GRAPH-BASED EXTRACTION OF SHAPE FEATURES FOR LEAF CLASSIFICATION

Andrew Zhao, Denis Tsygankov, Peng Qiu

Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology, Atlanta, GA, USA
Email: peng.qiu@bme.gatech.edu

ABSTRACT

Conventional approaches to feature extraction based on shape typically focus on using mathematical descriptors such as aspect ratio, rectangularity, area, etc., or the Centroid Contour Distance Curve (CCDC) to describe the shape based upon edge points. Such descriptors are able to achieve modest success when classifying leaf species with example-limited datasets. In this paper, we propose the use of Medial Axis Transformation (MAT) to convert the complex shapes of leaves to graph structures that describe the topological skeleton of the leaves, and utilize the topological skeleton to extract features. From a data set composed of 99 different leaf classes with 10 samples per class, we show that 18 features extracted from the topological skeleton outperforms 64 features obtained from CCDC with '10-fold' cross validation accuracies of 73.84% and 58.08% respectively when using random forest as a classifier. This result suggests that MAT-graph-based features are able to more succinctly distinguish shapes when compared with conventional approaches.

Index Terms — Medial Axis Transformation, Topological Skeleton, feature extraction, leaf classification

1. INTRODUCTION

Classifying and cataloguing plants is a tedious process that requires trained experts to ensure proper identification. However, this task is necessary to properly catalogue plant species to track the presence of invasive species and for use in the agricultural management of crops. An automated process would be essential to accurately catalogue plants and help manage invasive populations to prevent their spread. Leaves can serve as a prime source of identification and are an ideal source of features for machine classification due to their ubiquity, ease of collection, and unique identifying features.

Previous work has been done to classify leaves using leaf margins [1] and veins [2] with common classifiers such as linear discriminate analysis, KNN, and random forest. Other works used probabilistic neural networks [3] and neural network based ontological approaches [4]. However, recent results from a Kaggle competition have shown that commonly used machine learning algorithms such as random forest and logistic regression can achieve 100% accuracy without the use of neural networks that may require extensive training. This places a high emphasis on proper feature extraction.

Shape based features have been extracted based upon the edge through mathematical descriptors or Centroid Contour Distance Curve (CCDC). Du et al. [5] used various mathematical descriptors such as aspect ratio, rectangularity, area ratio of convex hull, perimeter ratio of convex hull, sphericity, etc. to create features. However, these features do not completely describe the shape and certain descriptors are sensitive to translation and rotation. An orientation in-variant approach utilizes the CCDC to describe the shape by starting at an arbitrary point along the edge and obtaining the distance from the point to the centroid, from which features can be extracted [4], [6]. Mallah et al. [7] utilized the CCDC to describe leaf shape by interpolating and down scaling the sampling to 64 points. Combined with other features for leaf texture and margin, they found that shape (~62%) was the poorest predictor compared to texture (~72%) and margin (~75%) for a data set of 100 different species with 16 samples of each species, leaving room for improvement in shape-based feature extraction.

2. FEATURE EXTRACTION

We use the data set and the image set from Mallah et al. [7] as a framework to compare our extracted features to those obtained using the CCDC. In this paper, we perform Medial Axis Transformation (MAT) to convert complex shapes of leaves to graph structures that describe their topological skeletons [8], and extract robust features from the graphs as shape descriptors.

2.1 Data Set

All image data sets were obtained from the Kaggle leaf classification challenge (https://www.kaggle.com/c/leaf-classification), which used the image set collected by Mallah et al. [7]. The image sets were composed of binary images of varying sizes and also included 192 pre-extracted features

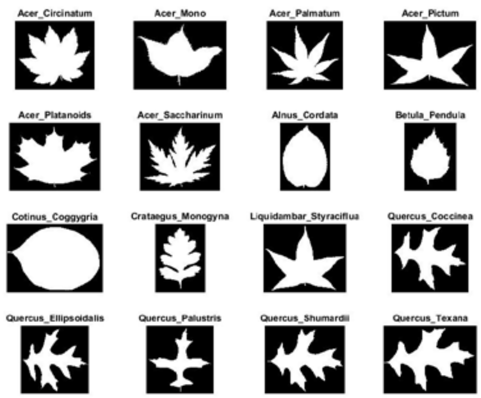

Figure 1. A small subset of different leaf species within the dataset, one image from each species, illustrating both the similarity and differences in appearance of leaves.

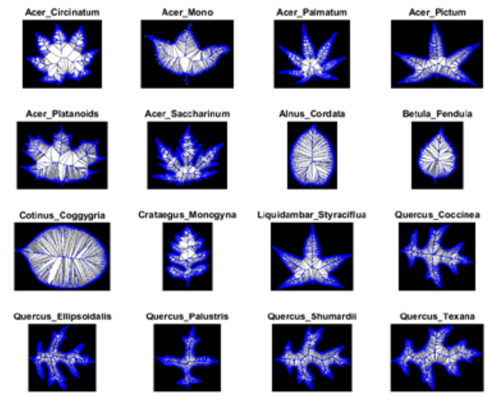

Figure 2. MAT-based graph representation of the images shown in Figure 1.

based upon shape, margin and texture, each contributing 64 of the 192 pre-extracted features [7]. Since the images are only available in binary, only the shape-based features can be re-created from the image sets. The image sets contained a training set of 99 different species with 10 examples of each species. An example of the diversity and similarity of the species is shown in Figure 1.

2.2 Medial Axis Transform

Medial axis are formed by a set of points that are equidistant from at least two edge points. This concept was originally presented by Blum to characterize biological shapes [9]. MAT specifically is defined as a set of centers of inscribed circles that are tangential to the edge at least at two points along the boundary of a closed shape. MAT are commonly used for shape identification and are sensitive to small perturbations of the edge, and as a result, many MAT-based algorithms included pruning to remove excess branches [10], [11]. Although the robustness and sensitivity needs to be carefully considered, MAT has the potential to derive a comprehensive representation of complex shapes with minimal information loss. For example, Tsygankov et al. [8] developed a computational framework, CellGeo, which takes advantage of this sensitivity to track cell protrusions and shape changes during cell migration and morphogenesis. Given the points on the edge of a cell forming a closed polygon, CellGeo combines Medial Axis and Voronoi transformations, and constructs a directed graph to connect the medial axis (internal points) generated by the transformations, as well as each point representing the edge of the shape. This process creates a directed tree graph, with the root at the center of the largest circle that can be inscribed in the shape, and the terminal nodes corresponding to the edge points. This tree graph is able to capture shape feature of any scale, including fine structures such as narrow protrusions of the cells membrane as they form, extend, and retract during cell migration and morphogenesis [8].

Motivated by CellGeo's ability of tracking arbitrarily complex shapes of cells, we used the shape-to-graph transformation in CellGeo to convert the binary images of leaves to tree graphs. Similar to the cell shape analysis, the tree graph generated from a leaf contained both edge points and internal points. The structure of the tree graph approximated the topological skeleton of the leaf, which in some cases resembled the veins of the leaf, although the graph was derived solely based on binary images that only contain shape information. Examples are shown in Figure 2.

2.3 Graph-Based Feature Extraction

The MAT-based graph is used to establish three sets of features. The first set of features are obtained by binning graph nodes based upon their distance to the edge. For each node corresponding to an internal point, we compute the graph distances to edge points reachable from this internal point. The shortest distance (distance to the closest reachable edge node) is defined as the distance of this internal point to the edge. For each leaf, we computed the distance-to-edge for all internal points, summarized the distribution into a 6-bin histogram, and normalized by the total number of edge points. The normalized counts in each bin are used as features. Example histograms are shown in Figure 3. These histograms described how internal points are distributed with respect to the edge, and show significant difference among many of the 99 classes of leaves.

The second set of features are obtained based upon the graph distances between the edge points and the root node. We ordered the edge-to-root distances by tracing the contour, and generated graph-based contour profiles similar to the CCDC [5]. One example is shown in Figure 4. Figure 4a shows the binary image of a leaf, overlaid with the MAT-

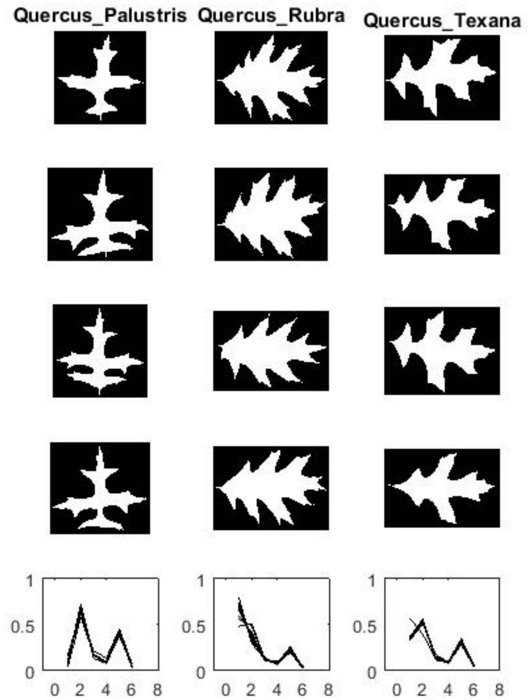

Figure 3. The distribution of internal nodes with respect to distance-to-edge. Example images of three related species, and the histograms of distances-to-edge of internal nodes. Each bottom panel contains 10 histograms corresponding to all 10 images belonging to the species. The histogram is highly consistent within species, and show decent difference even for related species.

based graph representation. Figure 4b shows the graph-based contour profile. Using graph-distance rather than Euclidian distance allows for a more meaningful descriptor that matches the geometry of the leaf. From the graph-based contour profile, we defined 6 features (max, min, max/min, sum, std, mean/max), which described the major and minor axes, aspect ratio, size, and ruggedness of the leaf.

The third set of features are derived from thresholding the graph-based contour profiles, describing large and small leaflets based upon graph-distance from the root node. We applied quartile thresholds to the graph-based contour profiles. At each threshold level, we counted the number of times that graph-based contour profiles crossed over the threshold, and computed the area under the parts that exceed the threshold. The number and area of leaflets at each threshold are used as features. In the example in Figure 4b, the horizontal lines show the quartile thresholds. In the top panel of Figure 4c, with the lowest quartile threshold, the 4 colored pieces of the leaf exceeded the threshold. In the middle panel with a more stringent threshold at the second quartile, the leaf still has 4 leaflets but smaller area. With the most stringent quartile, the number leaflet changed to 5, corresponding to the tips that most distant from the center of the leaf. These quartile thresholds allowed for definition of features at different scales.

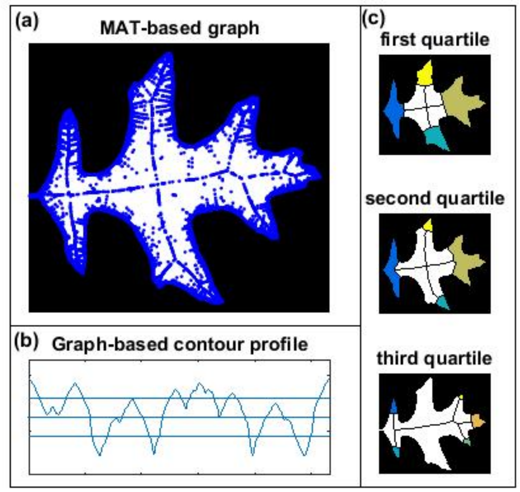

Figure 4. Thresholding of graph-based contour profile defines features at different scales. (a) A *quercus coccinea* leaf overlaid with MAT-based graph. (b) Graph-based contour profile with three levels of thresholds. (c) Areas of the leaf that exceed the thresholds are highlighted with different colors. Lower threshold levels highlight larger leaflet, while higher thresholds highlight smaller protrusions of the leaf.

In total, we derived 18 features from the MAT-based graph representation of the leaf shapes. 6 were obtained from the point distribution in reference to their distance from the edge. 12 were derived from the graph-based contour profiles.

3. RESULTS

Using the Random Forest classifier, we compare the features extracted from the MAT-based graph representation to the shape, margin, and texture features obtained by Mallah et al. [7]. The results of '10-fold' cross-validation of each set of features alone and together are shown in Table 1.

The shape features obtained using CCDC achieved a 58.08% accuracy, while the shape features derived from the MAT-based graph attained a 73.84% accuracy. Since we are considering a 99-class classification problem, this improvement in accuracy is non-trivial. The improved accuracy suggests that the features derived from the MAT-based graph were more meaningful than those derived from CCDC.

Comparing the shape, texture and margin features considered in the original analysis of this dataset by Mallah et al [7], shape attained the lowest accuracy, while texture and margin were significantly more accurate. This is consistent with the previous results. Although the CCDC-based shape features performed relatively poorly, combing with the texture and margin features achieved classification accuracy of 97.47%, indicating that texture and margin features are more informative in classification of leaves. When the shape features derived from MAT-based graphs were combined with texture and margin, with or without CCDC-based shape features, the performance accuracy improved by only 1~2%.

Overall, when the texture and margin features are included, the MAT-graph based shape features do not provide significant improvement over the CCDC shape features in leaf classification. However, in applications where the shape features are the most important and informative, the MAT-graph based shape features (accuracy 73.84%) are more informative than the CCDC-based shape descriptors (accuracy 58.08%). One such example is our ongoing project on characterizing complex shapes of cell mesh during vascular tube formation, where the imaging data does not provide enough resolution to accurately quantify the margin features [8].

Features	Accuracy (%)
CCDC Shape (Mallah)	58.08
MAT-based graph	73.84
Margin (Mallah)	82.93
Texture (Mallah)	79.70
Margin, Texture	95.76
CCDC, Margin, Texture	97.47
MAT-based graph, Margin, Texture	98.38
All Features	99.19

Table 1. Classifications results of '10-fold' cross-validation of the leaf dataset using different combinations of features obtained from MAT-based graph representation and those from Mallah et al. [7]

4. CONCLUSION AND FUTURE WORKS

Here we described shape features that can be defined using the MAT-based graph representation. Using a binary image set of leaves from 99 classes, we showed that shape features defined by the MAT-based graph representation resulted in significantly higher classification accuracy compared to shape features derived from a more conventional CCDC approach. In the context of leaf classification, when more informative features (texture and margin) are present, the MAT-based graph representation does not provide significant benefit over conventional approach. However, in other applications where the primary focus is the shape features, the MAT-graph based shape features are more informative than the CCDC-based shape descriptors.

This paper only considered 18 features derived from the MAT-based graph representation, which of course do not fully capture the richness of information encoded in the MAT-based graph. There are more possible features beyond the ones considered here, such as those derived from spectral analysis. Moreover, if one can develop a distance metric to directly and robustly evaluate the similarly between two MAT-based graphs, the distance metric can be used as the kernel of many machine learning algorithms that lead to higher shape-based classification accuracy.

5. ACKNOWLEDGEMENT

This work is supported by the Marcus Center for Therapeutic Cell Characterization and Manufacturing collaborative grants, and the National Science Foundation grant CCF 1552784.

6. REFERENCES

[1] C. Kalyoncu and Ö. Toygar, "Geometric leaf classification," *Comput. Vis. Image Underst.*, vol. 133, pp. 102–109, Apr. 2015.

[2] M. G. Larese, R. Namías, R. M. Craviotto, M. R. Arango, C. Gallo, and P. M. Granitto, "Automatic classification of legumes using leaf vein image features," *Pattern Recognit.*, vol. 47, no. 1, pp. 158–168, Jan. 2014.

[3] S. G. Wu, F. S. Bao, E. Y. Xu, Y.-X. Wang, Y.-F. Chang, and Q.-L. Xiang, "A leaf recognition algorithm for plant classification using probabilistic neural network," in *Signal Processing and Information Technology, 2007 IEEE International Symposium on*, 2007, pp. 11–16.

[4] Hong Fu, Zheru Chi, Dagan Feng, and Jiatao Song, "Machine learning techniques for ontology-based leaf classification," 2004, vol. 1, pp. 681–686.

[5] J.-X. Du, X.-F. Wang, and G.-J. Zhang, "Leaf shape based plant species recognition," *Appl. Math. Comput.*, vol. 185, no. 2, pp. 883–893, Feb. 2007.

[6] Z. Wang, Z. Chi, D. Feng, and Q. Wang, "Leaf image retrieval with shape features," in *International Conference on Advances in Visual Information Systems*, 2000, pp. 477–487.

[7] C. Mallah, J. Cope, and J. Orwell, "Plant Leaf Classification using Probabilistic Integration of Shape, Texture and Margin Features," 2013.

[8] D. Tsygankov, C. G. Bilancia, E. A. Vitriol, K. M. Hahn, M. Peifer, and T. C. Elston, "CellGeo: A computational platform for the analysis of shape changes in cells with complex geometries," *J. Cell Biol.*, vol. 204, no. 3, pp. 443–460, Feb. 2014.

[9] H. Blum, "A Transformation for Extracting New Descriptors of Shape," in *Models for Perception of Speech and Visual Form*, MIT Press, 1967, pp. 362–680.

[10] O. Aichholzer, W. Aigner, F. Aurenhammer, T. Hackl, B. Jüttler, and M. Rabl, "Medial axis computation for planar free–form shapes," *Comput.-Aided Des.*, vol. 41, no. 5, pp. 339–349, May 2009.

[11] H. Tek and B. B. Kimia, "Symmetry maps of free-form curve segments via wave propagation," *Int. J. Comput. Vis.*, vol. 54, no. 1–3, pp. 35–81, 2003.