A CONTENT BASED IMAGE RETRIEVAL SYSTEM FOR BIODIVERSITY SYSTEM

Jyoti Madake,Rohit Agrawal, Vineet Pawar, Shripad Bhatlawande
Department of Electronics and Telecommunication Engineering
Vishwakarma Institute of Technology
Pune, India
{jyoti.madake, rohit.agrawal.20, Vineet.pawar.20, shripad.bhatlawande}@vit.edu

***Abstract* - The variety of life on Earth means biodiversity, In plants particularly there are various Species. And these varieties should be properly documented according to their characteristics. As a response to this need the paper proposed a CBIR system retrieves images based on the content that uses aoproches like computer vision and Machine learning. In the recent scenario, the use of CBIR has increased as the need to retrieve images from multimedia databases. This system proposes a CBIR system for identifying and retrieving images from the database. the database consists of 15 categories of medicinal leaves from the Segmented Medicinal Leaf dataset of images from the Flavia dataset [8]. Different feature extraction techniques are used on the training dataset image to extract features like Color, Shape, and Texture. The extracted features are then concatenated to get the final features matrix and the feature matrix is used to train machine learning classification models.**

***Keywords* - *(*CBIR, Biodiversity, Shape features, texture features, color features, medicinal leaves.)**

I. Introduction

The CBIR system retrieves images based on the content from a large dataset from a query image. Biodiversity systems in plants particularly have a variety of Species, To identify the different plant species according to their features is hence a very important need, which is very useful while doing surveillance work in the field. There is a key difference among the plant leaves of different plant species, but also have various features as similar. Therefore it is very important to consider different types of features to have more discriminating features for all the species of leaves. Therefore the proposed system extracts various features from a plant leaf including Shape features, color features, texture features, and key points in the leaf structure. where the shape features contain the area ratio values of the leaf which are the pixels occupied by the leaf object by total pixels. color features contain the RBG value of the color which is present mostly in the leaf images and the texture features are extracted by calculating the harlicks features from GLCM matrics which are calculated with different values of relation pixel distance and angle. the key points of the leaf structures are then extracted by finding SIFT features of the gradient magnitude of the leaf image. All the extracted features are then fused to construct the final feature matrix. The feature matrix is then used to build a classification model to classify different species of medicinal plant leaves based on the leaf feature. which will then identify the species of the leaf and then retrieve the images of the respective species from the dataset.

II. Literature Survey

CBIR techniques are used in many applications with increased multimedia data. Therefore analysis of different approaches used for CBIR and different classification approaches to classify leaf species by their visual contents is done. For the classification of medicinal plant leaves [1] A Classification model on total 6 varieties of leaf datasets of medicinal plants with species like Ocimum sanctum, Stevia rebaudiana, Nepetacataria, Melissa officinalis, Aegle marmelos, and Mentha balsamea. is trained based on machine learning approaches. using Sobel filter seed-intensity-based line/edge detection is done. [2] Also, compact convolution Neural network models with 15 layers of deep learning network using transfer learning with the AlexNet model which compare to AlexNet alone, proved to consume less computation time more give higher accuracy.[3] Systems also use various classification models such as BOF(Bag of features) which is a very powerful classification model and DPCNN which has been proven for texture features to have good ability. For the classification of leaves, a BOF_DP technique based on DPCNN BOF is also suggested. Form feature extraction, a BOF_SC technique enhanced from a bag of contour fragments is applied. For extracting textual features and decreasing the dimensionality of the features, respectively, BOF_DP and LDA techniques are used. Finally, a SVM with linear kernel uses both characteristics of classification, and the new method outperformed existing methods on various common leaf datasets. Other methods such as (CPAM) The colored pattern appearance model is used to capture the color and texture features of an image were 4X4 grids are divided in patches. In order to include spatial information, color correlograms are utilized. The LBP feature, which may describe patterns in images, is utilized for texture descriptors[4]. In another approach, The outer and inner edges of an image are found and examined using the Gabor wavelet approach. To depict color characteristics of an image in Hue Saturation Value (HSV), a colour histogram is utilized. Applying the HSV and color moment methods yields the edge gradient feature[5]. while considering color features [6] On the Corel-5000 dataset, techniques like the histogram of color, CAC, DCD,EHD, BVLC, CDH, and BDIP were applied and examined in 3 different spaces of color i.e RGB, HSV. Both the benefits and drawbacks of the system are highlighted.

Texture features analysis is a very important step to conclude the best method which can be used for this specific application in the system [7] Gabor wavelet characteristics are employed for texture analysis as well as for thorough experimental evaluation. A comparison of various multiresolution texture features is carried out using the Brodatz texture database. Some real-time approaches considering various shape features give brief information on how the shape features are a major discriminating factor for leaf species. In [8] The system extracts the feratures from the leafs by focusing on the leaf tooth. The system extracts the count of the teeth present in the leaf, also it calculates the sinus value of the tooth structure and its shape. with that it is calculating the order of the teeth, this all features of tooth are very useful to discriminate between the leafs according to there tooth structure. In [9] discussion of two image retrieval techniques was followed by a discussion of shape descriptors and how and on what grounds they are categorized in content-based image detection. There have been discussions of later mixes of global and local elements. This entire conversation is focused on the fish's size and shape. Color, shape, spatial arrangement, and texture of regions of interest are among the visual components of images that CBIR uses. study of several CBIR system techniques. Also, leaf counters can be used to get various features of the leaves. In [14] There have been two morphological leaf descriptors utilized. Periodic patterns within leaf edge profiles could be observed using contour covariance. The venation of leaves is described by a covariance histogram. Grayscale photos were the only ones used in the experiments. A clever edge detector performs the initial step of edge detection. Shapes are said to be independent of changes in scale, translation, rotation, and viewing angle. MPP is utilized for shape representation.

After going to various approaches to classify the leaf category or to retrieve images from the datasets and to extract different features from the leaves like shape, color, texture, pattern, edge tooths, and many more. it can be concluded that to classify more species accurately by discriminating amongst their key features and less computation one should consider the maximum type of features, else only a particular feature will become similar for some species and will lead to wrong classification. With this conclusion, the paper further proposes the use of feature fusion of maximum types of features to classify more species with less computation.

III. METHODOLOGY

Feature fusion approach for the recognition and retrieval of medicinal leaves involves using various features of the leaf such as Shape features like Area, Perimeter, circularity, Color features like RGB values and color histogram, Edge patterns, Texture features using GLCM matrix (Gray-Level Co-occurrence Matrix). These features would then be used to develop a ML model, such as a R. FOREST or D. Tree classifier. To evaluate the performance of the model, a test dataset was used, which was not including the images that were used during model building phase. The model has achieved a accuracy that was then measured by comparing the predicted class labels with the true class labels. Below Fig 1 depicts the system's intended total flow.

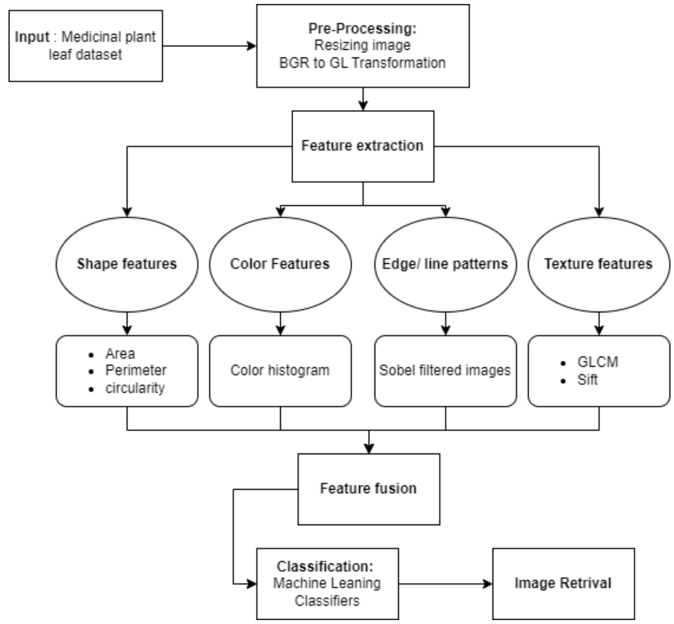

Fig. 1. System's intended total flow.

A. Dataset and Preprocessing

System uses the Segmented Medicinal Leaf Image dataset. The dataset comprises fifteen species of medicinal plant leafs such as Santalum album, Plectranthus amboinicus / Coleus amboinicus, Muntingia calabura, Brassica juncea, and many more. The dataset consists with 1500 images of fifteen species. Fig. 2 shows some types of leaf images from the dataset.

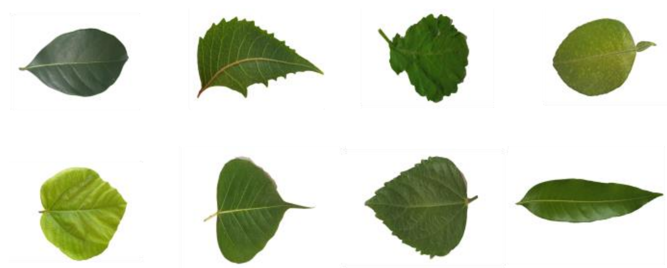

Fig. 2 Dataset images

For extracting maximum detail of leaf structure the system extracts the leaf's shape features, color features, edge line patterns, and texture features.

B. Leaf shape Feature extraction

The images are masked for extracting Brown, yellow, and green color HSV values to get the leaf pixels from the image. All pixels with brown, green, and yellow color ranges are then added to get the final leaf image extracted as shown in the below Fig. 3 .

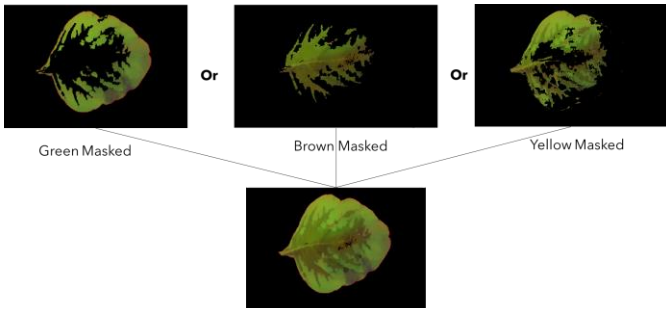

Fig. 3. contour detection

All the pixels with values other than zero contribute to being part of the leaf. The area is thus calculated by counting all of the pixels with values higher than 0.
and Therefore, the leaf's area to total area (i.e., total number of pixels) is equal to the leaf's area.

$$\textbf{Area ratio} = \frac{Total\ number\ of\ pixels\ with\ value\ greater\ than\ 0}{Total\ number\ of\ pixels\ in\ the\ image} \quad (1)$$

With Area, the perimeter is also very important information about the leaf which will help to understand the circularity of the leaf. So the system is extracting the leaf perimeter from the figure(picture) contour as depicted in Fig. 4


Fig. 4 Extraction leaf area and perimeter

By counting the pixel's count in the second image that are having value above zero, the arclength (perimeter) can be determined in Fig.4

$$\textbf{Perimeter} = Number\ of\ pixel\ greater\ than\ 0 \quad (2)$$

The circularity of the leaf can then be calculated using Leaf area and Leaf perimeter.

$$\boldsymbol{Circularity} = \frac{Area}{Perimeter} \quad (3)$$

Hence shape features consist of its Area, perimeter, Or Circularity.

C. Leaf color Feature extraction

The Color features consist of the RGB value of the color which is present mostly in the image of the leaf.

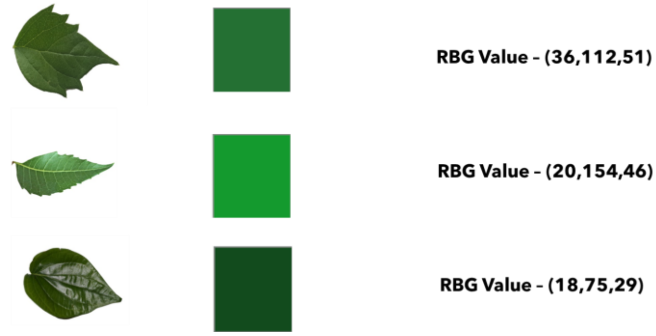

Fig. 5 Figure showing the variation of RBG values for different leafs

The Fig. 5 demonstrates how for different leafs the value of RGB respectively are different. as shown the three leaves have R values as [36,20,18], G values as [112,154,75] and B values as [51,46,29]. which shows that each color has a major contribution to 1 or 2 of the colors from RGB which helps to discriminate between the color of different leafs.

D. Leaf texture Feature extraction

The texture features are extracted using the GLCM method. The co-occurrence matrix is calculated which gives details about the relationship between the pixels as shown in Fig. 6.

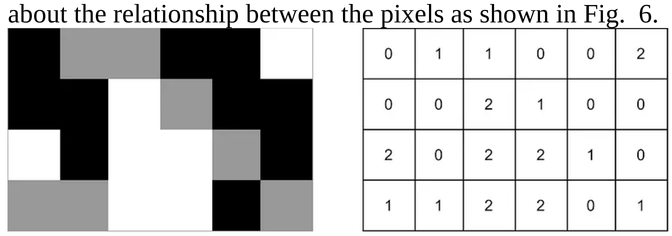

Fig. 6 Depicts the co-occurrence matrix value of an image [7]

The co-occurrence matrix value can be calculated with different combinations of distance i.e the pixels' pair distance offset and angle i.e pixel pair angles according to the table. 1.

Table. 1. The combinations of distances and angles considered for calculating co-occurrence matrix

Angle	**Distances**
0 rad	[1,2,3,4,5]
ŧ/6 rad	[1,2,3,4,5]
ŧ/4 rad	[1,2,3,4,5]
ŧ/3 rad	[1,2,3,4,5]
ŧ/2 rad	[1,2,3,4,5]
3*(ŧ/4)	[1,2,3,4,5]

The GLCM method gives the texture properties of the images which can be calculated through the following formulas.

homogeneity : $\sum_{i,j=0}^{levels-1} \frac{P_{i,j}}{1+(i-j)^2}$ (4)

contrast : $\sum_{i,j=0}^{levels-1} P_{i,j}(i-j)^2$ (5)

dissimilarity : $\sum_{i,j=0}^{levels-1} P_{i,j}|i-j|$ (6)

correlation : $\sum_{i,j=0}^{levels-1} P_{i,j} \left[\frac{(i-\mu_i)\ (j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right]$ (7)

ASM : $\sum_{i,j=0}^{levels-1} P_{i,j}^2$ (8)

energy: $\sqrt{ASM}$ (9)

The GLCM feature extracts these properties of the co-occurrence matric as shown below with the help of sklearn functionality greycometrix and greycoprops for calculating GLCM features.

Algorithm 1: Harlicks feature extraction from GLCM matrix

Input: distance array, angels array, image dataset.
Output: 5 haralick features array.

Initialization:
GLCM_features(dataset)

for an image in dataset
  distances = [1,2,3,4,5]
  angles = [0, ŧ/6, ŧ/4, ŧ/3, ŧ/2, 3*ŧ/4]

  for distance in distance
    for angle in angles
      GLCM = greycometrix(image, [distance], [angle])
      H1= greycoprops(GLCM, "energy")[0]
      H2 = greycoprops(GLCM, "correlation")[0]
      H3 = greycoprops(GLCM, "dissimilarity")[0]
      H4 = greycoprops(GLCM, "homogeneity")[0]
      H5 = greycoprops(GLCM, "contrast")[0]
    end for
  end for
  entropy = shannon_entropy(image)
  Haralick_Features.add({H1, H2, H3, H4, H5, entropy})
end for
return Haralick_Features

The GCM properties such as energy, correlation, dissimilarity, homogeneity, contrast as well as shannon_entropy with the combination of different values of distance and angle as mention in the table. 1. gives a total of 150 haralicks features. as the texture features of the leaf.

To get more variety and detailed information on the leaf structure the system then extracts the edge pattern. Sobel filter is used to extract the gradient_magniture of the image and SIFT detector is used on the gradient_magniture of images to extract the details of the leaf structure patterns.

Algorithm2:SIFT feature extraction from image gradient magnitude

Input: gradient magnitude image dataset, filter_X, filter_Y
Output: SIFT descriptor array

Initialization:
Sift_Descriptor(dataset)

for an image in dataset
  image_horizanal= convolution(image, filter_X)
  image_vertical = convolution(image, filter_Y)

  grad_image = squareroot(image_horizantal^2 + image_vertical^2)

  sift_descriptors = detectAndCompute(grad_image)
  Fsift = Kmean(sift_descriptor)
end for
return Fsift

Horizontal gradient Vertical gradient Gradient Magnitude

Fig. 7. Sobel filtered images and image gradient magnitude

Fig. 7. shows the Sobel filtered output of Gradient magnitude using horizontal edge and vertical edges of the leaf.
Shift detector is then used on the output images from the Sobel filter which detects the specific pattern that the leaf structure has as shown in Fig. 8.

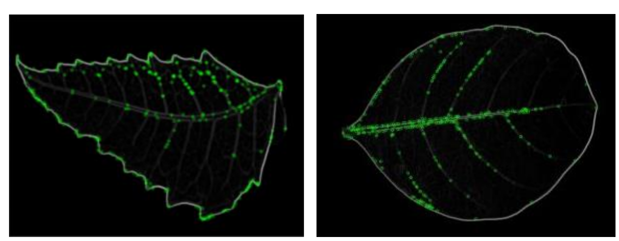

Fig. 8. Descriptors detected by SIFT detectors on filtered images

The SIFT detector detects 128 key points for each leaf object. KNN clustering is used to select 6 out of 128 feature descriptors which finally provide 6 features for the leaf structure pattern.
All the features include Shape features, color features, texture, and structure patterns features. contributes to 1, and 3,150,6 features respectively. which gives a features matrix of [876 X 160] features in total as shown in Fig. 9.

Algorithm 3: Feature fusion of all the extracted features

Input: HiR, HiG, HiB, Area, Fharalick, Fsift.
Output: Final feature matrix

Initialization:
Feature_fusion(Fcolor, Fshape, Fharalick, Fsift.)

for an image in dataset
  Fcolor = {HiR, HiG, HiB}
  Fshape = {(Area/Total number of pixels)}
end for

Ffinal = {Fsift, Fcolor, Fshape, Fharalicks}
return Ffinal

IV. RESULTS AND DISCUSSIONS

All the extracted features are split in a standard ratio of 80:20 into train and test data respectively. The training data is then used to train different ML models like R. forest, D. tree, and KNN. All the models worked well and gave an accuracy of classifying a leaf image to its correct class are depicted in the table. 2.

Table 2: Accuracy of various ML models on the feature matrix

Classifier name	Accuracy
Random forest	92.52
Decision tree	91.37
KNN	81.03

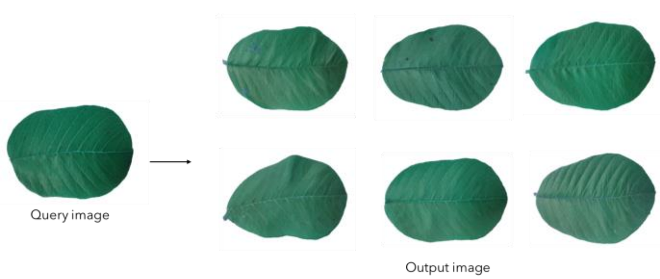

Fig. 9 retrieved images Output of the system with respect to a query image.

Fig. 10 Output of the system classifying the species of the query image plant leaf..

Fig. 9 shows the output of 6 such similar images to that of query images from the database from the same class.
Fig. 10. shows the classification of the query image leaf's plant species which identifies the species of the plant correctly as Psidium Guajava(Guava) of the query image.

V. CONCLUSION

It can be concluded that to classify more species accurately by discriminating amongst their key features and less computation one should consider the maximum type of features, else only a particular feature will become similar for some species and will lead to wrong classification. With this conclusion, the paper further proposes the use of feature fusion of maximum types of features to classify more species with less computation. The system extracts features like Shape, Color, texture, and edge pattern keypoints. which results in matrices with a number of features as 1,3,150,6 respectively and combined giving 160 features of a leaf.

Leaf contour and color histogram are used to Shape and color feature extraction respectively and for texture features by calculating GLCM matrices the Haralicks features utilising combinations with various values of pixel distances and orientation angle the extraction of texture feature is done. also to get the details of the edge keyoints the gradient magnitude of the image is calculated and SIFT descriptors are then extracted.

With extracted features on 15 different species of plants, the system has given a high accuracy of 92.52 % with very less computation. The system thus is successfully able to classify the query image species and retrieve similar images from the dataset.

REFERENCES

[1] SAMREEN NAEEM,. (2021). USING A MACHINE LEARNING APPROACH, MEDICINAL PLANT LEAVES ARE CATEGORISED ACCORDING TO MULTISPECTRAL AND TEXTURE FEATURES. 10.3390/AGRONOMY11020263. AGRONOMY. 11. 263.

[2] MOHAMMAD FASEEHUDDIN, SAWAL HAMID MD. ALI, SHIVALI AMIT WAGLE, AND HARIKRISHNAN. UTILISING NEW COMPACT CONVOLUTIONAL NEURAL NETWORK MODELS TO CLASSIFY PLANT LEAVES

[3] RECOGNIZATION OF LEAF IMAGE USING BAG OF FEATURES BY YAONAN ZHANG, ZHAOBIN WANG, AND YUFANG MIN

[4] A CBIR SYSTEM FOR IMIGES OF MARINE LIFE," 15TH ISCE, 2011 IEEE, SINGAPORE, 2011, PP. 29–33, DOI: 10.1109/ISCE.2011.5973777. A. R. SHEIKH,

[5] "CBIR USING HYBRID FEATURE EXTRACTION FROM QUERY IMAGE," 2018 ICICET, PUNE, INDIA, PP. 1-4, N. M. VARMA AND A. RIYAZ.

[6] K. P. KAUR, "ON COMPARATIVE ANALYSIS OF PERFORMANCE OF COLOR, EDGE AND TEXTURE BASED HISTOGRAMS FOR CBIR," ICRITO, NOIDA, INDIA, 2014

[7] "RETRIEVAL OF IMAGE DATA FOR BROWSING USING TEXTURE FEATURES," BY B. S. MANJUNATH AND W. Y. MA, IEEE TPAMI, VOL. 18,

[8] LI C, TIAN Y, ZHOU H, TIAN D, AND WANG H (2019) EXTRACTION OF A PLANT LEAF'S TEETH CHARACTERISTIC. HTTPS://DOI.ORG/10.1371/JOURNAL.PONE.0204714 PLOS ONE 14(2)

[9] A REVIEW ON CBIR REPRESENTATION AND DESCRIPTION FOR FISH," 4TH ICACSAT, KUALA LUMPUR, MALAYSIA, 2015, PP. 39–44, DOI: 10.1109/ACSAT.2015.15, N. S. OSMAN AND M. R. MUSTAFFA.

[10] METHODS FOR CBIR EVALUATION A. O. ADETUNMBI2, B. A. OJOKOH2, J. O. OLALEKE1, IROJU OLARONKE1, AND 1THE ADEYEMI COLLEGE OF EDUCATION IN ONDO, NIGERIA

[11] USING VARIOUS MACHINE LEARNING ALGORITHMS, DIFFERENT PLANTS CAN BE IDENTIFIED THROUGH IMAGE PROCESSING. VIKAHO Z SWU, IBAPHYRNAISHISHA KHARIR, AND DIBYA JYOTI BORA NOVEMBER-DECEMBER (2020) VOL. 43, NO. 3 (III) ISSN: 2249-6661 SAMBODHI, PAGE 173 OF THE UGC CARE JOURNAL.

[12] IN IEEE ACCESS, VOL. 8, 2020, J. HUIXIAN, "USING DEEP LEARNING ARTIFICIAL NEURAL NETWORK ANALYSIS OF PLANTS IMAGE RECOGNITION.

[13] LEAF BASED PLANT RECOGNITION USING MORPHOLOGICAL FEATURES," 2013 IEEE ICIP , MELBOURNE, VIC, AUSTRALIA, E. APTOULA AND B. YANIKOGLU.

[14] Shape based leaf image retrieval," IEE Proceedings in visp, vol. 150, February 2003. Z. Wang and D. Feng.

[15] CBIR using Scale Invariant Feature Transform and Moments", IEEE UPCON. Prashant Srivastava & Ashish Khare.

[16] A novel multistage method for identifying plant leaves based on deep CNN was developed by Ali Beikmohammadi, and Ali Motallebi and published in Expert Systems with Applications.

[17] Li C, Tian Y, Zhou H, Tian D, and Wang H (2019) Extraction of a plant leaf's teeth characteristic.