# Image Recognition and Classification Based on Cascading SVM and improved ResNet

Shizhuo Sun, Zehao Wang, Heng Zhao and Zhichao Pan

## Abstract

Image classification and recognition are the two most common tasks in computer vision processing. The traditional machine learning (ML) method and the popular deep learning (DL) method are widely used in the classification and recognition of pictures and both have achieved good results in different degrees. In this paper, the traditional machine learning method based on cascading Support Vector Machine (SVM) and the deep learning method based on residual network, which has the leading effect at present, are respectively employed to classify and identify the data set of 40 kinds of images created by ourselves. There are three main points in our contribution: (1) We compare and analyze ML methods and DL methods based on our self-built dataset. (2) The improved ResNet Network based on residual was proposed and achieved good results. (3) The cascading SVM model was created and used for image classification and recognition. In the experiment, the results and method are compared and analyzed. And the results of classification based on DL had achieved a high accuracy rate of 93% while the SVM classifier based on ML achieved the highest accuracy rate of 69 %, which verified the effectiveness of DL method and ML method.

Key words: machine learning, deep learning, image recognition and classification

## 1 Introduction

Image classification and recognition is a basic problem in the field of computer vision. In order to solve this problem, the traditional ML method uses manual extraction of features and then combines them with classification model for recognition. In recent years, with the extensive application of DL method and the rapid development of Convolutional Neural Networks (CNN), the recognition accuracy has been continuously improved and good results have been achieved.

However, both the traditional ML method and the rapidly developing DL method based on CNN have their own advantages and disadvantages in the task of image recognition and classification. In this paper, we build an image data set by ourselves, and adopt ML method based on cascading SVM classifier and DL method based on residual network to classify and identify the self-built image data set respectively.

In the experimental part, Histogram of Oriented Gradient (HOG), HSV, Local Binary Patterns (LBP), Grey-Level Co-occurrence Matrix (GLCM) and Scale-invariant feature transform (SIFT) features are extracted from the data set and spliced according to weight based on the ML method and the recognition results are obtained by

combining the Cascading SVM classifier. And based on the DL method, we improved the traditional ResNet, added the attention module and combined with the idea of Inception Net[1], classified the data set, and finally reached the recognition accuracy of 93% on the test set. In this paper, there are three main points in our contribution: (1) We compare and analyze ML methods based on SVM and DL methods based on improved ResNet. (2) We improved the ResNet Network based on residual and achieved good results. (3) The cascaded SVM model is employed for image classification and recognition and the weighted feature combination is used as the new feature.

This paper is organized into five sections: Section I introduces the background of the problem mentioned in this paper. Section II discusses about the works and researches related to ML and DL in the field of image classification and recognition. Section III presents the process of ML method and network architecture of DL in detail. In Section IV, we described details about the setup of the database as well as experiments and analyzed the final results. Eventually, we came to our conclusion in Section V.

# 2 Related Works

## 2.1 Image classification and recognition based on Machine Learning

In recent years, with the advent of the era of artificial intelligence, image recognition has made great strides in development. Based on traditional machine learning, image recognition is mainly carried out from the following aspects: information acquisition, preprocessing, feature extraction, selection, classifier design and classification decision. Feature extraction refers to the use of mathematical methods to extract features of the image itself. Moreover, images are classified according to the features extracted by the classifier, so feature extraction is very important for image recognition [1]. The common image feature extraction algorithm includes SIFT, HOG, LBP, etc. Classifier design refers to the classification of recognized photos by using the training data with tags and the recognition rules of training, that is, the prediction label of unknown types of images. Currently, classification algorithms used in image recognition mainly include convolutional neural network (CNN), support vector machine (SVM), bag of word model (BoW), etc.

At present, there are many image recognition algorithms. In 1979, Nobuyuki Otsu proposed the most classical method of maximum interclass variance [2], which is also known as Otsu method through self-adjustment to determine the threshold value. In 1967, Mac Queen. J proposed k-means algorithm [3]. L.incen proposed the watershed algorithm [4] for image edge extraction. Maria Frucc improved on the watershed algorithm. This algorithm is divided into two stages, the goal of which is to reduce the number of regions divided by watershed algorithm. In the first stage, the seed set is reduced by mining, while in the second stage, some regions meeting the requirements are merged [5]. Stephen Gang Wu first converted the images into grayscale images, and then used principal component analysis (PCA) to normalize 12 features, and then used PNN as a classifier to classify 32 species of plants, and finally achieved over 90%

accuracy [6]. Hoshang Kolivand et al. proposed a new method for leaf shape classification and plant species identification based on venation detection [7].

## 2.2 Image classification and recognition based on Deep Learning

Traditional image recognition technology is mainly based on shallow hierarchical structure model, which requires artificial preprocessing of images, resulting in reduced accuracy of image recognition [8]. To improve the accuracy of image recognition, deep learning model structure is proposed. In recent years, deep learning model has made great breakthroughs in extracting high-level feature representation of images. Like the hierarchical processing of information in the visual cortex of the brain, deep learning model can form high-level feature representation from pixels to targets, which is not available in the traditional shallow learning structure. However, the multi-layer network structure can not complete the training efficiently. It was not until 2006 that Hinton et al. proposed to solve the training problem of deep structure by using unsupervised greed layer by layer training algorithm area, which attracted people's attention to deep learning.

Since then, deep learning has developed. It uses CNN to learn features directly from the original data. In 2011, the Relu activation function was proposed to effectively suppress the gradient disappearance. In 2012, Hiton and his students joined ILSVRC and won the championship, since which deep learning has become dominant in computer vision [9]. Hiton and his students took part in ILSVRC and won the championship, thus establishing a solid position in deep learning [6]. Later, deep learning became more and more popular. In the 2014 Image Net competition,Karen Simonyan et al. [10] proposed that VGG achieved the second place. In 2015, Kaiming He [11] proposed ResNet in the ILSVRC classification task and obtained the first prize.

As a result of the success of deep learning, people began to classify images of large orders of magnitude [12][13][14][15][16] , and achieved good results.

# 3 Method

## 3.1 The self-built dataset of Images

To ensure the reliability of the experimental data set, instead of using the open source data set, we randomly crawled 40 kinds of images **shown in Table I** from Baidu and Google websites, with about 500 images in each of these categories and totally 20000 images. We designed a series of scripting tools to help us screen out the data that can be trained and expand the data set through inversion, mirroring, adding noise, left-right translation, etc., finally making the data set number of each group unified at 500.

Table I 40 categories images

| 40 kinds of images | | | | | | | |
|---|---|---|---|---|---|---|---|
| giraffe | train | car | zebra | fox | plane | puppy | whale |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| sheep | bear | monkey | statue | cow | tiger | lion | tower |
| horse | elk | bird | koala | rose | chrysanthemum | cockscomb | carnation |
| anthurium | tulip | irises | lotus | lilium | pear flower | peony | violet |
| gardenia | azalea | jasmine | ling | callas | morning glory | red flower | osmanthus |

## 3.2 Image classification and recognition based on cascading SVM

### 3.2.1 The preprocess of data

For machine learning, the robustness of the data set can directly determine the final classification result and the robustness of the model. The high quality of data sets plays a crucial role in the extraction and optimization of subsequent feature projects. Considering that the Grab-cut method is computation-intensive and not conducive to user interaction, we chose another method of image significance enhancement based on local contrast [17] to enhance the image features and remove the useless background. As **Shown in Figure 1**, the features of the picture become clearer after significant enhancement.

Fig.1. The image of lily after significant enhancement

### 3.2.2 Feature Engineering

we extract various features of the image, including SIFT, GLCM, HOG, HSV and LBP, which can be used for training after stitching and dimensionality reduction.

SIFT, the scale-invariant feature transform, is a description of the field of image processing. This description has scale invariance and can detect key points in the image. The direct SIFT feature descriptor of each image is not fixed and cannot be used as feature vector. In this paper, the BOW model is used to generate feature vector of SIFT. The specific measures are as follows: The first step is to use SIFT algorithm to extract visual terms from each type of image and to gather all visual terms together **shown in Figure 2**. Second, the k-means algorithm is used to construct the word list. This algorithm takes K as the parameter and divides N objects into K clusters, so that the similarity between clusters is higher and the similarity between clusters is lower. The third step is to use the words in the word list to represent the image. By counting the occurrence times of each word in the word list in the image, the image can be

represented as a k-dimensional numerical vector.

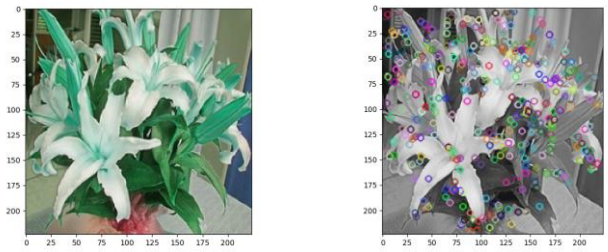


Fig.2. SIFT feature

GLCM, is a common method to describe texture by means of spatial correlation of grayscale. Haralick et al. extracted 14 eigenvalues from the grayscale symbiosis matrix, among which 4 eigenvalues were unrelated, which were easy to calculate and had high classification accuracy. Energy is the sum of squares of each element value in the grayscale co-occurrence matrix, which reflects the texture thickness and grayscale uniformity of the image. Entropy is a random measure of image information, which reflects the disorder degree of image texture gray distribution. Contrast describes the distribution of the metric matrix value and the number of local changes in the image, which reflects the depth of the groove of the image texture and the clarity of the image. Correlation is used to measure the similarity between row or column elements in a matrix.

LBP, Local Binary Pattern, is an operator used to describe the Local texture features of an image, which has significant advantages such as rotation invariability and gray invariability. The formula of LBP characteristic operator is as follows :

$$LBP_{P,R} = \sum_{P=0}^{P-1} s(g_p - g_c)2^P \ , \ s(x) = \begin{Bmatrix} 1, & x > 0 \\ 0, & x < 0 \end{Bmatrix} \tag{3.1}$$

Where g_c is the gray value of the central pixel and g_p is the gray value of its neighborhood pixel. By comparing the size of the neighborhood pixel and the center pixel around the center pixel, which is greater than the center pixel and is equal to 1, and less than the center pixel and is equal to 0, binary coding is finally carried out in a certain order.For the extraction of LBP features, two methods are adopted. The first method is to use the LBP feature extraction **shown in Figure 3** interface provided by Scikit-Image, and the second method is to use the self-implemented combination of LBP operator and dimension reduction function to extract LBP features. In the aspect of controlling LBP feature dimension, due to the different sizes of images in the training set, a function getLBPH unified training set LBP feature size which can dynamically change the operator size of LBP histogram according to the size of each image is designed.

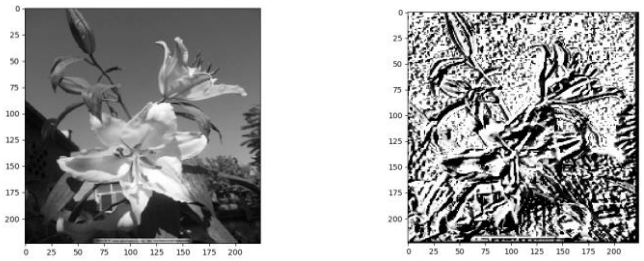


Fig.3. LBP feature

HSV features and HSV features based on image segmentation. Swain and Ballard first proposed the color histogram as the representation method of color features. After color quantization, the color histogram can be obtained by counting the number of pixels falling in each color interval. The specific calculation formula is as follows:

$$H(K) = n_k / N(\ k = 0, 1, \cdots, L - 1) \tag{3.2}$$

L is the number of quantized color intervals, nk is the number of pixels falling in the color interval K, and N is the total number of pixels. Due to the particularity of flower images, the background color of most flower images is mainly green. Therefore, if the color histogram is directly extracted, these large-area similar background colors will weaken the ability of distinguishing the colors of flowers themselves. In order to avoid the uncertain influence of segmentation effect, the significance graph was used to calculate the weighted color histogram, and the significance value of pixels was used to replace the number of pixels. In this way, background information was not completely discarded, while the weight of background information was reduced, which provided better robustness. The specific calculation formula is as follows:

$$H(K) = \sum SM(K) / \sum_{i=0}^{L-1} SM(i)(k = 0,1,..L - 1) \tag{3.3}$$

HOG, Histogram of Oriented Gradient, a feature descriptor used in computer vision and image processing for object detection. The working principle of HOG algorithm is to create the histogram of gradient direction distribution in the image and then normalize it in a very special way. This special normalization allows A HOG to effectively detect the edges of an object, even when the contrast is very low. With the strong recognition and classification capabilities of SVM and HOG feature combined in pedestrian detection and face recognition **shown in Figure 4**, we intend to try the recognition effect of feature under SVM classifier.

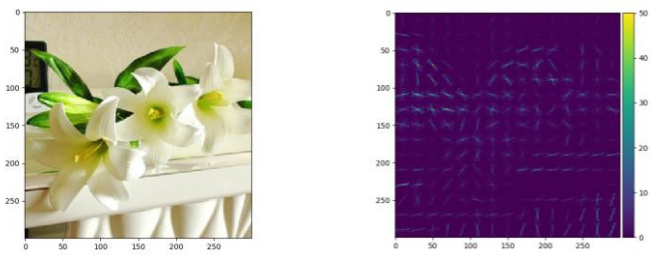


Fig.4. HOG feature

After extracting the features, we directly splicing them into an array and normalize the data, then using PCA method for dimension reduction. The normalization formula is as follows :

$$Y=(x-min)/(max-min+0.00000001) \quad (3.4)$$

Where min is the minimum value of data, max is the maximum value of data, x is the original data, and Y is the normalized data. PCA projects data into a low-dimensional subspace to achieve dimension reduction **shown in Figure 5**. For example, dimensionality reduction of a two-dimensional data set is the projection of points into a line, and each sample of the data set can be represented by one value instead of two. A three-dimensional data set can be reduced to two dimensions, which maps variables into a plane. In general, nn dimensional data set can be mapped down to kk dimensional subspace, where k ≤n.

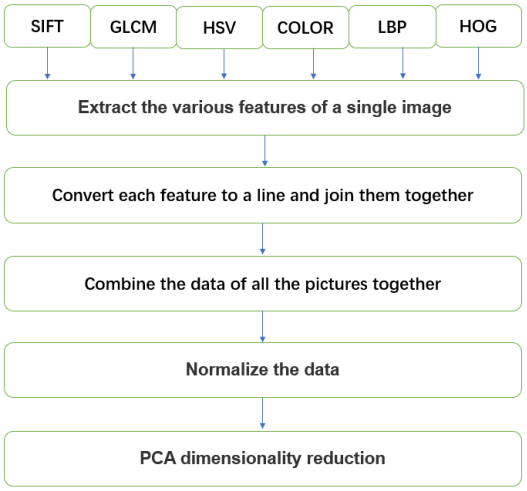



Fig.5. feature extraction process

### 3.3 The SVM classifier and optimization

SVM (Support-vector Machine) has been one of the most popular classical machine learning models for recent years and has often been used as a baseline in many tasks of computer vision and other area. A support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection.

Theoretically, one single SVM model mentioned can only be adapted to binary classification problem, yet multiple SVM models combined by OVR (one versus rest) or OVO (one versus one) strategy, which commonly referred to SVC, can easily handle multi-category classification task. As is observed, classification task on the self-built dataset described in 3.1 can be considered as 2 subtasks in order: Firstly predict whether the photograph belongs to flower or one of the other 20 categories; Secondly if it's the former, predict which kind of flower it is otherwise output the result directly. The cascading SVM model described in the present paper consists of two independent SVC models, and they were trained separately on different dataset.

**Coarse-svm** the model applied to predict whether the photograph belongs to

flower or one of the other 20 categories was denoted by *coarse-svm*. It's trained on 21-category dataset contains 20 kinds of non-flower-type objects and a new category named *flower*, which is formed by 400 pictures subsampled (identically distributed) from all kinds of flower photographs. This model is applied to handle the first subtask in cascading model described above.

**Grained-svm** the model applied to predict which kind of flower the photograph should be was denoted by *grained -svm*. It's trained on the dataset contains 20 variety of flowers photo graphs. This model is applied to handle the second subtask (is coarse-svm predicts the photograph to be flower) in cascading strategy described above.

**Cascade** To output correct label on complete category dataset, a dictionary mapping from labels of 2 subsets is necessary. Since the dictionary is built before splitting the training dataset, we can cascade two SVC model to predict correct label (or index) on the complete category datasets.

## 3.3 Image classification and recognition based on improved ResNet

### 3.3.1 Data preprocessing

Since the size of photographs in self-built dataset is not very inconsistence, resizing and cropping is applied to the dataset, along with normalization. Also, data argument such as translation and rotation can be applied to extend the dataset

### 3.3.2 ResNet with split-attention mechanism

The improved model based on ResNet contains two key parts: 1. divide feature map into several sub groups 2. apply split attention across these groups. These two parts build up for the basic block in the model.

**Feature map dividing** The feature maps are divided into R mini-groups and again each mini-block is separated to K micro-groups, hence the total number of groups is G=KR, also each group is denoted as U$i,\ i \in \{1,2,\dots G\}$.

**Split Attention** Element-wise summation across multiple splits is applied to each mini-groups and produces a combined representation of transformed features according to **equation 3.5** as follows:

$$\hat{U}^k = \sum_{j=R(k-1)+1}^{Rk} U_j,$$
$$\hat{U}^k \in \mathbb{R}^{H\times W\times C/K} \text{ for } k \in 1,2,\dots K, \quad (3.5)$$

Global contextual information can be gathered with global average pooling across multiple dimensions .The c component is calculated according to **equation 3.6**:

$$s_c^k = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\hat{U}_c^k(i,j). \quad s^k \in \mathbb{R}^{C/K} \quad (3.6)$$

Then a channel-wise soft attention is applied to the feature to produce a weighted fusion of mini-group representation. The representations of mini-groups are simply concatenated along channel dimension, the c channel is calculated as **equation 3.7**, where $a$ is given by **inequation 3.8**:

$$V_c^k = \sum_{i=1}^{R} a_i^k(c) U_{R(k-1)+i}, \tag{3.7}$$

$$a_i^k(c) = \begin{cases} \frac{exp(\mathcal{G}_i^c(s^k))}{\sum_{j=0}^{R} exp(\mathcal{G}_j^c(s^k))} & \text{if } R > 1, \\ \frac{1}{1+exp(-\mathcal{G}_i^c(s^k))} & \text{if } R = 1, \end{cases} \tag{3.8}$$

The block architecture (bottleneck) is depicted as below **shown in Figure 6**:

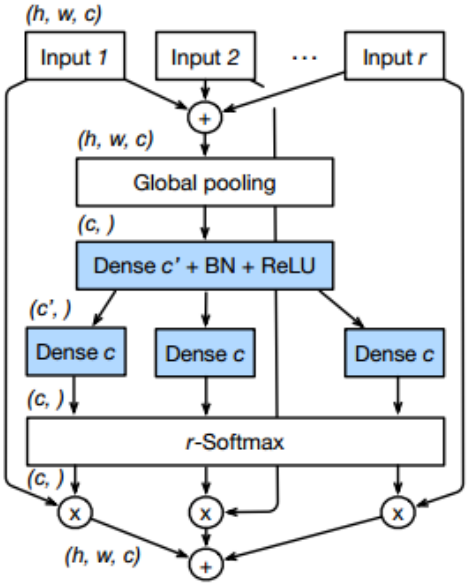



Fig.6. The block architecture of the network

**Residual connection** Similar to standard residual blocks, the block described in the present paper outputs a final representation using a shortcut connection $Y = V + X$.

**Squeeze and extract** As in SE-net, the model squeeze the two-dim feature channel into one scalar which further captures the global distribution of channel wise responses, then a gate mechanism is applied to produce channel wise weights (so-called extraction). The model finally reweight feature maps back to origin dimension and send features into next block.

**Output Layer** At the bottom of the model, two fully connected layers with ReLU and softmax activation function are added to output the final predictive probability of each label, the final output dimension is set to 40 which is the number of categories in the task.

# 4 Experiment

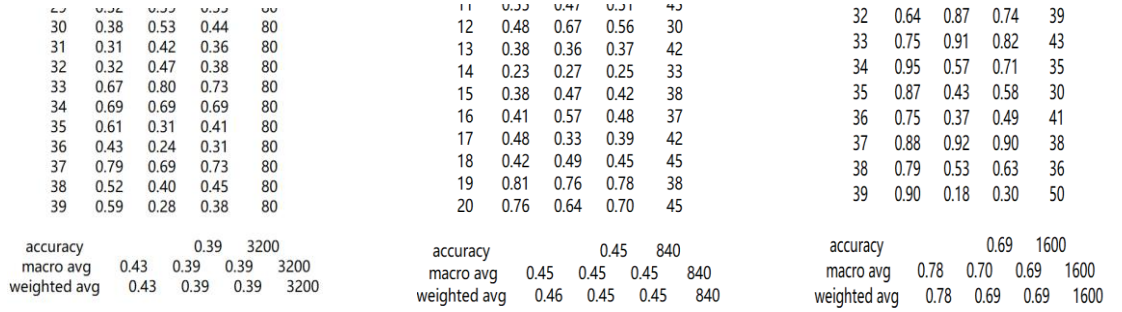
## 4.1 Experimental setup and results based on cascading SVM

A picture can extract a variety of features, and there are many combinations of these features **shown in Table II**. We tried all the possible feature combinations to improve the final accuracy **shown in Figure 7**. We splice the picture features into an array, and weight each feature when splicing, so as to avoid that the dimension of one feature is too large to affect the effect of other features. The features of all pictures are

put together to form a two-dimensional array, and the extracted features are normalized and then reduced by PCA. In dimensionality reduction, to avoid excessive data loss and improve the efficiency of SVM training, we reduce the dimensionality to half of the original one, which means that the accuracy of the trained model will be higher.

Table II characteristics of the combination

| IDX | feature | Before PCA | After PCA | Accuracy |
|---|---|---|---|---|
| 1 | HOG + COLOR + HSV | 1280 | 600 | 0.38 |
| 2 | COLOR + GLCM + HSV | 1034 | 500 | 0.39 |
| 3 | HOG + HSV + LBP + COLOR | 5056 | 2000 | 0.42 |
| 4 | GLCM + HOG + HSV + LBP | 4554 | 2000 | 0.42 |
| 5 | GLCM + COLOR + HOG + LBP | 4554 | 2000 | 0.45 |
| 6 | GLCM + HSV + COLOR + HOG | 1290 | 600 | 0.48 |
| 7 | GLCM + HSV + COLOR + LBP | 4810 | 2000 | 0.69 |
| 8 | SIFT + HSV + GLCM + LBP | 5066 | 2000 | 0.52 |




(a) IDX 2 (b) IDX 5 (c) IDX 7

Fig.7. training results

## 4.2 Experimental setup and results based on improved ResNet

The model contains 8 improved ResBlock and total number of pamaeters is about 26M. Adam algorithm is applied as optimizer with learning rate 0.001, cross-entropy as loss function and batch-size set to 32. It is trained on complete category datasets with 2 GPU device (1080ti) for around 6 hours. The model correctly classifies 3489 samples in total 3900 test photographs and 93 samples in 100 online test photos. And as **shown in Figure 8**, after 20 epochs of training, the model converges.

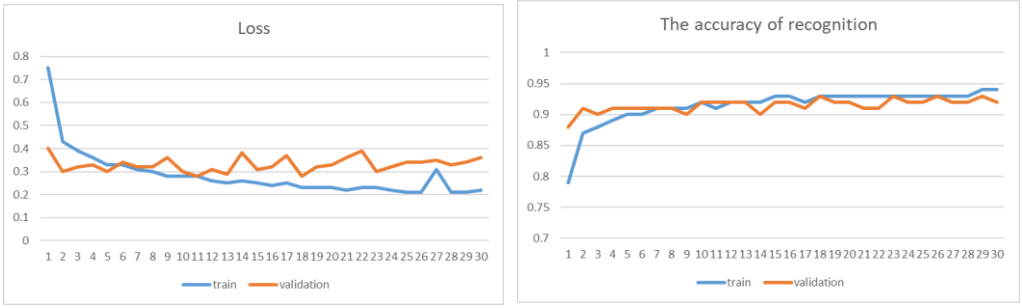



Fig.8. Training curve of CE-loss and accuracy on both training and validation datasets.

## 4.3 Analysis of ML method and DL method

In the case of using the same data set, the traditional ML method of training speed faster, but the recognition accuracy is hard to continue to ascend, the SVM model for binary classification problems of treatment effect is better than the effect of multiple classification problems. for DL methods, to build simple residual network, identification accuracy can reach 70%, through optimizing the network structure and parameters adjustment to better improve accuracy, but consume computing resources a lot and the over-fitting phenomenon occurs easily when the number of network layers is too deep.

# 5 Conclusion

In this paper, the traditional ML method and DL method based on CNN are respectively adopted to classify the 20,000 data sets created by ourselves in terms of image classification and recognition tasks. In the part of ML, we enhanced the significance of the data and extracted six different features for multiple combinations. Finally, we combined the cascading SVM model for classification and recognition. In terms of DL, we designed an improved ResNet for image classification based on the idea of residual and the inspiration of ResNet network structure according to the characteristics of our own data set. In the experiment, the recognition accuracy of the cascade SVM classifier can reach 69% at best, and the highest accuracy of the improved ResNet model on the test set reaches 93%, which verifies the effectiveness of these different method. Finally, we analyze the advantages and disadvantages of ML methods and DL methods, and conclude that traditional ML methods are tedious in process but require relatively small amount of computation. However, DL methods can achieve better accuracy but require more computing power and data volume.

# 6 Future Work

In the field of image classification and recognition, traditional ML methods can be further improved for feature engineering. For different hair objects, features that can express their own characteristics are not the same. For flowers, SIFT features and GLCM features may have better effects, while for other objects such as sculptures, HOG feature descriptor may have better effects. The weight of each feature can be

improved according to different objects so as to achieve adaptive feature extraction. In the deep learning part, the data set can be expanded by means of data argument, so as to achieve higher recognition accuracy. The introduction of YOLOv3 model and the method of random feature extraction can better process images to achieve better results.

## References


[1]Gonzalez，Rafael C，Woods，Richard E． Digital image processing［M］Digital Image Processing． 2010．

[2] Otsu N. A Threshold Selection Method from Gray-Level Histograms[J]. IEEE Transactions on Systems, Man and Cybernetics. 1979 ,9（1）: 62-66.

[3] Macqueen J . Some Methods for Classification and Analysis of Multi Variate Observations[C]. In Proc of Berkeley Symposium on Mathematical Statistics &Probability. 1965.1965:281–297.

[4] Vincent L, Soille P. Watersheds in digital space: An efficient algorithms based onimmersion simulation Watersheds in digital space: An efficient algorithms based onimmersion simulation[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1991, 13(6) : 583-598

[5] Frucci M., di Baja G.S, A new algorithm for image segmentation via watershed transformation. [C] International Conference on Image Analysis & Processing ,Racenna, Italy ， 2011 ： 168-177.

[6] Wu SG,Bao FS,Xu EY. A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network[C]. IEEE International Symposium on Signal Processing and Information Technology.2007:11-16.

[7] Kolivand H. Fern, BM.Saba T. A New Leaf Venation Detection Technique for Plant Species Classification. Arabian Journal for Science and Engineering[J].2018： 1-13

[8] Zheng Yuan-pan, Li Guang-yang, Li Ye. Review of the application of deep learning in image recognition [J]. Computer Engineering and Application,2019,55(12):20-36.

[9] Alex Krizhevsky，I Sutskever，G Hinton. Image Net Classification with Deep Convolutional Neural Networks[J], in Neural Information Pocessing Systems.2012.25(2): 1097–1105

[10] Simonyan K， Zisserman A. Very Deep convolutional Networks For Large-Scale Image Recognition[J]. Computer Science.2014.

[11] He KM, Zhang Xy, Ren XQ, and J. Sun, Deep residual learning for image recognition[C]. The IEEE conference on computer vision and pattern recognition,2016.770-778.

[12] Sun X, Qian H, Deng Zl. Chinese Herbal Medicine Image Recognition and Retrieval by Convolutional Neural Network. PLo S ONE 11(6):e0156327-.

[13] Dyrmann M , Karstoft Hn , Midtiby H S . Plant species classification using deep convolutional neural network[J]. Biosystems Engineering.2016,151:72-80.

[14] Lin CW , Ding Q , Tu W. et al. Fourier Dense Network to conduct plan tclassification using UAV-based optical images[J]. IEEE Access,2016.4

[15] Jose CR , Herve G , Pierre B. et al. Going deeper in the automated identification of Herbarium specimens[J]. BMC Evolutionary Biology.2017,17(1):181-194

[16] Lee SH, Chan CS, Mayo SJ. et al. How deep learning extracts and learns leaf features for plant classification[J] .Pattern Recognition. 2017.vol. 71, pp: 1–13.

[17] N. Bruce and J. Tsotsos, “Saliency, attention, and visual search: An information theoretic approach,” Journal of Vision, vol. 9, no. 3, pp. 5:1–24, 2009.