# Research on the Fine-grained Plant Image Classification

Zhifeng Hu[1, a], Yin Zhang* [1, b], Liang Tan[1, c]

[1]Department of Computer Science and Technology, Zhejiang University, Hangzhou, 310027, China

[a]email: huyangc@zju.edu.cn, [b]email: yinzh@zju.edu.cn, [c]email: tanliang519@163.com

* Corresponding Author

**Keywords:** fine-grained classification, convolutional neural network, SIFT, bag of word

**Abstract.** The similarity between different subcategories and scarce training data due to the difficulties of Fine-grained recognition. Even in the same subcategories, there can be some differences due to the distinct color and pose of objects. We propose some models for fine-grained plant recognition by taking advantage of deep Convolutional Neural Network (CNN) and traditional feature based methods including SIFT [1], Bag of Word (BoW) [2]. We evaluate our method on Oxford 102 Flowers dataset [3], our results show that the CNN method achieves higher accuracy than the traditional feature based methods. Our results demonstrates state-of-the-art performances on the Oxford 102 Flowers with 88.40% (Acc.).

## Introduction

Object recognition is one of the major focuses of research in computer vision. Most of existing recognition tasks are on basic-level: distinguishing between table, human, computer, car and so on. Categories differ greatly from each other on basic-level recognition. On the contrary, fine-grained recognition concentrates on differences between subcategory (breeds, species or product models), for example, recognition of different species of birds or species of flowers, which means similarities existing across categories and subtle differences needed to be found.

Scale-invariant feature transform (SIFT) is an algorithm for local features detection and description. SIFT and its variants are frequently used in image matching and image retrieval to extract features. Since Sivic et al. [2] introduced the BoW method from natural language processing to computer vision and achieved great success on many public datasets, including 15-Scenes [4], Caltech-256 [5], PASCAL VOC [6] etc.

CNN first was popularized by LeCun [7] to use in digit recognition, but fell out of fashion because of the requirement for strong computing power and large amounts of training data. With the development of parallel computing and the construction of large image databases, CNN goes to front stage again and achieves high success in many computer vision tasks. For instance, Krizhevsky et al. [8] achieved an impressive result using a CNN in ILSVRC2012 [9] with two GPUs to accelerate the computation of CNN parameters. Inspired by Krizhevsky et al., many groups proposed CNN architectures to solve the classification problems. In order to get a better performance, many CNNs ([10] [11] [12]) are first pre-trained on a large image set, ImageNet [9] for example, followed by domain-specific fine-tuning. Girshick et al. [10] proposed a model applied CNN to bottom-up region proposals and generalized the CNN classification results on ImageNet to Pascal VOC. N Zhang et al. [12] fine-tuned the ImageNet pre-trained CNN for the 200-way bird classify using the ground truth bounding box crops of the original images.

In recent years, a variety of methods about find-grained classification have been proposed. We divide these methods into two parts. One is traditional feature based methods, usually using some methods to extract hand-made features and then using a classifier for classification. Another is CNN based methods, usually using a deep convolutional neural network to extract features and obtain the classification result automatically.

In this paper, we propose our methods in both traditional hand-made features based and CNN. We combine the SIFT and BoW for image classification. Then we use CNN for image classification to compare to the method mentioned before. Our results show that CNN method can achieve higher

accuracy than the traditional method.

**Model for multi-features fusion with BoW**

We first use SIFT and its variants to extract the local features, we densely sample from the image with stride 4. We sample the image block size 8x8, 16x16, 24x24 and 32x32 each time. We conduct our experiment to find out the block size influence to the results. After local feature extraction, we use k-means [13] to generate the vision dictionary. Considering the influence of dictionary's size to the performance and accuracy of our model, we conduct our experiment on different dictionary sizes: 256, 512, 1200 and 2000. The purpose of encoding is to use the visual word in vision dictionary to represent the image local features. Encoding methods can be grouped into two categories according the differences of expressions: 1) local features are expressed as a linear combination of the visual words, 2) records of the differences between local features and the visual words. First category including: quantization coding [2], soft quantization coding [14], sparse coding [15], locality-constrained linear coding [16], etc. The second category including: Fisher vector coding [17] and super vector coding [18]. We choose the locality-constrained linear coding for a best result in our experiments. We use max-pooling to change the local feature coding vector to the whole image feature coding vector.

$$q_i = \max\{q_{1i}, ..., q_{Ni}\} \quad i = 1, ..., \mathrm{K}$$

$q_1, ..., q_N$ means the feature vectors of image feature $x_1, ..., x_N$, q means the image feature vector after pooling. After pooling, we use the feature vector to represent the image, then we use linear-kernel SVM for classification.

Table 1 shows the specific method used in our experiment during each stage.

Table 1. Specific method used in experiment during each stage.

| | |
|---|---|
| Feature Extraction | SIFT + Opponent-SIFT |
| Image Block Size | 16x16 |
| Stride of Dense Sample | 4 |
| Size of Vision Dictionary | 2000 |
| Method to Generate Vision Dictionary | k-means |
| Encoding Method | locality-constrained linear coding |
| Pooling Method | max-pooling |
| SPM [19] Layer | 1 |
| SVM Kernel | linear |

**Model for Convolutional Neural Network**

Krizhevsky et al. demonstrated the excellent result of CNN in ILSVRC2012, after that, many groups proposed CNN architectures to solve the classification problems. In order to settle the lack of training data, many of them will first trained the CNN on a large dataset and then fine-tune the CNN to the specific domain. We follow the procedure of first pre-training on ILSVRC dataset and then fine tuning to the Oxford 102 flowers dataset. We name the CNN architecture of Krizhevsky et al. AlexNet.

The architecture of AlexNet shows in Fig 2. Which contains 5 convolutional layers and 3 full connected layer. We first resize the image to 256x256 and then crop to 224x224. In order for translation invariance. We then flip the image for rotation invariance. Each layer's parameters show in table 2.

Besides AlexNet, we use GoogLeNet [20] for experiment as well. GoogLeNet is the best performance architecture in ILSVRC 2014. We first pre-train on ILSVRC 2014 and find-tune to Oxford 102 flowers. We keep the parameters of GoogLeNet as the same as Christian et al. except the output change to 102.

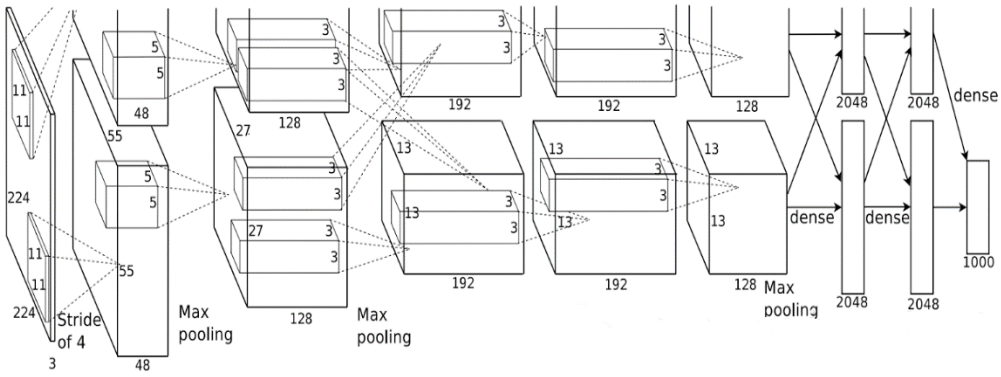


Fig 2. AlexNet architecture

Table 2. Parameters of AlexNet. Kernels means convolutions kernels, 11x11x3 means the size of each kernels.

| | |
|---|---|
| Layer 1 | 96 kernels: 11x11x3 stride: 4 |
| Layer 2 | 256 kernels: 5x5x48 stride: 1 |
| Layer 3 | 384 kernels: 3x3x256 stride: 1 |
| Layer 4 | 384 kernels: 3x3x192 stride: 1 |
| Layer 5 | 256 kernels: 3x3x192 stride: 1 |
| Layer 6 | full connected: 4096 neurons |
| Layer 7 | full connected: 4096 neurons |
| Layer 8 | full connected: 102 neurons |

Inspired by Girshick et al. [10], we consider that we can make use of the specific parts in image other than the whole image for a more classification. So we follow the Grishick et al. to introduce the region proposal to fine-grained plant classification. We first use selective search for region proposal and fine tune the CNN. Here, we use AlexNet and GoogLeNet for experiment.

## Test results

We show our result in table 3. In our results, we find out that the model of CNN obtains a better result than the traditional hand-make feature method. And the multi-feature fusion's result is better than one feature. However, the region with CNN methods achieve highest accuracy in our experiments, it is because the fine-grained plant classification's characteristic. Fine-grained classification forces us to focus on specific part of image such as leaf, flower and stem other than the whole image. So after region proposal, we recognize the image with the help of all regions, which augment the training set and get a best result in the end.

Table 3. Experiment results.

| Method | SPM | Accuracy |
|---|---|---|
| SIFT | 1 | 54.3% |
| SIFT | 2 | 61.9% |
| WSIFT | 1 | 56.8% |
| WSIFT | 2 | 65.8% |
| Opponent-SIFT | 1 | 68.7% |
| Opponent-SIFT | 2 | 70.2% |
| SIFT+WSIFT+OpponentSIFT | 1 | 71.7% |
| SIFT+WSIFT+OpponentSIFT | 2 | 72.2% |
| AlexNet without region proposal | / | 83.35% |
| GoogLeNet without region proposal | / | 75.85% |
| AlexNet with region proposal | / | 86.65% |
| **GoogLeNet with region proposal** | **/** | **88.40%** |

## Conclusion

We compare two kinds of different image recognition methods, one is hand-crafted feature methods, and another is CNN methods. Our result shows that the CNN methods can achieve higher accuracy than the other. In the meantime, our method utilizes the bottom-up region proposals and pre-trained CNN to boost the accuracy of fine-grained recognition. Our experiments show that it is highly beneficial to force CNN to focus on significant parts of object other than the whole image by bottom-up region proposals. In future extensions of this work, we will consider using CNN features of proposals to classify the object other than the CNN classify results. We also plan to investigate the influence of features from different CNN layer, because we think that we should always focus on significant subtle parts rather than the whole image in fine-grained recognition. Finally, we will explore the way to accelerate the recognition procedure to obtain both accuracy and speed.

## Acknowledgement

This work is supported by Zhejiang Provincial Natural Science Foundation of China (No.LY14F020027) and China Knowledge Centre for Engineering Sciences and Technology (No. CKCEST-2016-1-14).

## References

[1] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.

[2] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos[C]//Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE, 2003: 1470-1477.

[3] Nilsback M E, Zisserman A. Automated flower classification over a large number of classes[C]//Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on. IEEE, 2008: 722-729.

[4] http://www.cs.unc.edu/~lazebnik/research/scene-categories.zip/, 2006.

[5] http://www.vision.caltech.edu/Image-DataSets/Caltech256/, 2013.

[6] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88(2): 303-338.

[7] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[8] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[9] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[10] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[11] Sharif Razavian A, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: an astounding baseline for recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014: 806-813.

[12] Zhang N, Donahue J, Girshick R, et al. Part-based r-cnns for fine-grained category detection[C]//European Conference on Computer Vision. Springer International Publishing, 2014:

834-849.

[13] Hartigan J A, Wong M A. Algorithm AS 136: A k-means clustering algorithm[J]. Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979, 28(1): 100-108.

[14] Van Gemert J C, Geusebroek J M, Veenman C J, et al. Kernel codebooks for scene categorization[M]//Computer Vision–ECCV 2008. Springer Berlin Heidelberg, 2008: 696-709.

[15] Yang J, Yu K, Gong Y, et al. Linear spatial pyramid matching using sparse coding for image classification[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 1794-1801.

[16] Wang J, Yang J, Yu K, et al. Locality-constrained linear coding for image classification[C]//Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 3360-3367.

[17] Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification[M]//Computer Vision–ECCV 2010. Springer Berlin Heidelberg, 2010: 143-156.

[18] Zhou X, Yu K, Zhang T, et al. Image classification using super-vector coding of local image descriptors[M]//Computer Vision–ECCV 2010. Springer Berlin Heidelberg, 2010: 141-154.

[19] Lazebnik S, Schmid C, Ponce J. Spatial pyramid matching[J]. Object Categorization: Computer and Human Vision Perspectives, 2009, 3(4).

[20] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.