Feature Evaluation in Fine-gain of Leaf

Zhanhao Chen[1, a], Shan Xu[1,b], Yixiong Zou[1,c], Hualong Zhang[1,d], Zhu Zhang[1,e], Yue Li[1,f], Wei Wang*[1,g]

[1]NanKai University, Tianjin, China

[a] nkdxczh@mail.nankai.edu.cn, [b] xu_shan@mail.nankai.edu.cn, [c] zouyixiong@mail.nankai.edu.cn, [d]zhanghualong@mail.nankai.edu.cn, [e]zhangzhu@mail.nankai.edu.cn, [f]liyue80 @ nankai.edu.cn, [g]kevinwangwei @ nankai.edu.cn

Keywords: leaf recognition; random forest; feature evaluation

Abstract. In order to compare the value of several features involving leaf retrieval, we design an approach to evaluate 37 features about leaf's contour, content and texture. Random forest algorithm is employed to rank these features' contribution to leaf categorization. To forming the optimum features combination, we get the highest retrieval accuracy by gradually adding the most valuable features and depict the relationship between accuracy and feature number. Combined with the time analysis, different features group could be adopted for efficiency requirement. The leaf samples are from Taiwan and ICL database.

1. Introduction

In recent years, computer vision has drawn more and more attention to researchers as it is capable to help us handle a large quantity of statistics with high accuracy. Technologies such as human face recognition, vehicle license recognition and so on have been applied in everywhere around us. Among this topic, there is a subtopic that may be the most attractive one: fine-grained categorization. Many researchers choose to find a good search algorithm or use the effect of big data so as to establish a good categorization, in which a great many of excellent research have been carried out such as TED speech made by Feifei Li in March this year. In Li's research, they build a giant neural network in which there are tens of millions neurons with over one hundred million parameters trained using one billion pictures categorized by nearly 50,000 workers from 167 countries since 2007. Of course she constructs an excellent fine-grained categorization which can amazingly recognize many difficult pictures for ordinary categorizations.

However, many as these researchers are, only few focus on the features themselves they choose that can translate the language of a vivid picture into the digital signals. But how can a categorization be good enough without appropriate features being chosen? To tackle this problem, in this paper we chose 37 features used in [1-10] containing leaf texture, contour as well as color in recent years to form a framework using random forest algorithm [11] to construct a fine-grained categorization of leaves. As a consequence, on the one hand, the framework categorizes the leaves in the given database well applying the whole 37 features. On the other hand, it gives the importance score of each features. Then we used an accuracy-time consumption loss function shown in the following papers, so we reduced the features we chose ordered by the importance score and then used the same framework to categorize the database again. After the reduction, the categorization took only a rather short time compared with the original one, increasing the efficiency of categorizing without noticeable loss of accuracy.

The balance between accuracy and efficiency is always significant in every field of our lives. Although the experiment is carried out among the database of leaves, the result of this research can also be applied to different fields of computer vision in terms of fine-grained categorization so as to strike a balance between accuracy and efficiency.

2. Research on Leaf Feature Extraction

Different from traditional image recognition such as scene or object recognition, fine-grained categorization deals with images with subtle distinctions, which usually involves the classification of subclasses of objects belonging to the same class like birds[1],dogs[2], planes[3], plants[4], etc. Plant recognition, based on fine-grained leaf categorization is very important and necessary to agricultural information, ecological protection, and automatic plant recognition systems. Popular works on plant leaf recognition focused on shape features [19–28]. They developed edge detectors or used the existing edge detection methods to extract the contour of a leaf, which was matched directly or represented in other formats such as curvature scale space or deformable templates for matching. Kumar et al.[24] extracted curvature features from the binarized leaf images, and selected a nearest neighbor classifier with histogram intersection as the distance metric for classification. Color is another remarkable characteristic of image [24] and there are different kinds of color spaces, such as RGB color space, HSV color space, Luv color space and Lab color space [25]. For attaining color feature, color histogram, color set and color moment are comprehensively used [26].

Referring to the past researches domestic and international [30], plant leaf recognition at early stage is processing mainly based on single-kind feature. As a result, multiple-kind features have increasingly become the mainstream. In 2003 [32], Zhang Ling and eta carried on a research taking samples of 100 kinds of plants among Beijing area, and worked out that the shape feature together with the vein feature is better than using either one of them. In 2014 [27], Zhu Haodong and etc. conducted a research making use of Flavis data set and drew a conclusion that the vein feature with the addition of Hu invariant moment would have a higher success rate compared with using either one of the two. In 2015[31], Wang Lijun and etc. did a research of 50 samples and figured out that the shape feature contributes most for the recognition success rate; using the combination of feature kinds is superior to using a single kind; and the vein feature contributes more than the color one.

3. Feature and Algorithm

3.1. Features Description.

Refer to the experience of formal researchers, we select 37 representative features. Table 1 lists the detail information of these features. The background of content features is light grey and that of texture features is dark grey. In the equation about texture features, S represents elements of Gray-level Co-occurrence Matrix which has N rows and columns. μ and σ represents the average value and deviation value of Gray-level Co-occurrence Matrix's elements.Hu_1 to Hu_7 represent the invariant moments which are put forward by Hu.M.K in 1962.

Table 1 Feature Collection from Previous Researches

Feature	Description	Reference
Aspect_Radio	maximum length of the mini-mum bounding rectangle/minimum length of the mini-mum bounding rectangle	[24][43]
A_Convexity	area of broad leaf/area of convex hull	[24][44]
Centroid_ Radii	average distance between centroid and boundary pixels（normalized）	[39]
Coarseness	perimeter of leaf contour/length of internal border	[36]
Complexity	square of broad leaf's perimeter/area of leaf	[24]
Curvature	mean of contour curvature	[40]
Dcur	standard deviation of contour curvature	[40]
DNRL	standard deviation of distances between centroid and boundarypixels（normalized）	[38]
Eccentricity	shortest axis/longest axis.	[24][44]
Eccentricity1	the longest distance between centroid and boundarypixels/the shortest distance between centroid and boundary pixels	[24]

Solidity	internal area connecting the valley points/external area connecting the top points	[38]
Sphericity	radius of the ex-circle of the leaf/radius of in-circle of the leaf	[24][43]
Hu_1-Hu_7	Hu invariant moment of contour	[35]
Lobation	the shortest distance between centroid and boundary pixels/short axis	[32]
LPR	Major axis of the best fit ellipse/Number of boundary pixels	[37]
Mshape	Leaf's area/leaf's morphological-closing area.	[38]
P_Convexity	broad leaf's perimeter/convex hull's perimeter	[24][44]
RA	change rate of distance between centroid and pixels of contour	[38]
Rectangularity	area of leaf/area of bound-box	[24][43]
Roughness	Average variation of distance between centroid and boundary pixels（normalized）	[24]
SaturationAvg	Saturation average	[41]
SaturationDev	Saturation variance	[41]
GrayAvg	gray average	[41]
GrayDev	gray variance	[41]
HueAvg	Chromaticity average	[41]
HueDev	Chromaticity variance	[41]
texture0	$\sum_{i=1}^{N}\sum_{j=1}^{N}\{S(i,j)\}^2$	[42]
texture1	$\sum_{i=1}^{N}\sum_{j=1}^{N}[(i-j)^2 * S(i,j)]$	[42]
texture2	$\frac{1}{\sigma^2}\{[\sum_{i=1}^{N}\sum_{j=1}^{N} i \cdot j \cdot S(i,j)] - \mu^2\}$	[42]
texture3	$\sum_{i=1}^{N}\sum_{j=1}^{N}[(i-\mu)^2 * S(i,j)]$	[42]
texture4	$\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{S(i,j)}{1+(i-j)^2}$	[42]

3.2. Random Forest Algorithm.

We adopt Random Forest to process the data we have now so as to get the importance of each feature. First of all, Random Forest uses bootstrap method to get sample sets at the size of n. For each sample, it will select m features randomly from the whole feature sets to generate a tree.

Each tree is a binary tree. When a tree is being built, it's split from the top to the bottom, following the Gini rule. That is, supposing P(wj) is the frequency of the samples belonging to class wj in node n, the impurity level I(n) can be expressed as

$$I(n) = \sum_{i \neq j} P(\omega_i)P(\omega_j) = 1 - \sum_j P^2(\omega_j) \quad (1)$$

Stop splitting one node only if the samples in it belong to the same class.

Repeat for times and get trees as a forest so as to form a strong classifier. When there comes a new input to classify, the forest will let each decisive tree to classify respectively and vote for the class it output. The result of the forest is the class that owns the most votes.

As Breiman describes in [1], we use out-of-bag data to estimate the error. Given a specific training set T, form bootstrap training sets Tk, construct classifier h(x, Tk) and let these vote to form the bagged predictor. For each y,x in the training set, aggregate the votes only over those classifier for which Tk does not containing y,x. Call this the out-of-bag classifier. X is the input vector and Y is the class that X ought to belong to. Then the out-of-bag estimate for the generalization error is the error

rate of the out-of-bag classifier on the training set. For each tree in the forest, the misclassification rate is computed using OOB(out-of-bag) data, denote it errOOB1. Then we randomly cast some disturb on the feature in the whole OOB data. That is, the values of the selected variable in the out-of-bag examples are randomly permuted while the values of others' stay the same. For each feature, we denote the misclassification rate after being disturbed as errOOB2. Given N trees, the importance of the feature is

$$i = \sum \frac{errOOB2 - errOOB1}{N} \tag{2}$$

We use databases from Taiwan and ICL, randomly picked as training set and treat the rest as testing set. There are kinds of trees and samples in sum. We have 37 features in total, which are all rotation independent. We take m features from the feature set randomly as the features to generate trees. As Random Forest is not sensitive to m, we always take m as $\sqrt{M}$, in which M stands for the size of the total feature set. So, we get the importance of each feature, time consumed in training and testing, classification rate and so on as illustrated below. Having done these, we pick top5, 7, 10, 15, 20, 25, 26, 27, 30, 37 features order by the importance as the whole feature set to form a forest and get the responding classification rate respectively. The results are attached below.

In the experiment, we encountered that the classification rate didn't increase as we add features into the forest in order of importance. As illustrated by [1], say there are two variables x1 and x2 which are identical and carry necessary predictive information. Because each gets picked with about the same frequency in a Random Forest, noising each separately will result in the same increase in error rate. But once x1 is entered as a predictive variable, using x2 in addition will not produce any decrease in error rate. So it does not add predictive accuracy when combined with the second.

4. Experiment and Analysis

4.1. Experiment Design.

The leaves pictures our experiment based on are from the database of Taiwan (102 species, 50 samples for each species) and ICL, the Intelligent Computing Lab of Chinese Academy of Sciences (220 species, 39-260 samples for each species). Each picture has white background and the single leaf in it has been randomly rotated.

In order to cut the leaf out, we first convert the picture to a binary one and then apply morphological closing operation to it, which could avoid the influence of the petioles and tiny hackle. Then we regard the longest contour as the leaf's contour, within which we can get the information about content and texture. The contour would save as a set of points. The second step is to divide the whole samples into two parts: 25 of them will serve as training samples and the others would be testing samples. We extract the 37 features from these samples and save them as txt files, while recording the time consumed to extract in the same time. The data was used to process the following experiments:

1) Feature valuation: use random forest algorithm to train, recognize the samples and score each feature. Get the recognition accuracy. The number of variables randomly selected at node is 5, max depth is 50 and the max number of trees in the forest is 200.

2) Optimal feature number: Respectively select different number of most valuable features of the upper experiment. Use them for training and testing, get the accuracy of each experiment. The parameters of random forest algorithm are same as the upper experiment.

3) Sorted feature valuation: Respectively select different kind of features (contour, content and texture) to process three experiments to get the accuracy. Parameters are same as the upper two experiments.

4.2. Experiment Analysis.

Table 2 exhibits the contribute percentage of each feature for the two database separately and the sum of them. The features are ranked in the descent order of the total percentage. The correlation coefficient of the two databases' results is 0.633793928. Analyzing the specific contribute percentage

of every feature, we can get the following conclusion. First of all, some features which are usually be used together perform serious divergence in the experiments. The most distinct ones are the Hu of contour: the highest one, Hu_2, ranks fourth while Hu_7 get the 36 place, and the diversity of other Hu features is apparent, too. Thus, putting all of these features in the algorithm without different weights will be unadvisable. Besides, we are acknowledged that the most valuable features are about contour since the 6 highest percentages are from contour's information.

Table 2 The Contribution of the top 10 features

Feature	**TW**	**ICL**	**TW+ICL**
LPR	0.0579408	0.0394955	0.0974363
P_Convexity	0.0395499	0.0443795	0.0839294
A_Convexity	0.0331583	0.044061	0.0772193
Hu_2	0.0309221	0.0410406	0.0719627
Curvature	0.0376713	0.0305487	0.06822
Hu_1	0.0282752	0.0363234	0.0645986
GrayDev	0.0413601	0.0231632	0.0645233
Hu_3	0.034813	0.028985	0.063798
texture0	0.0321755	0.0312939	0.0634694
Rectangularity	0.0281197	0.0312056	0.0593253

To get features owning stable performance, we divide the features of two database respectively into three groups (12+12+13) according to their contribute percentages. Since we don't need so many features to process leaf recognition, we should only pay attention on the feature group owning the highest contribute percentage. We can see that for the most valuable group, the two experiments present greatly overlap. The phenomenon that the value of features about content and texture would be largely influenced by the quality of pictures could explain the unstable performance of GrayAvg, GrayDev and texture1. Table 3 shows general groups. For example, if the feature appears in the first group in both of the two experiments, it would be put into the group whose color is the darkest. The number of each group show amazing similar: each group owns 7-8 features. Thus, this table gives us advice for practical application: it recommends us to use the features of dark color to represent a leaf.

Table 3 Features' overall performance

P_Convexity	Hu_1	Roughness	Solidity	Sphericity
A_Convexity	Rectangularity	HueAvg	Hu_4	Mshape
Hu_2	Complexity	Eccentricity1	RA	Coarseness
LPR	Centroid_Radii	texture2	Lobation	Eccentricity
texture0	DNRL	Dcur	Hu_6	Hu_5
Curvature	texture4	GrayAvg	SaturationDev	texture3
Hu_3	SaturationAvg	GrayDev	HueDev	Hu_7
		texture1		Aspect_Radio

Figure 1 shows the relationship of accuracy and the number of features. The accuracy is calculated by dividing the number of test samples by the number of the samples whose classifications are accurately predicted by the experiment. From the figures, both of two experiments reach stable accuracy when we use about 20 features. For TW database, the accuracy reaches 83.8% while using 26 features. The other experiment's accuracy keeps fluctuating around 70% after the feature number gets higher than 20. And when we use only 5 features, the accuracy will drops sharply to below 20%.

Figure 2 shows the relationship of time consumed in training models and the number of features. The proportional relation is apparent. Exceptional deviation might be caused by the erratic conditions of computer.

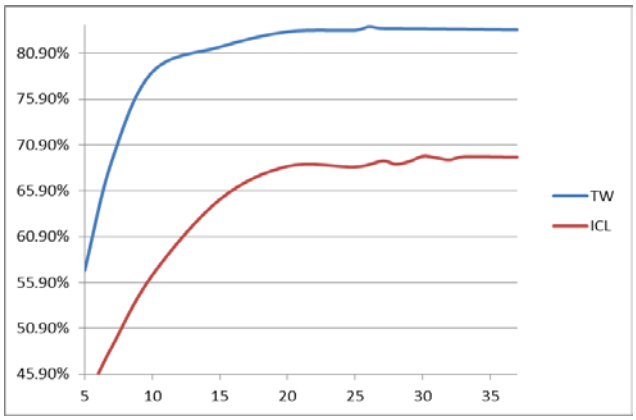

Figure 1 The accuracy improvement according to feature inserting on the two datasets

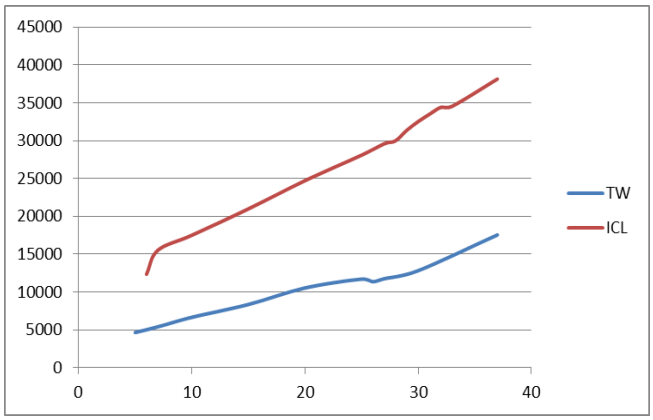

Figure 2 The training time increase according to feature inserting on the two datasets

REFERENCES

[1] TEIXEIRA P R F, AWRUCH A M,"Numerical simulation of fluid-structure interaction using the finite element method," Computersand Fluids .2005

[2] Lynn DE, Waldren S,"Morphological variation in populations of Ranunculus repens from thetemporary limestone lakes (Turloughs) in the west of Ireland," Annals of Botany. 2001

[3]]Nero J C, Meyer G E, Jones D D, et al, "Plant species identification using elliptic Fourier leaf shape analysis," Journal of Computers and Electronics in Agriculture.2006

[4] Ojala T, Valkealahti K, Oja E, et al, "Texture discrimination with multidimensional distributions of signed gray-level differences," Journal of Pattern Recognition.2001

[5] He Peng,Huang Lin,"Feature extraction and recognization of leaves," Journal of Agricultural Mechanization Research. 2008(06)

[6] Xiao Peng, Xu Jun, Chen Shaochong, "Texture extraction methods," Electronic Technology ,2010.

[7] Zhang Xin, Wen Xianbin, Meng Qinxia,"Image retrieval based on color feature," Compture Science.2012

[8] Song Yan, Liu Fangai,"Image retrieval based on color and texture feature," Computer Engineering and Design,2007.

[9] Wang zhirui, Yan Cainiang,"Image feature extraction method review," Journal of Jishou University,2011.

[10] Zhu Haodong, Shen Zhen,"Recognition of plant leaves based on the law of cosines and K-means," Journal of Central China Normal University,2014.

[11] Zhang Zao, Yang MingCang, He Dongjian,"Research on the classification of plant leaves based on PCA and SVM," Journal of Agricultural Mechanization Research，2013.

[12] Ingrouille M J, Laird S M, "A quantitative approach to oak variability in some north London woodlands," Nat. 1986

[13] He Peng,"Identification of broadleaf machine based on an integrated feature of the blade," Journal of Northwest A&F University,2008.

[14] Zhang Lijun, Huai Yongjian, Peng Yuecheng,"Plants species identification based on multi-feature fusion.," Journal of Beijing Forestry University,2015.

[15] Zhang Ning, Liu WenPing,"Plant leaves Recognition based on clonal selection algorithm and K-neighbors," Computer application,2013.

[16] Du Jixiang,"Plant species machine identification technology," Journal of University Of Science And Technology Of China, 2005.

[17] Wang Xiaofeng, Huang Deshuang, Du Jixiang., "Research on image feature extraction and leaves recognition," Computer Engineering and Applications,2006.

[18] Chen C C, "Improved Moment Invariants for Shape Discrim Ination," Pattern Recognition .1993

[19] Qingfeng W, Kunhui L, Changle Z, et al,“Feature extraction and automatic recognition of plant leaf using artificial neural network[J],” Advances in Artificial Intelligence, 2007.
[20] Arribas J I, Sánchez-Ferrero G V, Ruiz-Ruiz G, et al,“Leaf classification in sunflower crops by computer vision and neural networks[J],” Computers & Electronics in Agriculture, 2011.
[21] WC P, AV A, AF I, et al,“A non-linear morphometric feature selection approach for breast tumor contour from ultrasonic images[J],” omr n Bology and Mdn, 2010.
[22] Chaki J,“Plant Leaf Recognition using Shape based Features and Neural Network classifiers[J].”International Journal of Advanced Computer Sciences & Applications, 2011.
[23] Zhao X, Yang B, Gao S, et al,“Multi-contour registration based on feature points correspondence and two-stage gene expression programming[J],” Neurocomputing, 2014.
[24] Li Guohui, Liu Wei, Cao Lihua, “Image retrieval method based on color feature,” Journal of Image and Graphics,1999.
[25] Tan Ju, Zhang Youzhong,“Scene recognition based on texture feature GLCM,” Journal of Chongqing University of Arts and Sciences(Nature Science),2010.
[26] Du J, Wang X, Zhang G, “Leaf shape based plant species recognition.[J],” Applied Mathematics & Computation,2007.
[27] Lin H, Peng H, “Machine Recognition for Broad-Leaved Trees Based on Synthetic Features of Leaves Using Probabilistic Neural Network,” Computer Science & Software Engineering International Conference on. IEEE Computer Society, 2008.
[28] Ho, Tin Kam (1995), "Random Decision Forest". Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition, Montreal, Canada, 1995.
[29] Bangpeng Yao, Gary Bradski, Li Fei-Fei,“A Codebook-Free and Annotation-Free Approach for Fine-Grained Image Categorization,”In CVPR, 2012