FINE-GRAINED PLANT CULTIVAR RETRIEVAL VIA TWO-BRANCH SECOND-ORDER POOLING-BASED FEATURE EXTRACTION AND FUSION

/

基于双分支二阶池化特征提取与融合的细粒度植物育种检索

Pengrui XI [1)], Jie WANG [1)], Shu FENG [*2)]

[1)] College of Software, Shanxi Agricultural University, Taigu, Shanxi / China

[2)] College of Basic Sciences, Shanxi Agricultural University, Taigu, Shanxi / China

Corresponding author: Shu Feng; Tel: +86-15903440196; E-mail: fengshu@sxau.edu.cn

DOI: https://doi.org/10.35633/inmateh-78-59

Keywords: *Fine-grained Plant Cultivar Retrieval, CNN; Second-order Pooling, Feature Fusion*

ABSTRACT

The highly similar visual appearance among different cultivated plant species makes fine-grained plant cultivar retrieval a challenging task. Considerable efforts have been devoted to this problem, and significant progress has been achieved in recent decades. This paper proposes a simple and effective method for fine-grained plant cultivar retrieval. The main contributions are threefold. First, experimental analysis indicates that image resolution plays a crucial role in fine-grained plant retrieval, with 896×896 pixels representing the most cost-effective resolution. Second, a radial basis kernel function is employed to capture the nonlinear channel correlation of the feature map, enabling the extraction of more discriminative features. In addition, Log-TiedRank is applied to improve robustness to noise and to obtain a more compact representation. Finally, two types of deep features extracted from two convolutional neural networks are fused to further enhance retrieval performance. Compared with state-of-the-art methods, the proposed approach improves the retrieval rate by 15.71%, 15.95%, 14.02%, and 8% on the SoyCultivar200 dataset, 4.68% on PeanCultivar100, and 4.01% on the Mulberry dataset, demonstrating the effectiveness and superiority of the proposed method.

摘要

不同栽培品种植物在视觉外观上的高度相似性，使得精细植物栽培品种检索成为一项具有挑战性的任务。过去几十年中，研究者们投入了大量努力并取得了显著进展。本文提出了一种简洁有效的精细植物栽培品种检索方法，主要贡献体现在三个方面。首先，通过实验发现图像分辨率在精细植物检索中起着至关重要的作用，而 896×896 像素可能是最具成本效益的分辨率选择。其次，我们采用径向基核函数捕捉特征图的非线性通道关联信息，旨在提取更具判别力的特征。同时引入 Log-TiedRank 方法提升紧凑表征对噪声的鲁棒性。最后，通过融合两种卷积网络提取的深度特征进一步优化检索性能。实验结果表明，本方法在 SoyCultivar200 数据集上检索率较现有最优方法提升 15.71%、15.95%、14.02%和 8%，在 PeanCultivar100 和 Mulberry 数据集上分别提升了 4.68%和 4.01%，充分证明了该方法的优越性与有效性。

INTRODUCTION

Thousands of plants, such as soybean, peanut, apple tree and strawberry, play a significant role in our daily life, because they can provide human beings with sufficient food, protein, vitamins, medicine and industrial raw materials. Therefore, accurately recognizing different plant species is very important. Moreover, plant breeding is a key and vital technology in modern agriculture, with the goal of increasing production, improving quality, protecting the ecological environment, strengthening disease resistance (*Yang et al., 2023*) and enriching food nutrition. Evidently, different plant breeders of the same species have different effects and properties. However, their visual appearance, shape, and texture are highly similar. Even the human visual system and agricultural experts have difficulty in distinguishing the subtle differences between various plant cultivars, so the fine-grained plant cultivar retrieval and recognition (*Wang et al., 2022*) is an active and challenging research topic in precision agriculture. Over the past two decades considerable works have been put forward to increase conventional and fine-grained plant image recognition accuracy, they can be categorized into handcrafted features and deep learning feature based methods.

Handcrafted Features: Plant leaves contain rich texture information, so texture plays a key role in plant recognition. The LBP is a classical texture descriptor, since the co-occurrence pattern counts for the frequency of any two LBP patterns, so its descriptive power is stronger than the original LBP.

The rotation invariant co-occurrence among adjacent LBPs (RICLBP) (*Nosaka et al., 2013*) and pairwise rotation invariant co-occurrence LBP (PRICoLBP) (*Qi et al., 2014*) are two renowned image descriptors. *Chen et al., (2022)* proposed the informative SBT descriptor to learn the multiple scale co-occurrence texture patterns. Moreover, a new K nearest neighbour based feature fusion fashion was developed. Recently, *Chen et al. (2023)* proposed fan-beam binarization difference projection (FB-BDP) method to describe the leaf image texture patterns, which can measure the inner structure information of objects from many orientations and suppress image noise. Shape is also an important visual cue of plant leaves. *Yang (2021)* proposed to employ multi-scale triangle descriptor (MTD) for capturing leaf shape information and local binary pattern histogram Fourier (LBP-HF) for extracting texture feature, combing such two features via weighted distance measurement can obtain excellent leaf classification and retrieval results. *Hu et al. (2012)* proposed a shape geometry descriptor via multi-scale distance matrix (MDM) with good invariance properties against rotation, symmetry, scaling, and translation. *Chen and Wang (2020)* proposed a histogram of Gaussian convolution vectors (HoGCV) for plant species retrieval, where the contour vectors are convolved with Gaussian functions under various widths. *Wang et al. (2020)* improved the sliding chord matching via multi-scale scheme (MSCM) for plant cultivar recognition, where the chord is able to capture the interior appearance of leaves and features of exterior shape. *Chen and Wang (2024)* developed a symmetry-constrained linear sliding co-occurrence LBP (SCLS-CoLBP) method with multi-scale, multi-position and multi-orientation strategies under bag-of-words framework. Although handcrafted methods get great progress in plant species retrieval rate, however their performance in plant cultivar recognition is still unpromising, and their robustness against complex background and compound leaves is also still needed to be improved.

Deep Learning Features: Due to the high-level semantic information learning ability, deep convolutional network has been proved to be more discriminative and robust against many degradations than the low-level handcrafted features. A number of deep neural network based image retrieval approaches have been proposed, representative ones are region maximum activation convolution feature (RMAC) (*Tolias et al., 2016*), cross-dimensional weighting (CROW) aggregated convolutional features (*Kalantidis et al., 2016*), first-order generalized-mean (GeM) pooling (*Radenovic et al., 2019*). Research suggests that fusing CNN features of three leaf image patterns and summing the support vector machine decisions is able to enhance soybean cultivar identification accuracy. In summary, these deep features can be concluded as first-order features. To extract higher-order features, *Feng (2022)* explored the use of a radial basis kernel function to capture second-order channel correlation information, achieving leading results in plant species identification and retrieval. In addition, combining deep features with conventional handcrafted features - such as texture, shape, and triangle features - has been considered a possible strategy for further improving the discriminative power of the extracted features. *Wu et al., (2023)* proposed the RCCF+R-MAC method that combines multi-scale first-order region maximum activation convolution features and second-order regional convolution covariance feature, which offers outstanding plant cultivar recognition performance. *Deng et al., (2019)*, proposed a lightweight and shallow convolutional network to learn face and object image features via pre-defining random-field eigen-filters and scattering network structure. *Pang et al., (2018),* proposed the selection and weighting informative deep convolutional features by replicator equation (ReSW) based on pre-trained and fine-tuned VGG16 from siaMAC (*Radenovic et al., 2016*). *Liu and Yang, (2021),* proposed the deep-seated feature histogram (DSFH) for content-based image retrieval, where the fully connected layer feature histogram (HFCLF) under pre-trained VGG16 is first extracted via ranking techniques and L2 normalization, and then whitening and normalization are performed to get the DSFH descriptor.

However, the retrieval rate for fine-grained plant cultivar retrieval remains unsatisfactory, and there is considerable room for improvement. Moreover, 224×224 has been recognized as the standard input size for neural networks. But for plant cultivar leaves from the same species, they look very similar and share many visual characteristics in common, their differences are even smaller at low resolution. To the best of current knowledge, few studies have investigated the effect of image resolution on plant cultivar recognition. To evaluate the impact of image resolution on the accuracy of fine-grained plant cultivar retrieval, extensive experiments were conducted on the PeanCultivar100 dataset (*Chen et al., 2022*) and the SoyCultivar200 dataset (*Wang et al., 2020*). Although multi-scale or multi-resolution approaches are widely recognized as indispensable in the image recognition literature, the present study focuses only on the influence of image resolution rather than on multi-resolution fusion in plant cultivar retrieval. Four image resolutions were evaluated: 224×224, 448×448, 896×896, and 1120×1120 pixels. Three feature extraction techniques were tested: first-order generalized mean (GeM), second-order pooling, and fully connected layer features. In addition, two off-the-shelf pre-trained convolutional neural networks (CNNs), ResNet and VGG, were applied.

All comparison results are presented in Fig. 1. The results indicate that plant leaf retrieval accuracy can be improved more effectively by increasing the input image resolution than by designing more complex feature learning techniques.

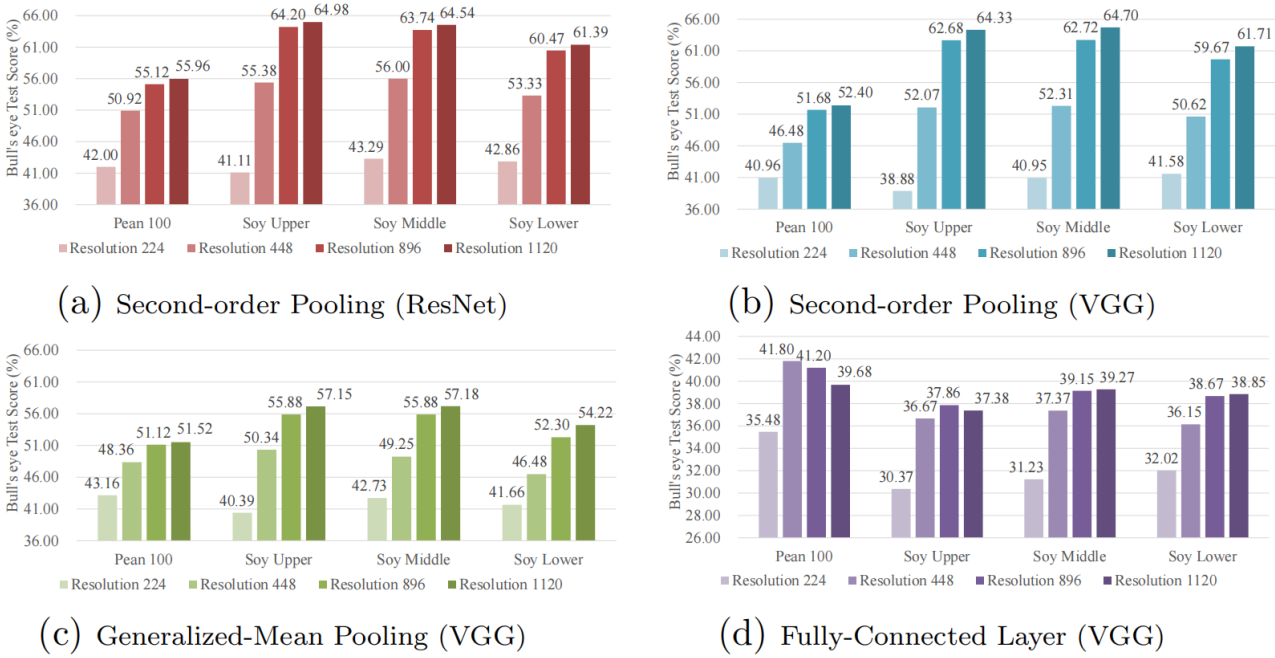


Fig. 1 – Retrieval results under different image resolutions

Motivation

It is well known that higher image resolution contains more information than lower resolution. Two plant cultivar images, 003FZ and 004FZ, from the lower part subset of the SoyCultivar200 dataset are shown in Fig. 2. The corresponding image patches exhibit similar structures at low resolution; however, when the images are enlarged, additional details become clearly visible. These observations suggest that image resolution may be a key factor in fine-grained plant cultivar recognition.

Additionally, different neural network models generally exhibit complementary characteristics. Therefore, investigating feature fusion from diverse neural networks, such as VGG and ResNet, can be considered a plausible approach to improving plant leaf retrieval performance. For simplicity, pre-trained CNN models are employed without fine-tuning the network parameters. It should be noted that the CNN architectures can be easily replaced by other models, such as pre-trained Vision Transformers (*Liu et al., 2021*).

This paper presents a simple and effective method for fine-grained plant cultivar retrieval. The main workflow is illustrated in Fig. 4. The key components include leaf image input, extraction of a specific feature map, second-order feature extraction, and computation of the similarity or distance matrix. All these components are straightforward and not entirely novel. Instead, the main contributions lie in investigating their behaviour at large input resolutions and in exploring feature fusion across different neural network architectures.

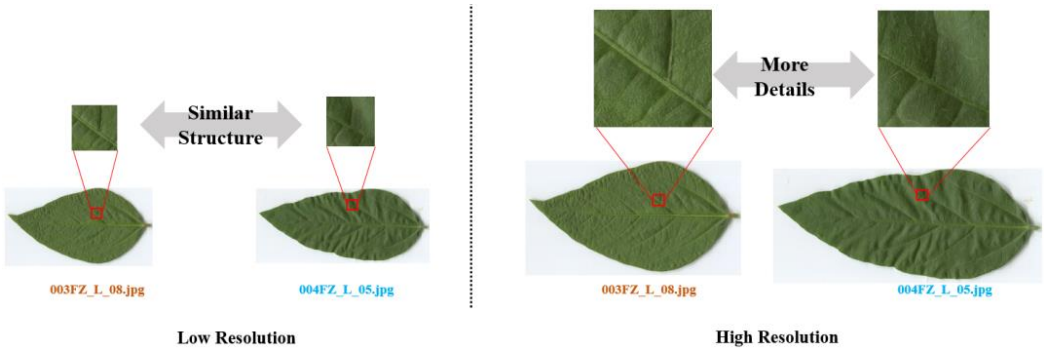


Fig. 2 – Comparison between the local image patches of two images under low- and high-resolution

Contributions

In summary, the main contributions of the proposed method can be summarized as follows:

(1) The impact of high image resolution on fine-grained plant cultivar retrieval was investigated. The results indicate that higher resolution generally leads to a significant improvement in retrieval performance. However, the retrieval rate does not continue to increase proportionally with resolution, and the growth rate gradually stabilizes. Considering both retrieval performance and computational cost, a resolution of 896×896 pixels appears to be the most suitable image size.

(2) Feature fusion from different convolutional neural networks was examined to exploit the complementary characteristics between them. A simple fusion strategy based on k-nearest neighbours (k-NN) resulted in a 3–5% improvement in retrieval rate.

(3) Extensive experiments conducted on three representative plant cultivar datasets demonstrate the effectiveness and superiority of the proposed approach compared with 21 state-of-the-art methods.

MATERIALS AND METHODS

Dataset Description

As shown in Fig. 3, three widely used plant cultivar datasets - SoyCultivar200 (*Wang et al., 2020*), PeanCultivar100 (*Chen et al., 2022*), and Mulberry (*Chompookham and Surinta, 2021*) - were used in this study. **SoyCultivar200** is the first plant cultivar dataset employed to evaluate the performance of the proposed method and other competing approaches. The dataset contains 6000 plant images from 200 soybean cultivars. In other words, each soybean cultivar is represented by 30 images captured from three plant parts: upper, middle, and lower. Each part includes 2000 images belonging to 200 cultivars. The **PeanCultivar100** dataset contains 500 plant leaf images from 100 peanut cultivars, with five leaves per cultivar. Since all cultivars belong to the same species (peanut), their leaves share similar shapes, veins, and overall appearance. For example, the leaf images of cultivars U_001 and U_010 are extremely similar. Consequently, retrieving a peanut leaf image from this dataset is also a challenging task. The **Mulberry** dataset contains images with multiple leaves at different scales and orientations. The background is complex, and the leaves often overlap, making this dataset more challenging for plant cultivar recognition. The dataset consists of 5262 images of 10 mulberry cultivars, including four from Thailand, three from Australia, two from Taiwan, and one from Turkey. All plant images were captured in natural environments using DSLR cameras and smartphones.

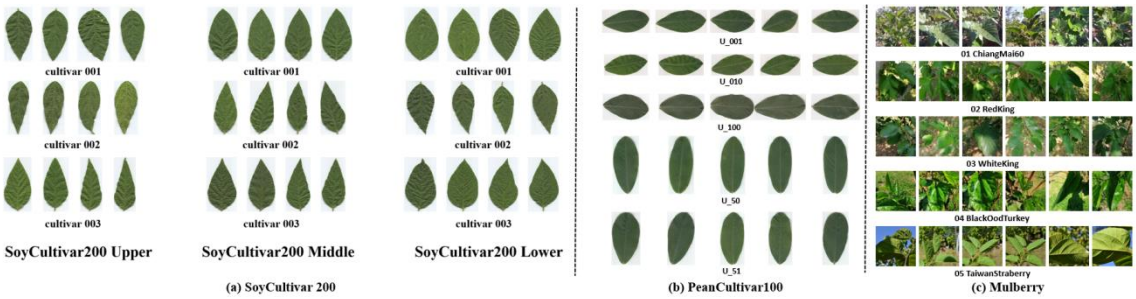


Fig. 3 - Example images for (a) SoyCultivar200, (b) PeanCultivar100 dataset, (c) Mulberry dataset

Our Method

In this section, the second-order deep feature extraction process is first described, followed by the presentation of the feature fusion strategy for features extracted from two CNN models. The overall workflow of the proposed method is illustrated in Fig. 4.

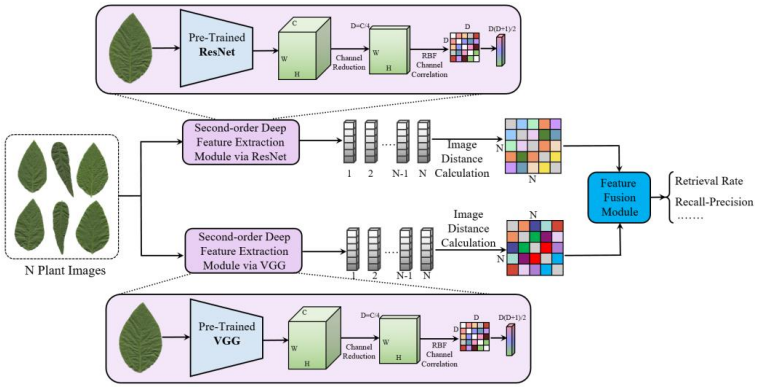


Fig. 4 - The pipeline of our proposed method

Second-order Deep Feature Extraction

Given a plant cultivar leaf image *X*, the image is fed into a pre-trained CNN model, such as VGG or ResNet. When the input image size is 224×224, the feature map at the *res4bx* layer of ResNet has a size of 14×14×1024. If the input image resolution is increased fourfold to 896×896, the feature map is also enlarged four times in the spatial dimension, resulting in a size of 56×56×1024. Consequently, more detailed spatial information, such as lines and textures, can be captured when learning the second-order channel correlation matrix. Therefore, the input image is first resized to 896×896.

Considering the potential overfitting issue when training a CNN model from scratch using only 500~6000 plant leaf images, as well as the simplicity and effectiveness of pre-trained CNN models, the ImageNet pre-trained networks ResNet and VGG are employed. Both networks are denoted as CNN models in the following formulation. Let l denote the layer index used in this study. The feature map at layer l can then be obtained as follows:

$$F = CNN(X,\ l) \in R^{H\times W\times C} \tag{1}$$

Since the number of channels is large (e.g., 1024 at the res4bx layer of ResNet), the resulting second-order pooling matrix would have a size of 1024×1024, and the feature vector dimension would reach 524,800. Therefore, the number of channels is directly reduced using an averaging operation instead of principal component analysis (PCA)-based dimensionality reduction. The averaging operation is performed along the channel dimension for each sliding tensor *F*(:, :, 4*i*-3:4*i*), where *i*=1,…,⌊C/4⌋. Consequently, the size of the feature map tensor *F* is reduced to *H*×*W*×*D*, where D=⌊C/4⌋. Afterward, second-order pooling ϕ is applied to *F* to learn the correlation information between channels, as follows:

$$P = \phi(F) \in R^{D\times D} \tag{2}$$

Since Rank(*P*)≤min(*D*, *HW*-1), a singular matrix problem may occur in *P* so when *D*>*HW*-1. Therefore, the radial basis kernel function (*Feng, 2022*), rather than the traditional covariance matrix, is employed to capture the non-linear second-order information of F. Moreover, since *P* lies on a Riemannian manifold, the logarithmic mapping function logm is applied, following *Ng et al. (2018)*, to project it back to the Euclidean tangent space at point Q, as follows:

$$S = Q^{\frac{1}{2}} \log m\left(Q^{-\frac{1}{2}} P Q^{-\frac{1}{2}}\right) Q^{\frac{1}{2}} \tag{3}$$

Since matrix P is symmetric, only the upper triangular elements and the main diagonal elements of S are retained. The feature vector *f* can be obtain with a dimensionality of *D*(*D*+1)⁄2. To obtain a compact feature representation, Log-TiedRank (*Ng et al., 2018*) is subsequently applied to f.

Feature Fusion of Different Convolutional Networks

Different network models tend to focus on diverse feature representations of plant leaf images; therefore, complementary information often exists among the features extracted by different networks. After obtaining feature representations for the *N* images in a plant cultivar dataset, a distance matrix of size *N*×*N* can be computed. Accordingly, two distance matrices, M1 and M2, are obtained for the two convolutional neural networks. The objective is to improve plant cultivar retrieval accuracy using the k-nearest neighbour (k-NN)-based feature fusion scheme proposed by *Chen et al. (2022)*. The main idea can be summarized as follows. For each position (*i*, *j*), where *i*, *j*=1,...,*N*, the following operations are performed. For each query image *i*, the first *K* smallest distances among its *N* distances are selected. For each leaf image *j*, the first 2~*K*+1 smallest distances among its *N* distances are selected. Then $M1_{ij}$ is normalized by dividing it by the mean value of the 2K smallest distances. Similar operations are applied to $M2_{ij}$. Finally, feature fusion is achieved by summing the normalized values of $M1_{ij}$ and $M2_{ij}$ at position (*i*, *j*). Experimental results indicate that combining multiple deep features yields better performance than combining handcrafted and deep features, as reported in *Chen et al. (2022)*. Therefore, different deep features extracted from ResNet and VGG are fused in the proposed approach.

RESULTS

Experiment Setup and Evaluation Measure

The retrieval measure Bull's eye score is applied. Assuming a plant dataset with N images, for each image x_i in class k with C_k samples, it is compared with all the N samples using cosine distance metric.

Then the first nearest $2C_k$ samples are selected, so the average precision can be defined: $AP_{x_i} = r/C_k$, where r is the number of samples relevant to x_i. Therefore, the mean average precision (MAP) score can be obtained: $MAP = \sum_{i=1}^{N} AP_{x_i}/N$. Moreover, precision–recall curves are also plotted, for each sample image x_i, the N distances between x_i and all the samples are calculated, for $1 \le j \le N$, $P_j = \sum_{i=1}^{N}(r_{x_i}/j)/N$, where r_{x_i} is the number of samples relevant to x_i among j retrieval results. $R_j = \sum_{i=1}^{N}(r_{x_i}/C_{x_i})/N$. After having ($R_j$, P_j), the precision-recall curve can be drawn. For simplicity, the default value K=50 as in (*Chen et al., 2022*) is used for the K nearest neighbours based feature fusion module.

To fully demonstrate the advantages of the proposed approach, the method is compared with 21 other approaches, including nine handcrafted methods: MDM (Hu et al., 2012), RICLBP (*Nosaka et al., 2013*), PRICoLBP (*Qi et al., 2014*), HoGCV (*Chen and Wang, 2020*), MSCM (*Wang et al., 2020*), MTD+LBP-HF (*Yang, 2021*), SBT (*Chen et al., 2022*), FB-BDP (*Chen et al., 2023*) and SCLS-CoLBP (*Chen and Wang, 2024*); and 12 neural network based methods: VGG (*Simonyan and Zisserman, 2015*), ResNet50 (*He et al., 2016*), DenseNet121 (*Huang et al., 2017*), RMAC (*Tolias et al., 2016*), CROW (*Kalantidis et al., 2016*), GeM (*Radenovic et al., 2019*), ReSW (*Pang et al., 2018*), siaMAC+ReSW (*Pang et al., 2018*), SCBP (*Deng et al., 2019*), HFCLF (*Liu and Yang, 2021*), DSFH (*Liu and Yang, 2021*) and RCCF+M-RMAC (*Wu et al., 2023*).

To ensure a fair comparison, the input resolution was set to 896×896 for the methods based on VGG, ResNet50, GeM, RMAC, and CROW. Following the RCCF+M-RMAC method, the feature map at layer 28 was used for Ours (VGG), GeM, RMAC, and CROW. Following the SBT method, the last fully connected layer was used as the deep feature representation for ResNet50, VGG, and DenseNet121. For the RICLBP and PRICoLBP methods, the original image size (which may reach 2000×3000) was retained in order to preserve more texture information. For SCBP, the original parameter settings were adopted. For the remaining methods, since the source codes were not available, the experimental results were directly cited from the literature (*Chen et al., 2022; Wu et al., 2023*) or from the original publications.

Impact of network layer

Fig. 5 illustrates the performance of the proposed method under different image resolutions and ReLU layers of ResNet. For a resolution of 224×224, the retrieval performance decreases as the network depth increases, even though the feature dimensionality becomes larger. For a resolution of 2240×2240, the retrieval rates are relatively low at shallow layers but improve at deeper layers. For the other resolutions, as the network depth increases, the retrieval performance initially improves slightly, then fluctuates and gradually declines. This trend suggests that intermediate-layer features play a more important role in plant image recognition. The feature dimensionalities corresponding to different stages or layers of the proposed method are 2,080, 8,256, 32,896, and 131,328, respectively. Considering the trade-off among computational cost, retrieval performance, and feature dimensionality, a resolution of 896×896 and layer 100 appear to provide the most suitable parameter configuration.

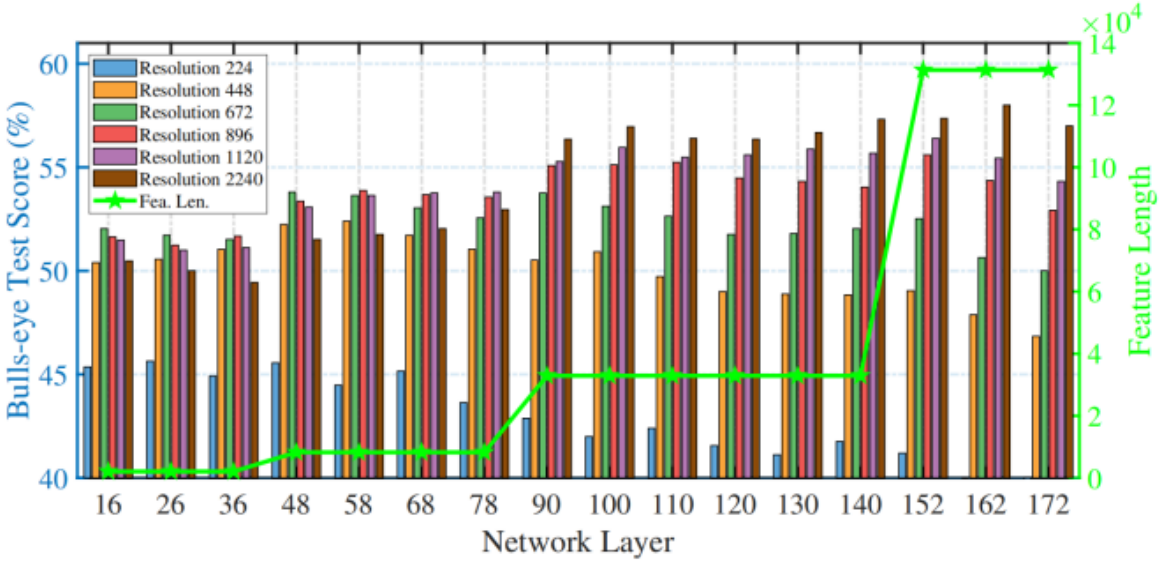


Fig. 5 - Retrieval performance under different image resolutions and ReLU activation layers

Table 1

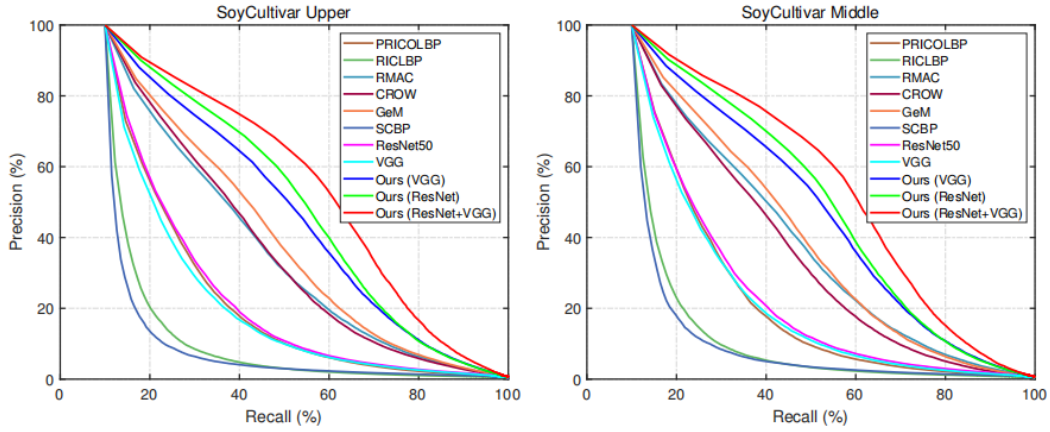
Bull's eye retrieval score (%) on SoyCultivar200 dataset

Method	Upper	Middle	Lower	Joint	Method	Upper	Middle	Lower	Joint
PRICoLBP	38.67	38.50	35.28	45.80	MTD+LBP-HF	40.79	43.13	41.51	76.65
RICLBP	25.09	26.20	25.37	39.74	HFCLF	38.35	38.50	36.13	66.14
RMAC	53.25	55.26	52.85	85.15	DSFH	41.52	41.82	40.35	69.90
CROW	53.02	52.52	48.36	79.05	ReSW	45.12	46.33	43.78	80.97
GeM	55.88	55.88	52.30	82.79	siaMAC+ReSW	47.32	48.35	44.25	78.58
SCBP	22.15	24.24	25.69	28.35	SBT	47.57	48.40	47.92	81.69
ResNet50	39.51	40.42	40.38	70.69	FB-BDP	49.12	51.76	50.59	84.20
VGG	37.86	39.15	38.67	68.13	RCCF+M-RMAC	53.71	52.36	50.47	84.42
DenseNet121	36.58	38.18	36.95	71.03	SCLS-CoLBP	50.04	51.37	51.01	83.34
MDM	22.94	24.44	25.69	43.26	Ours(VGG)	62.68	62.72	59.67	89.70
HoGCV	29.04	31.48	32.48	59.81	Ours(ResNet)	64.20	63.74	60.47	89.40
MSCM	34.07	36.62	37.40	64.53	Ours(ResNet+VGG)	**69.42**	**68.31**	**64.49**	**92.42**

Results on SoyCultivar200

The quantitative comparison results on the SoyCultivar200 dataset are presented in Table 1. The following observations can be drawn: (1) The proposed method (ResNet + VGG) achieves the highest retrieval rates compared with other competing methods. The improvements over the second-best method are 15.71%, 15.95%, 14.02%, and 8%, respectively. (2) Even when using only a single pre-trained network (VGG), the proposed method outperforms the other methods. This result indicates that a large input resolution combined with second-order pooling can extract more discriminative information. (3) Since RCCF+M-RMAC uses only covariance pooling, its performance is inferior to that of the proposed method, which employs radial basis function kernel pooling to avoid the singular matrix problem. It should also be noted that the RCCF+M-RMAC method combines four M-RMAC features extracted at resolutions of 512x512, 448x448, 336x336 and 224x224. (4) The relatively low retrieval rates of RMAC, CROW, and GeM may be attributed to the fact that these methods extract only first-order pooling features, rather than higher-order pooling features. (5) The handcrafted texture feature methods PRICoLBP, RICLBP, and MTD+LBP-HF do not achieve satisfactory results, even though feature combination and co-occurrence techniques are employed. (6) By applying a feature fusion strategy combining two pre-trained convolutional neural networks (ResNet and VGG), the retrieval performance is significantly improved, confirming the effectiveness of the feature fusion module. (7) The proposed method is the only approach that achieves a retrieval rate exceeding 90% under the joint matching protocol.

The recall–precision curves of the comparative methods on the SoyCultivar200 dataset are shown in Fig. 6. The closer a recall–precision curve is to the upper-right corner, the better the performance of the corresponding method. The blue, green, and red curves represent the proposed methods, showing superior recall–precision performance. The results of the curve comparison are consistent with the quantitative results presented in Table 1.



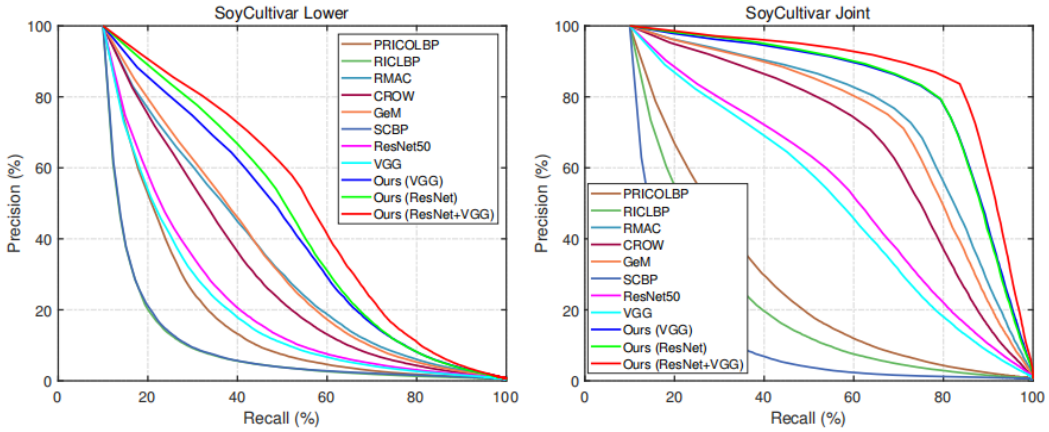


Fig. 6 - Recall-Precision Curves of the comparative methods on SoyCultivar200 dataset

Results on PeanCultivar100 and Mulberry

The comparison results on the PeanCultivar100 dataset are presented in Table 2. The following observations can be made: (1) The proposed method (ResNet) achieves the highest retrieval rate, which is 1.72% higher than that of the feature-fusion-based method RCCF+M-RMAC and 5% higher than the state-of-the-art handcrafted method SBT. (2) When a feature fusion strategy combining VGG and ResNet is applied, the retrieval rate increases to 58.08%, outperforming the state-of-the-art method RCCF+M-RMAC by a margin of 4.68%. (3) The superior performance of the proposed approach can be attributed to two main factors: (i) second-order pooling captures channel correlation information; and (ii) high-resolution input images provide more detailed visual information, such as texture and vein structures. The recall–precision curves of the comparative methods on the PeanCultivar100 dataset are shown in Fig. 7(a).

It can again be observed that the three recall–precision curves of the proposed method are closer to the upper-right corner, indicating its effectiveness in peanut cultivar retrieval and its advantage over the other methods.

The retrieval rate comparison on the Mulberry dataset is reported in Table 2. The proposed method (VGG) and the proposed method (ResNet) achieve retrieval rates of 49.58% and 49.77%, respectively, which are very close to the 50.38% achieved by the relatively complex method RCCF+M-RMAC. When features from VGG and ResNet are fused, the proposed method outperforms the state-of-the-art method RCCF+M-RMAC by a margin of 4.01%. Under the same input image resolution, the improvements of the proposed method over the original VGG and ResNet50 models are approximately 16–18%, while the gains over CROW, GeM, and RMAC are about 6–8%. This observation highlights the importance of second-order pooling in capturing informative leaf features. In contrast, the handcrafted feature methods MTD+LBP-HF, SBT, PRICoLBP, and RICLBP achieve relatively low retrieval rates. This is mainly because these methods struggle to extract useful texture features from leaf images with complex variations. The recall–precision curves of the competing methods on the Mulberry dataset are shown in Fig. 7(b). The three recall–precision curves of the proposed method are consistently closer to the upper-right corner than those of the other methods, demonstrating its robustness against complex backgrounds and leaf overlapping. In other words, the proposed method remains effective for recognizing plant cultivars with large variations, such as tangled backgrounds, viewpoint changes, and compound leaf structures.

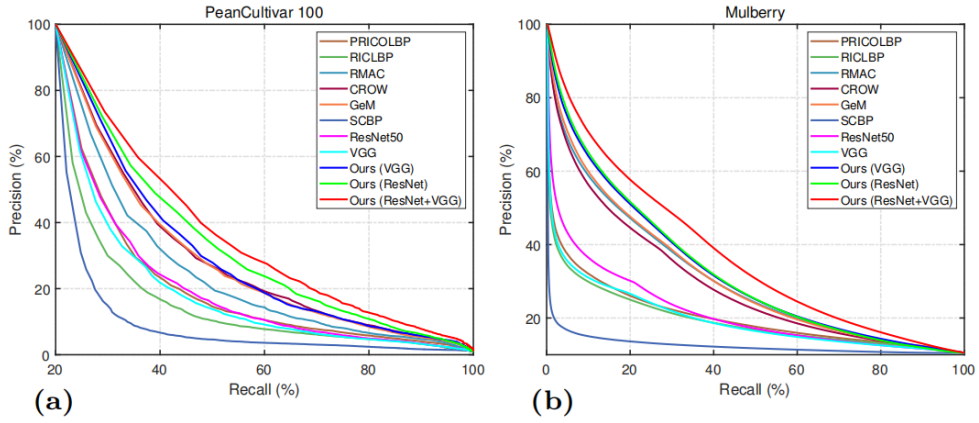


Fig. 7 - Recall-Precision Curves on (a) PeanCultivar100 dataset and (b) Mulberry dataset.

Table 2

Bull's eye retrieval score (%) on PeanCultivar100 and Mulberry datasets.

Method \ Dataset	PeanCultivar100	Mulberry	Method \ Dataset	PeanCultivar100	Mulberry
PRICoLBP	42.04	39.21	MTD+LBP-HF	44.52	34.43
RICLBP	37.60	37.93	HFCLF	39.48	38.13
RMAC	46.88	48.57	DSFH	37.84	32.66
CROW	51.40	46.55	ReSW	42.72	41.20
GeM	51.12	48.24	siaMAC+ReSW	37.80	42.58
SCBP	29.92	25.91	SBT	50.12	35.94
ResNet50	43.24	38.77	FB-BDP	51.80	-
VGG	41.20	36.81	RCCF+M-RMAC	53.40	50.38
DenseNet121	39.24	48.15	Ours(VGG)	51.68	49.58
MDM	28.48	22.59	Ours(ResNet)	55.12	49.77
HoGCV	35.16	22.84	Ours(ResNet+VGG)	**58.08**	**54.39**
MSCM	42.80	31.88			

Retrieval Visualization

To illustrate the retrieval results, five plant images were randomly selected as query images from the SoyCultivar200 Upper dataset. For each query image, the nine most similar retrieved plant leaf images are presented in Fig. 8, since each plant cultivar contains only 10 images. It can be observed that the shape and texture characteristics of the retrieved leaves are similar to those of the corresponding query image. Moreover, at least the first five retrieval results are correct for each query image. These results indicate that the proposed plant recognition method is reliable, particularly for plant identification tasks in which the label of the most similar sample is assigned to the test image.

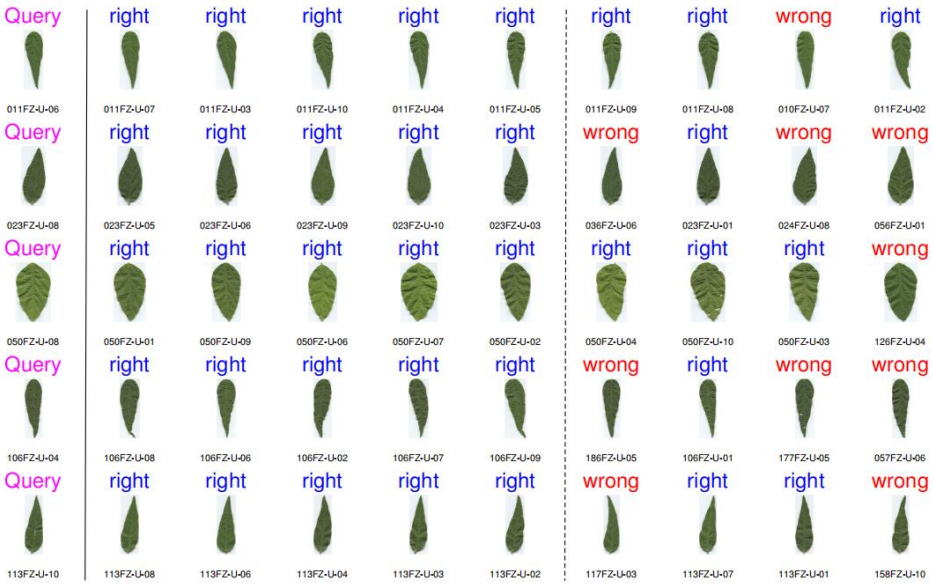


Fig. 8 - Top 9 retrieval results on SoyCultivar200 Upper dataset using Ours (ResNet) features. Five query images are shown here. The description under each image is its name including class label information.

CONCLUSIONS

This study demonstrates that image resolution has a significant impact on fine-grained plant cultivar retrieval. Higher image resolutions generally lead to notable improvements in retrieval performance; however, the rate of improvement gradually decreases as the resolution increases. The results indicate that a resolution of 896×896 may represent the most suitable choice when considering the trade-off between model complexity and retrieval performance, which differs considerably from the commonly used default resolution of 224×224. In addition, a straightforward feature extraction framework based on second-order pooling and feature fusion of two pre-trained convolutional neural networks was presented under an input resolution of 896×896.

Experimental results show that the proposed method achieves retrieval rate improvements of up to 15.71%, 15.95%, 14.02%, and 8% on the SoyCultivar200 dataset, as well as improvements of 4.68% and 4.01% on the PeanCultivar100 and Mulberry datasets, respectively, compared with 21 existing state-of-the-art approaches.

ACKNOWLEDGEMENT

The work described in this paper was partially supported by the Key Research and Development Programs of Shanxi Province (202402020101008), Fundamental Research Program of Shanxi Province of China (20210302124543).

REFERENCES

[1] Chen, X., Wang, B. (2020). Invariant leaf image recognition with histogram of Gaussian convolution vectors. *Computers and Electronics in Agriculture*, 178, 105714.

[2] Chen, X., Wang, B., Gao, Y. (2022). Symmetric Binary Tree Based Co-occurrence Texture Pattern Mining for Fine-grained Plant Leaf Image Retrieval. *Pattern Recognition*, 129, 108769.

[3] Chen, X., Wang, B., Gao, Y. (2023). Fan-Beam Binarization Difference Projection (FB-BDP): A Novel Local Object Descriptor for Fine-Grained Leaf Image Retrieval. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11068–11077.

[4] Chen X., Wang B. (2024). Symmetry-constrained linear sliding co-occurrence LBP for fine-grained leaf image retrieval. *Computers and Electronics in Agriculture*, 218, 108741.

[5] Chompookham, T., Surinta, O. (2021). Ensemble methods with deep convolutional neural networks for plant leaf recognition. *ICIC Express Letters*, 6, 553–565.

[6] Deng, W., Hu, J., Guo, J. (2019). Compressive binary patterns: Designing a robust binary face descriptor with random-field eigenfilters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41, 758–767.

[7] Feng, S. (2022). Kernel pooling feature representation of pre-trained convolutional neural networks for leaf recognition. *Multimedia Tools and Applications*, 81, 4255–4282.

[8] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

[9] Hu, R., Jia, W., Ling, H., Huang, D. (2012). Multiscale distance matrix for fast plant leaf recognition. *IEEE Transactions on Image Processing*, 21, 4667–4672.

[10] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. (2017). Densely connected convolutional networks. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269.

[11] Kalantidis, Y., Mellina, C., Osindero, S. (2016). Cross-dimensional weighting for aggregated deep convolutional features. In: *Computer Vision–ECCV 2016 Workshops*, pp. 685–701.

[12] Liu, G.H., Yang, J.Y. (2021). Deep-seated features histogram: A novel image retrieval method. *Pattern Recognition*, 116, 107926.

[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002.

[14] Ng, C.J., Low, C.Y., Toh, K.A., Kim, J., Teoh, A.B.J. (2018). Orthogonal filter banks with region Log-TiedRank covariance matrices for face recognition. *Journal of Visual Communication and Image Representation*, 55, 548–560.

[15] Nosaka, R., Suryanto, C.H., Fukui, K. (2013). Rotation Invariant Co-occurrence among Adjacent LBPs. In: *Computer Vision-ACCV 2012 workshops*, pp. 15–25.

[16] Pang, S., Zhu, J., Wang, J., Ordonez, V., Xue, J. (2018). Building discriminative CNN image representations for object retrieval using the replicator equation. *Pattern Recognition*, 83, 150–160.

[17] Qi, X., Xiao, R., Li, C.G., Qiao, Y., Guo, J., Tang, X. (2014). Pairwise rotation invariant co-occurrence local binary pattern. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36, 2199–2213.

[18] Radenovic, F., Tolias, G., Chum, O. (2016). CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. In: *Computer Vision – ECCV 2016*, pp. 3–20.

[19] Radenovic, F., Tolias, G., Chum, O. (2019). Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41, 1655–1668.

[20] Simonyan, K., Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations (ICLR)*, pp. 1–14.

[21] Tolias, G., Sicre, R., Jegou, H. (2016). Particular object retrieval with integral max-pooling of CNN activations. In: *International Conference on Learning Representations (ICLR)*, pp. 1–12.
[22] Wang, B., Gao, Y., Yuan, X., Xiong, S., Feng, X. (2020). From species to cultivar: Soybean cultivar recognition using joint leaf image patterns by multiscale sliding chord matching. *Biosystems Engineering*, 194, 99–111.
[23] Wang, B., Li, H., You, J., Chen, X., Yuan, X., Feng, X. (2022). Fusing deep learning features of triplet leaf image patterns to boost soybean cultivar identification. *Computers and Electronics in Agriculture*, 197, 106914.
[24] Wu, H., Fang, L., Yu, Q., Yang, C. (2023). Deep convolutional feature aggregation for fine-grained cultivar recognition. *Knowledge-Based Systems*, 275, 110688.
[25] Yang, C. (2021). Plant leaf recognition by integrating shape and texture features. *Pattern Recognition*, 112, 107809.
[26] Yang, L., Yu, X., Zhang, S., Long, H., Zhang, H., Xu, S., Liao, Y. (2023). GoogLeNet based on residual network and attention mechanism identification of rice leaf diseases. *Computers and Electronics in Agriculture*, 204, 107543.