

A Crop Disease Image Retrieval Method Based on the Improvement of Inverted Index

Yuan Yuan, Lei Chen^(✉), Miao Li, and Na Wu

Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China
{yuanyuan, chenlei, mli}@iim.ac.cn, wuna@mail.ustc.edu.cn

Abstract. According to the characteristics of crop leaf disease images, we proposed a new image retrieval method based on the improvement of inverted index to diagnose crop leaf diseases. First of all, the input crop disease images were preprocessed, including compression, denoising, enhancement, etc. And then the features of disease in the whole image were extracted. Meanwhile, in order to reduce the storage space of inverted index feature vectors, the Hash method was adopted to map the inverted index feature vectors to binary values. Hamming distance was used in the similarity calculation between the obtained features data and the lesion features from the constructed disease images indexes. According to the ranking of similarities, top 5 images were selected as the candidate diagnostic results list of the input crop disease image. And the results were evaluated by some standard criteria, such as precision, recall, etc. The experiments were conducted on cucumber disease images, including: downy mildew, powdery mildew and target spot disease, and rice disease images, including: rice blast, leaf spot and sheath blight. The results showed that the proposed method can achieve the higher retrieval accuracy than traditional SVM method both of cucumber and rice disease images.

Keywords: Image retrieval · Crop disease diagnosis · Inverted index
Image processing

1 Introduction

In recent years, in the field of crop disease prevention and control, the application of computer vision, digital image processing and artificial intelligence technology provided some new ways and ideas for non-destructive detection and intelligent diagnosis of crop diseases. Since 1980s, image processing technology was used to identify crop diseases [1]. And with the development of machine learning technologies, Support Vector Machine (SVM) and Artificial Neural Network (ANN) have become the most commonly used methods in the studies of crop diseases identification [2–6]. The results of these two methods are more dependent on the original training samples. In the conditions of simple image background or under laboratory environment, these two methods can achieve good results. However,

the actual crop disease images captured in the field are easily affected by the surrounding environment, such as uneven illumination, color distortion, image noise and so on. which may easily cause the great differences among samples. For these crop disease images, SVM and ANN can not often achieve the desired classification results.

Due to the good fault tolerance of the samples, the content-based image retrieval [7] can avoid the above problems of SVM and ANN. However, when the training data is not large enough, the effect of this method is still not ideal. In recent years, the development of big data technology has overcome this shortcoming. The content-based image retrieval method has got a new attention and development in crop diseases classification. Especially with the popularity of smart phones, the real-time crop diseases diagnosis is expected to achieve. When a farmer or an agricultural technician takes a picture of crop disease and uploads it to the server, the crop disease image retrieval system can quickly match the image features of the disease and return the diagnosis result of the most similar disease images in time. Li et al. [8] proposed a wheat pest image retrieval system, which focused on image feature extraction, image similarity measurement and users' feedback technology. Song [9] and Pu [10] presented a method of disease image retrieval based on color and other spot features for eggplant and tobacco respectively.

These works have been successful in some crop disease image retrieval. However, we found that when the training data of crop disease images is large, there are two key issues need to be solved in the content-based image retrieval method, given as follows:

- How to select and extract the appropriate features of crop disease images is critical.
- How to set up the index to improve the retrieval efficiency is also a problem that needs to be further studied and solved.

In order to solve these two problems, this paper proposed a crop disease image retrieval method based on the improvement of inverted index. On the one hand, both global and local characteristics of crop diseases were considered for selecting and extracting the features of images. On the other hand, the inverted index approach was introduced to construct the index of feature vectors. Meanwhile, in order to reduce the storage space of inverted index feature vectors, a Hash function was adopted to map the inverted index feature vectors to further compressed binary values. In the experiments, the disease images of cucumber and rice were used to validate this method.

2 The Method

2.1 Method Overview

As Fig. 1 showing, the whole process of the proposed method mainly consists of the following two parts:

1. Constructing the index of features for the training set, the procedure is as follows:
 - Using the feature matching algorithm to extract the feature of each image in training set to form the feature database.
 - Constructing and purifying the feature descriptors and compressing the feature vectors.
 - Using Hash function and tree structure to build hierarchical index information.
2. Searching the candidate diagnostic results list of the input crop disease image, the procedure is as follows:
 - Extracting and compressing the features of the input crop disease image.
 - Using the fast matching technology to search the matching indexes in the feature database.
 - Calculating the Hamming similarity between the feature vectors and returning the top 5 images, as the final classification results according to the similarity ranking.

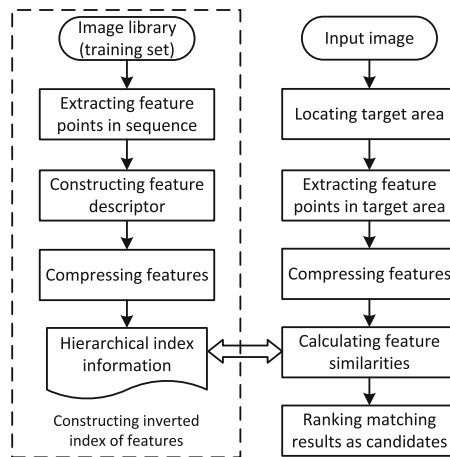


Fig. 1. The overview of system flow

2.2 Feature Extraction and Compression

Since some features of crop disease images, such as color, texture, etc., are easily affected by the natural environment, this paper mainly focuses on the extraction of shape features. Besides, considering the characteristics of crop disease images, the extracted shape features should have the good stability under the conditions of rotation, uneven illumination, color distortion and affine transformation, which often occur during the process of image acquisition. Therefore, we choose Scale-invariant Feature Transform (SIFT) [11] as the main method for

extracting features. As a local feature detection algorithm, the SIFT method can quickly find some interest points or corner points of the given image in spatial scale with their scale and orientation descriptors to get features.

In this paper, according to the SIFT method, we extract features of crop disease images as follows:

- Compressing the original captured image to the resolution $600 * 400$ pixels, denoted by i_0 .
- Adopting the Different of Gaussian to find interest points or corner points in scale space of i_0 .
- Locating the key points by comparing the adjacent layers of each Different of Gaussian space in the same group.
- Determining the distribution of gradient directions by the gradient histogram.
- Counting and describing gradient information of Gauss image in the neighborhood of feature points.

After feature extraction, the Gauss normalization method is adopted to normalize the obtained features. And then the feature database can be constructed.

Since the feature vectors takes up a lot of space while the image database is large, the feature vector should be compressed to improve the retrieval efficiency. Given a feature vector x_i , the difference matrix C of the specified crop disease type can be calculated. And according to the distribution pattern of C , the feature vector x_i can be converted to the binary code of length l , which is denoted by $B = (b_0, b_1, \dots, b_{l-1})$. Since the value of l in SIFT is 128, the storage space of binary codes is very large and the computation efficiency of feature matching is low. So a compression method of feature descriptors is proposed. Given a feature descriptor $B = (b_0, b_1, \dots, b_{l-1})$ and a threshold α , the compression is conducted by the following formula:

$$C(i, j) = \begin{cases} 11, & \text{if } b_j - b_i > \alpha \\ 10, & \text{if } -\alpha < b_j - b_i \leq \alpha \\ 00, & \text{otherwise} \end{cases} \quad (1)$$

After the feature compression, the storage space of the binary codes is significantly reduced. Meanwhile, the computational efficiency of feature matching can be improved. Figure 2 shows an example of feature extraction and compression.

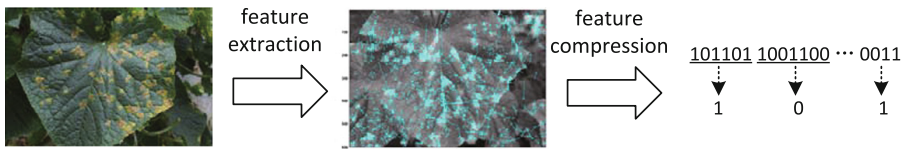


Fig. 2. Feature extraction and compression

2.3 Image Retrieval

Since the inverted index method can meet the actual retrieval requirement, which finds records according to the value of the selected property, the image retrieval method based on inverted index has been widely studied in recent years [12–14]. The inverted index consists of a lexicon and corresponding inverted lists. The lexicon is a set of strings made up of all the words that appear in the document set. Each index entry records some information about the word itself and a pointer to the inverted list. And the inverted list uses a data structure of the form $(docid, wf, \langle p_1, \dots, p_f \rangle)$ to record the hit information of a word item in a document, where $docid$ is the ID of the document, wf is the word frequency and p_i indicates each position offset of the word appearing in this document. For each word item in a query, it is necessary to find out whether it appears in the lexicon, and the disk location of the corresponding inverted list. Since the speed of locating a word directly determines the retrieval efficiency, the efficient data structures including Hash table and tree structure are used to build the word lexicon and conduct the search. Besides, in the image retrieval tasks, the index storage space is usually larger because of the large number of image features. Besides, the retrieval efficiency is low when there are much more candidate retrieval results. Therefore, it is worthy of further study on index compression [15, 16] and result filtering [17, 18].

In this paper, in order to reduce the storage space, the following Hash function is trained to map the feature vector of the original space to binary codes,

$$h_k(x_i) = \text{sgn}(w_k^T x_i + a_k), \quad (2)$$

where x_i denotes the content feature vector of the image, w_k denotes the mapping vector that needs to be learned, a_k is the average vector of all content features that need to be mapped. And then the hierarchical index is constructed in tree structure. In the inverted index structure, each inverted list is represented by a visual word. Each list is equivalent to a cluster. And the corresponding cluster center is the visual word. When constructing an inverted index for image retrieval, the visual features of the image are inserted into the inverted list corresponding to the nearest visual word. The image retrieval procedure usually finds several visual words which have the nearest distance to the visual features of the query image. And the query results are all the visual features of the inverted list corresponding to these visual words. If the final query results only need KNN similar visual features, some distance such as Euclidean, Hamming, etc. needs to be calculated and sorted between the query feature and all the query results. Hamming distance is adopted in this paper. Given two feature vectors x_i and x_j with the length l , the Hamming distance between binary codes is calculated to measure the similarity of features as the following formula:

$$S(H(x_i), H(x_j)) = \sum_{n=1}^l (\text{xor}(H(x_i), H(x_j))). \quad (3)$$

In order to further reduce the search space, only the index values of the first half of the Hash table are used to quickly classify the query object. And then the

retrieved files are ranked according to the final statistics of their hit frequencies. Hence, according to the similarity ranking, five candidate images with the highest similarities of the input image in the relative database can be obtained.

3 Experiments

3.1 Experimental Setup

We conducted experiments both on small scale and large scale samples. When the experimental sample scale is small, the proposed image retrieval method was compared with the traditional SVM method, which can achieve good performance on small scale samples. The traditional SVM method contains some main steps, such as image segmentation, feature extraction and classification recognition, etc. The experimental data includes six diseases of cucumber and rice which were captured in the field. More concretely, the images used in this paper were collected from 10:00am to 15:00pm on sunny days in Hefei, using the digital SLR camera of the model Canon EOS 6D, with the lens EF 17–40 mm f/4L USM for capturing the disease images of cucumber leaves in greenhouse and the macro lens EF 100 mm f/2.8L IS USM for capturing the disease images of rice leaves in field. The original image resolution is $5472 * 3648$ pixels. The cucumber diseases include target spot, powdery mildew and downy mildew. The rice diseases include rice blast, sheath blight and brown spot.

In image preprocessing, first we selected and clipped the main diseased area from the original images. And then the clipped images were compressed to the resolution of $600 * 400$ pixels for the consideration of time and space efficiency while preserving more image details. Some examples are shown in Fig. 3.

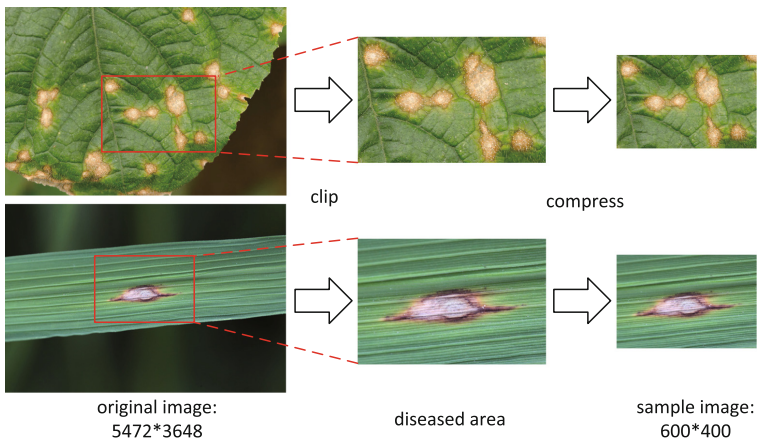


Fig. 3. Examples of main diseased area clipping and compressing

In the traditional SVM method, we adopted graph cuts fusing multiple features to segment the disease images [19]. And we conducted feature extraction and classification according to the early work [20].

In the proposed method, the training set was used to construct the inverted index. And the test set was used to verify the method. Hamming distance was adopted to measure the feature similarity. According the ranking of similarities, the top 5 images were selected as the candidate diagnostic results list of the input test image. And the results were evaluated by some standard criteria, such as precision, recall, that is: $P = h_1/S_t * 100\%$, $R = h_5/S_t * 100\%$, where h_1 is the number that the ranked first image in the candidate images is the correct matching, h_5 is the number that the top five candidate images contains the correct matching, S_t is the size of test set.

Besides, we further evaluated our method in large scale samples. In order to verify the robustness of this method, the inverted index is trained by using the original images without any preprocessing.

The experimental environment are as follows: Intel Core i7-4790@3.66 GHz, 4 GB RAM, Microsoft Windows 10 64-bit, Microsoft Visual Studio 2013, Open Source Computer Vision Library (OpenCV) 2.4.8, Matlab R2014b.

3.2 Experiments on Small Scale Samples

In the experiments on small scale samples, each disease set contains 60 images, where 50 images are randomly selected as the training set and the remaining 10 images are taken as the test set. Table 1 shows the experimental results, where SVM denotes the classification accuracy of the traditional SVM method, IR-P and IR-R are the evaluation criteria of the proposed method mentioned above.

Table 1. The experimental results on small scale samples

(a) The classification of cucumber diseases			
	Target spot	Powdery mildew	Downy mildew
SVM (%)	37.9	93.5	23.6
IR-P (%)	70	40	50
IR-R (%)	100	90	90
(b) The classification of rice diseases			
	Rice blast	Sheath blight	Brown spot
SVM (%)	89.6	37.9	63.9
IR-P (%)	70	90	10
IR-R (%)	100	100	30

It can be seen that our method is better than the traditional SVM method. Especially in the identification of cucumber target spot, downy mildew and rice

sheath blight, the percentage that the first image of five candidates is the correct matching is obviously better than the traditional SVM method. For the classification of other diseases, although the proportion that the first candidate image is the correct matching is not high, the percentage of the correct matching occurring in five candidates is still better than SVM. However, in the identification task of rice brown spot, the effect of our method is not good, which indicates that our method needs to be further optimized. Figure 4 shows the more intuitive comparison of experimental results.

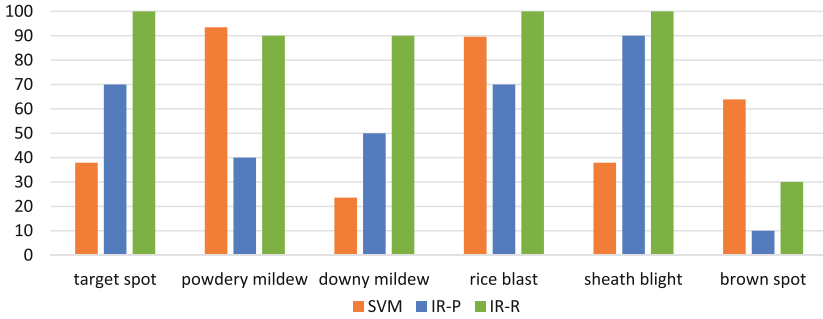


Fig. 4. Comparison between the traditional SVM method and the proposed method

3.3 Experiments on Large Scale Samples

In the experiments on large scale samples, the image database consists of 4338 images of cucumber and rice diseases. More details are shown in Table 2.

Table 2. The sizes of experimental training set and test set

Crop diseases		Training set	Test set	Total
Cucumber	Target spot	755	50	805
	Powdery mildew	794	52	846
	Downy mildew	1000	70	1070
Rice	Rice blast	471	31	502
	Sheath blight	266	17	283
	Brown spot	780	52	832
Sum		4066	272	4338

The training set of cucumber diseases contains 2549 images, where the test sets contain 50, 52, 70 images of target spot, powdery mildew and downy mildew respectively. The training set of rice diseases contains 1517 images, where the

Table 3. The experimental results on large scale samples

(a) The classification of cucumber diseases			
	Target spot	Powdery mildew	Downy mildew
P (%)	84	50	38.6
R (%)	96	92.3	75.7
(b) The classification of rice diseases			
	Rice blast	Sheath blight	Brown spot
P (%)	74.2	90	5.8
R (%)	100	100	57.7

test sets contain 31, 17, 52 images of rice blast, sheath blight and brown spot respectively. Without any preprocessing, all images in training sets and test sets are original. Table 3 gives the experimental results on large scale samples.

Figure 5 shows the comparison of the experimental results on small scale and large scale samples, denoted by IR1 and IR2 respectively. It can be seen that our method can still achieved the good results even if the inverted index was trained with a large scale of the original image without any preprocessing, which shows that the proposed method has good robustness.

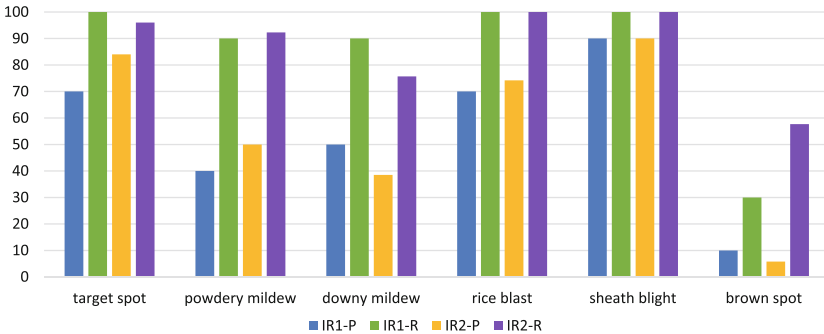


Fig. 5. Comparison between small scale and large scale samples

In addition, we conducted the experiment that the training set contains 4066 hybrid images of cucumber and rice diseases, where the test sets of cucumber and rice are the same as the above experiments. The experimental results are shown in Table 4. The stable results show that our method has good universality. For a variety of crop diseases, this method does not need to train multiple classifiers. All kinds of crop disease images can be trained together to improve computational efficiency.

Table 4. The experimental results of hybrid diseases

(a) The classification of cucumber in hybrid diseases			
	Target spot	Powdery mildew	Downy mildew
P (%)	80	42.3	28.6
R (%)	96	86.5	51.4
(b) The classification of rice in hybrid diseases			
	Rice blast	Sheath blight	Brown spot
P (%)	77.4	17.6	3.8
R (%)	100	76.5	19.2

4 Concluding Remarks

The paper proposed a content-based image retrieval method for the classification of crop disease images. Focusing on how to improve the construction of inverted index of feature vectors, the Hash function was adopted to map the inverted index feature vectors to further compressed binary values to reduce the storage space and improve the efficiency of feature matching.

We validated this method by using cucumber disease images and rice disease images. The experimental results show that:

1. Comparing with the traditional SVM method, our method can achieve the better results. And the accuracy that the correct matching image occurs in five candidates is higher than the accuracy that the first candidate is the correct matching image.
2. Since the cucumber leaves occupy a larger proportion in the whole image and the interference of the complex background is less, the experimental results of cucumber disease image retrieval are better than the results of rice.
3. The method has good robustness on large scale of original image without any preprocessing. And the accuracy decrease of hybrid disease image retrieval is not obvious.
4. The proposed method is general and it can also be applied to other crop diseases.
5. The method of feature extraction and similarity calculation in this paper still need to be further optimized to improve the efficiency and accuracy of image retrieval.

The proposed method needs to be further improved. Comparing with other related algorithms and validating on more crop disease images will be concerned in next works.

Acknowledgments. The authors would like to thank the anonymous reviewers for their helpful reviews. The work is supported by the National Natural Science Foundation of China under No. 31501223.

References

1. Wolfe, R.R., Sandler, W.E.: An algorithm for stem detection using digital image analysis. *Trans. ASAE* **28**(2), 641–644 (1985)
2. Rumpf, T., Mahlein, A.K., Steiner, U., Oerke, E.C., Dehne, H.W., Plümer, L.: Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Comput. Electron. Agric.* **74**(1), 91–99 (2010)
3. Omrani, E., Khoshnevisan, B., Shamshirband, S., Saboohi, H., Anuar, N.B., Nasir, M.H.N.M.: Potential of radial basis function-based support vector regression for apple disease detection. *Measurement* **55**(9), 512–519 (2014)
4. Zhang, J., Kong, F., Li, Z., Wu, J., Chen, W., Wang, S., Zhu, M.: Recognition of honey pomelo leaf diseases based on optimal binary tree support vector machine (in Chinese). *Trans. Chin. Soc. Agric. Eng.* **30**(19), 222–231 (2014)
5. Zhu, L., Feng, Q., Yang, M., Zhang, Z.: Study on segmentation and diagnosis of wine grape disease based on image processing (in Chinese). *J. Chin. Agric. Mech.* **36**(1), 111–115 (2015)
6. Ma, X., Guan, H., Qi, G., Liu, G., Tan, F.: Diagnosis model of soybean leaf diseases based on improved cascade neural network (in Chinese). *Trans. Chin. Soc. Agric. Mach.* **48**(1), 163–168 (2017)
7. Liu, Y., Zhang, D., Lu, G., Ma, W.: A survey of content-based image retrieval with high-level semantics. *Pattern Recogn.* **40**(1), 262–282 (2007)
8. Li, Z., Liu, Y., He, D., Long, M., Liu, Q.: Investigation and implementation of content-based retrieval system for wheat pest images (in Chinese). *Trans. Chin. Soc. Agric. Eng.* **23**(11), 210–215 (2007)
9. Song, J.: Eggplant disease image search method based on color characteristics (in Chinese). *J. Anhui Agric. Sci.* **39**(19), 11920–11921, 11977 (2011)
10. Pu, Y.: Image searching method of tobacco disease based on disease spot feature fusion (in Chinese). *J. Henan Agric. Sci.* **44**(2), 71–76 (2015)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
12. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: 9th IEEE International Conference on Computer Vision (ICCV 2003), Nice, France, 14–17 October 2003, pp. 1470–1477 (2003)
13. Chen, Y., Guan, T., Wang, C.: Approximate nearest neighbor search by residual vector quantization. *Sensors* **10**(12), 11259–11273 (2010)
14. Jégou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(1), 117–128 (2011)
15. Anh, V.N., Moffat, A.: Inverted index compression using word-aligned binary codes. *Inf. Retrieval* **8**(1), 151–166 (2005)
16. Yan, H., Zhang, X., Shan, D., Mao, X., Zhao, X.: SIMD-based inverted index compression algorithms (in Chinese). *J. Comput. Res. Dev.* **52**(5), 995–1004 (2015)
17. Hwang, Y., Han, B., Ahn, H.: A fast nearest neighbor search algorithm by non-linear embedding. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), Providence, RI, USA, 16–21 June 2012, pp. 3053–3060 (2012)
18. Ai, L., Yu, J., Guan, T., He, Y.: Adaptively filtering query results for large scale image feature retrieval (in Chinese). *Chin. J. Comput.* **38**(1), 122–132 (2015)

19. Wu, N., Li, M., Chen, S., Yuan, Y., Zeng, X., Chen, L., Sun, X., Bian, C.: Automatic segmentation of plant disease images based on graph cuts fusing multiple features (in Chinese). *Trans. Chin. Soc. Agric. Eng.* **30**(17), 212–219 (2014)
20. Yuan, Y.: Research and application of image-based intelligent recognition for cucumber leaf diseases (in Chinese). Ph.D. thesis, Anhui Agricultural University (2013)