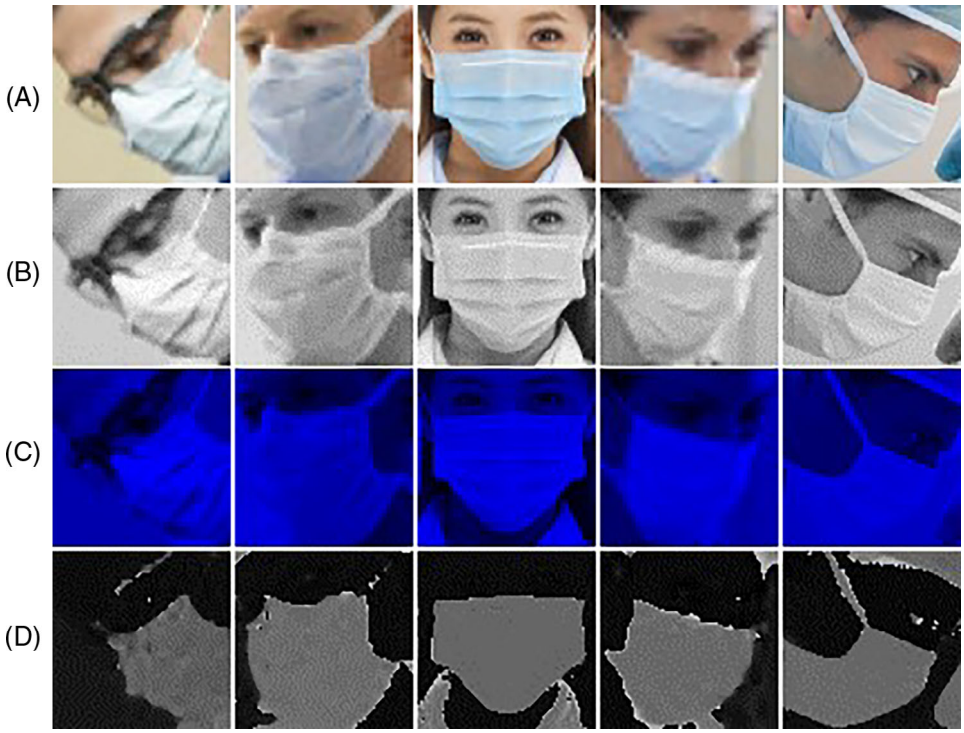

FIGURE 4 (A) RGB face images in different poses; (B) grayscale images; (C) B channel of RGB color space images; (D) H channel of HSV color space images

We found that the H channel of the HSV color space can distinguish the face and the mask area well. Therefore, we propose an image processing method based on color texture analysis. We process the image on the H-HSV color space to highlight the color difference between the enlarged face and the mask.

We need to first convert the image from RGB color space to HSV color space, because the final processing only needs the hue channel of the image. So we only summarize the conversion process from RBG to H-HSV.[27] The conversion formula is as follows:

$$R' = R/255, \tag{1}$$

$$G' = G/255, \tag{2}$$

$$B' = B/255, \tag{3}$$

where R, G, and B, respectively, represent the pixel value of the corresponding color channel, so the value range of R′, G′, and B′ is [0, 1].

$$C_{max} = \max(R', G', B'), \tag{4}$$

$$C_{min} = \min(R', G', B'), \tag{5}$$

$$\Delta = C_{max} - C_{min}. \tag{6}$$

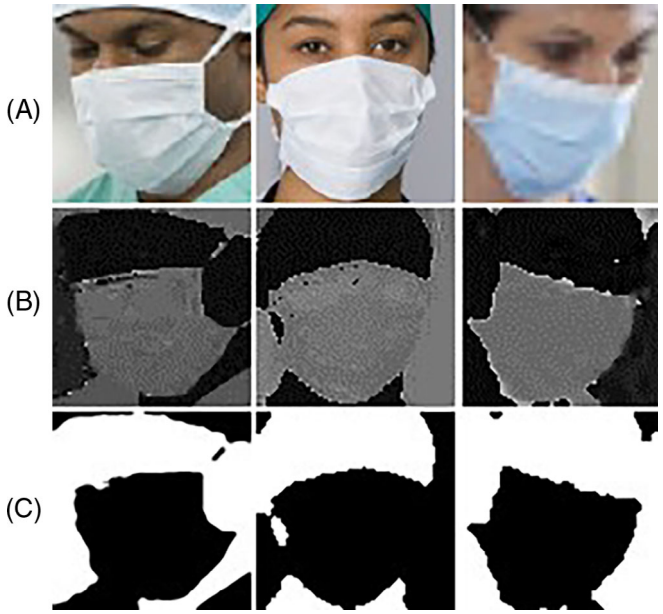

FIGURE 5 (A) RGB face image; (B) H channel image; (C) processed image

C_{max} and C_{min} are the maximum and minimum values of R′, G′, and B′, respectively. And Δ is the difference between the maximum and minimum. Finally, the expression of the pixel value of the H channel is:

$$H = \begin{cases} 0^\circ, \Delta = 0 \\ 60^\circ \times \left(\frac{G'-B'}{\Delta} + 0\right), C_{max} = R' \& G' \geq B' \\ 60^\circ \times \left(\frac{G'-B'}{\Delta} + 6\right), C_{max} = R' \& G' < B' \\ 60^\circ \times \left(\frac{B'-R'}{\Delta} + 2\right), C_{max} = G' \\ 60^\circ \times \left(\frac{R'-G'}{\Delta} + 4\right), C_{max} = B' \end{cases} \quad (7)$$

the pixel value can be calculated through five formulas corresponding to different conditions, and the final value range of the H channel of the HSV color space is [0°, 360°).

After obtaining the H channel image, we further process the obtained image. Because H channel only focuses on the hue information of the image, some small noise on RGB will be amplified and presented. First, we adjust the size of the face image to a uniform size, and then perform mean filtering[28] and binarization[29] on the image. In order to make the network pay more attention to the face area, we need to make the pixel value of the mask area smaller than the pixel value of other face areas such as eyes. Therefore, we divide the image into two parts with the same upper and lower area, and calculate the pixel values of the two parts, respectively. We know that the position of the mask is in the lower half of the face frame, so if the pixel value of the lower half of the image is greater than the upper half of the image, the pixel value of the image is reversed. And if the pixel value of the lower half of the image is smaller than the upper half, it will remain unchanged. The final processed image is shown in Figure 5.

3.2 | Line portrait generation of face contour

Due to face images are affected by light, image quality, and other aspects. The method of classification directly by extracting the features of the original image often does not work well.[3] In response to this problem, we propose a method based on line portrait, which filters out redundant information that interferes with detection, and highlights line features such as facial contours to provide more targeted image data for subsequent classification.

The algorithm for extracting lines based on the line detector, the most representative one is the line detection algorithm with the Gaussian kernel function as the line detector. Gooch et al.[30] proposed a method based on difference of Gaussians (DoG) to extract lines. But more noise in the image affects the overall effect. Kang et al.[31] improved the Gaussian difference operator in order to improve the continuity of the lines, and proposed the flow-based difference of Gaussians filtering (FDoG) algorithm based on feature flow. FDoG can automatically generating lines. This method can effectively capture and display the significant edge information in the image, thereby extracting a set of coherent and smooth stylized lines. We apply this algorithm to face images to extract lines that highlight the contour of the face and facial features.

FIGURE 6 RGB image and line portrait

First, the researcher proposed a new method to construct a smooth direction field to preserve the important features of the image. Then filter the image through the DoG model. This step is used to detect high-coherence lines and suppress noise. Finally, the algorithm extracts the main feature lines of the image, such as facial contours. The specific realization process is that FDoG constructs the edge tangential flow field of the input image by defining $t^{new}(x)$, its formula is:

$$t^{new}(x) = \frac{1}{k}\sum_{y\in\Omega(x)} \phi(x,y)t^{cur}(y)w_s(x,y)w_m(x,y)w_d(x,y), \tag{8}$$

where $t^{new}(x)$ is the edge tangent vector perpendicular to the image gradient $g(x) = \nabla I(x)$. And $\Omega(x)$ represents the field of x, which is a square field with a radius of r that satisfies the radial distribution, and k is the normalized value of the vector. w_s represents the spatial weight function, which is defined as:

$$w_s(x,y) = \begin{cases} 1, & if\|x-y\| < r \\ 0, & otherwise \end{cases} \tag{9}$$

the other two weighting functions w_m and w_d are the amplitude weighting function and the direction weighting function, which play a key role in ensuring the accuracy of the features of the input image.

After that, the image is filtered through the DoG model, and finally the main lines of the image features are extracted, including the face and contours.

The method used in this article is the line portrait extraction algorithm proposed by Dong et al. The algorithm is improved on the basis of the FDoG algorithm and incorporates a three Gaussian model that is more suitable for the animal visual system.[32] It enhances the extracted line information and better output the feature information of the face contour. Similarly, in order to make the network pay more attention to the information of the line part, we set the pixel value of the obtained image line part to 255, as shown in Figure 6.

3.3 | Pixel fusion and CNN-based feature extraction

Through the above method, we get the H channel of HSV color space and line portrait of the image. We need to merge them with the grayscale image at the pixel level. First, we use the square kernel to average filter the line portrait. Then we merge the pixels of the three images, the formula is as follows:

$$P_{xy} = W_g G_{xy} + W_l L_{xy} + W_h H_{xy}, \tag{10}$$

where G_{xy} is the value of the pixel in column x and row y of the grayscale image, L_{xy} is the value of the pixel in column x and row y of the line portrait, and H_{xy} is the value of the pixel in column x and row y of the H channel image. W_g, W_l, and W_h are the weights of the three images in the pixel fusion process, and the sum of the three weights is equal to 1. The image after pixel fusion is shown in Figure 7.

We use CNN to extract facial features. Due to the purpose of the experiment is to classify the head pose, and some interference information has been filtered out in the previous image processing. We can use a smaller network for feature extraction, and a smaller network has obvious advantages in terms of time-consuming. Since we will use the stacked generalization algorithm later, we need to prepare three networks to extract

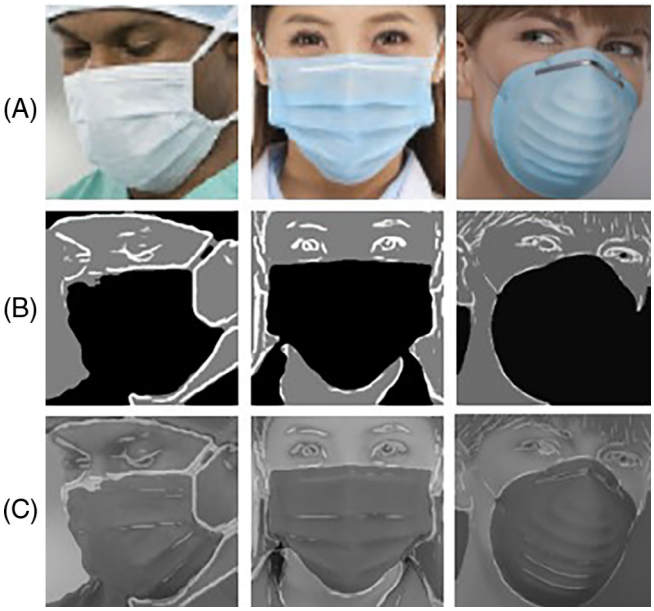

FIGURE 7 (A) RGB face image; (B) H channel image and line portrait pixel fusion; (C) input image

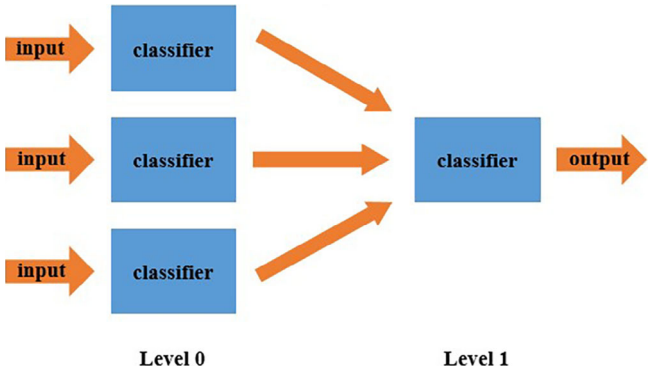

FIGURE 8 Stack generalization flowchart

features from the image. For the first network, we use the AlexNet[33] model that reduces the input size of the image and the convolution kernel size. The second network is the optimized AlexNet model. We not only modified the network parameters, but also subtracted a fully connected layer. For the third network, we use the ResNet-18[34] model that reduces the input size of the image and the convolution kernel size. All three models[35] use the Softmax classifier to classify and score the extracted features.

3.4 | Multi-classifier fusion based on stack generalization

Stacked generalization[24] is a type of ensemble learning. It uses high-level models to combine low-level models to achieve higher prediction accuracy. It is a scheme that can minimize the generalization error rate of multiple classifiers. The characteristic of stacking generalization is that when multiple classifiers are merged, a more complex strategy is used to combine them. In this way, we can learn more accurately which classifiers are more reliable. As shown in Figure 8, we divide the training set into three parts and use three classifiers for training and classification. The classifier used in this stage is called the base-learner (level 0 classifier). After that, we use the predicted results of the three base classifiers as the input data for the next stage and use a higher-level learning algorithm for classification. The classifier used in this stage is called a meta-learner (level 1 classifier). The characteristic of this method is to use the result of the first stage base-learner as the input of the next stage meta-learner. Compared with independent classification models, this can provide stronger nonlinear expression capabilities and reduce generalization errors.

In the head pose classification problem, the non-front face image contains the state in multiple directions in space, so the image set is more complicated and the difference within the class is large. Therefore, we use the model mentioned in the previous section as the level 0 classifier to extract deeper features.

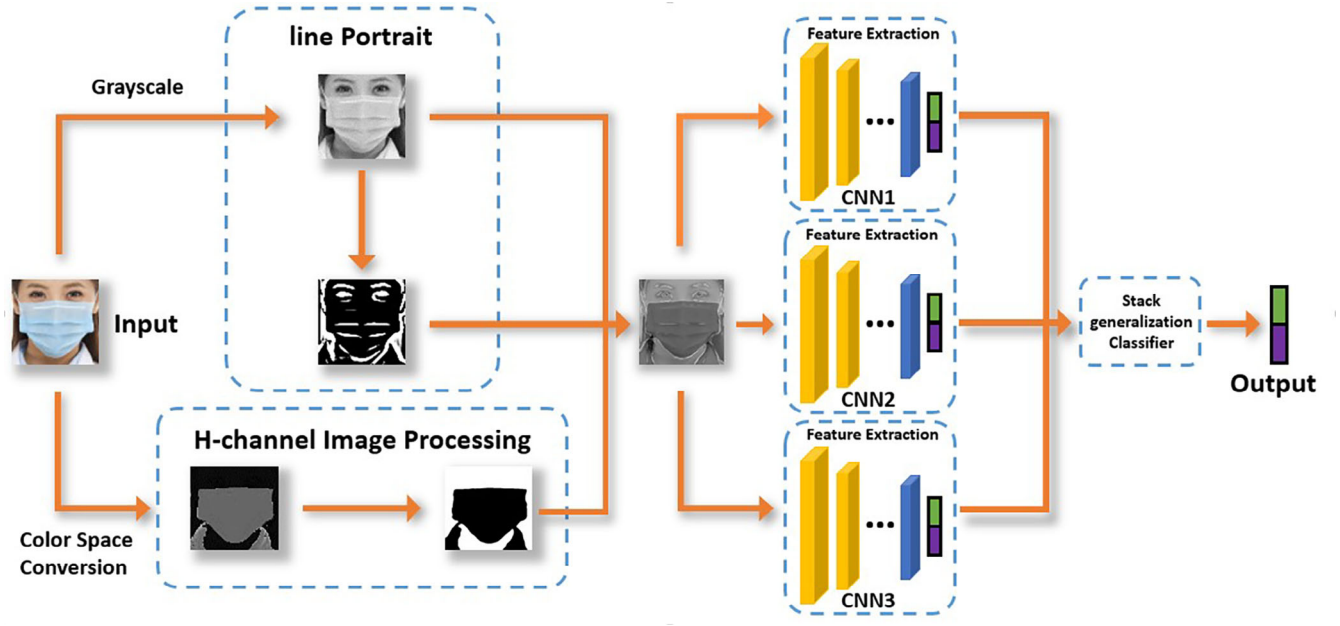

FIGURE 9 Algorithm flowchart

Before the second stage of classification, we use the output of the 0th class classifier as the input data of the 1st classifier. We accumulate multiple classification results and set a threshold.[36] When the score of the front face is greater than this threshold, the image is judged to be a front face, otherwise it is judged to be a side face.[37] The flowchart of the algorithm is shown in Figure 9.

4 | EXPERIMENT

4.1 | Datasets

MAFA:[29] The images in the MAFA[38] dataset are from the Internet, and there are 30,811 face images in total. The face information in almost all images is occluded, and most faces are occluded by masks. The images in the data set have been divided into two categories by the author and the training set and the test set are specified. These data are saved in folders named train and test. Each image is marked with necessary information, such as the position and size of the face frame. There are many other attributes marked in the data set, including the location of the eyes, the position of the mask, the posture of the head, and the degree of occlusion. Since ours want to solve the head pose classification problem, we need to use head pose information. The head pose of the data set is divided into five categories, including the left "1," the more left "2," the front face "3," the more right "4," and the right "5."

Since not all faces in the MAFA dataset have masks, we need to manually filter this data set and eliminate the data without masks. In the end, we obtained 23,845 face images with masks, including 20,139 in the train set and 3,706 in the test set. For the label of the data, we use the head pose provided by the data set as the classification standard, and divide the data into three types: front, more side, and side. From the readme file, we know that the face images with the head pose label "3" in the data set are the front, and the face images with the head pose label "2" and "4" are the more side faces, and the face images with the head pose label "1" and "5" are the side faces. In the process of classifying the data, we find that the labels provided by the dataset have obvious errors in the head pose classification of some face images. For example, some relatively large-angle side head pose images are classified as front. For these images with incorrect labels,[39] we manually selected them and corrected them to the correct labels. This is done to ensure the accuracy of the data set.

4.2 | Experimental details

Most face datasets with masks do not have the information of the calibration pose, and the faces in the datasets with pose information do not wear a mask. So our experiment is only conducted on the MAFA dataset. Since we only use a face data set, and the amount of training data provided by the MAFA data set is not enough, we need to expand the existing data. First, we use the face frame position information provided by the dataset to

crop the image. We expand the size of the face frame to 1.2, 1.4, 1.6, 1.8, and 2.0 times the original size and adjust the size of these face images to 80×80, when cropping the image. Then, we convert the color space of the image to HSV and extract the H channel. And we use a filter window with a size of 4×4 to perform average filtering on the image. After getting the filtered image, we calculate the average pixel value P of the image, and set P as the threshold to binarize the image. If the pixel value is greater than P set to 255 and less than P set to 0. Then we calculate the sum of the pixel values of the upper and lower parts of the image. If the pixel value of the lower half of the image is greater than the upper half of the image, the pixel value of the image is inverted. We fuse the H channel image with the portrait line drawing and grayscale image, and their weights are 0.25, 0.25, and 0.5, respectively. We send the obtained images to three CNNs. And after two stages of classification, we finally get the classification results of the head pose. It should be noted that for the test image, we only expand the size of the face frame to 1.4 times of the original image for testing.

We set the input size of the data layer of the CNNs mentioned in the previous section to 80×80. The optimized AlexNet has a total of seven layers, including a five convolutional layers and two fully connected layers. The fully connected layer is in the final stage of the network. We set the number of iterations in the training process to 250,000. The learning rate is set to 0.01, and after every 40,000 iterations, the learning rate is multiplied by 0.1. On NVIDIA RTX 5000 GPU, the training process of the entire experiment takes about 3.5 h.

4.3 | Experimental results

We compare the method proposed in this article with five commonly used and better-performing methods. Among them, EPnP-LAB is a method that uses LAB[11] face landmark detection combined with EPnP.[12] EPnP-LAB, Line,[17] Hope-Net,[23] and FSA-Net[19] algorithms all obtain the best classification accuracy by setting a set of thresholds with the best classification effect.

Table 1 shows the head pose classification performance of different methods for face images wearing masks. The frontal accuracy rate refers to the proportion of accurately classified data labeled "front face" on the test set, and the more side accuracy rate is the rate of accurate classified data labeled "more side face," and the side accuracy rate is the rate of accurate classified data labeled "side face."

From the test results of the MAFA data set, the EPnP-LAB algorithm is affected by the occlusion of the face by the mask, and the classification accuracy is poor. Because the performance of the LAB face landmark detection algorithm is affected, and the face wearing a mask has a greater impact on the accuracy of side positioning. LAB is a high-precision face landmark detection algorithm, which has better performance for error-prone samples than the DAN algorithm. If LAB[11] is replaced by other algorithms such as DAN,[10] it can be predicted that the performance of head pose classification will further decrease.

The output result of FSA-Net is the Euler angle of the head pose. We set an optimal threshold to classify the pose. The algorithm regresses head pose information through multi-level classification, and does not use face landmark detection in the calculation process. Therefore, the algorithm needs to obtain more information in the face image. But most of the information is lost in the face image wearing a mask, so the performance of FSA-Net is affected.

The reason for the performance degradation of the Hope-Net algorithm is the same as the FSA-Net algorithm. It is because the image of the face wearing a mask is missing a lot of information. The main reason for the performance degradation of the line portrait algorithm is that the texture information of the mask is incorrectly represented as the contour feature information of the human face. But because the extracted lines contain most of the face contour information, the algorithm performance does not drop too much.

The results show that the method proposed in this article is better than other algorithms in the front, more side, and side accuracy of classification. The front accuracy of the algorithm is 1.47% higher than the optimal accuracy of other algorithms, the more size accuracy is 3.2% higher than the optimal accuracy of other algorithms and the size accuracy is 2.36% higher than the optimal accuracy of other algorithms. The reason for this result is that images processed through the H channel can provide more targeted information to the network. And the algorithm can better complete the head pose classification problem by combining the contour information of the face in the line portrait and the pixel contrast in the gray image.

TABLE 1 Comparison of different algorithms on the MAFA data set

Method	Front accuracy	More side accuracy	Side accuracy
Line[17]	92.67%	83.40%	88.53%
FSA-Net[19]	74.97%	76.64%	71.20%
EPnP-LAB[12]	90.04%	68.92%	50.13%
Hope-Net[23]	72.16%	76.53%	73.87%
Ours	94.14%	86.58%	90.93%

Note: Through the accuracy of the three classification to show the performance of each algorithm.

Testing set	Front accuracy	More side accuracy	Side accuracy
Optimized AlexNet	93.39%	87.00%	90.40%
Not use H-channel	91.04%	84.36%	88.53%
Not use line portraits	90.87%	86.79%	89.33%
Line[17]	92.67%	83.40%	88.53%
Ours	94.14%	86.58%	90.93%

TABLE 2 Ablation study for different aggregation methods and the results are the MAFA of the front, more side, and side accuracy

4.4 | Ablation study

We used ablation study to determine the effects of individual component, including the use of a single network (optimized AlexNet), do not use H-channel image, and do not use line portraits. Table 2 reports the results. Since our method includes line portrait, its performance is also listed as a reference. The experimental results show that the algorithm with only one element does not always get good results, but the algorithm with all elements has the best overall performance. And showing that complementary information is learned in different elements.

5 | CONCLUSION

In this article, we propose a new method to solve the problem of multi-angle head pose[40] classification of faces wearing masks during the COVID-19 coronavirus epidemic. This method uses image gray information, color texture information, and face contour information based on line portraits. By extracting and processing the H channel pixels in the HSV color space, it is used to distinguish the face and mask area in the image. And use line portraits to highlight the contour information of the face in the image. The obtained multiple images are fused by the pixel fusion method. Finally, the training results of the three CNNs are processed through the stacked generalization algorithm to obtain the pose information output by the algorithm. Experimental results on the MAFA dataset show that our method is superior to the other methods. Solving the head pose classification problem also provides help for studying other face-related problems wearing masks. In practical applications, we can add a task when executing the face detection algorithm, which is to distinguish between wearing a mask and not wearing a mask. If the face in the image is wearing a mask, we can use the method proposed in this article for classification. And if it is a normal face image, we can use FSA-Net, line portrait, and other algorithms for classification.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Nos. 61901436 and 61972040) and the Premium Funding Project for Academic Human Resources Development in Beijing Union University (No. BPHR2020AZ03).

ORCID

Shuang Li https://orcid.org/0000-0002-3089-7221

REFERENCES

1. Ning X, Li W, Tang B, et al. BULDP: biomimetic uncorrelated locality discriminant projection for feature extraction in face recognition. *IEEE Trans Image Process*. 2018;27(5):2575-2586.
2. Guan WJ, Ni ZY, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. *New Engl J Med*. 2020;382(18):1708–1720.
3. He F, Zhao Q. Head pose estimation based on deep learning. *Comput Technol Dev*. 2016;026(011):1-4.
4. Mukherjee SS, Robertson N. Deep head pose: gaze-direction estimation in multimodal video. *IEEE Trans Multimed*. 2015;17(11):2094-2107.
5. Feng Y, Wu F, Shao X, Wang Y, Zhou X. Joint 3D face reconstruction and dense alignment with position map regression network. arXiv, Proceedings of the European Conference on Computer Vision. Munich, Germany; 2018:557-574.
6. Zhang M, Hou X, Ren W, Xu J, Wang S. Face pose estimation based on nonlinear least squares method. *J Zhejiang Univ Technol*. 2016;44(1):34–38.
7. Bai X, Yan C, Yang H, et al. Adaptive hash retrieval with kernel based similarity. *Pattern Recognit*. 2017;75:136-148.
8. Lu C, Ye K, Chen W, et al. ADGS: Anomaly Detection and Localization Based on Graph Similarity in Container-Based Clouds. 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS). Tianjin, China; 2019; IEEE.
9. Ning X, Duan P, Zhang S. Real-time 3D face alignment using an encoder-decoder network with an efficient deconvolution layer. *IEEE Signal Process Lett*. 2020;27:1944-1948.
10. Kowalski M, Naruniec J, Trzcinski T. Deep alignment network: a convolutional neural network for robust face alignment. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, HI; 2017.
11. Wu W, Qian C, Yang S, et al. Look at boundary: a boundary-aware face alignment algorithm. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE; 2018.
12. Lepetit V, Moreno-Noguer F, Fua P. EPnP: an accurate O(n) solution to the PnP problem. *Int J Comput Vis*. 2009;81(2):155-166.

13. Jourabloo A, Liu X. Pose-invariant face alignment via CNN-based dense 3D model fitting. *Int J Comput Vis*. 2017;124(4):187-203.
14. LeiZhoua XB, Liu X, Zhou J, Hancock ER. Learning binary code for fast nearest subspace search. *Pattern Recognit*. 2019;98:107040.
15. Ning X, Gong K, Li W, et al. Feature refinement and filter network for person re-identification. *IEEE Trans Circuits Syst Video Technol*, 2020;99:1–1.
16. Ning X, Gong K, Li W, Zhang L. JWSAA: joint weak saliency and attention aware for person re-identification. *Neurocomputing*. 2020. https://doi.org/10.1016/j.neucom.2020.05.106.
17. Li S, Sun L, Ning X, Shi Y, Dong X. Head pose classification based on line portrait. 2019 International Conference on High Performance Big Data and Intelligent Systems. ShenZhen, China; 2019:186-189.
18. Li S, Ning X, Yu L, et al. Multi-angle head pose classification when wearing the mask for face recognition under the COVID-19 coronavirus epidemic. 2020 International Conference on High Performance Big Data and Intelligent Systems. ShenZhen, China; 2020; IEEE.
19. Yang TY, Chen YT, Lin YY, Chuang YY. FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation From a Single Image. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE; 2019.
20. Deng Z, Zhao Q, Hu C. Face pose classificati on method based on deep learning. *Comput Technol Dev*. 2016, 2016;(7):11-13.
21. Chang FJ, Tuan Tran A, Hassner T, Masi I, Nevatia R, Medioni G. FacePoseNet: making a case for landmark-free face alignment. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Venice, Italy; 2017.
22. Hsu HW, Wu TY, Wan S, Wong WH, Lee CY. QuatNet: quaternion-based head pose estimation with multiregression loss. *IEEE Trans Multimed*. 2019;21(4):1035-1046.
23. Ruiz N, Chong E, Rehg J M. Fine-grained head pose estimation without keypoint. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA; 2018; IEEE.
24. Puertas E. Generalized stacked sequential learning. *Elcvia Electron Lett Comput Vis Image Anal*. 2015;14(3):24–25.
25. Chen Y, Mensink T, Gavves E. 3D neighborhood convolution: learning depth-aware features for RGB-D and RGB semantic segmentation. International Conference on 3D Vision (3DV). Canada; 2019.
26. Boulkenafet Z, Komulainen J, Hadid A. Face spoofing detection using colour texture analysis. *IEEE Trans Inf Forens Sec*. 2017;11(8):1818-1830.
27. Boulkenafet Z, Komulainen J, Hadid A. Face anti-spoofing based on color texture analysis. International Conference on Image Processing (ICIP). Quebec city, Canada; 2015; IEEE.
28. WangNo N, Chiewchanwattana S, Sunat K. An improved K-means clustering filter for mixed noise removal in RGB color. ECTI-CON. The State of Israel; 2018.
29. Saddami K, Munadi K, Muchallil S, Arnia F. Improved thresholding method for enhancing jawi binarization performance. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto, Japan; 2017.
30. Gooch B, Reinhard E, Gooch A. Human facial illustrations: creation and psychophysical evaluation. *Acm Trans Graph*. 2004;23(1):27-44.
31. Kang H, Lee S, Chui CK. Coherent line drawing. 5th International Symposium. California, USA; 2007:43-50; ACM Press.
32. Chao-Yi L, Xing P, Yi-Xiong Z. Role of the extensive area outside the X-cell receptive field in brightness information transmission. *Vis Res*. 1991;31(9):1529-1540.
33. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Neural Inf Process Syst (NIPS)*. 2012;25:1097-1105.
34. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. *IEEE Computer Society*. 2016;1:770–778.
35. Li R, Liang H, Shi Y, et al. Dual-CNN: a convolutional language decoder for paragraph image captioning. *Neurocomputing*. 2020;396:92-101.
36. Chen Z, Ahn H. Item response theory based ensemble in machine learning. *International Journal of Automation and Computing*. 2020;17(5):3-18.
37. Ning X, Li W, Wei M, et al. Face anti-spoofing based on deep stack generalization networks. 7th International Conference on Pattern Recognition Applications and Methods. The Portuguese Republic; 2018.
38. Ge S, Li J, Ye Q, et al. Detecting masked faces in the wild with LLE-CNNs. IEEE Conference on Computer Vision and Pattern Recognition. USA; 2017.
39. Li R, Zhang X, Chen G, et al. Multi-negative samples with generative adversarial networks for image retrieval. *Neurocomputing*. 2020;394:146-157.
40. Ning X, Nan F, Xu S, Yu L, Zhang L. Multi-view frontal face image generation: a survey. *Concurr Comput Pract Exper*. 2020. https://doi.org/10.1002/cpe.6147.

How to cite this article: Li S, Dong X, Shi Y, Lu B, Sun L, Li W. Multi-angle head pose classification with masks based on color texture analysis and stack generalization. *Concurrency Computat Pract Exper*. 2021;e6331. https://doi.org/10.1002/cpe.6331