

# Feature Engineering for Tree Leaf Classifier

Daria Korotaeva and Maksim Khlopotov

ITMO University, Saint-Petersburg, Russia,  
daria.korotayeva@yandex.ru

**Abstract.** The article presents an approach to classify trees based on the image of a leaf. In this work, 1464 images of 18 species of trees that are typical for Russian flora were used, and the method of k-nearest neighbors was used for classification model. The process of feature retrieval includes image verification, binarization and removal of the leaf petiole. The selected 22 features are based on the analysis of the image moments and distances from the centroid to the boundary coordinates. An accuracy of more than 95% is obtained on the testing set.

**Keywords:** tree recognition, leaf image classification, feature engineering, machine learning.

## 1 Introduction

Trees are an important part of terrestrial ecosystems and have a great impact on the environment. Some species of deciduous trees can represent a source of allergens for people prone to pollinosis, and the ability to determine such a tree is necessary for them. Trees are also of interest to collectors of herbariums and parents with children. Despite the importance of the ability to determine the tree type, most people do not have the necessary skills for this. The proposed method may be used for creating a tree recognition application.

Despite the large variety of plant organs that can be used for determining the type of a plant, a leaf is the most convenient object in terms of image analysis. Unlike fruits and inflorescences that are present on the tree for a short period of time, leaves can be collected for a long period throughout the year, and can also be used in a dried form.

In this article, feature engineering for the k-NN classifier is presented, as well as the method of image preprocessing. Section 2 deals with the existing research in this area. Section 3 is focused on data collection and the method of image verification. The processing of the image, removal of the petiole and computation of the one-dimensional array of distances, is described in section 4. The final set of features is presented in section 5, section 6 deals with the achieved results, conclusion and further work are described in section 7.

## 2 Related Work

The existing solutions offer different approaches to leaf image analysis. Most authors prefer to analyze leaf shape for extracting basic features. For example,

the Flavia Plant Leaf Recognition System [1] is based on a probabilistic neural network to classify leaves based on the 12 features describing the shape of a leaf. The authors achieved an accuracy classification of more than 90% for 32 plant species growing in China. J.-X. Du et al. [2] also extracted features from the leaf shape, but a new method is proposed for classification, referred to as a hypersphere classifier. The authors compare this method to the k-NN method for classifying 20 plant species. The study of Prof. M. Kumar et al. [3] is focused on various classification techniques for plant classification task. The authors have concluded that the simplest method is the k-NN classifier, the main disadvantage of which is sensitiveness to noise.

Various ways of solving the classification problem on one dataset are presented in the study by H. Goëau et al. [4]. The 55 species of plants in a dataset are represented by both scans and photos with a natural background. Participants from different countries presented their ways of solving the problem, the most popular solution was the shape boundaries analyses. At the same time, the authors note the perspectives of using metadata, in particular, geo-tags. The best result was achieved by INRIA [5], the authors of which used a contour-based shape descriptor called Directional Fragment Histogram. The essence of the method is to represent the leaf shape as groups of elementary components having the same direction. The Swedish tree leaf dataset [6][7], was used to implement and test the descriptor. The mentioned dataset was also partially used in this study. B. Wang et al [8] also investigate the leaf shape focusing on the convexity and concavity properties of the leaf arches as the major features. The achieved accuracy is estimated more than 96% for the Swedish tree leaf dataset. A similar approach was used by the authors of the application LeafSnap [9] to identify leaves of the 184 tree species of the Northeastern United States. The dataset [10] collected specifically for this task is estimated as the largest leaf dataset for today. The curvature-based shape descriptor is used for extracting features, high results are obtained within the top 5 results shown to the user. For the correct work of the application, the authors implemented a verification of the image uploaded by the user. The method of removing the petiole described in the study was applied in this work for the correct extraction of features.





A research by P. Novotný and T. Suk [11] suggests applying a Fourier descriptor to a leaf shape. The accuracy of more than 88% was obtained on a dataset called Middle European Woody Plants containing 153 species [12]. The analyses of the distances from centroid to boundaries described in this paper was also presented in a study by J. Chaki and R. Parekh [13]. The feature vector was obtained by describing the 36 radii and the evaluation method differs from the one presented in this article. The classification of 3 plant species is described in the paper, which does not allow to fully evaluate the efficiency of the method. Despite the fact that most authors do not consider leaf color analysis a reliable method of classification, T. Munisami et al. [14] included the color histogram as one of the features, resulting in a classification accuracy of more than 87% for 32 species of plants. Accuracy of more than 97% was obtained on the same dataset by implementing leaf venation analyses described by K.-B. Lee and K.-S. Hong

[15], although the study [9] noted that most mobile phone cameras are unable to detect leaf venation. The leaf texture analysis is presented in a recent study by Vijayashree T. and A. Gopal [16] with an obtained accuracy of 89% for 50 images.









### 3 Collecting Data







The dataset<sup>1</sup> used in this research was formed basing on the following publicly available datasets: LeafSnap [10], MEW 2012 [12] and the Swedish tree dataset [7]. The choice of species was made regarding their specificity for the territory of Russia. Although species with different leaf configurations are present in the list, the leaf image dataset consists only of simple leaves and terminal leaflets of compound leaves, as suggested in the MEW 2012 dataset. The species considered in this article are listed in Table 1.

Table 1: Tree species represented in the dataset (# stands for the number of images in the dataset)

Latin name	Example Image	#	Latin name	Example image	#
Acer platanoides		119	Ilex aquifolium		73
Aesculus hippocastanum		63	Populus nigra		75

<sup>1</sup> <https://goo.gl/3zBxzR>

Anus incana		62	Populus tremula		64
Betula pendula		62	Prunus padus		68
Betula pubescens		127	Quercus robur		101
Corylus avellana		55	Salix alba		75

Crataegus monogyna		66	Syringa vulgaris		62
Fraxinus excelsior		60	Tilia cordata		134
Ginkgo biloba		79	Ulmus laevis		59

To ensure the correct work of the assumed application and to avoid errors during the processing of images from the dataset, a binary classifier was created to check whether the image meets the following criteria:

- one leaf must be present on the image;
- at least 10% of the image area must be occupied by a leaf;
- a leaf should not touch the image borders (except for thin parts, such as petiole);
- the image must be taken on a light and neutral background.

An image which doesn't meet any of these criteria will not be proceeded for further evaluation unless it contains more than one object: then the algorithm is applied to the largest object on an image.

## 4 Image Processing

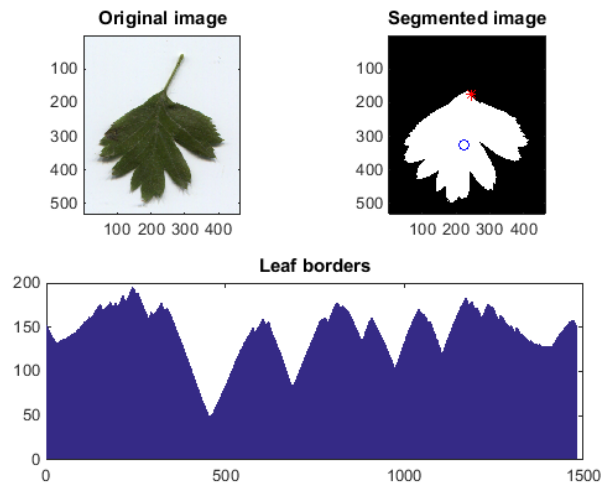
To retrieve the characteristics, the image that was successfully verified is passed through several processing steps. First, the threshold binarization using the Otsu method is applied to the image [17]. This allows us to highlight the objects on the image. The next step is to eliminate the noise, since small objects are present in most in-situ photos of leaves, not excluding the images used in this research as part of the dataset.

In most cases, photos of leaves contain petiole, since its removal requires additional manipulation. However, when analyzing the contour of the leaf, the petiole can seriously affect the extracted characteristics, for example, eccentricity or convex hull. The solution to this problem is to remove the petiole during the image processing. For this, the top-hat operation is applied to the binary image. The top-hat transformation of a binary leaf image ( $L$ ) is  $L$  minus its opening:

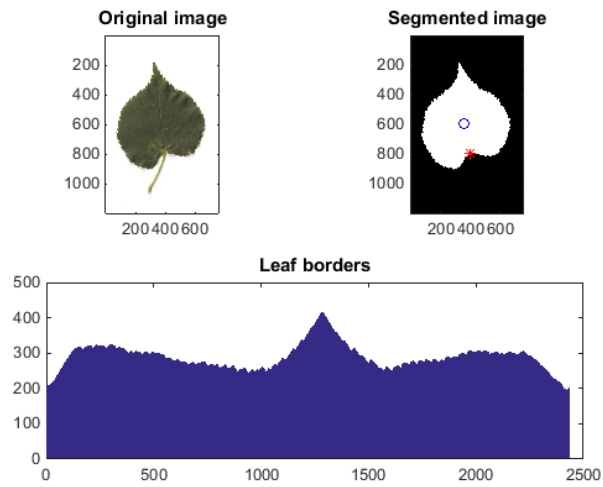
$$T_{hat}(L) = L - (L \circ SE), \quad (1)$$

where  $L \circ SE$  is an opening operation of  $L$  by structuring element  $SE$ , defined as the erosion of  $L$  by  $SE$ , followed by a dilation of the result by  $SE$  [18]. This method was proposed in [9] and is effective for most images of leaves. The shape thus obtained corresponds to the shape of a leaf plate without a petiole and can be used for further analysis. In case the petiole is absent on the image, the longest object remained after the top-hat operation will be removed from the image.

To evaluate the primary characteristic of a leaf, we search for centroid of the obtained shape, and then the coordinates of the points on the contour boundaries. The distances from the centroid to each of these points form a one-dimensional array, which we investigate for further feature extraction. However, since leaf position and orientation on an image may vary, in order to obtain informative data, it is necessary to start calculating the distances at the same point for all leaves. The most convenient point in this case may be the base of the petiole, which we removed, so the area occupied by the petiole is examined for proximity to the main object. The closest point is considered the base of the petiole and the distance to it is set as the first during the formation of a one-dimensional array of distances (AOD). If more than one point of the petiole border with the main figure (petioles on most leaf images have a thickness of several pixels), the first one found is considered the petiole base. The algorithm is thus invariant to leaf rotation. The processed image and the bar graph of AOD are shown in Fig. 1 and Fig. 2.



**Fig. 1.** Processing of a leaf of *Crataegus monogyna*



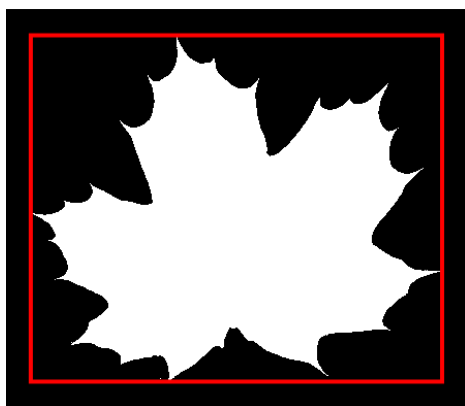
**Fig. 2.** Processing of a leaf of *Tilia cordata*

## 5 Extracting features

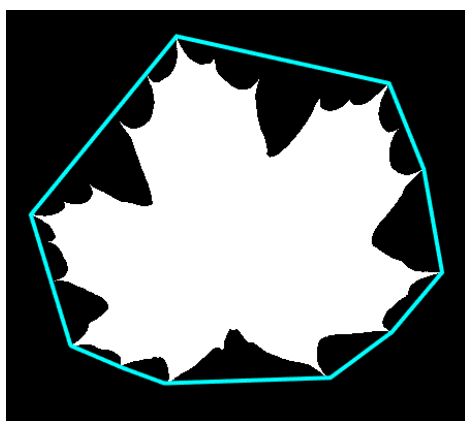
All the retrievable features are invariant to rotation. They can be divided into two groups: the proposed in this paper features extracted from AOD, and described

by many authors image moments obtained after analyzing the binary image of a leaf after petiole removal. The following features were used for building the learning model:

1. Eccentricity of the ellipse that has the same second-moments as the leaf shape.
2. Extent: ratio of area of the leaf to the smallest rectangle (bounding box) containing the leaf (as shown in Fig. 3).
3. Solidity: ratio of area of the leaf to its convex hull (see Fig. 4).
4. Diameter equivalent: the diameter of a circle with the same area as the leaf.
5. Ratio of the leaf area to a circle with a radius of minimal centroid-boundary distance (shown in Fig. 5).

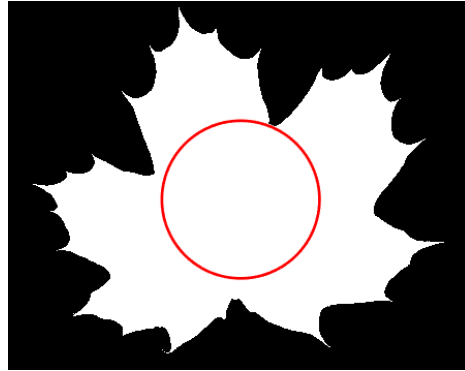


**Fig. 3.** Bounding box of a leaf



**Fig. 4.** Convex hull of a leaf





**Fig. 5.** Feature #5

6. Expectation: mean value of AOD
7. Variance of AOD
8. Median of AOD
9. Mode of AOD
10. Vertical symmetry: ratio of areas of leaf halves divided vertically.
11. Horizontal symmetry: ratio of areas of leaf halves divided horizontally.
12. Minimal distance: ratio of the minimal value of the AOD to its mean value.
13. Maximal distance: ratio of the maximal value of the AOD to its mean value.
14. Length ratio: ratio of length of the AOD to its maximum.  
For features 15-22 local maximums and minimums were analyzed (see Fig. 6).
15. Number of peaks: number of local maximums of the AOD.
16. Peak width: mean of peak width of the AOD.
17. Peak prominence: mean of peak prominence of the AOD.
18. Minimal peak: the minimal value in the array of the local maximums of the AOD.
19. Number of valleys: number of local minimums of the AOD.
20. Valley width: mean of valley width of the AOD.
21. Valley prominence: mean of valley prominence of the AOD.
22. Maximal valley: the maximal value in the array of the local minimums of the AOD.

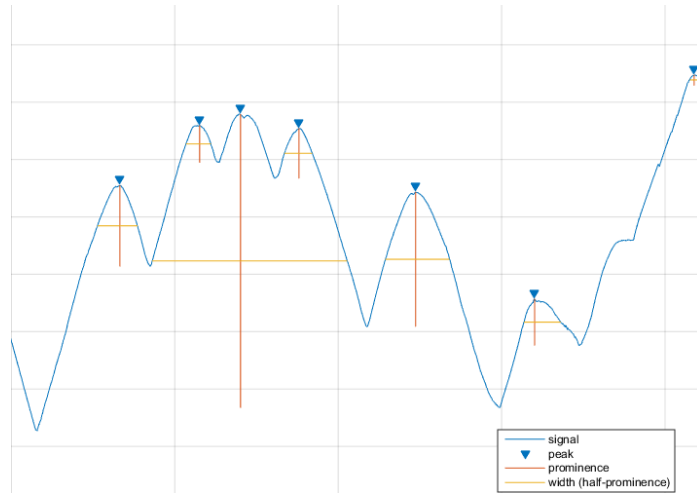


Fig. 6. Peaks, peak width and peak prominence (AOD of a Quercus robur leaf)

## 6 Results

1-	28	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2-	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3-	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
4-	0	0	0	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5-	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	
6-	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	
7-	0	0	1	0	0	0	15	0	0	0	0	1	0	0	0	0	0	0	
8-	0	0	0	0	0	0	0	17	0	0	0	0	1	0	0	0	0	0	
9-	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	
10-	0	0	0	0	0	0	0	0	0	15	1	0	0	0	0	2	0	0	
11-	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	
12-	0	0	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	
13-	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	
14-	0	0	0	0	0	0	1	0	0	0	0	0	0	12	0	0	0	0	
15-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	
16-	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	29	0	0	
17-	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	11	0	
18-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Fig. 7. Confusion matrix for classification results of the testing set

The k-nearest neighbors classifier was implemented, with the dataset divided into training and testing sets in a 3:1 ratio. A  $k = 1$  was chosen based on the model performance. The number of extracted features was reduced to 22 according to the achieved results, as many features extracted during the study proved to be uninformative. A learning model based only on the features extracted from the AOD showed the classification accuracy up to 90,5% on the testing set, while using only the features 1-5 allowed to obtain an accuracy of 80,4%. Using a complete set of features allows to achieve 95.5%, depending on the partitioning of the dataset, the confusion matrix for this result is shown in Fig. 7. For the Swedish tree dataset [7] the accuracy of 94% is obtained for the testing set. Other machine learning algorithms, including random forest algorithm and bagging, allowed to achieve the same results.

## 7 Conclusion and Future Work

In this article, a method for extracting features for classifying trees based on a leaf image has been described. A dataset was formed based on the images from MEW 2012, LeafSnap the Swedish leaf dataset. The developed verification algorithm allows to exclude errors during image processing. For the 18 classes considered, the classification accuracy of more than 95% was obtained with a 1-NN method, based on 22 features. The method was also applied to the Swedish tree dataset and showed the accuracy of 94%. The learning model has showed resistance to increasing the number of classes. The considered method can be used in combination with the previously described methods of leaf image analysis to develop an application focused on a larger number of tree species.

The additional features may be extracted from the AOD to achieve better results on a larger dataset, for example, the fast Fourier transform may be applied to the AOD. It is also expected that the Russian species that are not represented in the existing datasets will be added after a collaboration with the botanists.

## References

1. Wu, S.G., Bao, F.S., Xu, E.Y., Wang, Y., Chang, Y., Xiang, Q.: A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network. In: IEEE International Symposium on Signal Processing and Information Technology, pp. 11-16 (2007).
2. Du, J.X., Wang, X.F., Zhang, G.J.: Leaf shape based plant species recognition. *Applied Mathematics and Computation* 185, pp. 883-893 (2007).
3. Kumar M., Kamble M., Pawar S., Patil P., Bonde N.: Survey Techniques for Plant Leaf Classification. *International Journal of Modern Research (IJMER)*, Vol.1, Issue.2, pp-538-544 (2011).
4. Goëau, H., Bonnet, P., Joly, A., Boujemaa, N., Barthelemy, D., Molino, J.F., Birnbaum, P., Mouysset, E., Picard, M.: The CLEF 2011 plant images classification task. In: CLEF (Notebook Papers/Labs/Workshop) (2011).
5. Yahiaoui, I., Hervé, N., Boujemaa, N.: Shape-based image retrieval in botanical collections. In: *Advances in Multimedia Information Processing - PCM 2006*, vol. 4261, pp. 357-364 (2006).

6. Swedish Leaf Dataset, <http://www.cvl.isy.liu.se/en/research/datasets/swedish-leaf/>, last accessed 2017/04/20.
7. Söderkvist, O.J.O.: Computer Vision Classification of Leaves from Swedish Trees. Master's thesis, Linköping University, SE-581 83 Linköping, Sweden (2001).
8. Wang, B., Brown, D., Gao, Y., La Salle, J.: Mobile plant leaf identification using smart-phones. In: 2013 20th IEEE international conference on image processing (ICIP), pp 4417-4421 (2013).
9. Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V.B.: Leafsnap: A Computer Vision System for Automatic Plant Species Identification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 502-516. Springer, Heidelberg (2012).
10. Leafsnap Dataset — Leafsnap: An Electronic Field Guide, <http://leafsnap.com/dataset/>, last accessed 2016/09/15.
11. Novotný, P., Suk, T.: Leaf recognition of woody species in central europe. *Biosyst Eng* 115(4):444-452 (2013).
12. Download Middle European Woods (MEW 2012, 2014) — Department of Image Processing, <http://zoi.utia.cas.cz/node/662>, last accessed 2017/01/15.
13. Chaki, J., Parekh, R.: Plant Leaf Recognition using Shape based Features and Neural Network classifiers. In: *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 2, no. 10, pp. 41-47. (2011).
14. Munisami, T., Ramsurn, M., Kishnah, S. and Pudaruth, S.: Plant leaf recognition using shape features and colour histogram with k-nearest neighbour classifiers. In: *Procedia Computer Science*, vol. 58, pp. 740-747 (2015).
15. Lee, K.-B., Hong, K.-S.: An Implementation of Leaf Recognition System using Leaf Vein and Shape. In: *International Journal of Bio-Science and Bio-Technology*, vol. 5, No. 2, pp. 58-66 (2013).
16. Vijayashree, T., Gopal, A.: Authentication of Leaf Image Using Image Processing Technique. In: *ARPJN Journal of Engineering and Applied Sciences*, vol. 10, No. 9, pp. 4287-4291 (2015).
17. Otsu, N.: A threshold selection method from gray level histograms. *IEEE Trans. Systems, Man and Cybernetics* 9, pp. 6266 (1979)
18. Gonzalez, R., Woods, R.: *Digital image processing*. Pearson/Prentice Hall (2008)