https://doi.org/10.20965/jaciii.2021.p0040

Paper:

Helmet Detection Based on Deep Learning and Random Forest on UAV for Power Construction Safety

Guobing Yan[†], Qiang Sun, Jianying Huang, and Yonghong Chen

Guangdong Power Grid Corporation
757 Dongfeng East Road, Guangzhou, Guangdong 510600, China
E-mail: 896681727@qq.com
[†]Corresponding author

[Received June 1, 2020; accepted October 27, 2020]

Image recognition is one of the key technologies for worker's helmet detection using an unmanned aerial vehicle (UAV). By analyzing the image feature extraction method for workers' helmet detection based on convolutional neural network (CNN), a double-channel convolutional neural network (DCNN) model is proposed to improve the traditional image processing methods. On the basis of AlexNet model, the image features of the worker can be extracted using two independent CNNs, and the essential image features can be better reflected considering the abstraction degree of the features. Combining a traditional machine learning method and random forest (RF), an intelligent recognition algorithm based on DCNN and RF is proposed for workers' helmet detection. The experimental results show that deep learning (DL) is closely related to the traditional machine learning methods. Moreover, adding a DL module to the traditional machine learning framework can improve the recognition accuracy.

Keywords: safety helmet, deep learning, convolutional neural network, random forest

1. Introduction

The use of unmanned aerial vehicle (UAVs) has enabled increased opportunities for monitoring in the power industry. Traditional image recognition technology has difficulty with workers' helmet detection [1, 2]. At present, smart substations and some unattended substations are gradually implementing HD video monitoring, infrared thermal imaging and other intelligent monitoring technologies [3, 4]. Traditional manual inspection also results in a large number of visible light, infrared, ultraviolet, and other detection images. These massive media data streams provide a database for deep learning (DL) regarding workers' helmet detection [5].

In [6], a DL approach for accurate safety helmets wearing detection, employing a single-shot multi-box detector, was proposed. In [7], an innovative and practical safety helmet wearing detection method based on image processing and machine learning was proposed. In [8], a novel system with image processing and deep convolutional neural networks (CNNs) was proposed for motorcyclists' helmet detection. In [9], a novel and practical safety helmet detection framework based on computer vision, machine learning, and image processing was proposed. An automated system based on DL was proposed for the detection of biker helmets [10]. Paper [11] proposed a novel solution based on DL to recognize a helmet.

In summary, there are three main problems associated with workers' helmet detection: (1) the helmet identification rate fluctuates greatly, (2) it is easily disturbed by the environment, and (3) the color of safety helmets is not uniform.

To improve the traditional recognition methods for workers' helmet detection, a DL method is introduced in this paper. For feature extraction, this paper proposes two independent CNN models to extract the features of workers with safety helmets based on the AlexNet model. For image reorganization, an intelligent recognition algorithm based on CNN and random forest (RF) is proposed in this paper.

2. Feature Extraction

2.1. AlexNet Model

AlexNet is a typical CNN, which can be regarded as a feature engine. The convolutional layer in the middle accurately describes the local features of the image, while the second and third layers from the bottom are fully connected layers, which can describe the global features of the image.

The execution process of AlexNet is as follows:

Step 1: Set the basic parameters of the CNN, such as the size of the input data, number of convolution and down sampling layers, size of the convolution kernel, and size of the pooling window.

Step 2: The values of the convolution kernel, bias, etc., are initialized.

Step 3: Train the CNN; the training data are divided into batches to realize forward propagation (FP) of the training. The FP process is as follows:

(1) Take a sample (X, X_p) from the sample set, where X is the input image and X_p is the category of X and take X as the input value of the whole network.

(2) Calculate the corresponding output value O_p based on X.

$$O_p = F_n\left(\cdots F_2\left(F_1\left(X_p W(1)\right) W(2)\right) \cdots W(n)\right). \quad (1)$$

Then, backward propagation (BP) is realized:

(1) Calculate the difference between O_p and X_p.

(2) An appropriate BP algorithm is used to adjust the value of the weight matrix so that the difference between O_p and X_p is minimized.

The error and gradient value of the neural network can be calculated by BP, and the modified model value can be obtained by adding the calculated gradient to the original model.

Step 4: After the CNN model parameters are obtained through training, the testing data set can be used to test the accuracy of the current model.

In the training phase, to reduce the time spent in the training process and improve the generalization ability of the model, AlexNet initiates the following optimization process:

(1) For the input image, the image is normalized to $256 * 256$ pixels before being input to the network. This has the advantage of increasing the translation invariance and rotation invariance of the image, as well as increasing the number of images in the test set.

(2) In the process of BP, AlexNet does not use the sigmoid function or tanh function, but instead chooses a rectified linear unit (ReLU) to calculate the gradient, which has the advantage of reducing the time complexity of the gradient descent algorithm and the convergence time of the algorithm.

(3) Dropout technology and a non-overlapping pooling window are used to avoid overfitting, which can improve the generalize ability of the model.

AlexNet's convolution depth is five layers, and the number of convolution cores in each layer is 96, 256, 384, 384, and 256, respectively. Moreover, ReLU accelerates the convergence. AlexNet can extract rich image features at a faster speed because dropout technology and the use of non-overlapping pooling windows prevent overfitting. However, because the design and training of AlexNet relies on experience and skills, for a given deep model, it is usually only applicable to a given task, and without applying modification directly to other tasks, it can lead to inefficiency. Therefore, the structure of the AlexNet model should be modified and adjusted for different tasks. Although the use of ReLU can reduce the time complexity of the algorithm, AlexNet requires a large number of test images for training, so a GPU is needed to improve the training speed. To improve the recognition accuracy and reduce the training time of the model, this study introduces the principle of human vision into CNN modeling, and proposes and designs a DCNN model. On the basis of AlexNet, DCNN uses a double-channel data stream to extract two sets of deep CNN eigenvalues, and then combines the two sets of eigenvalues at the last layer of the model to obtain 512-dimensional eigenvectors.

2.2. DCNN Model

This paper proposes a double-channel CNN model, which is achieved by analyzing and testing a variety of current mainstream CNN models. This model obtains two sets of equipment features through two independent CNN models, and the final image features can be obtained after cross-mixing of the two sets of features at the top.

2.2.1. Structure of the DCNN Model

The DCNN model proposed in this study is shown in **Fig. 1**. The DCNN is composed of two independent and parallel convolutional neural network models, CNNa and CNNb. DCNN learns the features of images through CNNa and CNNb, and finally carries out cross-mixing on the learned features at the top of the model to obtain the final image features. The last fully connected layer of CNNa and CNNb outputs 512 neural units.

After the feature data of the last fully connected layer of CNNa and CNNb are obtained, DCNN adopts a secondary crossover operation for the two groups of feature data. That is, the output of the two fully connected layers is first cross-connected, and the result is regarded as the input of the next fully connected layer. Then, the entire connection layer is divided into two parts, the two parts of the data are mixed and connected, and the resulting feature vector is the final image feature.

2.2.2. Design of DCNN

In this study, a double-channel convolutional neural network structure is designed based on the AlexNet network structure for expansion and modification. The fully connected layer of AlexNet is studied. If there are too many feature vectors in each set of CNN features, it is not suitable for feature mixing operation and feature redundancy, which is not conducive to classification processing. Therefore, the full-link layer of AlexNet is expanded and improved, and an 11-layer deep CNN is proposed and designed.

For a CNN model, the parameters of the model reflect the fitting ability of the model. The number of neurons and parameters is related to the size of the convolution kernel, moreover, reduction of the convolution kernel and step size will lead to an increase in the feature map, leading to an increase in the number of meridional elements. To ensure that the features are extracted by the two groups of CNN, this paper carries out appropriate transformation on the input images to ensure that CNNa and CNNb have differences in input, specifically as follows: (1) the input of CNNa is an image with a size of $256 * 256$ pixels

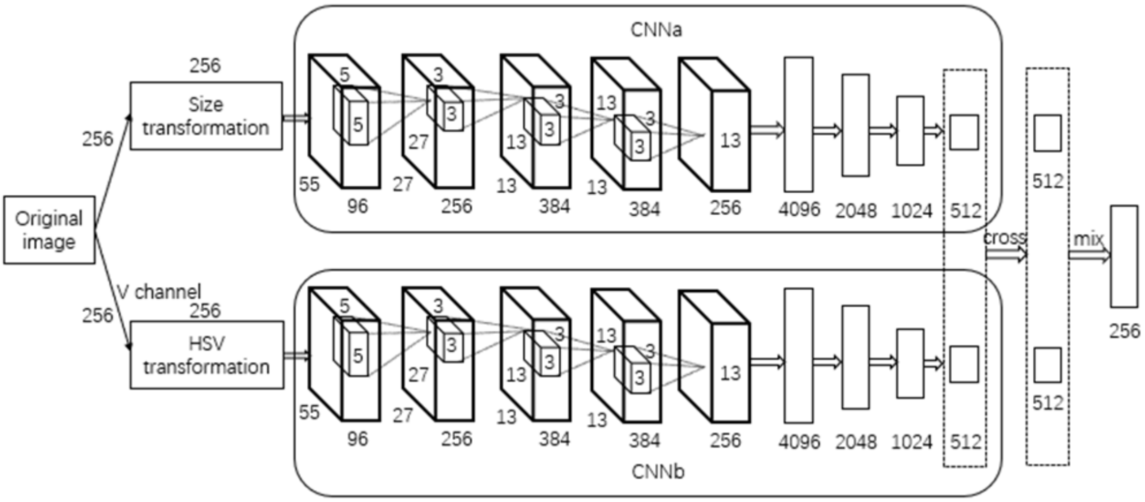

Fig. 1. Architecture of DCNN.

after the original image has been normalized; (2) the input of CNNb is the V channel component extracted after HSV (Hue, Saturation, Value) transformation of the original image; (3) CNNa and CNNb have the same structure, both of which are nine-layer neural networks, including five convolutional layers (CLs) and four fully connected layers (FCLs) (see **Fig. 1** for the specific structure).

The first–fifth layers are CL, and the numbers of convolution kernels are 96, 256, 384, 384, and 256; the sizes of the convolution kernels are $11*11*3, 5*5*48, 3*3*56, 3*3*192$, and $3*3*192$; and the steps of the convolution operation are 4, 1, 1, 1, and 1, respectively. The fifth convolutional network layer is followed by multiprogramming level (MPL). Layers 6, 7, 8, and 9 are all FCL, and their numbers of nerve units are 4096, 2048, 1024, and 512, respectively.

In the 10th layer, DCNN first makes a cross-connection between the output of the ninth layer of CNNa and the ninth layer of CNNb as the input of the 11th layer. In the 10th layer, the results of the cross-connection are decomposed into two pieces, each containing 512 neurons. Then, the CNN features extracted by the two transformation flows are mixed for the second time in the 11th layer to obtain the 256-dimensional feature vector, which is the DCNN and the depth eigenvalue.

2.2.3. Feature Extraction Algorithm

In this paper, the traditional back propagation (BP) mechanism is selected for both CNN, the error is propagated forward layer by layer, and the weights and bias values of the convolution kernel are then obtained based on the chain derivation rule. The proposed feature extraction algorithm in this study is as follows:

Step 1: Enter and set the data.

For images of the same length and width, the image is first scaled to 256 pixels long and 256 pixels wide. However, for images with different lengths and widths, the long side of the image is first fixed to 256 pixels, after which the other side is transformed according to the scaling ratio of the long side. The empty part after transformation was filled with 0. The image data obtained by this operation were used as the input for CNNa. This paper takes the V channel image transformed by HSV as the input for CNNb.

Table 1. Parameters for DCNN.

Type	***Size of the convolution kernel***	***Pooling step size***	***Feature map size***	***Number of feature image***
CL 1	$11*11$	4	$55*55$	96
PL 1 /LRN	$3*3$	2	$27*27$	96
CL 2	$5*5$	1	$27*27$	256
PL 2 /LRN	$3*3$	2	$13*13$	256
CL 3	$3*3$	1	13*13	384
CL 4	$3*3$	1	13*13	384
CL 5	$3*3$	1	13*13	256
PL 3	$3*3$	2	6*6	256
FCL 1	4096			
FCL 2	2048			
FCL 3	1024			
FCL 4	512			

Step 2: Set model parameters.

The structures of the two CNN models, CNNa and CNNb, are the same, and the relevant parameter settings are shown in **Table 1**.

Step 3: FP algorithm.

Let X_L be the output of the previous layer, w_L represent the weight matrix, and g_L represent the activation function. The output of FP can be expressed as:

$$z = g_L(X_L; w_L) \,. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

At the top of the network, the logarithmic loss error function is used to calculate the difference between the output result and the actual result. Then, the loss function of layer L can be expressed as:

$$l\left(x_i,y_i\right) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \ln\left(g_L\left(x_i;w_i\right)\right)\right) + \lambda\sum_{k=1}^{L}\sum_{j=1}^{k}\left\|w_j\right\|^2, \quad . \ . \ . \ . \ . \ . \ . \ (3)$$

where x_i is the input value, n is the number of images in the test set, y_i is the category of x_i, and λ is the regularization coefficient of L_2.

Step 4: BP algorithm

Suppose the objective function of a CNN network is expressed as:

$$O = \arg\min_{w_L}\frac{1}{n}\sum_{i=1}^{n} l\left(g_L\left(x_i;w_L\right),y_i\right). \quad . \ . \ . \ . \ . \ (4)$$

The value of w_L is updated by

$$\frac{\partial l}{\partial w_L} = \lambda w_L - \frac{1}{n}\left(Y_L - g_L\left(X_L;w_L\right)\right) \times g_L'\left(X_L;w_L\right). \quad . \ . \ (5)$$

In the process of obtaining the optimal solution of the objective function, the error between the output value of the model and the actual value can be converged by iterating with Eq. (5).

The value of the weight matrix of the CNN output layer can be updated using Eq. (6).

$$\begin{cases} \dfrac{\partial l}{\partial w_{L-1}^A} = \dfrac{\partial l}{\partial g_L}\dfrac{\partial g_L}{\partial g_{L-1}^A}\dfrac{\partial g_{L-1}^A}{\partial w_{L-1}^A}, \\ \dfrac{\partial l}{\partial w_{L-1}^B} = \dfrac{\partial l}{\partial g_L}\dfrac{\partial g_L}{\partial g_{L-1}^B}\dfrac{\partial g_{L-1}^A}{\partial w_{L-1}^B}. \end{cases} \quad . \ . \ . \ . \ . \ . \ (6)$$

The weights of the two cross-mixing layers of the DCNN are updated using Eq. (7).

$$\begin{cases} \dfrac{\partial l}{\partial w_{L-2}^A} \\ = \dfrac{\partial l}{\partial g_L}\dfrac{\partial g_L}{\partial w_{L-1}^A}\left(\dfrac{\partial g_{L-1}^A}{\partial g_{L-2}^A}\dfrac{\partial g_{L-2}^A}{\partial w_{L-2}^A} + \dfrac{\partial g_{L-1}^A}{\partial g_{L-2}^B}\dfrac{\partial g_{L-2}^B}{\partial w_{L-2}^B}\right), \\ \dfrac{\partial l}{\partial w_{L-2}^B} \\ = \dfrac{\partial l}{\partial g_L}\dfrac{\partial g_L}{\partial w_{L-1}^B}\left(\dfrac{\partial g_{L-1}^B}{\partial g_{L-2}^B}\dfrac{\partial g_{L-2}^B}{\partial w_{L-2}^B} + \dfrac{\partial g_{L-1}^B}{\partial g_{L-2}^A}\dfrac{\partial g_{L-2}^A}{\partial w_{L-2}^A}\right). \end{cases} \quad (7)$$

Here, g^A and g^B represent the transformation function of exchange flow A and B, respectively. w^A and w^A represent the weight matrix of exchange flow A and B, respectively.

According to the above introduction, DCNN includes CL, LRNL, MPL, FCL, and a cross-mixing layer (CML). Therefore, there is a difference between the error reverse transfer mode in the DCNN's BP algorithm and the traditional error reverse transfer mode.

For the convolutional layer L, if the next layer is MPL, then the error of the convolutional layer L is expressed by:

$$y_{L-1}^j = upsample\left(y_L^j\right) \cdot g_L'\left(x_L^j;w_L\right), \quad . \ . \ . \ . \ (8)$$

where y_L is the MPL error, g_L is the transformation function of MPL, and $upsample(\cdot)$ is the MPL function. When using the function $upsample(\cdot)$ to sample the maximum value, the position of the maximum value in the sampling block should to be found and recorded.

If the upper layer of the convolutional layer L is MPL, the error of the convolutional layer L should be used to calculate the MPL error. The calculation process is as follows.

If the output number of the MPL is N, the output number of the convolutional layer L is M, the error of the convolutional layer L is y_L, and the convolution kernel of the convolutional layer L is K_{ij}. Then, the error calculation formula for the j-th output of the MPL is expressed as follows:

$$y_{L-1}^j = \left(\sum_{i=1}^{M} y_L^i * K_{ij}\right) \cdot g_L'\left(x_L^j;w_L\right). \quad . \ . \ . \ . \ (9)$$

3. Target Recognition

As a supervised learning neural network, CNN is primarily composed of a CL, PL, and output layer (OL). Among them, feature extraction of the CNN mostly depends on CL and PL. The principle of adjusting parameters in a CNN is combined with the gradient descent method to correct the error by reverse transmission. Iterative training can improve the accuracy of the network. The classifier classifies the acquired depth features to achieve image recognition. At present, there are two types of classifiers used in deep CNN: namely logistic classifiers and softmax classifiers.

3.1. Target Recognition Based on DL and RF Classifier

There are many types of helmet colors. Moreover, the complex environment in a substation and other places leads to a complex background for the images obtained. Considering that a logistic classifier is generally used to realize binary classification problems, although the softmax classifier can solve multiple classification problems, softmax has a high classification error rate for complex and confusing targets. Therefore, logistic classifiers and softmax classifiers are not suitable for helmet recognition.

Although DL has made amazing breakthroughs with respect to speech, images, natural language processing, and other fields, many applications that were slow to progress have made leaps and bounds. However, the good effect of DL cannot displace traditional learning theory. In this study, deep learning and traditional machine learning theory are integrated, and a RF classification method combining DL is proposed. As a theory of statistical learning, RF primarily realizes the function of voting as a de-

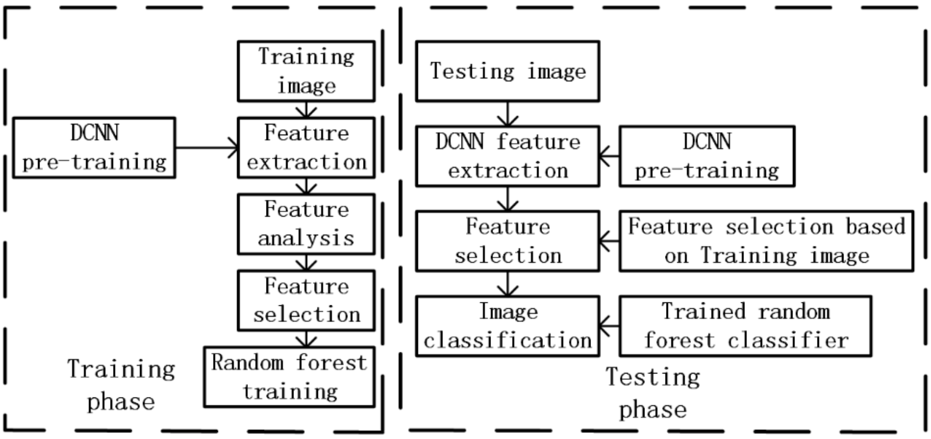

Fig. 2. Framework of the classifier.

cision to determine the prediction result, which has the characteristics of high prediction accuracy, strong anti-noise ability, and good fitting performance. Therefore, an RF classifier is adopted in this study to form a decision forest by generating multiple randomly selected sample subsets and decision trees generated by feature subspaces, and the classification results are obtained by voting in the classification stage. The classifier framework based on DL and RF is shown in **Fig. 2**.

As can be observed from **Fig. 2**, the classification method proposed in this paper consists of two parts: the training phase and the testing phase. During the training phase, first, the DCNN proposed in this paper is used to randomly select images from the image database and extract image features. Then, the learned features are analyzed based on the adaptability of the RF classification, and the features are selected based on the results of the analysis. Considering that the features extracted by the lower CL do not contain rich semantic information, and if these features are used, the dimensions representing image features will increase significantly, this paper only analyzes and selects the features of the DCNN's FCL and mixed layer. Finally, the selected characteristics are used to train the RF. In the testing phase, the DCNN is first used to calculate the depth features of the image. Then, the feature subset selected in the training phase is used as the final image feature, finally, the trained RF is used to classify the input images.

3.1.1. Random Forest (RF) Training

To train a tree in an RF, the input space is recursively divided into a set of disjoint partitions, starting with the root node corresponding to the entire input space. At each node, each partition needs to determine a set of segmentation rules and prediction models to minimize losses. Considering the high image feature dimension extracted by DCNN, a constant segmentation model is selected in this study for RF training.

Assuming that a random forest $F = \{T_i\}$ is a group of trees and T_i is a tree in F, training is conducted on a randomly selected training sample $S = \{s_i = (X_i, y_i)\}$. $X_i \in R^d$ is the feature vector of training sample s_i, and y_i is the category label of the corresponding image. Given feature X_i, at each node, the segmentation function is defined as:

$$\begin{cases} X_i^{(j)} \geq T_0, & \text{Send to the left subtree,} \\ \text{Other,} & \text{Send to the right subtree,} \end{cases} \quad . . (10)$$

where T_0 is a threshold. $X_i^{(j)}$ is the j-th dimension of the vector X_i. At each node, each sample in S is sent to the left or right subtree using the selected dimension and threshold, and S is split into S_l and S_r. The training continues to split the sample until all samples are tested.

For each node, multiple hypothesis tests are generated by randomly selecting some dimensions and thresholds. Given the Gini coefficient standard minimum score in the decision tree algorithm widely used in choosing segmentation attributes, the sample is accordingly divided into left or right child nodes.

Suppose the sample in S comes from an m different class $C_i\,(i = 1,\ldots,m)$. The Gini coefficient of the set S is defined as

$$Gini\,(S) = 1 - \sum_{i=1}^{m} p_i^2, \quad . . . . (11)$$

where p_i is the ratio of the number of samples in class C_i to the number of samples in set S.

The Gini coefficient is an impure measure. When all samples in the set belong to a class, the Gini coefficient reaches the minimum value. When all samples in the set are evenly distributed, the Gini coefficient reaches its maximum value. If, in the hypothesis test, the set S is divided into two subsets, S_l and S_r, then, the Gini coefficient can be expressed as:

$$Gini_{split}\,(S) = \frac{|S_l|}{|S|} Gini\,(S_l) + \frac{|S_r|}{|S|} Gini\,(S_r)\,. \quad . . (12)$$

At each node, the dimensions and thresholds are tested randomly, and the one that provides the smallest $Gini_{split}$ is selected to perform sample segmentation.

3.1.2. Feature Selection

For each node of the tree in the RF, depth features need to be tested to select the best feature as the feature representation of the image. If the image represents a large number of dimensions, the effective search space of the test will also be large, making it difficult to obtain a good segmentation effect. How can we characterize the helmet image with fewer dimensions and ensure discrimination of the features obtained? To solve this problem, an effective feature selection method is proposed in this paper to obtain more effective image features.

In the training of RF, the dimension used for sample segmentation is selected based on the Gini coefficient standard. In fact, the sample is divided into more pure dimensions to achieve the goal of easier discrimination of samples of different categories. In this study, linear discrimination is used to evaluate the depth feature.

In the selection of features, each dimension of the image representation is dealt with independently, and the validity of each feature is evaluated in a manner similar to the Filch standard. For each dimension, the calculation formula for the intra-class dispersion of all samples is given by:

$$S_W^{(k)} = \sum_{i=1}^{m} S_i^{(k)}, \quad \ldots\ldots\ldots\ldots\ldots (13)$$

$$S_i^{(k)} = \sum_{X \in D_i} \left(X^{(k)} - \mu_i^{(k)} \right)^2, \quad \ldots\ldots\ldots (14)$$

$$\mu_i^{(k)} = \frac{1}{n_i} \sum_{X \in D_i} X^{(k)}, \quad \ldots\ldots\ldots\ldots (15)$$

where k refers to the k-th dimension represented by the image, m represents the number of categories to be separated, X represents the image feature vector, D_i is the sample set from category i, and n_i is the number of samples from category i. Intra-class scattering provides the variance of samples in the same category in the test dimension. When testing the dimensions, the equations for the class dispersion are expressed as:

$$S_B^{(k)} = \sum_{i=1}^{m} n_i \left(\mu_i^{(k)} - \mu^{(k)} \right)^2, \quad \ldots\ldots\ldots (16)$$

$$\mu^{(k)} = \frac{1}{n_i} \sum_{X \in D_i} X^{(k)} = \sum_{i=1}^{m} n_i \mu_i^{(k)}. \quad \ldots\ldots\ldots (17)$$

where n is the total number of samples for all categories and D is the sample set. The inter-class dispersion matrix provides the dissimilarity of samples of different categories in dimension k.

When the intra-class variance of the samples is small and the inter-class dissimilarity is large, the samples are easily divided into purer subsets. Therefore, the criterion for assigning scores to each test dimension is given by:

$$f(k) = \frac{S_B^{(k)}}{S_W^{(k)}}. \quad \ldots\ldots\ldots\ldots\ldots (18)$$

The discriminant ability of dimension k can be determined by Eq. (18). By seeking the maximum value of $f(k)$, the dimensions of different categories can be best separated. As the score of a feature dimension increases, the more likely it is to be selected for image classification and recognition. However, if feature selection is performed directly based on the above criterion, the classification performance of RF may be unstable. To reduce correlation between the selected dimensions, this study performs feature selection sequentially to ensure that each newly selected dimension is least relevant to the previously selected dimension.

Let K be the set of selected dimensions. To add new dimensions to K, another subset of candidate dimensions L is selected with the probability of being weighted by scores from all unselected dimensions. For candidate dimension l in L, its correlation with dimension k in the previous selected K can be calculated by:

$$Cor(l,k) = \frac{1}{m^2} * \left| \sum_{i=1}^{m} \left(\mu_i^{(k)} - \mu_i^{(l)} \right) \right| * \left| \sum_{i=1}^{m} \left(\sigma_i^{(k)} - \sigma_i^{(l)} \right) \right|. \quad \ldots\ldots (19)$$

Then, by selecting dimension g to be added to j, Eq. (20) is obtained.

$$l^* = \arg\max_{l \in L} \min_{k \in K} Cor(l,k). \quad \ldots\ldots\ldots (20)$$

Based on the above approach, we can select the new dimension from subset L that is least relevant to the previously selected subset.

3.1.3. Category Forecast

After the RF $F = \{T_i\}$ training, each tree has a set of leaves. For the incoming image to be recognized, the DCNN is first used to extract the depth features. Second, the depth features are extracted from the sixth layer to the 11th layer of the DCNN and connected to the image representation. Then, the image representation is formed again by discarding the unselected dimensions. Moreover, the reconstructed image indicates that through each random tree, it descends from the root node and remains descending according to the splitter function until it reaches the leaf node. In addition, the predicted value of the image to be recognized can be obtained at the leaf node. Finally, the category prediction with the maximum number for votes from the RF is used as the label of the image to be recognized.

4. Performance Evaluation

In this study, two different types of scenarios were tested, including target scenarios and non-target scenarios. The target and non-target scenarios have five different scenes. The target scenario is an image of workers with safety helmets. The non-target scenario is an image of workers without safety helmets. In the experiment in this study, the target and non-target scenarios contain five different types of power equipment: a transformer

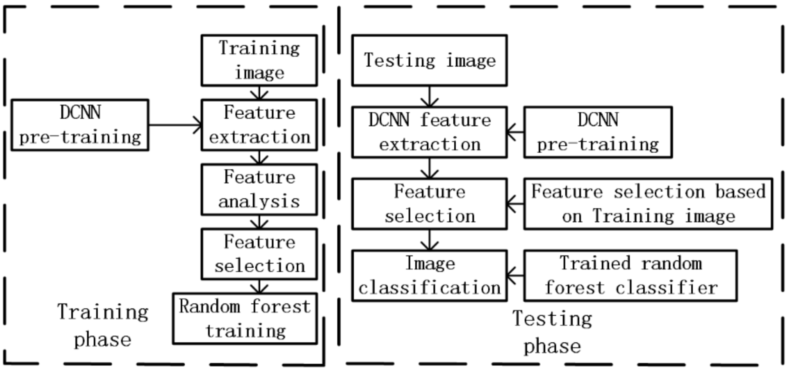

Fig. 3. The training process for single CNN and DCNN.

(scene 1), insulator (scene 2), circuit breaker (scene 3), pole (scene 4), and tower (scene 5).

The image database used for the test was Power Image. The image database is a safety helmet image database collected and sorted by the laboratory, with a total of more than 30,000 images. During the training, some images were randomly selected from the five types of images as training samples. However, all the remaining images were considered testing samples.

In the recognition stage, if the entire image of people with safety helmets is directly recognized, the accuracy of recognition will be affected. Therefore, before feature extraction, the image is segmented, and depth feature extraction and recognition are then performed.

During the testing, dropout was adopted after all the FCLs to prevent the DCNN from over-fitting. The dropout ratio is equal to 0.5. To speed up the convergence rate, the ReLU activation function was used in the nonlinear transformation, the initial learning rate was 0.005, and a polynomial reduction strategy was used to control the learning rate. The reduction value of the learning rate was set as 0.5. To speed up the training, the images were trained and tested by batch processing. The Batch size was 20 and the number of iterations was set to 200,000. After convergence, and 200,000 iterations, the final model parameters were obtained.

For the RF classifier, the number of random trees was set to 1,000, the dimension of each node was set to 15, and the dimension evaluated during feature selection was set to 10. The DL platform used in the testing is Caffe, the operating system was Ubuntu, the memory of the graphics workstation was 2G, the GPU was NVIDAK60, and the programming language was Python.

During testing, 3,000 pairs of scene 1 images, 2,000 pairs of scene 2 images, 3,000 pairs of scene 3 images, 2,000 pairs of scene 4 images, and 500 pairs of scene 5 images were used. During the training, 2,000 pairs of insulator samples, 1,500 pairs of transformer samples, 2,000 pairs of circuit breaker samples, 1,000 pairs of power line pole samples, and 500 pairs of power line tower samples were used. The remaining images were testing samples. During testing, the scene was first segmented based on the visual significance model, and the target segmented from the original image was further utilized.

(1) Consider an image with the same length and width. First, the image is zoomed confirm whether the length and width of the image are 256 pixels. For the image with different lengths and widths, the long side of the image is fixed to 256 pixels, and the other side is transformed according to the scale of the long side. After transformation, 0 is used to fill spare parts. The image data obtained by this operation mode are input as CNNa.

(2) HSV transforms the image to a size of 256×256 after the zooming operation, and takes the V-channel image after HSV transform as the input for CNNb.

Using the method described in this paper, the recognition results and analysis are as follows.

4.1. Comparison of Recognition Results Between Single-Channel CNN and DCNN

To test the effectiveness of features extracted by the DCNN mentioned in this paper, the feature extraction results of the DCNN are compared with those of the single-channel CNN. **Fig. 3** shows the misclassification rate curves obtained when single-channel CNN and DCNN are used for training.

As shown in **Fig. 3**, during the training of the model, when the number of iterations is 50, the misclassification rate of the single-channel CNN reaches the minimum value of 8%, and the misclassification rate of the DCNN reaches the minimum value of 5.4%. If the iterative operation continues, the misclassification rate of single-channel CNN and DCNN will be reduced. However, after two iterations, that is, at the 50th iteration, the misclassification rate of the CNN would increase again and there would be almost no fluctuation. After five iterations (i.e., at the 52th iteration), the misclassification of the DCNN rate will increase again and there would be almost no fluctuation.

After single-channel CNN and DCNN were used to recognize people with safety helmets in the training data and testing data, the classification accuracy is shown in **Table 2**, and the average accuracy and average time are shown in **Table 3**.

Table 2 and **3** show that: (1) Single-channel CNN and DCNN DL models can be used to classify images

Table 2. Comparison of accuracy rate for two models.

Model	The number of accuracy rate				
	Scene 1	*Scene 2*	*Scene 3*	*Scene 4*	*Scene 5*
Single-CNN	94%	89%	89%	92%	72%
DCNN	97%	92%	93%	96%	80%

Table 3. The average accuracy rate and computation complexity for two models.

Model	The number of accuracy rate		
	Average accuracy	*Average time in GPU* [s]	*Average time in CPU* [s]
Single-CNN	87.2	0.9	330
DCNN	91.6	1.2	452

with an average accuracy of greater than 85%. This result shows that the CNN features high abstraction level and strong expression ability, and it can obtain relatively high accuracy when identifying images of people with safety helmets. (2) Compared with DCNN, the recognition rate of a single-channel CNN decreased by 4.4%. This is because compared with single-channel CNN, the DCNN model has a wider "width" and can extract richer image features. However, because DCNN does not increase significantly in depth, the running time complexity of its GPU does not increase much compared with that of a single-channel CNN, which is only 0.3 s. (3) For the five scenes, scene 5 has the lowest recognition rate. This is because the tower's dataset is too small, with only 500 images, compared with more than 2,000 images for the other scenes. Therefore, more training samples can improve the generalization ability of the CNN model, reduce the risk of overfitting and obtain greater recognition accuracy. (4) The time spent running the DL algorithm on the CPU is much higher than that spent on GPU. This indicates that the DL algorithm is time-consuming. Therefore, in practical applications, if the DL algorithm is used, GPU support is required.

4.2. Comparison of the Recognition Results Between Non-Target Detection and Target Detection

To determine whether target detection will have an impact on the recognition results, we removed the step of image target detection and used the entire image directly as the input of the DCNN under the GPU. RF was then used for classification and recognition. The classification accuracy and time complexity obtained are shown in **Tables 4** and **5**, respectively. Non-target represents people without safety helmets, whereas target represents people with safety helmets.

Table 4. Accuracy rate for DCNN with target detection.

Model	Number of accuracy rate				
	Scene 1	*Scene 2*	*Scene 3*	*Scene 4*	*Scene 5*
DCNN	83%	72%	78%	88%	72%

Table 5. Average accuracy rate and computation complexity for DCNN with non-target detection.

Model	Accuracy rate	
	Average accuracy [%]	*Average time in GPU* [s]
DCNN	78.6	1.9

Tables 2 and **4** present that: (1) If target detection is not considered, the original image is directly classified with an average accuracy of 78.6%. The results fully illustrate the advantages of the convolution routing network regarding equipment recognition. Therefore, applying DL to the image recognition of safety helmets is feasible. (2) Compared with target detection, the average accuracy of non-target detection decreases. The main reason for this result is that the image background is complicated, and the target to be recognized is also complicated, which limits the classification and recognition ability of the classifier to some extent. (3) The classification accuracy of scenes 1, 2, and 3 decreased by more than 10%, to 14%, 20%, and 15% respectively. However, the accuracy of scenes 4 and 5 decreased by less than 8%. The main reason for the result is that the background of scenes 1, 2, and 3 is generally more complicated than scenes 4 and 5. Therefore, the classification accuracy is smaller than that of the other three scenes, regardless of whether there is target detection. (4) Among the five scenes, scene 5 has the lowest recognition rate. The reason for this result is that the Power Image dataset has the fewest samples for scene 5, far less than 1,000, while the number of sample images of other scenes all exceed 2,000. More training samples can improve the generalize ability of the CNN model, reduce the risk of overfitting, and lead to greater recognition accuracy.

It can be observed from **Tables 3** and **5** that the average time after target detection is 1.2 s, while the average recognition time of non-target detection is 1.9 s, showing a difference of 0.7 s. The main reason for the result is that the image background of the non-target detection is more complex, and more time is needed for target detection.

4.3. Comparison of Various Classification Methods

To test the effectiveness of the RF method proposed in this paper, the CNN SoftMax classifier (method 1), CNN + RF classifier (method 2) and traditional manual parameter extraction RF classifier (method 3) are compared. The results obtained in methods 1, 2, and 3 are shown in **Tables 6** and **7**, respectively.

Tables 2 and **6** show that: (1) Using methods 1 and 2 to classify the images, the average accuracy can reach greater than 80%. The results show that the image features extracted by CNN have a high abstraction level and strong expression ability. Moreover, high precision can

Table 6. Accuracy rate of method 1.

Type	Training recognition rate				
	Method 1 [%]	*Method* 2 [%]	*Method* 3 [%]	*Method* 4 [%]	*Method* 5 [%]
Training recognition rate [%]	90	84	85	83	74
Testing recognition rate [%]	91	86	85	85	76

Table 7. Accuracy rate of method 3.

Type	*Average time in GPU* [s]	Average time *in CPU* [s]
Training recognition rate [%]	1.0	404
Testing recognition rate [%]	0.2	87

be obtained when the device image is recognized. (2) The accuracy of method 1 is 7.2% lower than that of method 2. The main reason for this result is that the SoftMax classifier selects the depth features of the last hybrid FCL for classification processing. Although the last FCL represents the highest degree of semantic abstraction of the image, the most effective features are different for different scenes. Therefore, when designing the classifier, selecting the features of a deep CNN can further improve the classification performance. (3) Compared with methods 1 and 2, method 3 has the lowest recognition rate, with an average recognition rate of only 75.2%. The main reason for this result is that method 3 adopts manual features, such as color, texture, and direction, for classification and recognition. These features have a low degree of abstraction and cannot effectively describe the essential features of the target. The DL methods better than the traditional feature extraction method. Therefore, the final recognition rate is much higher than that of the traditional method. (4) For the four images of scenes 1, 2, 3, and 4, the recognition rates of methods 1 and 2 are over 85%, while the recognition rate of scene 5 is lower than that of method 3. The reason for this result is that scene 5 in the Power Image dataset has the smallest number of samples, far less than 1,000. For a small number of samples, the performance of DL is not as good as that of traditional feature extraction methods

The average time required for results to be obtained using methods 1 and 3 is shown in **Table 8**. It can be observed from the analysis and comparison of **Tables 3** and **8** that the traditional method is the fastest method for recognition. The method proposed in this paper has the highest time complexity, and the time complexity under GPU is 0.2 s and 1 s higher than that in methods 1 and 3, respectively. Therefore, the optimization of the algorithm should be considered.

Table 8. The computation complexity for methods 1 and 3.

Type	The number of average time	
	Average time in GPU [s]	*Average time in CPU* [s]
Method 1	1.0	404
Method 3	0.2	87

5. Conclusion

It is of great significance that unmanned aerial vehicles can replace manual power grid construction, providing considerable assistance to the power sector in formulating targeted maintenance measures, strengthening line operation and maintenance, and ensuring the safe operation of important national power grid lines. It is conducive to increasing the intensity of special patrols in key sections after heavy rainfall and increasing the number of equipment inspections under heavy load operation. Further, the participation of drones in the construction of the power grid is conducive to regular inspection and cleaning of trees and illegal buildings in the line channel to ensure the safety of the transmission channel. To accurately identify whether a worker is wearing a helmet, a DCNN model is proposed to extract image features of workers with safety helmets. First, the basic structure and feature extraction algorithm of the DCNN feature extraction model are introduced. Then, based on the CNN structure, the safety helmet recognition algorithm of an RF classifier combined with DL is studied. Finally, the proposed model is trained, and the performance of the algorithm is verified through experiments. The experimental results show that the recognition accuracy of the proposed method is much higher than that of other methods. Furthermore, the proposed method can effectively eliminate the effects produced by the complex background. Moreover, a higher recognition rate of workers with safety helmet can be obtained by adding a DL module to a traditional machine learning framework. Optimization of the algorithm will be considered in future work because the method proposed in this paper has higher time complexity than methods 1 and 3.

References:

[1] G. Feng, W. Liu, S. Li, D. Tao, and Y. Zhou, "Hessian-Regularized Multitask Dictionary Learning for Remote Sensing Image Recognition," IEEE Geoscience and Remote Sensing Letters, Vol.16, No.5, pp. 821-825, 2019.

[2] M. Chaa, Z. Akhtar, and A. Attia, "3D palmprint recognition using unsupervised convolutional deep learning network and SVM classifier," IET Image Processing, Vol.13, No.5, pp. 736-745, 2019.

[3] J. Pei, Y. Huang, W. Huo, Y. Zhang, J. Yang, and T.-S. Yeo, "SAR Automatic Target Recognition Based on Multiview Deep Learning Framework," IEEE Trans. on Geoscience and Remote Sensing, Vol.56, No.4, pp. 2196-2210, 2018.

[4] F. Liu and Z. Wang, "PolishNet-2d and PolishNet-3d: Deep Learning-Based Workpiece Recognition," IEEE Access, Vol.7, pp. 127042-127054, 2019.

[5] Y. P. Huang and H. Basanta, "Bird image retrieval and recognition using a deep learning platform," IEEE Access, Vol.7, pp. 66980-66989, 2019.

[6] X. Long, W. Cui, and Z. Zheng, "Safety Helmet Wearing Detection Based On Deep Learning," 2019 IEEE 3rd Information Technology,

Networking, Electronic and Automation Control Conf. (ITNEC), pp. 2495-2499, 2019.

[7] J. Li et al., "Safety helmet wearing detection based on image processing and machine learning," 2017 9th Int. Conf. on Advanced Computational Intelligence (ICACI), pp. 201-205, 2017.

[8] K. C. D. Raj, A. Chairat, V. Timtong, M. N. Dailey, and M. Ekpanyapong, "Helmet violation processing using deep learning," 2018 Int. Workshop on Advanced Image Technology (IWAIT), pp. 1-4, 2018.

[9] K. Li, X. Zhao, J. Bian, and M. Tan, "Automatic Safety Helmet Wearing Detection," 2017 IEEE 7th Annual Int. Conf. on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 617-622, 2017.

[10] C. A. Rohith, S. A. Nair, P. S. Nair, S. Alphonsa, and N. P. John, "An Efficient Helmet Detection for MVD Using Deep Learning," 2019 3rd Int. Conf. on Trends in Electronics and Informatics (ICOEI), pp. 282-286, 2019.

[11] N. Boonsirisumpun, W. Puarungroj, and P. Wairotchanaphuttha, "Automatic Detector for Bikers With No Helmet Using Deep Learning," 2018 22nd Int. Computer Science and Engineering Conf. (ICSEC), pp. 1-4, 2018.

Name:
Guobing Yan

Affiliation:
Senior Engineer, Guangdong Power Grid Corporation

Address:
757 Dongfeng East Road, Guangzhou, Guangdong 510600, China

Brief Biographical History:
1994- Guangdong Power Grid Corporation

Main Works:
- "Research on Influence of Groundwater Level on Deep Mixing Piles to Reinforcing Sea-Land Interphase Soft Soil Foundation," Construction Technology, Vol.48, pp. 39-42, 2019.

Name:
Qiang Sun

Affiliation:
Senior Engineer, Guangdong Power Grid Corporation

Address:
757 Dongfeng East Road, Guangzhou, Guangdong 510600, China

Brief Biographical History:
2004- Guangdong Electric Power Design Institute Co., Ltd.
2009- Guangdong Power Grid Corporation

Main Works:
- "Technical Analysis of Innovative Design on 500 kV Guishan Substation," Guangdong Electric Power, Vol.23, No.11, pp. 73-75, 2010.
- "An Object Detection Method and Optimization for Substation Video Surveillance Terminals," Guangdong Electric Power, Vol.32, No.9, pp. 62-68, 2019.

Name:
Jianying Huang

Affiliation:
Senior Engineer, Guangdong Power Grid Corporation

Address:
757 Dongfeng East Road, Guangzhou, Guangdong 510600, China

Brief Biographical History:
2004- Guangdong Power Transmission and Transformation Engineering Company
2011- Guangdong Power Grid Corporation

Main Works:
- "On key technologies in smart substations," Science Times, p. 62, 2011.
- "Prevention measures for insulation breakdown accident of 500 kV transformer," Guangdong Science and Technology, p. 96, 2011.

Name:
Yonghong Chen

Affiliation:
Senior Engineer, Guangdong Power Grid Corporation

Address:
757 Dongfeng East Road, Guangzhou, Guangdong 510600, China

Brief Biographical History:
2003 Graduated from North China Electric Power University
2003- Foshan Power Supply Bureau
2010- Senior Engineer, Foshan Power Supply Bureau
2013- Guangdong Power Grid Corporation

Main Works:
- "Influence of the Electricity and Magnetism from the Power Transmission Projects," Guangdong Power Transmission Technology, Vol.200801, 2008.
- "About the application of multi-loop narrow tower (steel billot) for large section wire," High Voltage Engineering, Vols.2010-2036, 2010.