A Novel Shape based Descriptor for Plant Identification

Komal Asrani
Department of Information Technology,
B.B.D.E.C.
Lucknow

Renu Jain
Department of Computer Science & Engineering
U.I.E.T.,
Kanpur

ABSTRACT
Content based image retrieval is one of the most challenging field in which widespread research is been carried out. In this paper, a modified and effective shape based leaf image retrieval system is presented for leaf identification. The method proposed is an extension of centroid distance method. The major drawbacks of the centroid distance approach are identified and resolved in our proposed approach. This approach, called Quad Centroid Distance Variation (QCDV) uses the concept of shape based image retrieval approach. Four values are computed and used as the feature vectors. The proposed approach effectively manages the size of the feature vectors, is computationally simple and affine invariant. Experiments were conducted on Swedish Leaf Image Database (SLID). The results justify the superior performance of the method even though the feature vector representing the image is very small.

General Terms
Image Retrieval, Pattern recognition

Keywords
Shape retrieval, plant identification, leaf recognition, image processing, Centroid distance.

1. INTRODUCTION
Plant identification has been an active and interesting field of interest for botanical based industry people and bio diversity conservation groups. This is because of the huge amount of plants (approximately 2,60,000) existing all around the world. Many plants are on the verge of extinction. So there is an urgent need to maintain the bio diversity database by managing the complete details of the plants. The identification of these plants presents a crucial scenario because of the complex techniques and terminology of cell biology, molecular biology and phytochemistry that forms a gap between the biologists and the layman. Also, the lack of availability to define a detailed and complete documentation of plants adds to the problem.

In the present scenario, enormous amount of development has been done in the field of image processing. Effective, advanced and sophisticated tools and techniques have been worked upon to ensure easy collaboration to bridge the existing gap to assist in easy plant identification process. Plant identification is based on various parts of plant like bark, stem, roots, flower, roots and leaf. Among all these parts, leaf is available in all seasons. Moreover, as leaves are defined as two dimensional, they can be easily captured for analysis purposes. As we are focusing on CBIR based on shape, the database created for leaves can be considered in the form of binary images instead of colored images. Various applications had been explored where the database is defined as black and white images like trademark, patent images, technical drawings, road signs, medical images. The pixels of these images are generated in the form of binary by defining a particular level of threshold for black and white images. Once the binary image is available, shape based image retrieval is done by extracting the contours which represents the shape of the image accurately. Some of the most important requirements for an effective shape based image retrieval system include confined yet complete feature vectors representation, invariance to geometric transformations, robust to noise and distortion, usefulness for wide applications.

Elaborate work has been done in the field of content based image retrieval for plant identification which use leaf for recognition purpose [1] [2][7][10]. Du et. al[11] extracted digital morphological features from he contour of the leaf which included geometrical and invariant moment features. The recognition process was done using move median centers (MMC) hypersphere classifier. Bottcher et. al.[14] presented a rather untypical approach CQQL (commuting quantum query language) and utilized the mathematical formalisms of quantum mechanics and logic eventually forming a probabilistic logic. This approach used color based low level features or GPS formula for recognition process. Huang et. al.[15] proposed computer aided plant species identification which was based on plant leaf images using a shape matching technique which used Douglas Pecker approximation algorithm to form a sequence of invariant attributes. Bylesja et. al.[16] suggested LAMINA as a tool for rapid quantification of leaf size and shape parameters. This tool used blade dimensions and area to study leaf shape and leaf serration traits. Yahiaoui et. al.[17] proposed directional fragment histogram to encode two kinds of information. At a local level, it coded the relative length of groups of elementary components. At a global level, it coded the elementary component frequency distribution. Wang et. al.[3] proposed to generate feature vectors using centroid-contour distance curve, eccentricity, angle code histogram and used fuzzy integral approach to combine the feature vectors. Im. et al. [5] used hierarchical polygon approximation representation of leaf shape to provide for identification of Acer plant variety. McLellan [8] used fractal dimension as single value descriptor to identify. Du et. al.[9] used polygonal approximation for representation of leaves. Belongie[4] used concept of Shape Context to identify the contours of the image. Adamek et. al[6] proposed multiscale convexity, concavity representation to understand the relative displacement of a contour point at different scales. Alajlan [12] used another multi-scale shape descriptor called triangle area representation (TAR).This approach used the boundary points to form triangle and the area for the corresponding triangle is computed for measuring the convexity/ concavity at different scales. Cerutti et. al.[13] proposed a didactic interaction with user, which evaluated high level characteristics and more generic shape features of the plant leaves under ImageCLEF Plant identification task.

Inspite of the elaborate existing work done in the field of plant identification, there is always a need to attain effectiveness with space limitation. Fulfilling these contradictory requirements poses great challenge for CBIR systems. However, an effort is done in order to introduce a novel technique for shape based image retrieval, which is referred to as Quad Centroid Distance Variation approach. This approach is based on contour based approach and provides concise representation of the image. In this approach, an effort is made to extend the centroid distance approach to effectively represent the image by removing the drawbacks identified in centroid distance approach. The major drawback in the centroid distance is the huge dimension of feature vectors, which represents the distances of contour points from centroid. This major drawback results in adding the time and space complexity. To resolve these problems, an extension of the centroid distance approach is made to provide concise yet complete representation. This is our contribution in this paper.

2. CENTROID DISTANCE METHOD

In this approach, the boundary points are identified representing the shape. The centroid for these points is calculated and is represented as (x_c, y_c). Then, the distances of the centroid from the boundary points are calculated using the equation 1.

$$d = \sqrt{(x - xc)^2 + (y - yc)^2} \quad (1)$$

The size of the array of the centroid distances is dependent on the number of points used for defining the boundary shape of the image. However, on an average, the array size is approximately 120-200.

3. QUAD CENTROID DISTANCE VARIATION (QCDV)

The motivation for introducing this approach is the drawback identified in the centroid distance method. This is resolved by incorporating strategies so that the size of the feature vectors can be reduced and thus the space and memory complexity can be reduced, thereby maintaining better levels of effectiveness. In this approach, the leaf image is processed to generate the boundary coordinates. The boundary points representing the complete image are represented as shown in Figure 1 where I={ P_1, P_2, P_3, P_4, P_5 ………. P_{n-1}, P_n}

Figure 1. Leaf (Binary Image)

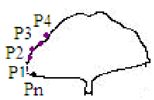

Figure 2.Leaf Image with boundary coordinates

where P_1 refers to the first boundary point defining the image. Here P_1 corresponds to specific value of (x, y) which defines the coordinate value in Cartesian system. Once the boundary coordinates are calculated, the centroid for the complete shape (Xc, Yc) is calculated using the equation (2) and (3):

$$Xc = \frac{1}{6A}\sum_{i=0}^{n-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \quad (2)$$

$$Yc = \frac{1}{6A}\sum_{i=0}^{n-1}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \quad (3)$$

Where

$$A = \frac{1}{2}\sum_{i=0}^{n-1}(x_i y_{i+1} - x_{i+1} y_i) \quad (4)$$

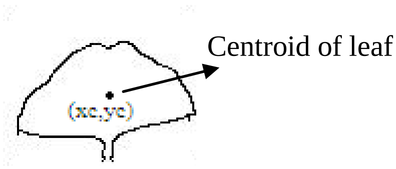

Figure 3.Leaf Image depicting the centroid of the shape

Then, the origin O of the coordinate system is shifted to the centroid of the leaf image which is defined as (x_c, y_c).

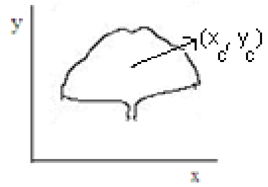

Figure 4.Leaf image in Cartesian System

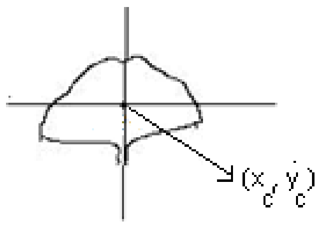

Figure 5. Leaf image with Centroid defined as origin

The leaf image is segmented into four quadrants with the centroid serving as the origin. Hereafter, the complete leaf image can be analyzed as four individual quadrants i.e. the image segment in each quadrant is considered individually and processed accordingly to compute values of coefficient of variation.

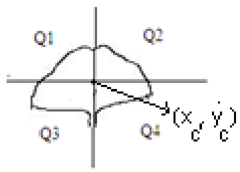

Figure 6.Leaf image with four quadrants defined

Coefficient of Variation: Also known as relative variability, it is a measure of dispersion of data points in a data series around the mean. It is calculated as:

$$\text{Coefficient of variation} = \frac{Standard\ Deviation}{Mean}\ X\ 100 \quad (5)$$

where standard deviation for a set of random values for a finite data set S is defined as :

$$S = \sqrt{\frac{\sum(X-\overline{X})^2}{N}} \quad (6)$$

where S is standard deviation of a sample, $\sum$ refers for "sum of", X refers for each value in the dataset, $\overline{X}$ is the mean of all values in the dataset and N is the number of values in the dataset.

Coefficient of variation is a useful statistical tool which is helpful for comparing one data series to another data series. Hence, coefficient of variation is computed from distances for each quadrant .

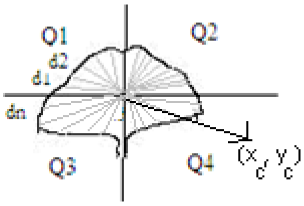

Figure7.Leaf image with distances computed for each quadrant

Algorithm:

1. Extract the boundary points of the contour which are represented as $D_{1=}\{(x,y)_1, (x,y)_2, (x,y)_3, (x,y)_4 \ldots\ldots\ldots (x,y)_n\}$.
2. Compute the centroid of these boundary points.
3. Referring the Cartesian coordinate system, relocate the centroid of the image so that it coincides with the origin of the Cartesian coordinate system.
4. Define the quadrants as Q1(left top), Q2(right top), Q3(left bottom) and Q4(right bottom) of the Cartesian coordinate system with centroid as origin. Hereby the image I is divided into four quadrants represented as

I= $\{Q_1, Q_2, Q_3, Q_4\}$
where Q_1 refers to the boundary points in first quadrant,
Q_2 refers to the boundary points in second quadrant and so on.

5. For each quadrant, identify the points of the image boundary belonging to each quadrant.
6. For each quadrant boundary points, calculate the distances from centroid of image.

$Q_{1=}\{ d_{11}, d_{12}, d_{13}, d_{14} \ldots\ldots\ldots\ldots d_{1n}\}$
$Q_{2=}\{ d_{21}, d_{22}, d_{23}, d_{24} \ldots\ldots\ldots\ldots d_{2n}\}$
$Q_{3=}\{ d_{31}, d_{32}, d_{33}, d_{34} \ldots\ldots\ldots\ldots d_{3n}\}$
$Q_{4=}\{ d_{41}, d_{42}, d_{43}, d_{44} \ldots\ldots\ldots\ldots d_{4n}\}$

7. For each quadrant, from these distance values, compute coefficient of variation and form feature vectors representing the complete image as

Image I=$\{Cov_1 , Cov_2, Cov_3, Cov_4\}$

The initial position of the centroid was at C and the distance of C from P is d.

$$d = \sqrt{(x-xc)^2 + (y-yc)^2} \quad (8)$$

If the object is shifted (Δx, Δy) along x- axis and along y-axis, then the point P will also be shifted by (Δx, Δy) to new location P_1. Then the coordinates of P_1 is defined as (x+Δx), (y+Δy). Then,

$$x_c^1 = \frac{m_{10}^1}{m_{00}}$$

$$y_c^1 = \frac{m_{01}^1}{m_{00}}$$

3.1 Properties of Quad Centroid Distance Variation

A shape based retrieval system is expected to have the general properties of translation, scale and rotation invariance. After performing some basic normalization, we are able to achieve these properties.

3.1.1 Translation Invariance Property

Though the computation for calculating distances is done with reference to the coordinates of the centroid, but as defined in the algorithm, the first important aspect identified is that once the centroid is computed, this point is later treated as origin for further calculations. So, even the image is translated by (Δx , Δy), the magnitude of the distances remains the same.

The object centroid (x_c,y_c) is defined as:

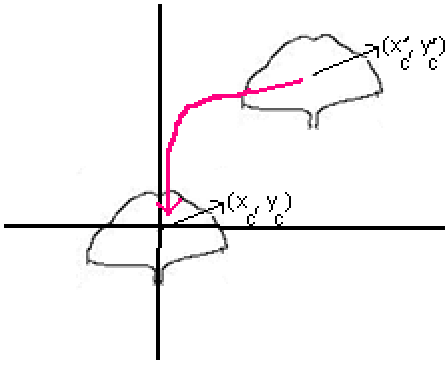

Figure 8.Leaf image translated from one specific location to origin

$$M_{pq}=\iint_R x^p y^q \ dx\, dy \quad (7)$$

$$x_c = \frac{m_{10}}{m_{00}} \quad , \quad y_c = \frac{m_{01}}{m_{00}}$$

$$x_c^1 = \frac{m_{10}^1}{m00^1} = \frac{\iint(x+\Delta x)dx\,dy}{\iint dx\,dy} = \frac{\iint(x)dx\,dy}{\iint \ dx\,dy} + \frac{\iint(\Delta x)dx\,dy}{\iint \ dx\,dy}$$

Hence, $x_{c1} = x_c + \Delta x$

When the initial position of the centroid was shifted by (Δx, Δy) to C_1 and then distance of C_1 from P_1 is defined as:

$$C_1P_1 = \sqrt{\{(x+\Delta x)+(x_c+\Delta x)\}^2 + \{(y+\Delta y)+(y_c+\Delta y)\}^2}$$

On simplification

$$C_1P_1 = \sqrt{\{x-x_c\}^2 + \{y-y_c\}^2}$$

$C_1 P_1 = C P = d$

Thus, from the above computation, we conclude that QCDV approach is translation invariant.

3.1.2 Scale Invariant

The distance for cent quad approach is always constant. For larger objects, there would be more points defining the boundary contour whereas for smaller objects, less number of points would be used to define the shape. For handling scaling aspect, the leaf image is processed before the feature vectors are calculated. Inspite of any dimension of the leaf image, the image is scaled to 100 pixels for square image and for other images, the dimension of the larger dimension of image is defined as 100 pixels and the image is resized by maintaining the aspect ratio property by using down-sampling or over-sampling technique

3.1.3 Rotation Invariant

This property defines that the values of the distances remains the same inspite of any orientation of the image. The property of rotation invariance can be managed effectively at the time of data acquisition or at the time of managing the database because the automatic identification of the tip and the base of the leaf is too difficult. For dealing with this problem, a convention needs to be followed such that the tip of the leaf is at the top and the base is at the bottom.

4. SIMILARITY MEASUREMENT

The feature vectors representing each leaf image is in the form of 4 values defining coefficient of variance for four quadrants. Thus, in total, there are four values representing the image. For measuring similarity, precision, recall and f- factor is identified. Precision is defined as follows:

$$Precision = \frac{Number\ of\ of\ true\ positives}{Number\ of\ true\ positives + false\ positives}$$

$$Precision = \frac{[\{relevant\ images\} \cap \{retrieved\ images\}]}{\{retrieved\ images\}}$$

Recall is defined as the ratio of ratio of number of relevant retrieved images to number of all relevant images.

$$Recall = \frac{[\{relevant\ images\} \cap \{retrieved\ images\}]}{\{relevant\ images\}}$$

$$Recall = \frac{Number\ of\ relevant\ retrived\ images}{Number\ of\ all\ relevant\ images}$$

F-measure is a measure that combines recall and precision. It is a harmonic mean of recall and precision and is defined as:

$$F\ measure = 2\ \frac{Precision.Recall}{Precision + Recall}$$

The recall, precision and f measure values are compared using Euclidean approach and the results are displayed with minimum distance from the query image as the best match.

5. EXPERIMENTS AND RESULTS

The experiments had been conducted on Swedish leaf image dataset (SLID). Before processing, each leaf is ensured to be kept upright. Then the edges representing the image are generated. The centroid of the image is computed and the image is translated with centroid at the origin. The distances from centroid are computed for each quadrant and coefficient of covariation is generated for each of the quadrant. The results generated by using QDCV are as follows:

Table 1. Leaf images with computed QCDV values

Leaf Image	Values of QCDV
	5.4189444, 13.4808, 4.292002, 11.392732
	23.636177, 6.265047,0.2206997,4.666754
	4.7570987, 7.6041913, 1.607199, 3.6954994
	0.17395079, 14.030577, 0.9699912, 2.6516309
	3.8530083, 10.080568, 2.630286,13.165314

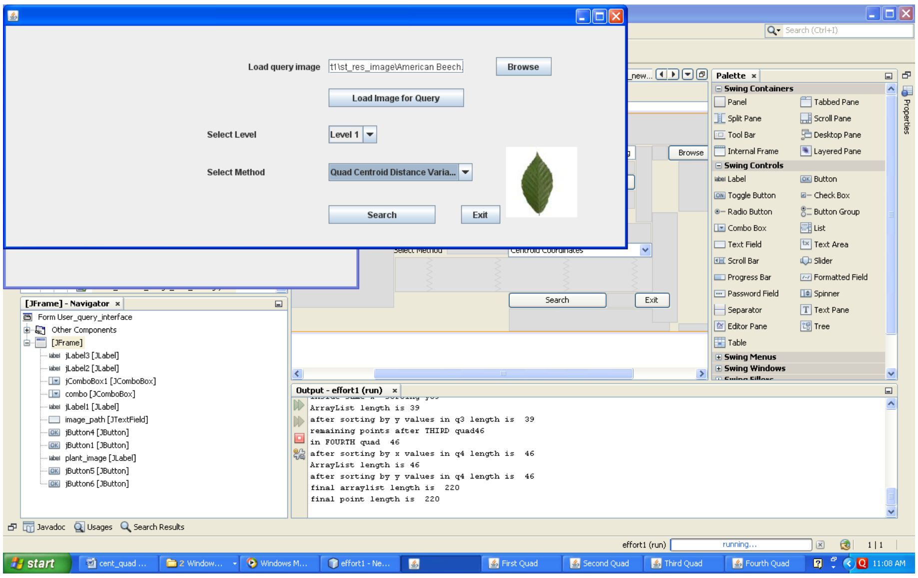

Figure 9. User Interface for providing query

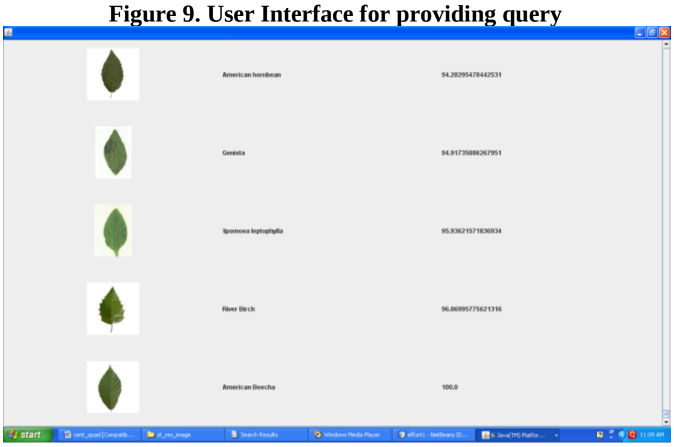

Figure 10. Results generated for query leaf with the leaf image, leaf name and matching percentage

Table 2. Precision, Recall, F-measure in %age for QCDV approach

Leafid	Precision(%)	Recall(%)	F-measure(%)
1	100	90	95
2	96	89	92
3	100	94	97
4	94	89	91
5	100	93	96
6	100	91	95
7	100	93	96
8	100	95	97
9	100	94	97
10	100	90	95
11	100	90	95
12	100	93	96
13	100	90	95
14	99	90	94
15	99	89	94
16	100	91	95
17	100	92	96
19	95	90	92
20	98	90	94

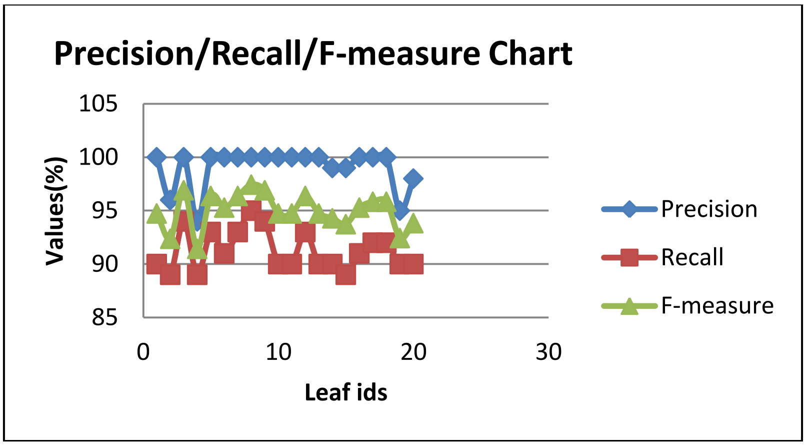

Figure 11. Precision/Recall/F-measure Chart

These results show that the proposed approach is able to provide good results and close to human perception. Analyzing the approach details, the size of the feature vectors used for defining the image are relatively very small as compared to centroid distance method. To add on, the computational complexity is reasonable when comparing the results. Inspite of the complexity of the image, every image is represented by 4 values and so there are no issues concerning matching of images with different dimensions and different array size. Moreover, the approach is made translation, scale and rotation invariant after proper normalization. Also, the computational complexity is also reduced as QCDV approach uses basic arithmetic operations for defining the feature vectors.

6. CONCLUSIONS AND FUTURE WORK

The paper reports automatic classification of plant species using plant images. The proposed approach QCDV is an extension of centroid distance method. QCDV effectively helps to represent the leaf image concisely by computing coefficient of variation for the four quadrants. It helps to overcome the large size of feature vector and provides simplified computation. The discussed approach is also closer to human perception. Presently, we have concentrated in measuring the effectiveness of our approach for different plants with distinct leaf shapes. The results have been presented in terms of recall, precision and F-measure. The results show an average precision, recall and f-measure of 99%, 91% and 94% respectively. The problem occurs when inter species similarity increases. In our future work, we would be looking into measuring its effectiveness for differentiating inter-species and intra-spices plants with respect to other content based image retrieval approaches based on shape.

7. REFERENCES

[1] S. Abbasi, F. Mokhtarian, and J. Kittler (1997) 'Reliable classification of chrysanthemum leaves through curvature scale space', in *Scale-Space Theory in Computer Vision* , volume 1252, pages 284–295. LNCS.

[2] J. Amores, N. Sebe, and P. Radeva(2007) 'Context-based object-class recognition and retrieval by generalized correlograms', *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 29(10):1818 –1833.

[3] Zhiyong Wang, Zheru Chi ; Dagan Feng (2002) 'Fuzzy integral for leaf image retrieval',Fuzzy Systems, *FUZZ-IEEE*'02. Proceedings of the 2002 IEEE International Conference on (Volume:1).

[4] Belongie, S., Malik, J., Puzicha, J(2002) ' Shape matching and object recognition using shape contexts', *IEEE Pattern Anal. Mach. Intell*. 24 509-522.

[5] C. Im, H. Nishida, T.L Kunii(1998) 'Recognizing plant species by leaf shapes—a case study of the Acer family', Proc.*Pattern Recognition* , Vol. 2, pp. 1171-1173.

[6] Adamek, T., O'Connor, N.E. (2004) 'A multiscale representation method for nonrigid shape with a single closed contour', *IEEE Trans. on Circuits and systems for video technology*. 14, 742-743.

[7] F. Mokhtarian and S. Abbasi(2004) 'Matching shapes with self-intersections: application to leaf classification', *IEEE Transactions on Image Processing* , 13(5):653 – 661.

[8] T. McLellan, J. A. Endler(1998) 'The relative success of some methods for measuring and describing the shape of complex objects', *Systematic Biology* 47, 264-281.

[9] J.-X. Du, D.-S. Huang, X.-F. Wang, X. Gu(2006) 'Computer-aided plant species identification (CAPSI) based on leaf shape matching technique', *Transactions Institute Of Measurement And Control* 28 ,275-284

[10] Z. Wang, Z. Chi, and D. Feng (2003) 'Shape based leaf image retrieval', *IEE Proceedings on Vision, Image and Signal Processing*, 150(1):34 – 43.

[11] J.-X. Du, X.-F. Wang, and G.-J. Zhang(2007) 'Leaf shape based plant species recognition', *Applied Mathematics and Computation*, 185(2):883 – 893.

[12] Alajlan, N., Rube, I.E., Kamel, M.S, Freeman, G.(2007) 'Shape retrieval using triangle-area representation and dynamic space warping', *Pattern Recognition* 40 1911-1920.

[13] Guillaume Cerutti, Violaine Antoine, Laure Tougne, Julien Mille, Lionel Valet, Didier Coquin, Antoine Vacavant(2013) 'ReVeS Participation - Tree Species Classification Using Random Forests and Botanical Features', *ImageCLEF* 2013.

[14] Thomas Böttcher, Sascha Saretz(2013)' BTU DBIS at *ImageCLEF*2013 Plant Identification Task'.

[15] D.S.Huang, X.-P. Zhang, G. –B. Huang(2005) 'Shape matching and recognition base on Genetic algorithm and

application to Plant Species Identification', *ICIC* , Part I, LNCS 3644, pp. 282-290.

[16] Max Bylesjö, Vincent Segura, Raju Y Soolanayakanahally, Anne M Rae, Johan Trygg, Petter Gustafsson, Stefan Jansson and Nathaniel R Street (2008) 'LAMINA: a tool for rapid quantification of leaf size and shape parameters', *BMC Plant Biology* , 8:82 doi:10.1186/1471-2229-8-82.

[17] Yahiaoui Itheri, Herve Nicolas, Boujemaa Nozha(2006) 'Shape Based Image Retrieval in Botanical Collections', PCM , *LNCS* 4261, pp. 357-364.