Ground-Based Cloud Image Recognition System Based on Multi-CNN and Feature Screening and Fusion

MA JINGYI\(^1\), TIEJUN ZHANG\(^2\), JING GUODONG\(^3\), YAN WENJUN\(^1\), AND YANG BIN\(^1\)

\(^1\)Gansu Branch, China Meteorological Administration Training Centre, Beijing 100081, China
\(^2\)Institute of Arid Meteorology, China Meteorological Administration, Beijing 100081, China
\(^3\)China Meteorological Administration Training Center, Beijing 100081, China

Corresponding author: Ma Jingyi (elose@126.com)

This work was supported by the Youth Science Fund Project: Research on Super-Resolution Reconstruction Method of Meteorological Satellite Cloud Image Based on Compressed Sensing under Project 61602486.

ABSTRACT The recognition of ground-based cloud images has rich application prospects in many aspects such as weather prediction, astronomical site selection, and meteorological observation. Affected by factors such as rotation and illumination, the traditional feature extraction method is difficult to accurately describe the features of cloud images, resulting in low accuracy of ground-based cloud image recognition, and cannot meet the requirements of practical applications. With the popularity of convolutional neural networks in image processing, ground-based cloud image recognition algorithms based on convolutional neural network have become a research focus. However, the features of the ground-based cloud image are relatively shallow, and the cloud texture and other features are seriously lost in the convolution process, and it is difficult to achieve a good recognition effect. This paper proposes a ground-based cloud image recognition system based on multi-scale convolutional neural network (Multi-CNN) and multilayer perceptron neural networks (MLP). The multi-level and multi-scale convolution feature extraction is performed through convolution layers of Multi-CNN, and the local features with strong resolving power are selected through the feature screening algorithm based on DP clustering. Finally, the local features are encoded and fused for cloud image classification based on MLP. Field test results show that our method was superior to other tested network models in terms of the recognition accuracy of 94.8% under 9 classification. In addition, ablation experiments show that the multi-scale feature extraction, screening and local feature coding in this paper have a significant effect on improving the algorithm’s ability to distinguish different cloud images.

INDEX TERMS Ground-based cloud image, CNN, cloud classification.

I. INTRODUCTION

Clouds play an important role in the conservation of the earth’s energy and is an important factor affecting global climate change. Cloud detection is an important part of meteorological observation. Accurately obtaining cloud information is of great significance to many fields such as flight support, weather forecasting, climate research, and astronomical site selection [1], [2].

In the past, the detection of clouds was generally based on manual observation. With the development of the manufacturing industry and the increasingly widespread application of automated observation technology in various fields such as medicine and agriculture, cloud observation methods have also shifted from manual observation to automated observation [3]. A variety of sensor devices are increasingly used in cloud detection. However, the main disadvantages of this type of instrumental measurement methods based on physical characteristics are that the observation results have low resolution, and are easily interfered by factors such as air impurities and moisture, resulting in large observation errors [4], [5].

Researchers began to explore cloud image observation methods based on machine vision. Cloud images can be obtained through satellite and ground observation. Large-scale cloud information can be obtained from satellites, but the low resolution of this information can easily cause misjudgments [6]. For example, limited by the viewing...
angle of the radiometer, small clouds are often ignored; low clouds, thin clouds, and the surface may be confused. The data obtained by ground-based observation equipment can reflect the microstructure information of the cloud, which can make up for the lack of satellite observation data and is an important cloud observation method [7].

At present, meteorological departments have achieved the estimation of cloud height and cloud volume through instrumental measurements, but the classification of ground cloud images can’t be achieved through instrumentation. Visual-based cloud image recognition algorithm has gradually become a hot spot [1], [2]. However, the shape and angle of clouds vary greatly, and it is difficult to distinguish clouds with large similarity based on traditional feature extraction algorithms. In addition, the annotation of satellite cloud maps is difficult to meet the training needs of deep convolutional networks, and most of the current network models can only distinguish cloud maps simply, and can’t achieve detailed classification. In order to meet the actual application needs, this paper proposes a ground-based cloud image recognition system based on convolutional neural network and feature selection and fusion. The main contributions of this paper are as follows.

1) A new ground-based cloud image feature extraction algorithm is proposed. Instead of directly using convolutional neural networks and fully connected layers for image recognition, the Multi-CNN in this article is only used to extract multi-scale features without participating in the image recognition process.

2) The Multi-CNN in this paper uses a pre-training model for feature extraction, and does not need to use a large number of ground-based cloud images to train the network, which enables the system to achieve good performance even in the absence of samples.

3) Instead of directly using convolutional features for image classification, this paper designs a feature selection and fusion network. The redundant local features are filtered, and the highly recognizable local features are encoded and fused, which improves the network’s ability to distinguish different cloud images.

### II. RELATED WORK

Since observation equipment based on physical characteristics can usually only obtain some low-resolution numerical results, it is difficult to compare with manual observations and can only be used as auxiliary data for observations. Therefore, various meteorological observation stations often adopt automatic observation equipment based on visible light images to obtain ground-based cloud images, and then analyze the conditions in the sky through image processing or manual image reading. Commonly used vision-based ground-based cloud observation equipment is divided into two types: one is a scanning-based all-sky imaging system; the other is a fish-eye-based all-sky imaging system.

Scanning-based all-sky imaging systems usually use low-cost cameras and lenses, mounted on a rotating pan/tilt that can rotate 360° and adjust the pitch angle within a range of 0° to 90°. This type of imaging system is more used for observation tasks that require high accuracy for a certain fixed azimuth observation but low requirements for sky panoramic observation, high requirements for cloud observation and low requirements for cloud amount estimation [8]. The all-sky imaging system based on fisheye lens mainly relies on the large field of view of fisheye lens (usually close to or even more than 180°) to achieve the purpose of imaging the sky panorama at one time. This type of imaging system is usually suitable for observation tasks that require high panoramic observation and emphasize cloud cover estimation [9].

The classification and recognition of cloud shapes is an important and difficult task in ground-based cloud observation tasks. The international cloud classification system developed by WMO (World Meteorological Organization) divides clouds into ten categories and 28 sub-categories [10]. In cloud observation, there are mainly 8 types of clouds that need to be observed and recorded: cumulus, stratus, stratocumulus, altocumulus, altocumulus, cirrostratus, cirrocumulus, and cirrus. These 8 kinds of clouds with the clear sky, a total of 9 categories, constitute the classification standard for manual observation. The classification and recognition of clouds is much more difficult than the recognition of other common targets. Therefore, in many cloud recognition methods of ground-based cloud images, the task of cloud recognition is simplified, and only some types of clouds are performed. Identify or simply reclassify clouds according to their thickness [5], [11], [12]. For example, Soumyabrata et al. [13] distinguished three types of sky, thin cloud, and thick cloud only on the basis of ground-based cloud detection. These methods are mainly applied to the simplified cloud classification standard, that is, some similar cloud shapes are combined as the same category for identification. The actual 9 types of cloud shape recognition are not satisfactory, and it is difficult to meet the application requirements.

In the ground-based cloud recognition method, one of the most important factors is the feature extraction of the ground-based cloud image. The stronger the expression ability of the extracted image feature, the better the recognition effect. Liu et al. [14] proposed a method integrates the high-level fusion and the output of low-level fusion with deep visual features and deep multimodal features. Shi et al. [15]...
believe that local rich texture information may be more important than global layout information, so based on the characteristics of shallow convolutional layers, a DCPS network is proposed to classify cloud images. Ye et al. [16] based on deep convolutional features to extract features of ground-based cloud image. With the application of convolutional neural network in image processing, the number of cloud image recognition algorithms based on convolutional neural network is increasing. However, the background, texture and other features of the ground cloud image are of low complexity and feature loss is easy to occur in the process of deep convolution. Therefore, how to appropriately extract the features of the cloud image through the convolutional neural network has become an urgent problem to be solved.

After obtaining the feature description of the ground-based cloud image, selecting a suitable classifier according to the extracted ground-based cloud image features also affects the final effect of ground-based cloud recognition to a certain extent. Chethan et al. [17] proposed a texture feature cloud classification method based on Gabor transform, and used the improved classifier Support Vector Machine (SVM) for classification. Alireza et al. [18] use multilayer perceptron (MLP) neural networks and support vector machine (SVM) capabilities for automatic cloud detection in whole-sky images. Although the two approaches generally generate similar accuracies, the MLP neural networks gave a better performance in some specific cases where the SVM generates poor accuracy.

III. PROPOSED METHOD

A cloud recognition system based on multi-scale convolution feature extraction, screening and fusion is proposed in this paper. The system consists of four parts. The first part is based on the CNN model for image feature extraction at different scales. The second part is the feature screening module, which screens out the more discriminative features from the extracted local features. The third part is feature coding fusion, and the last part is cloud classification. The system architecture diagram is shown in Fig. 2.

A. MULTI-SCALE LOCAL FEATURE EXTRACTION NETWORK

The main challenges of the refined cloud classification of ground-based cloud images are as follows:

(1). Different cloud shapes cannot be characterized by the same type of features. For example, the shapes of cumulus and altocumulus in Fig. 3 are quite special, while the shape of stratocumulus, cirrus and stratus are not obvious and can only be characterized by texture or color.

(2). Some different clouds sometimes have similar parts. For example, in the green box of cumulus and altocumulus in Fig. 3, there are half clouds and half the sky with obvious edges. At the same time, the yellow boxes in cirrus and stratus clouds also have similar textures. If the sampling happens to be this way, it is easy to cause confusion.

(3). Even if it is the same type of cloud, the scales are often different. For example, the red and yellow boxes of cumulus in Fig. 3 are typical cumulus clouds.

Therefore, in order to obtain the best ground-based cloud image features, it is necessary to be able to capture image features of multiple levels and multiple levels. Features are less affected by spatial position and angle. Convolutional neural networks have good multi-level feature extraction capabilities, which can meet the needs of this article.

The network model designed in this paper is shown in Fig. 4. Because the ground-based cloud map scene is relatively simple, it is not suitable for deep convolutional networks. Therefore, this paper constructs a neural network (Multi-CNN) with 5 convolutional layers and 2 fully connected layers. Due to the limited collection of ground-based cloud images and the difficulty of labeling, there is currently no training set that can meet the needs of this article. To solve this problem, this paper uses a pre-trained model to extract features. First, use a general large image data set such as ImageNet to train Multi-CNN, and use the trained model to extract multi-layer convolution features from ground-based cloud images. It should be pointed out that this article only uses the convolutional layers of Multi-CNN for feature extraction, and the fully connected layer for classification will not be used in this article.

Part of the features of the five-layer convolutional layers are randomly selected and visualized by deconvolution,
M. Jingyi et al.: Ground-Based Cloud Image Recognition System Based on Multi-CNN and Feature Screening and Fusion

FIGURE 4. Network structure of Multi-CNN.

as shown in Figure 5. It can be seen that the first layer of convolution filtering mostly responds to simple single textures or local gradients in different directions, the second layer characterizes some combined textures, and the third layer can characterize the structure and contours generated by the texture. By analogy, the deeper the convolutional layer can represent the higher-level and more abstract semantic features.

FIGURE 5. Visualization diagram of filter response of each convolutional layer.

In order to analyze the effect of the Multi-CNN network model on extracting features of ground-based cloud images, first extract the features of each convolutional layer through Multi-CNN, and then reconstruct the input image based on the features of each convolutional layer, as shown in Fig. 6. For shallow convolution features, more refined local details can be retained. For deep convolution features, some low-level detailed information is filtered or transformed into more abstract high-level semantic information. It can be seen that the convolutional layers of Multi-CNN for feature extraction can extract and retain multi-scale and multi-level rich features, which can meet the needs of subsequent classification.

FIGURE 6. The result of image reconstruction for each convolutional layer feature.

B. FEATURE SCREENING NETWORK

As mentioned earlier, one of the difficulties in cloud-like classification and recognition of ground-based cloud images is that different types of cloud images may have some familiar parts. Therefore, finding the key differences between cloud shapes is crucial for distinguishing different cloud shapes. It can be seen intuitively from Fig. 7 that the local areas in the green and yellow boxes in the cumulus, cirrocumulus, and cirrus clouds are very similar and difficult to distinguish. The local area in the red box is the key to distinguish these three foundation cloud maps. Therefore, in the recognition of different types of cloud images, it is not necessary to extract and retain all the local information of the image, and some local information may even weaken the difference of different cloud shapes and affect the final result. It is more important to extract key local information with distinguishing ability than to extract all local information as completely as possible.

FIGURE 7. Key local information between clouds.

In order to further verify this point of view, we randomly selected 40 ground-based cloud images for each cloud shape, a total of 360 images of nine cloud shapes. Multi-CNN is used for feature extraction. Taking the conv5 layer as an example, it contains $13 \times 13 = 169$ feature blocks, and each feature block corresponds to an image block on the ground-based cloud image. There are 360 pictures in total, so all the partial image blocks total $169 \times 360 = 60840$. Cluster the features corresponding to these local image blocks, and then count the distribution of each type of local feature on the 9 cloud images.

The results in Table 1 are the results of clustering the local image blocks into 5 categories. It can be seen that the fifth type of local image block after clustering is an untextured area close to the clear sky, and such local features appear in almost all ground-based cloud images. Obviously, the features of this type of local area are not very helpful for distinguishing different cloud-like categories. On the contrary, too much such local information will cause confusion and interference in feature classification. This type of local feature can be considered as redundant. It is necessary to remove such redundant local features before performing local feature encoding.

In order to get a better clustering effect, this paper adapted a density peak-based clustering method (DP). Compared with the common K-means clustering method, the clustering results of DP clustering are more deterministic.

In DP clustering, the selection of the cluster center point is mainly based on two considerations: 1) The density of the
sample point is greater than the surrounding points; 2) The
distance between the sample point and the point with greater
density is as large as possible. Based on this goal, the cut-off
kernel is first used in DP clustering to define the local density
function of the sample points:

\[ \rho_i = \sum_{j \in I_i} \mathcal{X}(d_{ij} - d) \]  \hfill (1)

\[ \mathcal{X}(a) = \begin{cases} 1, & a < 0 \\ 0, & \text{others} \end{cases} \]  \hfill (2)

where, \( d \) is a preset value, which means that the density of the
sample points is calculated statistically within the radius of
this distance. \( d_{ij} \) represents the distance between sample
point \( i \) and sample point \( j \). After obtaining the local density,
define the distance offset of each sample point:

\[ \delta_i = \begin{cases} \min_{j: \rho_j > \rho_i} (d_{ij}), & \text{condition1} \\ \max_j (d_{ij}), & \text{condition2} \end{cases} \]  \hfill (3)

where, \text{condition1} means that the sample point \( i \) is not the
point with the highest local density. \text{Condition2} means that
the sample point \( i \) is the point with the highest local density.
The meaning of distance offset is the minimum value of the
distance between point \( i \) with all the sample points with local
density greater than point \( i \). And the point with the largest
local density is treated separately as the maximum distance
from all points. After the local density \( \rho \) and distance offset
(\( \delta \)) are defined, the cluster center point can be determined by
the larger \( \rho \) and \( \delta \) values of the sample points. As shown in
Fig. 8, for the sample distribution in the left picture, make a
\( \rho - \delta \) coordinate map, and it can be found that No. 1 and
No. 10 are better than other points, and they are most suitable
as the cluster center points.

When statistically filtering the local convolution features
of each level of Multi-CNN, first cluster all the local features
of this layer in the sample image into \( M \) categories using the
DP clustering method, and each local feature will get a cluster
label \( e = \{1, 2, \ldots, M\} \). According to the cloud type of the
original image where each local feature is located, each local

\hspace{1cm} TABLE 1. The clustering and distribution of local patterns of
ground-based cloud images.

| Category after clustering | 1 | 2 | 3 | 4 | 5 |
|--------------------------|---|---|---|---|---|
| Examples                 |   |   |   |   |   |
| Altostratus              | 83| 4072| 2 | 0 | 2613 |
| Cirrostratus             | 2745| 701| 1693| 520| 1120 |
| Cirrus                   | 401| 2986| 3 | 0 | 3319 |
| Cumulus                  | 1697| 1831| 319| 121| 3006 |
| Stratus                  | 2187| 1265| 1048| 347| 1926 |
| Clear Sky                | 1315| 3018| 191| 42 | 2326 |
|                           | 41 | 2911| 1 | 0  | 3759 |

\hspace{1cm} FIGURE 8. Cluster center selection during DP clustering.

feature is given a corresponding cloud label \( l = \{1, 2, \ldots, L\} \),
where \( L \) is the total cloud type. Then count the distribution of
each type of local feature in each type of cloud, denoted as
\( D(c_i, l_j) \), and calculate as follows:

\[ D(c_i, l_j) = \frac{c_i \sim l_j}{c_{ALL}} \]  \hfill (4)

where, \( c_i \sim l_j \) represents the number of \( j \)-th cloud label in
the \( i \)-th cluster label. \( c_{ALL} \) represents the total number of
\( i \)-th cluster label. Based on this, the distribution of local features
corresponding to a class of cluster labels in various
ground-based cloud images is obtained. If the local features
corresponding to a cluster label are evenly distributed in the
images of each cloud type, it means that these local features
are not very helpful for cloud recognition. On the contrary,
if the local features corresponding to a cluster label are not
uniformly distributed in the images of various cloud-like
types, the discriminating ability of this type of local features
is strong. Here we define the discriminative ability of the \( i \)-th
cluster label as \( disc(c_i) \), which is calculated as follows:

\[ disc(c_i) = \text{std}(D(c_i, l_j)) = \sqrt{\frac{\sum_{l=1}^{L} (D(c_i, l_j) - \text{Davg}(c_i))^2}{L}} \]  \hfill (5)

\[ \text{Davg}(c_i) = \frac{\sum_{l=1}^{L} D(c_i, l_j)}{L} \]  \hfill (6)

Sort according to the discriminative ability \( disc(c_i) \) of
various clustering modes. Local features corresponding to
cluster label with the smallest discriminative ability should be
eliminated, and the other local features will remain. Fig. 9 and
Fig. 10 respectively show examples of local image blocks
that need to be eliminated and retained after screening in this
article, and their distribution in various ground-based cloud
maps.

It can be seen from Fig. 9 that the image blocks corre-
sponding to the eliminated local features are very flat and
almost have no texture features, and the colors are relatively
similar. It is widely distributed in various ground-based cloud
maps, and the distribution is relatively even, and the highest
distribution ratio does not exceed 14%. The image blocks
corresponding to the retained features shown in Fig.10 are
more than 30% concentrated in cumulus clouds, but hardly
exist in clear sky, stratus clouds, and high-level clouds. The
distribution in other cloud images is also different.
However, in the actual online test or observation process, images are often input one by one, so frequent clustering is time-consuming and the clustering results are not representative because only one image is input. Therefore, the cluster centers of all cluster labels during the training process will be recorded. During the testing process, the distance between the local features extracted from the image and the cluster centers will be directly compared and the category represented by the closest clustering center will be taken as the clustering category of each local feature point. The local features corresponding to the cluster labels with the weakest discriminative ability will be eliminated directly without participating in the subsequent feature encoding process.

C. FEATURE CODING AND FUSION NETWORK

After obtaining the filtered local features, the next step is to use these local features to form the global features of the ground-based cloud image for classification. When using local features of the convolutional layer to encode global features, due to the fixed structure of the convolutional neural network, the size of the input image will affect the size of the feature map output by the convolutional layer, so the encoding process cannot be affected by the number of local features in simple series. The non-linear changes of the fully connected neural network weaken the geometric space constraints to a certain extent, but it can only be done after the local features are arranged in a fixed order, which actually contains spatial location information.

This paper proposed another local feature encoding method based on Fisher Vector. Fisher Vector coding first assumes that all image information can be characterized by several feature points or feature descriptors. These feature points are considered to satisfy an independent random distribution and can be expressed by a Gaussian Mixture Model (GMM). The Gaussian mixture model can be expressed by (7).

$$p(x) = \sum_{k=1}^{K} \omega_k N(x | \mu_k, \delta_k)$$

where, $p(x)$ represents the probability of the feature point $x$, $K$ represents the number of Gaussian models, $\omega_k$ represents the weight of the kth Gaussian model, and $N(x | \mu_k, \delta_k)$ represents the kth Gaussian distribution, $\mu_k$ and $\delta_k$ respectively represent the mean and variance of the Gaussian distribution. Suppose that the probability of several feature points or local features $x_i$ of an image $X$ appearing in the mixed Gaussian space is $p(x_i | \lambda)$, where $\lambda = \{\omega_i, \mu_i, \delta_i, i = I, \ldots, K\}$ represents the parameter group of the Gaussian mixture model. The probability of image $X$ can be expressed as:

$$P(X | \lambda) = \prod_{i=1}^{N} p (x_i | \lambda)$$

The function of the fully connected layer in Multi-CNN is actually similar to a nonlinear feature encoding based on neural networks. Fig. 11 shows a schematic diagram of using a fully connected layer to re-encode the local feature map of the convolutional layer. Fig. 11(b) shows the local feature map output by any certain layer of convolutional layer. Each point in the figure is corresponding to a local area in the original input image, as shown in the three colors of red, green and yellow in the figure. The simplest local feature encoding method is shown in Fig. 11(b) by connecting the local features according to a certain rule, but obviously this method is relatively simple and inefficient. The fully connected neural network shown in Fig. 11(c) is a nonlinear mapping based on this, avoiding the disadvantages caused by the change in the number of local features in simple series.
where, \( N \) represents the number of feature points or local features contained in the image \( X \). For the convenience of calculation, usually take the logarithm of this expression before proceeding to the next step. The expression after taking the logarithm can become:

\[
L(X | \lambda) = \sum_{i=1}^{N} \log (p(x_i | \lambda))
\]

(9)

Since the Gaussian distribution is a weighted mixture of multiple Gaussian distributions, there is:

\[
p(x_i | \lambda) = \sum_{k=1}^{K} \omega_k p_k (x_i | \mu_k, \delta_k)
\]

(10)

\[
p_k (x_i | \mu_k, \delta_k) = \frac{1}{\sqrt{2\pi\delta_k}} e^{-\frac{(x_i-\mu_k)^2}{2\delta_k}}
\]

(11)

Find the partial derivative of \( L \) with respect to weight, mean and variance as shown in (12).

\[
\frac{\partial L(X | \lambda)}{\partial \omega_k}, \frac{\partial L(X | \lambda)}{\partial \mu_k}, \frac{\partial L(X | \lambda)}{\partial \delta_k}, k = 1, \ldots, K
\]

(12)

where, \( K \) is the number of Gaussian models in GMM, from which we can get \( K \) groups of such parameter partial derivatives. Assuming that the dimension of the feature point or local area feature \( x_i \) is \( D \)-dimensional, the parameters \( \mu_k, \delta_k \) and their corresponding partial derivatives are also \( D \)-dimensional vectors. In this way, these parameters can be concatenated into a \( K \times (2D + 1) \) dimensional feature vector, which is called the Fisher Vector. The feature encoding method based on the Fisher Vector can not only form a global feature independent of the number of local feature points, but also increase the feature dimension and make the feature more linearly separable.

In order to compare the sensitivity of the two feature encoding methods to image rotation and other changes, the same ground-based cloud image is rotated and transformed, as shown in Fig. 12. Then compare the global features after feature encoding using different feature encoding methods.

\[
\text{FIGURE 12. Original image and its various rotation transformations.}
\]

The two encoding methods are all based on the feature map output by conv5 in Multi-CNN for feature encoding. The result obtained through the fully connected layer coding can be directly represented by the output result of the FC7 layer fully connected layer. Since the dimension of the feature map output by the conv5 convolutional layer is 512, the number of Gaussian mixture models of the Fisher Vector is set to 4 to ensure that the dimension of the finally obtained Fisher Vector feature vector is \( 512 \times 2 \times 4 \). Taking the six images shown in Fig. 12 as input images, the image feature obtained by fully-connected network feature encoding is marked as \( f_i^{fc} = [x_1^i, x_2^i, \ldots, x_{4096}^i] \), and the image feature obtained by Fisher Vector encoding is marked as \( f_i^{FV} = [y_1^i, y_2^i, \ldots, y_{4096}^i] \). Then all feature vectors are normalized to \([0, 1]\) to calculate the difference between the features of different input images in the two encoding cases. The difference between features is calculated as shown in the following formula:

\[
D_{i,j}^{fc} = \frac{1}{4096} \sum_{k=1}^{4096} \left| x_k^i - x_k^j \right|, \quad i = 1, \ldots, 6, j = 1, \ldots, 6
\]

(13)

\[
D_{i,j}^{FV} = \frac{1}{4096} \sum_{k=1}^{4096} \left| y_k^i - y_k^j \right|, \quad i = 1, \ldots, 6, j = 1, \ldots, 6
\]

(14)

Therefore, the difference in feature expression caused by image flip and rotation in the two encoding methods can be obtained. As shown in Table 2 and Table 3, when the image changes, the feature change of Fisher Vector encoding is much smaller than that of fully connected layer encoding. Therefore, this article chooses Fisher Vector coding method for feature coding fusion.

\[
\text{TABLE 2. Feature changes of fully connected layer feature coding under different image transformations.}
\]

| Original Image | Horizontal flip | Vertical flip | Rotate 90° | Rotate 180° | Rotate 270° |
|----------------|-----------------|---------------|------------|-------------|-------------|
| 0.0211         | 0.0491          | 0.0354        | 0.0522     | 0.0431      |
| 0.0211         | 0.0421          | 0.0310        | 0.0428     | 0.0401      |
| 0.0491         | 0.0421          | 0.0411        | 0.0203     | 0.0298      |
| 0.0354         | 0.0310          | 0.0411        | 0.0396     | 0.0425      |
| 0.0522         | 0.0428          | 0.0203        | 0.0396     | 0.0352      |
| 0.0431         | 0.0401          | 0.0298        | 0.0425     | 0.0352      |

\[
\text{TABLE 3. Feature changes of Fisher Vector coding under different image transformations.}
\]

| Original Image | Horizontal flip | Vertical flip | Rotate 90° | Rotate 180° | Rotate 270° |
|----------------|-----------------|---------------|------------|-------------|-------------|
| 0.0093         | 0.0184          | 0.0253        | 0.0125     | 0.0217      |
| 0.0093         | 0.0210          | 0.0241        | 0.0192     | 0.0196      |
| 0.0184         | 0.0210          | 0.0267        | 0.0101     | 0.0243      |
| 0.0253         | 0.0241          | 0.0267        | 0.0195     | 0.0114      |
| 0.0125         | 0.0192          | 0.0101        | 0.0195     | 0.0221      |
| 0.0217         | 0.0196          | 0.0243        | 0.0114     | 0.0221      |

VOLUME 8, 2020
After the global features are obtained by feature encoding fusion, multilayer perceptron neural networks (MLP) [18] is selected for image classification.

IV. EXPERIMENTS AND ANALYSIS

A. DATASET

The data set in this article is mainly derived from the scanning all-sky imaging system developed by the Observation Center of the National Meteorological Administration of China. These images were taken at the Beijing Observatory. More than 1,300 images were selected by experienced meteorological observation experts and labeled into 9 categories: Clear sky, Cumulus, Stratocumulus, Stratus, Altostratus, Altocumulus, Cirroculumulus, Cirrostratus, Cirrus, each type of cloud images are in the range of 60 to 200. Due to the similarity of some clouds, it is difficult for non-professional meteorological observers to see the differences, such as altocumulus and cirrocumulus, cirrus and cirrostratus, stratus and altostratus. Therefore, this paper uses two classification standards to test the algorithms separately, as shown in Fig. 13 and Fig. 14. Among them, the low-level difficulty is classified into six categories, combining two very similar ground-based cloud images into one category. The high-level difficulty is classified into nine categories, each of which is independent of one category.

B. GROUD-BASED CLOUD IMAGE RECOGNITION EXPERIMENT

The algorithm in this paper is based on the Multi-CNN convolutional neural network, which extracts multi-layer and multi-scale convolution features. Coding and fusing the local features after screening, and then classifying them through MLP. In the process of testing the effect of ground-based cloud recognition, tests will be carried out according to two groups of six classification standards and nine classification standards. In order to reflect the robustness of the classification method, ten rounds of random experiments will be carried out for each group. In each round, some images are randomly selected as the training set, and the remaining images are used as the test set. Then take the average of ten rounds of results as the final test result. In addition, other advanced methods are also tested in the same data set.

Table 4 and Table 5 are the test results under the six classification standards and nine classification standards respectively. There are three cases for the method in this paper: ‘Ours(conv5)’ means that only conv5 convolution features are used in the feature extraction process, ‘Ours(conv345)’ means that multi-scale features of conv3, conv4 and conv5 are used, ‘Ours(conv-all)’ means that all 5 convolutional layers are used.

From Table 4 and Table 5, we can see that the method in this chapter is better than other methods under the two classification standards. Under the nine-category classification standard, the training samples of each category are fixed to 40 images. Comparing the results in the penultimate column of Table 4 with the last column of Table 5, it is found that the classification accuracy rate of the nine categories exceeds that of the six categories in the method of this paper. There

| Number of training samples per category | 5  | 10 | 20 | 40 | 80 |
|----------------------------------------|----|----|----|----|----|
| Chethan et al. [17]                    | 38.6 | 46.9 | 59.3 | 68.2 | 74.5 |
| (±3.1)                                 | (±4.4) | (±2.1) | (±1.2) | (±0.9) |
| Hu. et al. [5]                         | 36.9 | 44.1 | 53.0 | 60.1 | 66.1 |
| (±2.1)                                 | (±2.7) | (±1.8) | (±2.9) | (±1.1) |
| Liu. et al. [19]                       | 35.1 | 38.5 | 45.3 | 57.3 | 60.3 |
| (±2.7)                                 | (±3.2) | (±2.1) | (±1.4) | (±1.8) |
| Shi. et al.[15]                        | 28.9 | 29.8 | 35.7 | 38.9 | 43.5 |
| (±2.6)                                 | (±2.1) | (±1.4) | (±1.1) | (±2.1) |
| Alireza et al. [18]                    | 46.3 | 55.2 | 67.1 | 78.3 | 84.2 |
| (±3.9)                                 | (±2.6) | (±2.3) | (±1.4) | (±1.4) |
| Ours(conv5)                            | 61.7 | 71.7 | 74.9 | 81.8 | 85.7 |
| (±3.8)                                 | (±1.9) | (±2.5) | (±1.0) | (±1.5) |
| Ours(conv345)                          | 64.3 | 73.4 | 78.5 | 87.7 | 91.6 |
| (±3.5)                                 | (±2.1) | (±2.0) | (±1.4) | (±1.1) |
| Ours(conv-all)                         | 64.9 | 76.6 | 81.5 | 91.2 | 94.5 |
| (±3.4)                                 | (±3.3) | (±1.8) | (±1.2) | (±0.8) |
are two reasons. On the one hand, in the 6 classification standards, only 40 samples of the merged two types of clouds are selected as training samples, and a total of 240 images of the six types. Under the 9 classification standards, a total of 360 images participate in training. The total number of training samples affects the coverage of the training set and the generalization ability of the model training. On the other hand, for the merged two types of clouds, there may be an uneven distribution when randomly selecting training samples. When the feature distinguishing ability of the two types of samples is excellent, this unbalanced selection of training samples may cause interference to the classifier. The individual recognition effect of cirrus and cirrostratus, cirrocumulus and altocumulus, stratus and altostratus is very good. These phenomena indicate that the ground-based cloud image features obtained by the method in this paper can distinguish various cloud shapes well, and have strong characterization and discrimination capabilities.

Figure 15 shows the ROC curve of each network under 9 classification standards. It can be seen that the network in this paper has the highest recognition accuracy and the maximum AUC.

C. EFFECT EXPERIMENT OF LOCAL FEATURE CODING

Previously, this article roughly analyzed the sensitivity of the Fisher Vector feature encoding method to image rotation by comparing a set of sample pictures. In order to further verify that the performance of the Fisher Vector encoder is better than that of the fully connected layer (FC) encoder, a set of comparative experiments is designed based on the local features of the conv5 output. There are mainly three situations: 1. Use two fully connected layers to encode the local features of the conv5 layer (FC); 2. Connect the local features of the conv5 layer directly in sequence (Conv5); 3. The local features of the conv5 layer are coded with Fisher Vector using 4 Gaussian mixture models (FV).

Table 6 and Table 7 show the performance of the three conditions in the recognition of ground-based cloud images. It can be seen that the performance of using Fisher Vector for feature encoding is the best, and the performance of using fully connected layer features is almost the same as that of directly connecting the local features of the conv5 layer. It is proved that when the local convolution features are input to the fully connected layer, the arrangement of the features has been implicitly included in the output result of the fully connected layer. The characteristic of ground-based cloud images is weak spatial constraints, so local feature coding based on Fisher Vector is more suitable for ground-based cloud images.

D. MULTI-LEVEL AND MULTI-SCALE FEATURES EFFECT VERIFICATION EXPERIMENT

The algorithm in this paper uses a total of five convolutional layers of multi-scale and multi-level feature information. Each convolutional layer expresses different semantic information. In order to explore the performance of each level, the characteristics of each convolutional layer are used.
TABLE 7. Recognition accuracy rate of ground-based cloud image of 9 classification standards under different local coding methods (%).

| Cloud shape | Cumulus, | Cirrus | Cirrostratus | Cirrocumulus | Altocumulus | Clear sky | Stratocumulus | Stratus | Altostratus | Average |
|-------------|----------|--------|--------------|--------------|-------------|-----------|---------------|---------|-------------|---------|
| FC          | 76.8     | 67.1   | 69           | 91.2         | 61.4        | 95.5      | 63.4          | 97.3    | 74.6        | 77.4    |
| Conv5       | 73.5     | 64.2   | 70.2         | 89.3         | 61.2        | 94.6      | 58.7          | 94.5    | 71.8        | 75.3    |
| FV          | 80.1     | 76.8   | 74.3         | 95.2         | 69.5        | 92.7      | 64.1          | 98.4    | 79.3        | 81.2    |

In order to verify the effect of multi-scale and multi-level convolutional features on recognition, gradually increase the number of convolutional layers. As shown in Figure 18 and Fig. 19, from using only the fifth convolutional layer (conv5) to using all five convolutional layers (conv-all), each additional scale or level increases the classification effect. When conv2 was added, the network performance improved the most. The number of training samples per category in this experiment is 80. When all the convolutional layer features are used, the recognition accuracy of the network under the 9 classification criteria reaches 94.8%. In addition, as can be seen from Fig. 18 and 19, with sufficient training samples, as the number of convolutional layers used increases, the accuracy of the 9 classification standard will be higher than the accuracy of the 6 classification. This shows that the network has a strong ability to extract and distinguish the features of different cloud images.
E. FEATURE SCREENING EFFECT VERIFICATION EXPERIMENT

An important part of this article is the selection of local features, which is also the key to refined cloud recognition and classification. Therefore, a set of comparative experiments is set up to test the effect of local feature screening on cloud image recognition.

In this experiment, the features of the three conv3, conv4, and conv5 convolutional layers were tested for the independent feature screening effect and the feature screening effect after the three levels of fusion. The experimental results are shown in Table 8 and Table 9. Whether using single-level or multi-level convolution features, feature screening can significantly improve the effect of cloud classification and recognition.

| Cloud shape      | Cumulus, | Cirrus   | Cirrocumulus | Altocumulus | Clearsky | Stratocumulus | Stratus | Averag |
|------------------|----------|----------|--------------|-------------|----------|----------------|--------|--------|
| Conv3            | 82.9     | 82.6     | 67.7         | 94.1        | 70.1     | 88.7           | 68.4   | 91.3   |
| Conv3+screening  | 84.8     | 80.6     | 71.5         | 95.4        | 76.3     | 91.8           | 67.1   | 97.1   |
| Conv4            | 85.7     | 81.1     | 64.2         | 92.6        | 72.9     | 92.3           | 66.0   | 93.9   |
| Conv4+screening  | 84.9     | 82.3     | 72.1         | 95.6        | 74.3     | 95.1           | 66.3   | 98.1   |
| Conv5            | 83.9     | 85.1     | 67.2         | 91.8        | 73.4     | 91.9           | 68.5   | 96.8   |
| Conv5+screening  | 87.1     | 85.1     | 66.5         | 91.4        | 82.1     | 93.3           | 67.9   | 97.4   |
| Conv345          | 85.6     | 84.3     | 67.8         | 92.6        | 75.5     | 93.1           | 69.4   | 97.1   |
| Conv345+screening| 87.2     | 84.1     | 79.5         | 93.8        | 80.2     | 95.7           | 71.3   | 99.3   |

In the process of feature screening, the results of clustering may also have a very important impact on the effect of feature screening. This article uses DP clustering and the number of clustering categories is set to 10. The previous chapters have shown that DP clustering performs better than the commonly used K-means clustering. In order to further research the effects of different clusters and the influence of the number of cluster categories on recognition results, a set of comparative experiments was set up. Select the local features of three independent conv3 layers, conv3, conv4, and conv5 as the basis, and select the number of clusters 3, 5, 8, and 10 in both K-means and DP clustering methods for experiments. The test results are shown in the Fig. 20. It can be seen that the improvement effect of DP clustering is more obvious, and the K-means clustering method cannot even improve the final classification effect in some cases. When the number of clusters is 10, the feature filtering effect is the best, and too few clusters (when it is 3) will make the improvement effect decrease.

![FIGURE 20. The effect of different clustering methods and number of clusters on recognition.](image)

V. CONCLUSION

This paper proposes a ground-based cloud image recognition system based on convolutional neural network and feature selection and fusion. The multi-level and multi-scale convolution feature extraction is performed through the convolution layer of Multi-CNN, and the local features with strong resolving power are selected through the feature selection algorithm based on DP clustering. Based on Fisher Vector, the local features are encoded and fused to obtain global features, and MLP is used for classification prediction. Comparative experiments show that the algorithm has a strong ability to extract and distinguish the features of different cloud images.

REFERENCES

[1] Y. Wang, C. Wang, C. Shi, and B. Xiao, “A selection criterion for the optimal resolution of ground-based remote sensing cloud images for cloud classification,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 3, pp. 1358–1367, Mar. 2019, doi: 10.1109/TGRS.2018.2866206.

[2] M. Li, S. Liu, and Z. Zhang, “Dual guided loss for ground-based cloud classification in weather station networks,” IEEE Access, vol. 7, pp. 63081–63088, 2019, doi: 10.1109/ACCESS.2019.2916905.

[3] Y. Yuan and X. Hu, “Bag-of-words and object-based classification for cloud extraction from satellite imagery,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 8, no. 8, pp. 4197–4205, Aug. 2015, doi: 10.1109/JSTARS.2015.2431676.
L. Y. Liu, B. L. Lv, J. Y. Xu, S. C. Wang, C. H. Fu, and C. Z. Ye, “Automatic observation experiments of cloud amounts and cloud forms based on the image recognition,” in Proc. Int. Conf. Meteorol. Observat. (ICMO), Chengdu, China, Dec. 2019, pp. 1–4, doi: 10.1109/ICMO49322.2019.9025848.

X. Hu, Y. Wang, and J. Shan, “Automatic recognition of cloud images by using visual saliency features,” IEEE Geosci. Remote Sens. Lett., vol. 12, no. 8, pp. 1760–1764, Aug. 2015, doi: 10.1109/LGRS.2015.2424531.

S. S. Ali, A. Hammad, and A. S. T. Eldien, “MC2PS: Cloud-based 3-D place recognition using map segmentation coordinates points,” IEEE Commun. Lett., vol. 22, no. 8, pp. 1560–1563, Aug. 2018, doi: 10.1109/LCOMM.2018.2833122.

A. Pérez-Suay, J. Amorós-López, L. Gómez-Chova, J. Muñoz-Marí, D. Just, and G. Camps-Valls, “Pattern recognition scheme for large-scale cloud detection over landmarks,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 11, no. 11, pp. 3977–3987, Nov. 2018, doi: 10.1109/JSTARS.2018.2863383.

J. R. Amorós-López, M. Syrjäsuo, and K. E. A. van de Sande, “Automatic auroral detection in color all-sky camera images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 7, no. 12, pp. 4717–4725, Dec. 2014, doi: 10.1109/JSTARS.2014.2321433.

F. Su, W. Jing, J. Zhang, H. Wang, and M. Zhang, “A local features-based approach to all-sky image prediction,” IBM J. Res. Develop., vol. 59, no. 2–3, pp. 6:1–6:10, Mar./May 2015, doi: 10.1147/ JRDI.2015.2397772.

International Cloud Atlas Volume I: Manual on the Observation of Clouds and Other Meteors, World Meteorol. Org., Geneva, Switzerland, 1975.

C. Shi, Y. Wang, C. Wang, and B. Xiao, “Ground-based cloud detection using graph model built upon superpixels,” IEEE Geosci. Remote Sens. Lett., vol. 14, no. 5, pp. 719–723, May 2017, doi: 10.1109/LGRS.2017.2676007.

Q. Li, W. Lu, J. Yang, and J. Z. Wang, “Thin cloud detection of all-sky images using Markov random fields,” IEEE Geosci. Remote Sens. Lett., vol. 9, no. 3, pp. 417–421, May 2012, doi: 10.1109/LGRS.2011.2170953.

S. Dev, Y. H. Lee, and S. Winkler, “Multi-level semantic labeling of sky/cloud images,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Quebec City, QC, Canada, Sep. 2015, pp. 636–640, doi: 10.1109/ICIP.2015.7350876.

S. Liu, L. Duan, Z. Zhang, and X. Cao, “Hierarchical multimodal fusion for ground-based cloud classification in weather station networks,” IEEE Access, vol. 7, pp. 85688–85695, 2019, doi: 10.1109/ACCESS.2019.2926092.

C. Shi, C. Wang, Y. Wang, and B. Xiao, “Deep convolutional activations-based features for ground-based cloud classification,” IEEE Geosci. Remote Sens. Lett., vol. 14, no. 6, pp. 816–820, Jun. 2017, doi: 10.1109/LGRS.2017.2681658.

L. Ye, Z. Cao, Y. Xiao, and W. Li, “Ground-based cloud image categorization using deep convolutional visual features,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Quebec City, QC, Canada, Sep. 2015, pp. 4808–4812, doi: 10.1109/ICIP.2015.7351720.

H. K. Chethan, G. H. Kumar, and R. Raghavendra, “Texture based approach for cloud classification using SVM,” in Proc. Int. Conf. Adv. Recent Technol. Commun. Comput., Kottayam, India, 2009, pp. 688–690, doi: 10.1109/ARTCom.2009.43.

A. Taravat, F. Del Frate, C. Cornaro, and S. Vergari, “Neural networks and support vector machine algorithms for automatic cloud classification of whole-sky ground-based images,” IEEE Geosci. Remote Sens. Lett., vol. 12, no. 3, pp. 666–670, Mar. 2015, doi: 10.1109/LGRS.2014.2356616.

L. Liu, X. Sun, F. Chen, S. Zhao, and T. Gao, “Cloud classification based on structure features of infrared images,” J. Atmos. Ocean. Technol., vol. 28, no. 3, pp. 410–417, Mar. 2011.