IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 March 2024, accepted 1 May 2024, date of publication 9 May 2024, date of current version 22 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3399557

RESEARCH ARTICLE

Hand Function Rehabilitation Training Robot Based on Ycbcr and CNN

SHUZHI SHAN AND JING ZHOU
Kangda College, Nanjing Medical University, Lianyungang 222000, China
Corresponding author: Jing Zhou (ssz00123@126.com)

ABSTRACT Robot technology shows broad application prospects in rehabilitation medicine, especially in hand rehabilitation. Hand function plays an important role that cannot be ignored in daily life, and its key properties are reflected in multiple levels. From basic life skills to occupational needs, healthy hand function is indispensable. Hand function is the foundation for performing daily life skills, including activities such as self-care, eating, dressing, and grooming. The ability to freely use hand functions is directly related to an individual's quality of life and independence. This study proposed a gesture recognition algorithm by fusing Ycbcr color space and convolutional neural network. The method first converted gesture images and recognizes them through the converted images. Then, a hand function rehabilitation training robot based on Ycbcr and CNN was designed, which provided rehabilitation treatment for patients with impaired hand function. These experiments confirmed that when the data set size was 500, the signal-to-noise ratios of YOLOV3, YOLOV3-SPP, YOLOV4, and hybrid algorithms were 27.5dB, 32.7dB, 34.8dB, and 41.2dB, respectively. Their inter-section of union values were 0.53, 0.64, 0.77, and 0.89, respectively. These results confirm that the proposed hybrid algorithm model has good model performance in various algorithms.

INDEX TERMS Gesture recognition, hand function rehabilitation, Ycbcr color space, improved CNN.

I. INTRODUCTION

In daily life, hands play an important role, and many movements rely on the hands to complete. Therefore, hand injuries cause inconvenience to patients' lives [1]. In addition, the cost of rehabilitation for patients with hand injuries is high, which seriously affects the daily life of patients and their families [2], [3]. According to the rehabilitation theory and clinical research of medical experts, patients can actively recover their hand nerves and muscles by completing personalized rehabilitation training. In traditional rehabilitation treatment, doctors need to continuously massage the patient's fingers to quickly re-store the hand function of patients. However, this approach has some limitations. This process will result in a very heavy work-load for doctors. For patients, the cost is high and personalized treatment plans are difficult to achieve, and the treatment effect is not significant. Therefore, hand function rehabilitation training robots are designed to address this issue. However, traditional rehabilitation robots usually use fixed training modes, which cannot fully consider individual patient differences. The lack of personalized rehabilitation plans may lead to un-satisfactory treatment outcomes. Moreover, traditional robots have relatively weak capabilities in data recording and analysis, making it difficult for medical professionals to provide detailed reports on patient rehabilitation progress. So the study puts forward a Gesture Recognition (GR) algorithm that integrates the Ycbcr color space and Convolutional Neural Network (CNN). This method first converts gesture images and recognizes them through the converted images. The research aims to provide better rehabilitation treatment for hand function rehabilitation patients and compensate for the short-comings of traditional rehabilitation treatment. The contribution of the research lies in the fusion of YCbCr color space and CNN for patient GR, providing an efficient method for hand function recovery for patients with impaired hand function. The research content mainly includes four parts. Firstly, a brief introduction is provided to other scholars on the research topic of GR. Next is an explanation of the main methods used in this study. The third part is the model results obtained through the research

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

methods, and these results are analyzed. Finally, a summary of all the above studies and prospects for future research are presented.

II. RELATED WORKS

In early research, researchers studied GR using traditional methods. Ravender et al. believed that GR played an important role in under-standing meaningful movements in the human body. Although there have been studies in GR, the recognition rate of existing GR methods under dynamic gesture conditions in videos is still very low. They proposed a sign language recognition method based on CNN to address this issue. This method was divided into three stages, namely pre-processing, feature extraction, and classification. These experiments confirmed that this method achieved an accuracy of 94.83% in gesture data sets [4]. Tan et al. believed that GR could over-come language barriers and promote human-computer inter-action to enhance communication. Although neural networks are applied to GR, they are not able to encode the direction and position of hands in images. They proposed a visual converter model with attention mechanism for GR to address this issue. This model recognized the input gesture model. These experiments confirmed that the proposed recognition model had good model performance in various data sets [5]. Neethu et al. found that combination gestures were very complex and difficult for machines to classify. They proposed a gesture classification method based on support vector machines to address this issue. This method could convert red, green, and blue color gesture images into Ycbcr images during the pre-processing stage, and then segment the palm and finger regions through threshold processing. Then, the distance transformation method was applied to the segmentation image of palms and fingers. These experiments confirmed that the proposed method had good recognition speed and accuracy on various data sets [6]. Luo et al. designed a speech and gesture signal converter to achieve efficient communication between humans and machines. This converter could convert natural actions into electrical signals and achieve efficient communication in the human-machine interface. These experiments confirmed that in GR, the frictional electric translator could recognize simple gestures and determine the distance between the hand and the sensor based on the principle of electro-static induction [7]. Lv et al. proposed a remote GR system based on a multi-attention mechanism CNN frame-work to address the limitations of low performance of wearable devices. The system could utilize remote server hosts for gesture decoding. These experiments confirmed that the proposed algorithm had good recognition performance on different data sets [8].

Damaneh et al. found that traditional GR systems had low recognition accuracy. To address this issue, they proposed a deep learning-based recognition model to recognize static gestures in sign language. In this model, the proposed structures include CNN and non-intelligent feature extraction methods. These experiments confirmed that the proposed method had good model accuracy in various data sets [9]. Ultra-sonic technology can provide high-resolution solutions independent of light in non-contact human-computer interaction. Regarding this, Kong et al. proposed a GR system based on ultra-sonic frequency modulated continuous wave and ConvLSTM. The system consisted of a transmitter and three receivers, and the signals emitted by the transmitter were detected by the receiver after being manually reflected. Then the gestures were recognized. These experiments confirmed that the proposed system recognized gestures in motion well and had good accuracy [10]. Wang et al. found significant differences in gesture samples due to hand flexibility and personal habits. Traditional methods face great difficulties in recognizing gestures from un-known data sources. They proposed a data augmentation-based GR model to address this issue. These experiments confirmed that the proposed method model recognized complex information from different data sets and exhibited good model performance [11]. Mi et al. proposed an un-supervized fusion network to address the issues of blurred edge and texture features, as well as difficulties in extracting key features in glioma image fusion. The network first converted images in RGB space into Ycbcr space images, and then mapped the object features of the images. These experiments confirmed that the proposed method effectively improved the rich and natural color information of medical images [12]. Neethu et al. proposed a gesture classification method using Support Vector Machine (SVM) to address the low efficiency in existing gesture detection. This method converted red green blue color gesture images into Ycbcr images in the pre-processing stage, and then segmented the palm and finger regions through threshold processing. Finally, the image was classified using SVM. These experiments confirmed that the proposed algorithm had good recognition efficiency in various data sets, and it took less time [6].

In summary, many scholars have conducted research on gestures and achieved certain results, but no one introduces GR into hand function recovery. This study proposes a GR algorithm based on Ycbcr color space and CNN fusion, which converts gesture images and recognizes gestures through the converted images. Then, GR is applied to patients with hand dys-function to assist them in rehabilitation training.

III. HAND FUNCTION REHABILITATION TRAINING MODEL BASED ON YCBCR AND CNN

Section A of this study introduces Ycbcr and proposes a hand function rehabilitation training model based on Ycbcr and CNN to address the low detection rate of existing GR methods in complex environments. In the second section, a computer vision GR system based on deep learning is designed to test the application of GR in rehabilitation robots.

A. GESTURE DETECTION MODEL BASED ON YCBCR AND CNN

CNN mainly processes data with grid-like structures, such as images and videos. The core idea of CNN is to capture local features of input data through convolution operations

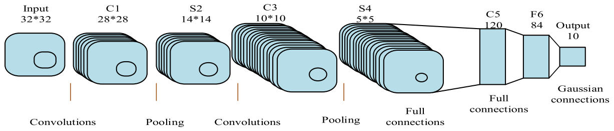


FIGURE 1. Structure diagram of convolutional neural network.

and reduce the spatial size of data through pooling operations, thereby reducing the number of parameters and improving computational efficiency. Figure 1 shows the structure of CNN.

Convolutional operation is the core of CNN. Convolutional layers gradually extract local features by sliding a convolutional kernel on the input data, which helps the network capture spatial structural information in the image [13]. Anon-linear activation function is usually applied to introduce non-linear characteristics after the convolutional layer. Pooling can reduce the feature map size while preserving the main information, mainly including maximum and average pooling. After the convolutional and pooling layers, there are usually one or more fully connected layers, which are used to map high-level features to output categories and complete the final classification or regression task. The study adopts a CNN structure with 12 convolutional layers and 4 pooling layers, using ReLu activation function and standardizing data at each convolutional level. 64 filters are used in the first and second convolutional layers, and then increased to 128, 256, and 512 filters in subsequent convolutional layers to further improve the feature extraction ability of images.

Ycbcr is a color coding method commonly used for the representation of digital images and videos. This encoding method separates color information and brightness information, making it more effective in compressing and processing images. Ycbcr represents a pixel with three components: brightness component Y, blue color difference component Cb, and red color difference component Cr. Y represents the brightness of the image, which is the gray-scale information, and the brightness component usually occupies most of the information encoded in the image [14], [15]. Cb represents the difference between the blue color and brightness of a pixel. Cr represents the difference between the red color of a pixel and its brightness. The RGB color space conversion of Y is represented by equation (1).

$$Y = 0.299R + 0.587G + 0.114B \tag{1}$$

In equation (1), Y represents the brightness component. R is the value of the red color component in RGB. G means the value of the green color component in RGB. B represents the value of the blue color component in RGB. Cb is represented by equation (2).

$$Cb = -0.1687R - 0.3313G + 0.5B + 128 \tag{2}$$

Cr is represented by equation (3).

$$Cr = 0.5R - 0.4187G - 0.0813B + 128 \tag{3}$$

A main advantage of Ycbcr is that this method optimizes the perceptual characteristics of the human eye, allowing for more effective compression while maintaining image quality. This fully utilizes the fact that the human eye is more sensitive to brightness changes and relatively less sensitive to color changes. Therefore, the study adopts the Ycbcr color space method to extract images of gesture skin regions [16]. Although Ycbcr is not an absolute color space, it is a method of encoding RGB information. Figure 2 shows the color space of Ycbcr.

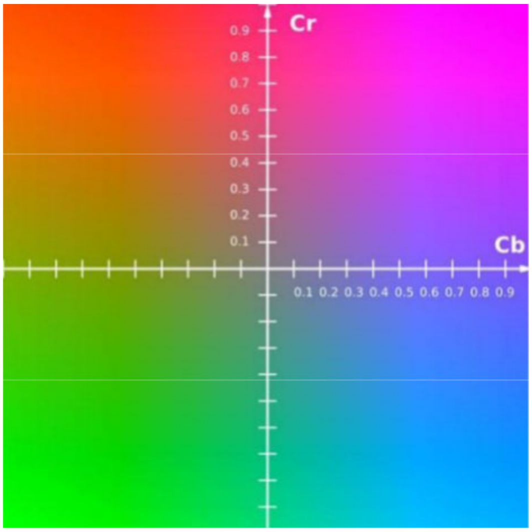


FIGURE 2. Ycbcr color space.

Because skin color is not affected by light and other factors in the color space, skin color detection can distinguish gestures and back-grounds. The key to skin color detection is the selection of thresholds. The gesture segmentation based on RGB images and the filtering method based on HSV images are both universal thresholds obtained from experimental data in specific contexts [17]. Although this method can have good performance in most cases, the threshold range cannot completely smooth the division of all gesture images. This study uses an OTSU segmentation method based on the Cr component of the Ycbcr color space to segment gestures. Figure 3 shows the main steps of OTSU segmentation method.

In Figure 3, the input image is first subjected to gray-scale processing. The quantity of pixels at each gray-scale level is counted to generate a gray-scale histogram. Then, the quantity of pixels in the histogram is divided by the total pixels in the image to obtain the probability distribution for each gray-scale level. The probability distribution for each gray-scale level is represented by equation (4).

$$p_k = \frac{n_k}{N} \tag{4}$$

In equation (4), p_k represents the gray-scale histogram. n_k is the quantity of pixels with a gray-scale value of k. N means the total number of pixels in the image [18]. Then, for each possible threshold t, the intra class variance of dividing the image into two categories is calculated, which is the weighted sum of the variances of pixel values within each category,

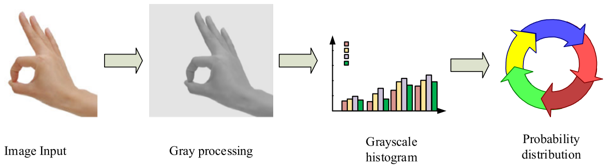


FIGURE 3. Main steps of OTSU segmentation method.

represented by equation (4).

$$\sigma^2(t) = w_1(t) \cdot w_2(t) \cdot (\mu_1(t) - \mu_2(t))^2 \tag{5}$$

In equation (5), w_1 and w_2 represent the probability distributions of two different categories of images, respectively. μ_1 and μ_2 are the average pixel values of two different categories of images, respectively. Then, among all possible thresholds, the threshold that minimizes the sum of intra class variances is selected as the optimal threshold. Finally, the image is binarized using the optimal threshold. Pixels below or equal to the threshold are classified as one category, while pixels above the threshold are classified as another category. Gesture images with sufficient information and smooth image contours can be obtained by using the OTSU segmentation method to segment gestures [19]. The Ycbcr method is combined with CNN to obtain the Ycbcr-CNN algorithm model. Ycbcr-CNN is a CNN based on the Ycbcr color space. Ycbcr is a color coding method used for digital image processing, which decomposes images into three components: brightness and chromaticity. The Y component represents the brightness information of the image, while the Cb and Cr components represent color information. The main purpose of Ycbcr-CNN is to utilize the characteristics of Ycbcr color space to improve the performance of image processing tasks. In traditional RGB color space, CNN may be affected by color information, leading to un-stable model performance or sensitivity to color changes. In contrast, the Ycbcr color space represents the brightness and color information of an image separately, which helps to improve the model's robustness to brightness changes and reduce sensitivity to color changes. The components of Ycbcr-CNN are similar to the traditional CNN, including convolutional layers, pooling layers, activation functions, and fully connected layers. However, the input image is usually first converted to the Ycbcr color space in Ycbcr-CNN, and then input to the network for processing. As a result, the network can better utilize the brightness and color information of the image, thereby improving the performance and robustness of the model. Figure 4 shows the proposed GR model based on Ycbcr and CNN.

In Figure 4, the gesture data set is first pre-processed by extracting hand region features, performing dimensionality reduction and gray-scale processing, and finally enhancing the image. Then, the pre-processed images are input into CNN for detection test, and the detection results are output [20]. Due to the large number of parameters and the complexity of the model, CNN is prone to overfit. To address this issue, Dropout is used to improve CNN. Dropout can randomly discard some of its neurons during training to avoid over-fit in Figure 5.

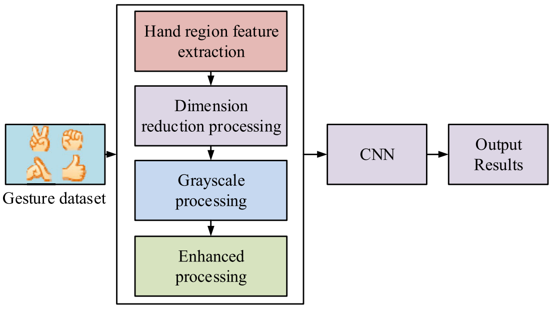


FIGURE 4. Gesture detection algorithm flow based on Ycbcr and CNN.

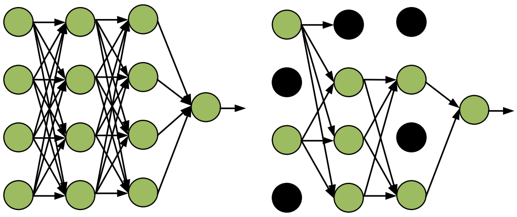


FIGURE 5. CNN with Dropout network application.

In Figure 5, Dropout adjusts certain neurons during learning to avoid participating in operations, thereby reducing parameters and avoiding over-fitting. When classifying different gestures, a cross-loss function can measure the error between the true and predicted values, and the discrete function is represented by equation (6).

$$H(p, q) = -\frac{1}{N}\sum_{i=1}^{n} p_i \ln q \tag{6}$$

In equation (6), p represents the true label in the sample. q is the predicted label in the sample. N means the size of the batch block. It is necessary to compare the predictive performance to better evaluate the ability of the model [21]. In deep learning-based object detection algorithms, common indicators include accuracy, recall, Inter-section of Union (IoU), and F1 score. This study evaluates the performance of the model using IoU, mean precision, and detection speed. In the object detection, IoU is generally used to represent the over-lap ratio between the predicted bounding boxes and the actual annotated bounding boxes. The IoU value is high, indicating a high accuracy of model detection, represented by equation (7).

$$IoU = \frac{AO}{AU} \tag{7}$$

In equation (7), AO represents the inter-section area, which refers to the area where two bounding boxes inter-sect. AU is the union area, which refers to the total area covered by two bounding boxes [22]. The range of IoU values is 0-1. 0 indicates no over-lap, 1 indicates complete over-lap. The average precision can measure the effectiveness of various types of detection, represented by equation (8).

$$mAP = \frac{\sum AP}{N} \tag{8}$$

In equation (8), AP is the average precision. N means the quantity of indicators. Detection speed refers to the quantity of frames processed per second in computer graphics and video processing, represented by equation (9).

$$FPS = \frac{1}{AT} \tag{9}$$

In equation (9), AT is the average detection time, which refers to the average object detection time calculated on a certain number of frames.

B. GESTURE DETECTION MODEL BASED ON YCBCR AND CNN IN REHABILITATION ROBOTS

The system structure of rehabilitation robots includes image acquisition, pre-processing, classification, and detection. Before under-going hand function rehabilitation training, it is necessary to evaluate the patient's condition. Brunnstrom is used to evaluate the finger movement function of patients for rehabilitation evaluation [23]. According to the recovery process of finger movement function, Brunnstrom can be divided into six stages in Table 1.

TABLE 1. Finger function performance at different stages.

Different stages	Embody
Phase 1	Fingers lose their motor function and cannot move
Phase 2	Fingers can be slightly bent and extended
Phase 3	Fingers into hooks, can't straighten hand.
Phase 4	Fingers are able to stretch slightly.
Phase 5	Fingers with basic grasping ability
Phase 6	Finger function is largely re-stored.

In Table 1, from stages 1 to 6, the finger movement ability gradually recovers. In stage 1, the fingers have almost no movement function. In stage 2, the fingers of patients can respond to commands. In stage 3, the hands of patients can be slightly moved, but cannot be straightened. In stage 4, the fingers of patients can be slightly extended and there is a significant reduction in convulsions and cramps. In stage 5, the fingers of patients have basic grasping ability [24]. In stage 6, the finger function of patients is basically re-stored.

In the image pre-processing, a camera is used to capture rehabilitation gestures, and then the gestures are segmented into multiple pixels to obtain a pixel map. In computer vision, different levels of gray-scale images can be obtained through cameras. However, when processing images, not much information is required, so an adaptive threshold binarization algorithm is used to segment the collected images. This method will gray the collected color images [25]. For images with significant changes in brightness, adaptive threshold binarization is generally used to determine the optimal threshold of the image based on the data information in the gray-scale histogram. The adaptive threshold for image binarization is calculated using the maximum inter class variance method, and the segmented image is represented by equation (10).

$$g(x,y) = \begin{cases} 1 & f(x,y) \le T \\ 0 & f(x,y) > T \end{cases} \tag{10}$$

In equation (10), 0 and 1 represent the back-ground and fore-ground, respectively. $f(x, y)$ means gray-scale level. T represents the threshold. When segmenting images, it is necessary to determine the optimal threshold for image processing. When the difference between the fore-ground and back-ground reaches its maximum, OTSU can be used to find the optimal threshold, and the total average gray-scale of the image is represented by equation (11).

$$U = W_0 * U_0 + W_1 * U_1 \tag{11}$$

In equation (11), U means the total average gray-scale. W_0 and W_1 are the fore-ground and back-ground points, respectively. U_0 and U_1 represent the average fore-ground gray-scale and the average back-ground gray-scale, respectively [26]. The variance of fore-ground and back-ground images is represented by equation (12).

$$S = W_0 * W_1 * (U_0 - U_1)^2 \tag{12}$$

In equation (12), S means the variance of the fore-ground and back-ground images. W_0 and W_1 represent the fore-ground and back-ground points, respectively. U_0 and U_1 are the average fore-ground gray-scale and the average back-ground gray-scale, respectively. When the variance value is maximum, the gray-scale value is the optimal threshold. It should process the image and reduce the irrelevant information such as noise to reduce the amount of information contained in the image. The noise in the image can be removed and blurred through filters, and a smooth average filter [27], [28] is used in this research. Smooth average filter is a filter used to reduce noise in signals or images. The basic principle of this method is to take the average of a certain number of data points in its neighbor-hood for each data point in the signal to reduce the influence of random noise and smooth the changes in the signal. For one-dimensional signals, the calculation of the smooth average filter is represented by equation (13).

$$y[n] = \frac{1}{N}\sum_{k=0}^{N-1} x[n-k] \tag{13}$$

In equation (13), $x[n]$ represents the value of the original signal at time n. $y[n]$ is the smoothed signal. N is the window size of the filter, representing the neighbor-hood size for taking the average. In two-dimensional image processing, smooth average filters typically use a rectangular or circular neighbor-hood to calculate the average value of each pixel [29], [30]. For two-dimensional images, the calculation of the smoothed image is represented by equation (14).

$$J(x,y) = \frac{1}{N \times N}\sum_{I=0}^{N-1}\sum_{J=0}^{N-1} I(x-i, y-i) \tag{14}$$

In equation (14), $J(x, y)$ is the value of the smoothed image at position (x, y). $N \times N$ represents the window size of the filter. Figure 6 shows the design of the rehabilitation system.

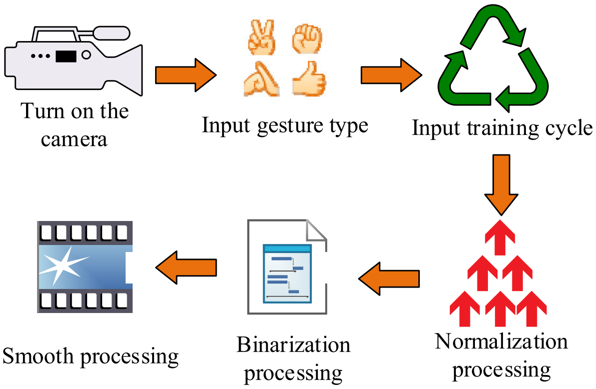


FIGURE 6. Design of rehabilitation system.

In Figure 6, different gesture training combinations are collected and defined to achieve the rehabilitation of hand function. The type of gesture collected is inputted, and the training cycle is set by calling the camera. Then, the image is processed using the Ycbcr color space algorithm to gray-scale the gesture image and obtain the corresponding gray-scale image [31], [32]. Then, noise is eliminated through a smooth mean filter to obtain labeled data. The standardized data are transformed into raw data through an improved CNN to verify the accuracy and stability of the model.

IV. PERFORMANCE ANALYSIS OF HAND FUNCTION REHABILITATION TRAINING ROBOT BASED ON YCBCR AND CNN

Section A introduced YOLOV3, YOLOV3-SPP, and YOLOV4, respectively, and the performance of these four algorithms was compared on different data sets. Section B introduced various algorithms into the system and analyzed the system performance.

A. PERFORMANCE ANALYSIS OF GESTURE DETECTION MODEL BASED ON YCBCR AND CNN

The data set used in this study is Ego Gesture and OUHANDs. Ego Gesture includes 83 gestures and 24161 gesture samples from different scenes, consisting of RGB images and depth images. OUHANDs is an open-source GR data set that contains a total of 3000 pieces of data. The training and testing sample data are 2:1, and the image resolution is 640*480. Most of the images are indoor images in RGB format. The experimental hardware uses Intel Core i5-8750H CPU, with NVIDIA Geforce GTX2080Ti GPU, 8GB graphics memory, and 16GB memory. In Figure 7, YOLOV3, YOLOV3-SPP, and YOLOV4 were compared with Ycbcr-CNN.

Figure 7 (a) shows the Signal-to-Noise Ratio (SNR) of four algorithms in Ego Gesture. Figure 7 (b) is the SNR of the four algorithms in OURHANDs. In Figure 7 (a), as the data set continued to increase, the SNR of these four algorithms also increased, indicating that the performance of each model was gradually improved. When the data set size was 500, the SNR of YOLOV3, YOLOV3-SPP, YOLOV4, and Ychcr-CNN algorithms was 27.5dB, 32.7dB, 34.8dB, and 41.2dB, respectively. In Figure 7 (b), as the data set continued to increase, the proposed algorithm's SNR also continued to increase, and its performance was stronger than that shown in Ego Gesture. When the data set size was 1000, the SNR of YOLOV3, YOLOV3-SPP, YOLOV4, and Ychcr-CNN was 29.7dB, 33.1dB, 37.2dB, and 43.7dB, respectively. These experiments confirm that the proposed Ycher CNN has better performance compared to other algorithms and performs well in different data sets. The IoU of these four algorithms was compared to further validate the model performance in Figure 8.

Figure 8 (a) shows the model IoU of the four algorithms in Ego Gesture. Figure 8 (b) shows the model IoU of the four algorithms in OURHANDs. In Figure 8 (a), as the data set increased, the IoU of these four models also continuously increased. IoU is generally used to represent the over-lap ratio between the predicted bounding boxes detected and the actual annotated bounding boxes. The IoU is high, indicating a high detection accuracy. When the data set size was 500, the IoU of YOLOV3, YOLOV3-SPP, YOLOV4, and Ychcr-CNN was 0.53, 0.64, 0.77, and 0.89, respectively. In Figure 8 (b), as the data set increased, the IoU of these four algorithms also increased, and the performance was better than that of Ego Gesture. When the data set size was 500, the IoU of YOLOV3, YOLOV3-SPP, YOLOV4, and Ychcr-CNN was 0.57, 0.72, 0.79, and 0.91, respectively. These experiments confirm that the proposed Ychcr-CNN performs the best among the four algorithms and shows good performance in both data sets. The running time of the models was compared in Figure 9.

Figures 9 (a) and (b) show the recognition time of different algorithms in Ego Gesture and OURHANDs. In Figure 9 (a), when the iteration was small, the performance of the model did not reach its optimal level, resulting in a longer recognition time. When the iteration was 90, the recognition time for YOLOV3-SPP, YOLOV4, and Ychcr-CNN was 1.3s, 2.1s, and 2.3s, respectively. In Figure 9 (b), the time taken by each algorithm on OHRHANDs was slightly longer than that of Ego Gesture. When the data set size was 160, the recognition time for YOLOV3-SPP, YOLOV4, and Ychcr-CNN was 3.6s, 3.8s, and 4.2s, respectively. These experiments confirm that the proposed algorithm performs well on two different data sets, and the recognition time for Ychcr-CNN is shorter than other algorithms. The overall performance of the four algorithms was compared in Table 2.

In Table 2, the P, R, IoU, F1, mAP, and loss function values corresponding to YOLOV3 were 0.87, 0.74, 76.52%, 0.72, 0.79, and 0.97, respectively. The P, R, IoU, F1, mAP, and loss function values corresponding to YOLOV3-SPP were 0.91, 0.77, 82.41%, 0.81, 0.82, and 0.82, respectively. The P, R, IoU, F1, mAP, and loss function values corresponding to the YOLOV4 method were 0.94, 0.82, 86.34%, 0.92, 0.81, and 0.076, respectively. The P, R, IoU, F1, mAP, and loss function values corresponding to Ycbcr-CNN were 0.96, 0.89, 88.97%, 0.96, 0.92, and 0.064, respectively. These experiments confirm that Ycbcr-CNN exhibits higher performance than other algorithms in all aspects.

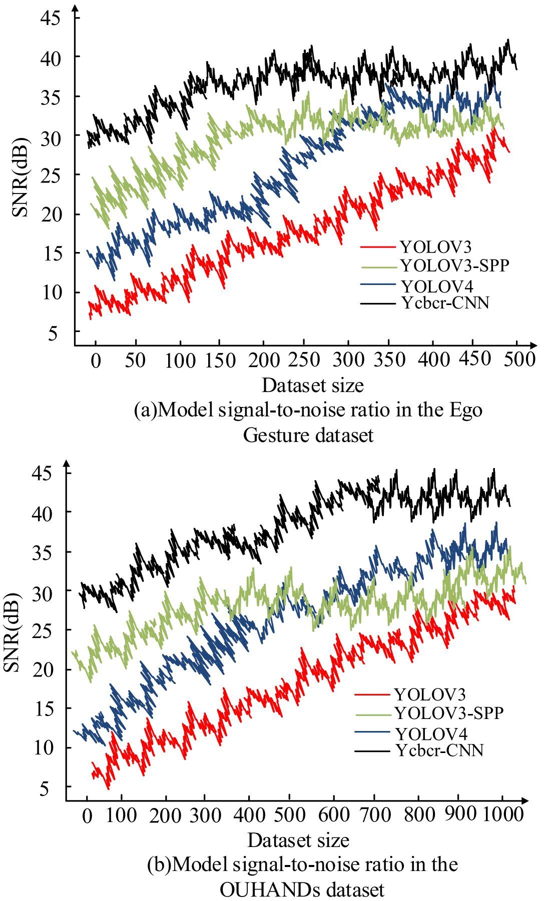


FIGURE 7. Model signal-to-noise ratio under two different data sets.

B. PERFORMANCE TESTING OF REHABILITATIONROBOTS BASED ON YCBCR AND CNN GESTURE DETECTION MODEL

By applying the proposed algorithm to rehabilitation robots, Figure 10 is the recognition accuracy of the model.

Figure 10 (a) is the accuracy of different algorithms on different data sets, while Figure 10 (b) is the accuracy of different algorithms at different iterations. In Figure 10 (a), as the validation set increased, the accuracy of each algorithm decreased, with the proposed Ycbcr-CNN having the smallest decrease in accuracy. When the data set was 900, the accuracy of YOLOV3, YOLOV3-SPP, YOLOV4, and Ychcr-CNN was 0.668, 0.678, 0.697, and 0.720, respectively. In Figure 10 (b), as the iteration increased, the recognition performance of each algorithm continuously improved, with the proposed Ycbcr-CNN having the highest accuracy. When the iteration reached around 50, the accuracy of YOLOV3, YOLOV3-SPP, YOLOV4, and Ychcr-CNN was 0.921, 0.934, 0.937, and 0.944, respectively. These experiments confirm that the proposed Ychcr-CNN has superior performance compared to other algorithms. By selecting different data sets to form a validation set, Figure 11 shows the analysis results of the algorithm's computation time.

Figures 11 (a), (b), (c), and (d) show the computation time of Ycbcr-CNN, YOLOV4, YOLOV3-SPP, and YOLOV3 in different data sets. The shaded areas in the figure represent the total time spent by each algorithm. In the figure, among these four algorithms, the proposed Ycbcr-CNN had lower computation time. These results indicate that the proposed method has better overall performance compared to other algorithms. The performance of four algorithms in different scenarios was compared in Figure 12.

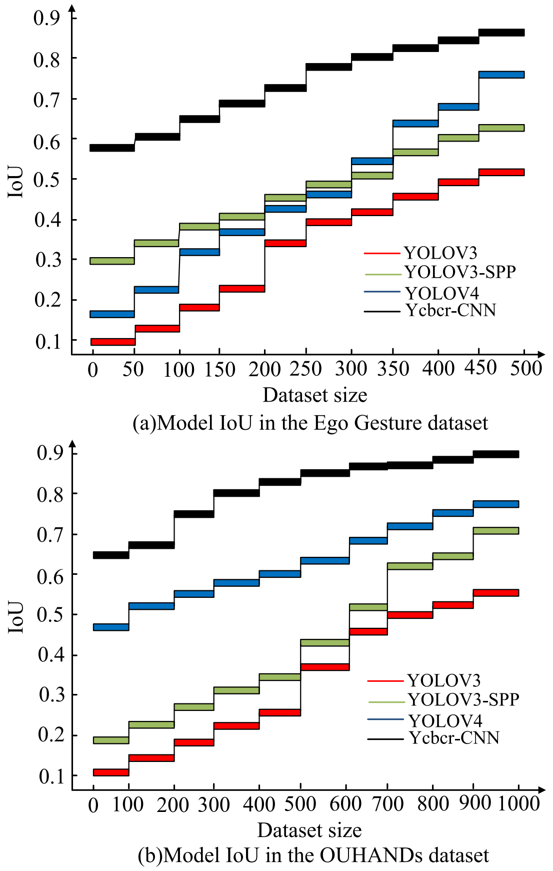
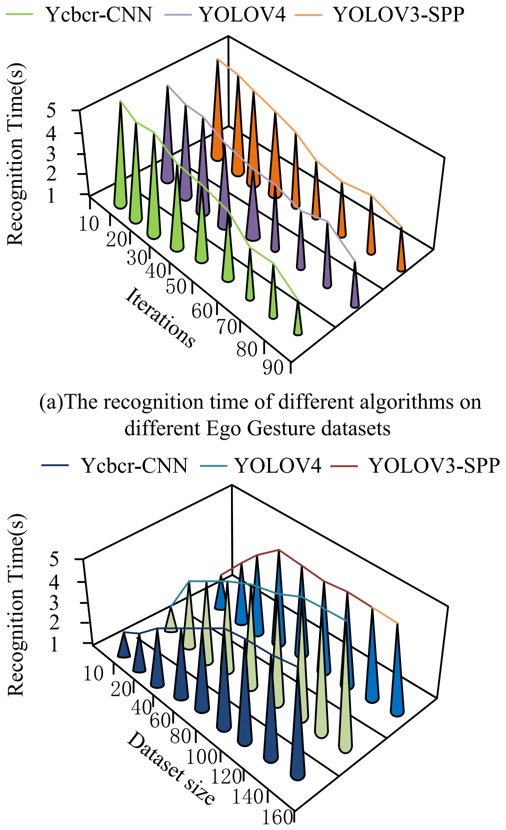


FIGURE 8. Inter-section and union ratio of models under two different data sets.

Figures 12 (a) and (b) are the recognition time and accuracy of different algorithms in different environments. In Figure 12 (a), among these four algorithms, the proposed Ycbcr-CNN had a relatively stable recognition time compared to other algorithms. The recognition time in different environments was 1.9s, 2.1s, 3.1s, and 2.9s, respectively. In Figure 12 (b), the accuracy of Ycbcr-CNN was around 0.8 in different environments. These experiments confirm that the proposed algorithm has high performance. Ablation experiments were introduced to further evaluate the performance of the model. The Ycbcr method was used to process the data, so the Ycbcr method was used as a variable in the ablation experiment. RGB images were converted to YCbCr color space. The Y component was used to correlate with the sensitivity of the human eye to brightness, while reducing the amount of data processed. The Y component usually contained most of the image information, so denoising was performed on the Y component at this stage. The chromaticity component usually had a small impact on image quality, so there was no need for denoising. The training data of the



(a)The recognition time of different algorithms on different Ego Gesture datasets

(b)The recognition time of different algorithms on different OURHANDs datasets

FIGURE 9. Recognition time of different algorithm models in different data sets.

TABLE 2. Overall performance comparison of algorithms.

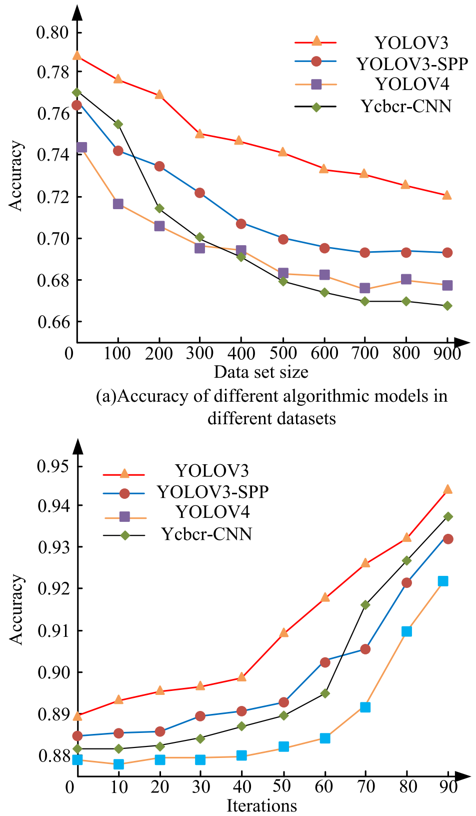
Model	P	R	IoU	F1	mAP	Loss
YOLOV3	0.87	0.74	76.52%	0.72	0.79	0.97
YOLOV3-SPP	0.91	0.77	82.41%	0.81	0.82	0.82
YOLOV4	0.94	0.82	86.34%	0.92	0.81	0.076
Ycbcr-CNN	0.96	0.89	88.97%	0.96	0.92	0.064

TABLE 3. Results of ablation experiment.

Model	P	R	IoU	F1	mAP	Loss
YOLOV3	0.84	0.72	76.52%	0.7	0.77	0.97
Ycbcr-YOLOV3	0.91	0.79	80.24%	0.77	0.79	0.24
YOLOV3-SPP	0.89	0.75	82.41%	0.79	0.8	0.82
Ycbcr-YOLOV3-SPP	0.92	0.81	85.62%	0.81	0.83	0.11
YOLOV4	0.92	0.81	86.34%	0.9	0.79	0.076
Ycbcr-YOLOV4	0.93	0.83	87.65%	0.93	0.88	0.067
Ycbcr-CNN	0.94	0.87	88.97%	0.94	0.9	0.064

denoising model were enhanced to improve the generalization ability. The ablation experiments were conducted on the model algorithms mentioned in this study. The results are shown in Table 3.

In Table 3, the performance of each algorithm model increased with the addition of Ycbcr processing after conducting ablation experiments on the models. The performance increased significantly compared to the previous algorithm model. The research results indicate that the proposed algorithm model has high performance. Fifty patients who needed to recover their hand function were randomly selected. After rehabilitation, Table 4 shows the scores of each patient on the machine.



(a)Accuracy of different algorithmic models in different datasets

(b)Accuracy of different algorithmic models at different number of iterations

FIGURE 10. Performance comparison of different algorithms.

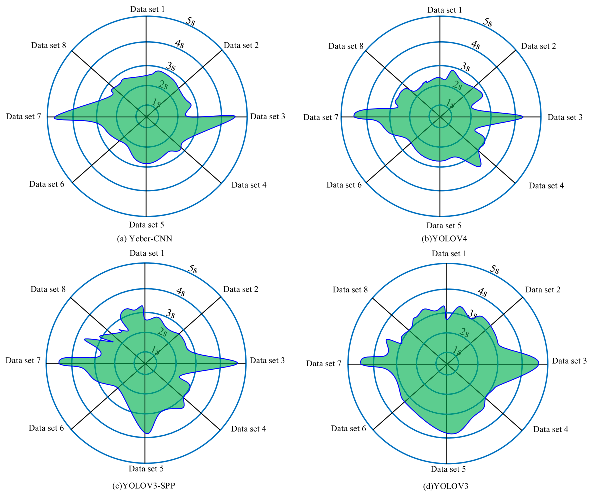
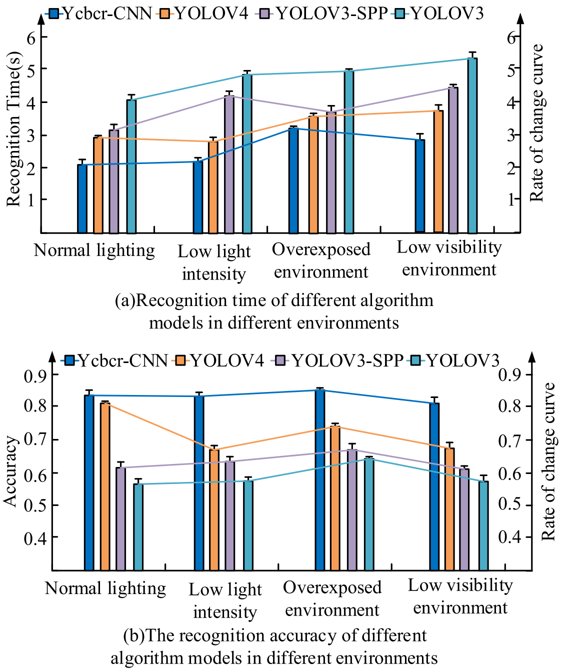


FIGURE 11. Operation time of four methods.

In Table 4, the evaluation score of Ycbcr-CNN robots for these five groups was 93.5, 94.2, 87.1, 88.5, and 87.9, respectively. The evaluation score for YOLOV4 robots in five groups was 87.1, 86.5, 84.3, 82.3, and 81.4, respectively. The



(a)Recognition time of different algorithm models in different environments

(b)The recognition accuracy of different algorithm models in different environments

FIGURE 12. Performance comparison of four algorithms.

TABLE 4. User evaluation form.

Model	Patient group 1	Patient group 2	Patient group 3	Patient group 4	Patient group 5	Rehabilitation level/%
Ycbcr-CNN	93.5	94.2	87.1	88.5	87.9	98.7
YOLO V4	87.1	86.5	84.3	82.3	81.4	87.9
YOLO V3-SPP	84.3	82.1	80.2	83.5	78.4	76.9
YOLO V3	78.5	79.8	78.5	76.6	75.2	75.4

evaluation score for YOLOV3-SPP robots in five groups was 84.3, 82.1, 80.2, 83.5, and 78.4, respectively. The evaluation score for YOLOV3 robots in five groups was 78.5, 79.8, 78.5, 76.6, and 75.2, respectively. Meanwhile, patients based on the Ycbcr-CNN model achieved a hand function recovery rate of 98.7%, which was relatively excellent compared to the other three models. These experiments confirmed that the proposed Ycbcr-CNN-based robot received positive feedback from patients.

V. CONCLUSION

The importance of hand function is not only reflected in the impact on physical health, but also in the profound impact on overall quality of life and social participation ability. For individuals affected by hand dys-function, rehabilitation training and technical intervention become particularly crucial. This study proposed a GR algorithm based on the fusion of Ycbcr color space and CNN. This method first converted gesture images and recognized gestures through the converted images. Then a hand function rehabilitation training robot based on Ycbcr and CNN was designed, which provided rehabilitation treatment for patients with impaired hand function. These experiments confirmed that in Ego Gesture, when the data set size was 500, the SNR of YOLOV3, YOLOV3-SPP, YOLOV4, and hybrid algorithms was 27.5dB, 32.7dB, 34.8dB, and 41.2dB, respectively. Their IoU was 0.53, 0.64, 0.77, and 0.89, respectively. The recognition time of the algorithm was 2.3s, 2.6s, 2.7s, and 2.9s, respectively. The P, R, IoU, F1, mAP, and loss function values corresponding to Ycbcr-CNN were 0.96, 0.89, 88.97%, 0.96, 0.92, and 0.064, respectively. When the data set size was 900, the accuracy of YOLOV3, YOLOV3-SPP, YOLOV4, and Ychcr-CNN was 0.668, 0.678, 0.697, and 0.720, respectively. The proposed Ycbcr-CNN had lower computation time and was relatively stable. The recognition time in different environments was 1.9s, 2.1s, 3.1s, and 2.9s, respectively. The evaluation scores for Ycbcr-CNN robots in five patient groups were 93.5, 94.2, 87.1, 88.5, and 87.9, respectively. The research results indicate that the proposed algorithm has good model performance, but there are still short-comings in research. The real-time performance of this model is still insufficient. To widely use GR-based human-computer interaction in social life, it is necessary to improve its accuracy and real-time performance to meet the needs of normal interaction. The proposed algorithmic model also has certain limitations. The Ycbcr color space divides the image into brightness and chromaticity components, but some detailed information may be lost in this process. The training data of Ycbcr-CNN are based on the Ycbcr color space, rather than directly using RGB images, which can lead to insufficient training data. Ycbcr-CNN may cause color distortion problems when processing images, especially in tasks sensitive to color information, where color distortion or offset may occur. Due to the separation of color information and brightness information by Ycbcr-CNN, the generalization ability of the model may be insufficient in handling certain specific color or brightness situations, resulting in poor performance on new data.

REFERENCES

[1] P. Gao, Z. Lv, Y. Song, M. Song, and P. Qian, "Evolution of morphology and microstructure of coarsened nanoporous gold studied by automatic thresholding and image recognition algorithms," *Scripta Mater.*, vol. 226, Mar. 2023, Art. no. 115256, doi: 10.1016/j.scriptamat.2022.115256.

[2] S. Wang, "Application of deep convolutional neural networks in image recognition and classification in library management," *Wireless Pers. Commun.*, vol. 29, pp. 236–249, Jun. 2023, doi: 10.1007/s11277-023-10571-5.

[3] J. Wang, C. Wang, Q. Lin, C. Luo, C. Wu, and J. Li, "Adversarial attacks and defenses in deep learning for image recognition: A survey," *Neurocomputing*, vol. 514, pp. 162–181, Dec. 2022, doi: 10.1016/j.neucom.2022.09.004.

[4] M. Ravinder, K. Malik, M. Hassaballah, U. Tariq, K. Javed, and M. Ghoneimy, "An approach for gesture recognition based on a lightweight convolutional neural network," *Int. J. Artif. Intell. Tools*, vol. 32, no. 3, pp. 185–199, May 2023, doi: 10.1142/s0218213023400146.

[5] C. K. Tan, K. Lim, Y. Chang, C. Lee, and A. Alqahtani, "HGR-ViT: Hand gesture recognition with vision transformer," *Sensors*, vol. 23, no. 5, pp. 274–286, 2023, doi: 10.3390/s23125555.

[6] P. S. Neethu, R. Suguna, and P. S. Rajan, "Performance evaluation of SVM-based hand gesture detection and recognition system using distance transform on different data sets for autonomous vehicle moving applications," *Circuit World*, vol. 48, no. 2, pp. 204–214, Mar. 2022, doi: 10.1108/cw-06-2020-0106.

[7] H. Luo, J. Du, P. Yang, Y. Shi, Z. Liu, D. Yang, L. Zheng, X. Chen, and Z. L. Wang, ''Human–Machine interaction via dual modes of voice and gesture enabled by triboelectric nanogenerator and machine learning,'' *ACS Appl. Mater. Interfaces*, vol. 15, no. 13, pp. 17009–17018, Apr. 2023, doi: 10.1021/acsami.3c00566.

[8] X. Lv, C. Dai, H. Liu, Y. Tian, L. Chen, Y. Lang, R. Tang, and J. He, ''Gesture recognition based on sEMG using multi-attention mechanism for remote control,'' *Neural Comput. Appl.*, vol. 35, no. 19, pp. 13839–13849, Jul. 2023, doi: 10.1007/s00521-021-06729-6.

[9] M. M. Damaneh, F. Mohanna, and P. Jafari, ''Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using ORB descriptor and Gabor filter,'' *Exp. Syst. Appl.*, vol. 211, Jan. 2023, Art. no. 118559, doi: 10.1016/j.eswa.2022.118559.

[10] F. Kong, J. Deng, and Z. Fan, ''Gesture recognition system based on ultrasonic FMCW and ConvLSTM model,'' *Measurement*, vol. 190, Feb. 2022, Art. no. 110743, doi: 10.1016/j.measurement.2022.110743.

[11] L. Wang, Z. Cui, Y. Pi, C. Cao, and Z. Cao, ''Low personality-sensitive feature learning for radar-based gesture recognition,'' *Neurocomputing*, vol. 493, pp. 373–384, Jul. 2022, doi: 10.1016/j.neucom.2022.04.035.

[12] J. Mi, L. Wang, Y. Liu, and J. Zhang, ''UEFSD: Unsupervised medical images fusion based on exclusive features and saliency detection for SPECT-MRI images of glioma,'' *Measurement*, vol. 216, Jul. 2023, Art. no. 112896, doi: 10.1016/j.measurement.2023.112896.

[13] W. Hu, Y. Huang, F. Zhang, R. Li, and H. Li, ''SeqFace: Learning discriminative features by using face sequences,'' *IET Image Process.*, vol. 15, no. 11, pp. 2548–2558, Sep. 2021, doi: 10.1049/ipr2.12243.

[14] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, ''Attention mechanism-based CNN for facial expression recognition,'' *Neurocomputing*, vol. 411, pp. 340–350, Oct. 2020, doi: 10.1016/j.neucom.2020.06.014.

[15] Q. Zhu, L. Gao, H. Song, and Q. Mao, ''Learning to disentangle emotion factors for facial expression recognition in the wild,'' *Int. J. Intell. Syst.*, vol. 36, no. 6, pp. 2511–2527, Jun. 2021, doi: 10.1002/int.22391.

[16] N. B. Kar, D. R. Nayak, K. S. Babu, and Y. Zhang, ''A hybrid feature descriptor with Jaya optimised least squares SVM for facial expression recognition,'' *IET Image Process.*, vol. 15, no. 7, pp. 1471–1483, May 2021, doi: 10.1049/ipr2.12118.

[17] X. Jin and Z. Jin, ''MiniExpNet: A small and effective facial expression recognition network based on facial local regions,'' *Neurocomputing*, vol. 462, pp. 353–364, Oct. 2021, doi: 10.1016/j.neucom.2021.07.079.

[18] R. Zhi, C. Zhou, T. Li, S. Liu, and Y. Jin, ''Action unit analysis enhanced facial expression recognition by deep neural network evolution,'' *Neurocomputing*, vol. 425, pp. 135–148, Feb. 2021, doi: 10.1016/j.neucom.2020.03.036.

[19] W. Yang, H. Gao, Y. Jiang, J. Yu, J. Sun, J. Liu, and Z. Ju, ''A cascaded feature pyramid network with non-backward propagation for facial expression recognition,'' *IEEE Sensors J.*, vol. 21, no. 10, pp. 11382–11392, May 2021, doi: 10.1109/JSEN.2020.2997182.

[20] N. Baruch, S. Behrman, P. Wilkinson, T. Bajorek, S. E. Murphy, and M. Browning, ''Negative bias in interpretation and facial expression recognition in late life depression: A case control study,'' *Int. J. Geriatric Psychiatry*, vol. 36, no. 9, pp. 1450–1459, Sep. 2021, doi: 10.1002/gps.5557.

[21] G. Lokku, G. H. Reddy, and M. N. G. Prasad, ''Optimized scale-invariant feature transform with local tri-directional patterns for facial expression recognition with deep learning model,'' *Comput. J.*, vol. 65, no. 9, pp. 2506–2527, Sep. 2022, doi: 10.1093/comjnl/bxab088.

[22] Y. Liu, X. Zhang, J. Zhou, and L. Fu, ''SG-DSN: A semantic graph-based dual-stream network for facial expression recognition,'' *Neurocomputing*, vol. 462, pp. 320–330, Oct. 2021, doi: 10.1016/j.neucom.2021.07.017.

[23] Y.-T. Li, J.-Z. Li, L. Ren, K. Xu, S. Chen, L. Han, H. Liu, X.-L. Guo, D.-L. Yu, D.-H. Li, L. Ding, L.-M. Peng, and T.-L. Ren, ''Light-controlled reconfigurable optical synapse based on carbon nanotubes/2D perovskite heterostructure for image recognition,'' *ACS Appl. Mater. Interfaces*, vol. 14, no. 24, pp. 28221–28229, Jun. 2022, doi: 10.1021/acsami.2c05818.

[24] K. Bhosle and V. Musande, ''Evaluation of deep learning CNN model for recognition of devanagari digit,'' *Artif. Intell. Appl.*, vol. 1, no. 2, pp. 114–118, Feb. 2023, doi: 10.47852/bonviewaia3202441.

[25] A. Olamat, P. Ozel, and S. Atasever, ''Deep learning methods for multichannel EEG-based emotion recognition,'' *Int. J. Neural Syst.*, vol. 32, no. 5, pp. 225–232, May 2022, doi: 10.1142/s0129065722500216.

[26] C. C. Liu, S. Ghosh Hajra, S. D. Fickling, G. Pawlowski, X. Song, and R. C. N. D'Arcy, ''Novel signal processing technique for capture and isolation of blink-related oscillations using a low-density electrode array for bedside evaluation of consciousness,'' *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 453–463, Feb. 2020, doi: 10.1109/TBME.2019.2915185.

[27] S. Phadikar, N. Sinha, and R. Ghosh, ''Automatic EEG eyeblink artefact identification and removal technique using independent component analysis in combination with support vector machines and denoising autoencoder,'' *IET Signal Process.*, vol. 14, no. 6, pp. 396–405, Aug. 2020, doi: 10.1049/iet-spr.2020.0025.

[28] Q. He and S. Pursiainen, ''An extended application 'Brain Q' processing EEG and MEG data of finger stimulation extended from 'Zeffiro' based on machine learning and signal processing,'' *Cognit. Syst. Res.*, vol. 69, pp. 50–66, Oct. 2021, doi: 10.1016/j.cogsys.2020.08.006.

[29] Q. Lin, S. Song, I. D. Castro, H. Jiang, M. Konijnenburg, R. van Wegberg, D. Biswas, S. Stanzione, W. Sijbers, C. Van Hoof, F. Tavernier, and N. Van Helleputte, ''Wearable multiple modality bio-signal recording and processing on chip: A review,'' *IEEE Sensors J.*, vol. 21, no. 2, pp. 1108–1123, Jan. 2021, doi: 10.1109/JSEN.2020.3016115.

[30] T. Xu, Y. Zhou, Z. Hou, and W. Zhang, ''Decode brain system: A dynamic adaptive convolutional quorum voting approach for variable-length EEG data,'' *Complexity*, vol. 2020, pp. 1–9, Mar. 2020, doi: 10.1155/2020/6929546.

[31] H. Wang, H. Guo, K. Zhang, L. Gao, and J. Zheng, ''Automatic sleep staging method of EEG signal based on transfer learning and fusion network,'' *Neurocomputing*, vol. 488, pp. 183–193, Jun. 2022, doi: 10.1016/j.neucom.2022.02.049.

[32] Y. Xu, C. Hu, Q. Wu, S. Jian, Z. Li, Y. Chen, G. Zhang, Z. Zhang, and S. Wang, ''Research on particle swarm optimization in LSTM neural networks for rainfall-runoff simulation,'' *J. Hydrol.*, vol. 608, May 2022, Art. no. 127553, doi: 10.1016/j.jhydrol.2022.127553.

SHUZHI SHAN was born in Lianyungang, Jiangsu, in February 1987. He received the bachelor's degree in stomatology from Nanjing Medical University, in 2012, and the master's degree in business administration from Guangxi Normal University, in 2018, specialized in rehabilitation medicine education and management. He was a Staff Member (2011–2016) and the Chief of the Academic Affairs Office (2016–2021) with the Kangda College, Nanjing Medical University, where he has been the Deputy Director of the Department of Rehabilitation Medicine, since 2021. He has published four academic papers related to his majors, presided over two university-level scientific research projects and one Lianyungang Social Science Fund project, and participated in the research of one national higher education teaching reform research project and two municipal and department-level projects.

JING ZHOU was born in Gaoyou, Jiangsu, in March 1990. She received the bachelor's degree in nursing from Kangda College, Nanjing Medical University, in 2013. She is currently pursuing the master's degree in public administration with Guizhou University of Traditional Chinese Medicine, specializing in public policy and traditional Chinese medicine business management. From 2013 to 2021, she was a Staff Member with the Kangda College, Nanjing Medical University, where she has been the Secretary of the Rehabilitation Medicine Department Training Center, since 2021. She has published three academic papers related to the profession, led one university level teaching and research project and one Lianyungang Association for Science and Technology soft research project, and has participated in research on one university level teaching and research project and in three municipal level projects.