IJCTA, 8(3), 2015, pp. 1025-1032

Classification of Flower Images using Clustering Algorithms

R. Ashwini and K. Roopa

Abstract: This paper presents a flower classification and identification system that takes a flower image as input and identifies it to be belonging to a particular category present in the database. It begins by performing pre-processing operations on the input image. A set of digital images are segmented using k-means clustering algorithm from which texture and color features are extracted. Texture features are extracted using Log Gabor filters and Color features are extracted by calculating Mean and Standard deviation of color distribution from R, G, and B color channels. Texture and color features are concatenated to form a feature matrix. Features are clustered using K-Means clustering algorithm and Prim's minimum spanning tree algorithm to obtain classes. Test image will undergo the same segmentation and feature extraction process. Finally test image is identified as belonging to particular class based on similarity measure. The algorithms are implemented in MATLAB using Image Processing toolbox.

Keywords: Clustering; K-Means; Prim's Minimum Spanning Tree (PMST); Texture features; Color features.

1. INTRODUCTION

There is a huge number of flower species in the world. It is impossible for someone to remember the names of all flowers. Most of the people who can identify the flowers are not specialists. Also because, there is similarity between the flowers it is hard for someone to not get confused when identifying flowers. A digital flower classification and identification system can be used for automatic recognition of flowers without requiring the expertise of floriculturist.

Many theories have been proposed for classification and identification of flower images. The next few paragraphs makes an attempt to go through few of them.

Chomtip Pornpanomchai et al. [1] have proposed a Herb Flower recognition system (HFRS) which uses Minimum Distance Method to recognize the thai herb flower. They used Herb flowers features such as color, size and edge of petal features. The herb flower images used for recognition were taken in real environment. 110 images were used for training set and 50 images were used for test data set. An accuracy rate of 94 percent was achieved.

M.E Nilsback and Zisserman [2] developed a visual vocabulary that explicitly describes the various characteristics (color, shape, and texture) of flowers. A total of 1360 images from 17 flower species are used in the dataset, experimental results have shown that the combined vocabulary outperforms each of the individual ones.

Pavan Kumar Mishra et al. [3] have presented a semiautomatic plant identification technique based on features extracted from flower and leaf images. Flower and leaf of a plant can give information about shape, color and texture. Since, a single feature is not sufficient for classification and recognition, color,

Sir MVIT, Krishnadevaraya Nagar, Hunasamaranahalli, International Airport Road, Bangalore 562 157, Bangalore, Karnataka, India
* e-mail- ashwinirmurthy@gmail.com

texture and cell volume features are extracted from the flower. Three stage comparisons are performed, Redness, Greenness and Blueness are compared in the first stage, second stage compares shape features and cell and volume features are compared in the last stage.

Tanakorn Tiay et al. [4] have proposed a Flower recognition system based on Image processing. The system makes use of edge and color features of flower images to classify flowers. Hu's seven moment algorithm is applied to get edge characteristics. Histograms are used to derive red, green, blue, hue and saturation characteristics. K-nearest neighbor is used to classify flowers.

Rodrigo Nava et al. [5] have proposed a system for Texture Image retrieval based on Log-Gabor filters. The authors claim that the system based on log-Gabor filters has a strong correlation with the Human Visual System (HVS). Log-Gabor filtering approach is said to have outperformed the image retrieval performance for the extracted texture in comparison with Gabor filters.

In this work an attempt has been made to automatically classify and identify flowers using K-means and Minimum Spanning tree algorithms. Flowers either downloaded from the World Wide Web or captured from digital camera are used.

The methodology followed is as follows:

- Segmentation using K-Means clustering algorithm.
- Feature extraction from training image set. Texture features are extracted using Log Gabor filters and Mean and Standard deviation from each color channel to form a feature matrix.
- Clustering the features obtained from training images using K-Means clustering and Prim's Minimum Spanning tree (PMST) clustering algorithms to form classes.
- Segmentation and feature extraction from test/query image.
- Identifying the query image as belonging to a particular class based on similarity measure.

The paper is organized as follows: Section 1 provides a brief introduction to the topic. Section 2 discusses the proposed system. Section 3 presents Results and Discussions. Conclusion is presented in section 4.

2. PROPOSED MODEL

Figure 1 shows the block diagram of the proposed system.

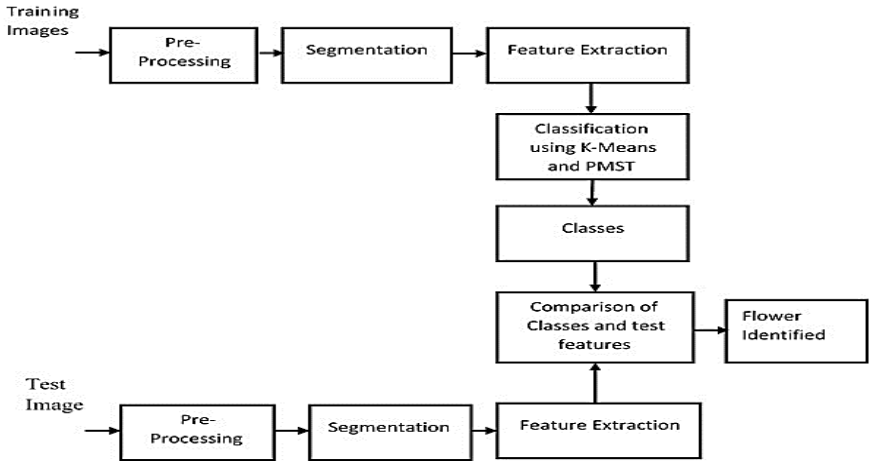

Figure 1: Block diagram of the proposed system.

2.1 Image Pre-Processing

The work starts by performing some pre-processing operations on the images, such as image resizing and gray scaling. Performing image processing on high resolution images makes the algorithm slower because of the large image size. Also, to avoid underutilization of high resolution images resizing is done. All the images are scaled to 512 × 512 size. This is done in order to increase efficiency. Because, texture and color features extracted from the image depend on probability distribution, image size should not change the result of comparison. It is also important at the same time to make sure that image is not resized to an extent such that important features in the image are missed out making it not suitable for segmentation.

2.2 SEGMENTATION

The input flower image which is resized and converted to gray scale is given as input to K-Means clustering algorithm. The K-Means algorithm is applied on image with K=2, the resulting clusters will have roughly separated foreground and background clusters.

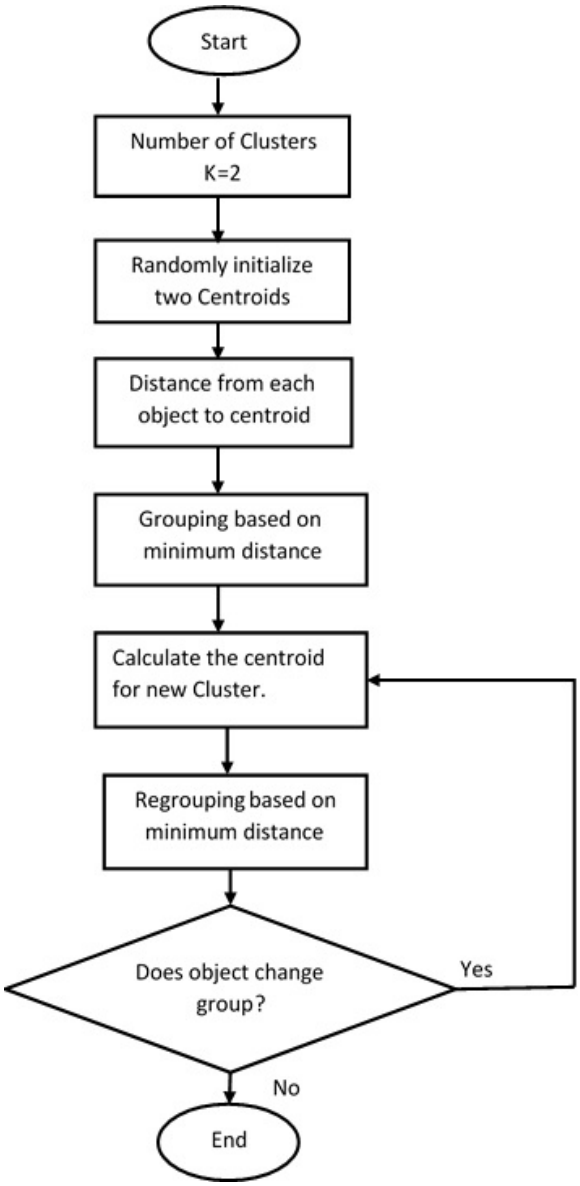

Figure 2: Flow chart for Segmentation using K-means Clustering algorithm.

A single segmentation technique will not give desirable results. Therefore, morphological operations like erosion and dilation are performed to remove small unwanted noise from the segmented image. The steps followed in K-Means clustering algorithm are listed below:

Step 1: Initialize the centroids using K random objects.

Step 2: Calculate the distance of each object using Euclidian distance measure to each centroid to obtain a distance matrix.

Step 3: Assign the objects to a particular cluster depending on minimum distance.

Step 4: Recalculate the centroid for each cluster.

Step 5: Repeat steps 2 and 3 till the point where no object move from one cluster to other cluster.

Flow chart for segmentation using k-means clustering algorithm is shown in figure 2.

The output of K-Means clustering algorithm will have two clusters, showing roughly separated foreground and background. Foreground image containing the flower is used for the feature extraction process.

2.3 FEATURE EXTRACTION

Features are extracted from segmented image. The quality of the segmented image determines the quality of features extracted. In this work texture and color features are combined together so that they give better information about the image.

Texture features are extracted using Log Gabor filters. FFT is applied to get the frequency representation of the image in the frequency domain. Gabor filter realized using equation (1) is used to process image before converting back to spatial domain using IFFT. The features extracted from the training images gives a feature matrix of size 25 × 512.

$$G(w) = e^{\left({-log\left(w/w_0\right)^2} \Big/ {\left(2log\left(k/w_0\right)^2\right)} \right)} \tag{1}$$

Mean and standard deviation of color distribution of each of the colors from the RGB color channels is extracted. The mean for the *i*th color channel at *j*th image pixel *p*ij can be defined as

$$E_i = \sum_{j=1}^{N} P_{ij} \frac{1}{N} \tag{2}$$

Mean is the average color distribution in the image. It is calculated using the formula in equation 2.

Where E_i is the Mean distribution of a color, N is the Total number of pixels of a particular color, P_{ij} is the ith color moment of jth image pixel.

The Standard deviation for the *i*th color channel at *j*th image pixel *p*ij can be defined as:

$$\sigma_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(P_{ij} - E_i\right)^2} \tag{3}$$

Where δ_i is the standard deviation, N is the total number of pixels, E_i is the mean of the color distribution.

A total of 6 features are extracted, 2 from each color channel. These 6 features are then concatenated with 512 texture features to give total 518 features per image.

2.4 CLASSIFICATION

Once the features are extracted from the training images they are classified as belonging to particular class using clustering algorithm. Here, classification is done using two algorithms namely K-Means clustering algorithm and Minimum Spanning Tree algorithm and the results are compared.

When using K-Means Clustering for classification, features extracted from the previous stage are given as input. The algorithm is explained in the segmentation section (2.2), the algorithm randomly chooses five initial centroids (K=5) and then calculates the Euclidean distance between each data points and the cluster. Based on minimum distance each data point is assigned to a cluster. The new cluster centroid is calculated by taking the mean of all the data points belonging to cluster. The process of assigning the data points to the cluster and recalculation of mean is repeated until no data point changes the group or the centroids do not change for two successive iterations. The clusters are labelled as class and the class is given a flower name, so that the test image can be identified as belonging to particular flower type.

Prim's algorithm was conceived by computer scientist Robert Prim in 1957[6]. The traditional PMST is very time consuming for large data set. An alternated form of PMST can be directly applied to Matrix specifying distance between the vertices. This form is called **Matrix form of Prim's algorithm** [7].

Steps followed in the Matrix form PMST are as follows:

Step 1: Label the column corresponding to the start vertex with a 1. Delete the row corresponding to that vertex.

Step 2: Ring the smallest available value in any labelled column.

Step 3: Label the column corresponding to the ringed vertex with a 2, etc. Delete the row corresponding to that vertex.

Step 4: Repeat steps 2 and 3 until all rows have been deleted.

Step 5: Write down the order in which edges were selected, this is the minimum spanning tree.

The flow chart of matrix form of PMST is shown in figure 3.

When applying PMST to feature matrix, the Euclidean distance between each feature vector is calculated. The resulting matrix is of size 25x25 which is called distance matrix. Distance correspond to the edges/weights in the minimum spanning tree. Matrix form of PMST is applied on this distance matrix which gives minimum spanning tree with 25 vertex and 24 edges. To get clusters from the PMST, edges with maximum weight are removed .Removing one edge results in two clusters, removing two edges results in three clusters and so on. Since, we want to form 5 clusters, 4 edges have to be removed. These clusters are labelled as a class and then a unique name is given to the class to identify the flower category. The classes formed here are used in recognition of test image, where the correlation between the set of features in the class and test feature is calculated to assign the test image as a particular flower type.

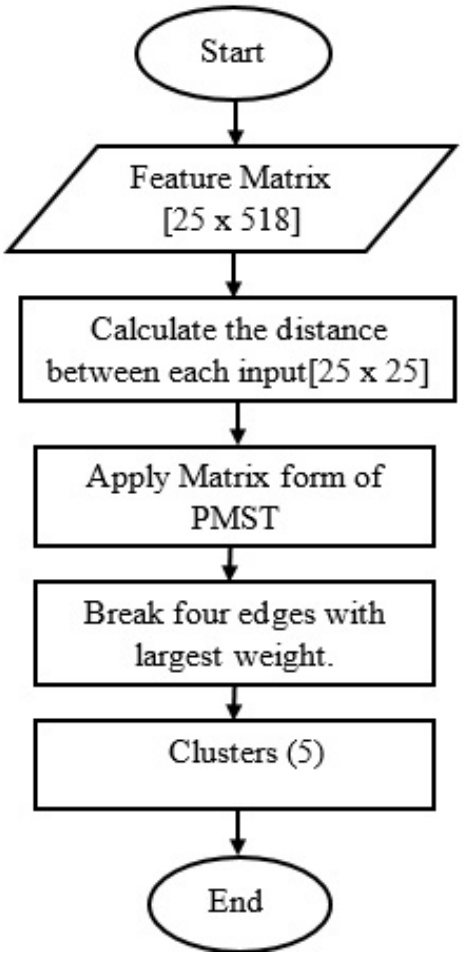

Figure 3: Flow chart for Matrix form of PMST

2.5 IDENTIFICATION OF TEST IMAGE

Test images are identified as belonging to particular category by calculating similarity between features extracted from images. In this paper two measures are used, that is Euclidian distance measure and Correlation.

Euclidean distance is considered as a measure when calculating the similarity between the Test features and Clusters centroids obtained using K-means clustering algorithm. When Euclidean distance between the test feature and centroid of that particular class is the least, image is identified as belonging to that particular class.

Correlation between the test feature and features in the database when considering the clusters obtained from PMST clustering is calculated. Closer the value of Correlation co-efficient to 1, stronger is the relation between the dataset.

3. RESULTS AND DISCUSSION

Fig. 5 shows few of the images used for classification. All the images used for the classification purpose, whether downloaded from World Wide Web or captured from the webcam are resized to standard size 512x512. The images then undergo the steps discussed in the previous section.

A total of 25 images were used in the training set and 10 images were used in the test set. Few of the images used in training and test set are shown below.

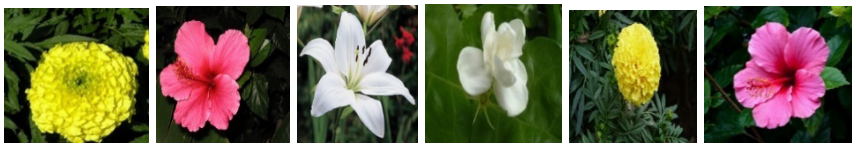

Figure 4: Images used for Classification in the training set.

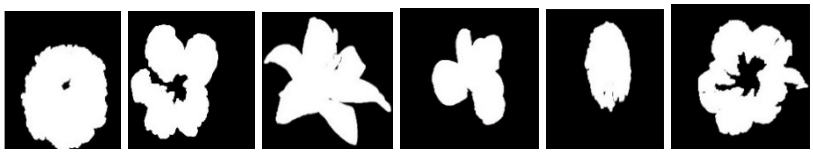

Figure 5: Corresponding Segmented image after noise removal

(a) (b) (c) (d) (e) (f)

Figure 6: Images used in the test set for identification before and after segmenatation and noise removal.

Figure 6 (a), (b), (c) shows the few of the images used in the test set for identification, Figure 6 (d),(e),(f) shows the corresponding segmented images in the test after noise removal.

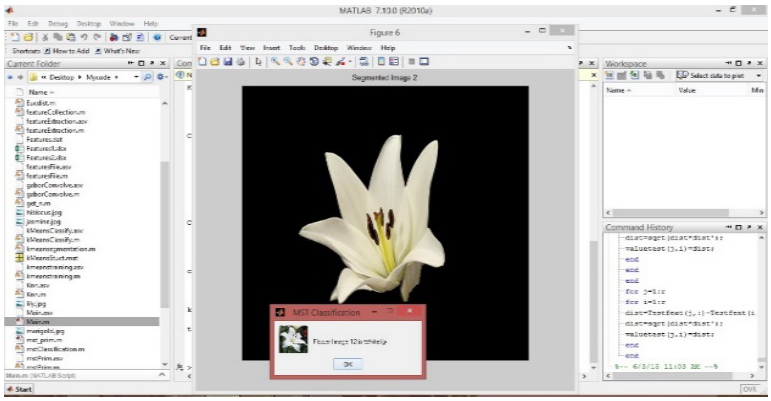

Figure 7: Test flower image identified as belonging to particular class.

Table 1 shows the comparison of accuracy of K-means clustering and PMST clustering algorithm on training images.

Table 1

Comparison of K-Means and PMST clustering algorithm for classification of images in the training set.

Number of Images in the training set	*% Accuracy of K-Means Clustering algorithm*	*% Accuracy of Minimum Spanning tree algorithm*
25	64	72

Table 2 shows the comparison of accuracy of identification of test images.

Table 2

Accuracy of recognition of test images

Total Number of flowers used for testing	*% Accuracy when using Euclidean distance measure*	*% Accuracy when using Correlation co-efficient*
10	60	70

Accuracy is on the lower side because of high dimensional data. However, reducing image size to 256 × 256 or 128 × 128 and reducing the size of feature matrix will increase the speed of execution along with saving memory.

4. CONCLUSION

A flower classification and recognition system based on K-Means and Prim's Minimum spanning tree algorithm has been presented. Both the algorithms are tested on real flower images downloaded from World Wide Web. Images were also captured from webcam and used for testing. The algorithms have proven to be very effective in multidimensional data clustering. The experimental results show that Prim's Minimum spanning tree based clustering algorithm outperforms classical K-Means clustering algorithm. The algorithms are written by using image processing toolbox in MATLAB.

Using image size of 512 × 512 may pose memory constraints for a large database. Hence image size can be reduced to 256 × 256 or 128 × 128 which increases the speed of execution along with saving memory. The size of features can be further reduced which might increase the speed of computation and also accuracy. The response time of the system to recognize a single flower image is observed to be 3.39s.

ACKNOWLEDGMENT

The authors would like to thank the Heads of the Departments of E&C and Telecommunication of Sir MVIT for the encouragement and constant support provided during the course of work.

References

[1] Chomtip Pornpanomchai, P. Sakunreraratsame, R. Wongsasirinart, N. Youngtavichavhart, "Herb Flower Recognition System (HFRS)", *International Conference on Electronics and Information Engineering,* 2010.

[2] M.-E. Nilsback and A. Zisserman "A visual vocabulary for flower classification". *In Proc. CVPR*, volume 2, pp. 1447–1454, 2006

[3] Pavan Kumar Mishral, Sanjay Kumar Maurya2, Ravindra Kumar Singh3, Arun Kumar Misral "A semi-automatic plant identification based on digital leaf and flower Images" IEEE-International Conference On Advances In *Engineering, Science and Management* (ICAESM -2012) March 30, 31, 2012.

[4] Tanakorn Tiay, Pipimphorn Benyaphaichit, and Panomkhawn Riyamongkol "Flower Recognition System Based on Image Processing" 2014 Third ICT International Student Project Conference (ICT-ISPC2014).

[5] Rodrigo Nava, Boris Escalante-Ramírez, Gabriel Cristóbal "Texture Image Retrieval based on Log-Gabor Features." *CIARP*, 2012, pp.414-421.

[6] Grygorash, O., Zhou, Y., Jorgensen, Z. "Minimum spanning tree based clustering algorithms", International Conference on Tools with Artificial Intelligence (2006).

[7] http://www.pearsonschoolsandfecolleges.co.uk/FEAndVocational/Mathematics/ALevel/AdvancingMathsForAQA2ndEdition/Samples/SampleMaterial/Chp-01%20001-022.pdf