CONTENT BASED IMAGE RETRIEVAL SYSTEM with MODIFIED KNN ALGORITHM

T. Dharani, Laurence Aroquiaraj

1Assistant Professor, Department of Computer Science, Periyar University Salem - 636011, Tamil Nadu, India
2Research Scholar, Department of Computer Science, Periyar University Salem - 636011, Tamil Nadu, India

dharanimca28@gmail.com laurence.raj@gmail.com

Abstract: An image retrieval system is a computer system for browsing, searching and retrieving images from a large image database. Content Based Image Retrieval (CBIR) is a technique which uses visual features of image such as color, shape, texture, etc…to search user required image from large image database according to user’s requests in the form of a query. MKNN is an enhancing method of KNN. The proposed KNN classification is called MKNN. MKNN contains two parts for processing, they are validity of the train samples and applying weighted KNN. The validity of each point is computed according to its neighbors. In our proposal, Modified K-Nearest Neighbor (MKNN) can be considered a kind of weighted KNN so that the query label is approximated by weighting the neighbors of the query. The procedure computes the fraction of the same labeled neighbors to the total number of neighbors. MKNN classification is based on validated neighbors who have more information in comparison with simple class labels. This paper also concentrates identifying the unlabeled images with help of MKNN algorithm. Experiments show the validity takes into accounts the value of stability and robustness of the any train samples regarding with its neighbors and excellent improvement in the performance of KNN method. This system allows provide label to unlabeled image as user input.

Keywords: CBIR, Image Classification, Graph Minicut, KNN, MKNN, Unlabeled image.

1. INTRODUCTION

An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images. Most traditional and common methods of image retrieval utilize some method of adding metadata such as captioning, keywords, or descriptions to the images so that retrieval can be performed over the annotation words. Applications of image retrieval are remote sensing, fashion, crime prevention, publishing, medicine, architecture, etc [1]. It is classifying two types of retrieval are Text Based Image Retrieval and Content Based Image Retrieval.

Initially it working based on text of image. Text Based Image Retrieval is having demerits of efficiency, loss of information, more expensive task and time consuming. Overcome these problems by using Content Based Image Retrieval (CBIR) system for image retrieval. CBIR working by using features of the content of the image is known as Content-Based Image Retrieval (CBIR) [2]. The image retrieval system acts as a classifier to divide the images in the image database into two classes, either relevant or irrelevant. In this sense, an annotated image can be represented by a feature vector x, e.g. a set of image features or eigen features, and its label y that is either relevant or irrelevant [3] and without label image is called unannotated or unlabelled image. These images are not considered during retrieval process. Unlabelled image is most useful to getting the better result of retrieval.

Unlabelled images are identified when classifying the image database by using classifiers. The classifiers are working based on various classification properties of image database. The begin stage of image database classification is supervised. The supervised classification is finding the labeled images very well. The unsupervised classification is used to finding the unlabelled images very well. The section of this paper presenting basics of Content Based Image Retrieval system. Section III of this paper is tried to explain various existing image classification methods with unlabelled image. Section IV is presenting proposed method of image classification of CBIR system. Section V is including the conclusion of the paper

2. CONTENT BASED IMAGE RETRIEVAL

Content based image retrieval is working with different types of image database. All databases are having two types of images like labelled and unlabelled clouds, sky, building, people without label of image Labeled image Unlabelled image
Feature extraction is refers that dimensionality reduction of that object. In general it defines when performing analysis of complex data one of the major problems stems from the number of variables involved. It plays an important role in image processing. Features are classified into three types in image processing, that is low, middle and high. Low level features are color, texture and middle level feature is shape and high level feature is semantic gap of objects. Feature vector is used to store the extracted features as matrix format in database for processing. Feature extraction is working with matrix, histogram, Transformation, Rotation, Segmentation, Region of Interest, Region of Center of the image

2.1 Color Feature Extraction

CBIR system is working with mainly one common feature of image is color. The color images are having the standard color is RGB color. Here using the coloring model of image database. First of all separating the original image as RGB color form by using MATLAB code for best retrieval process.

```matlab
I1=imread('Image01.jpg');
Red color: r = I1(:,:,1);
Green color: g = I1(:,:,2);
Blue color: b = I1(:,:,3);
```

Figure 3: MATLAB code for RGB color model

2.2 Texture Feature Extraction

An image texture is a set of metrics calculated in image processing designed to quantify the perceived texture of an image. Image Texture gives us information about the [16] spatial arrangement of color or intensities in an image or selected region of an image. Texture analysis attempts to quantify intuitive qualities described by terms such as rough, smooth, silky, or bumpy as a function of the spatial variation in pixel intensities. In this sense, the roughness or bumpiness refers to variations in the intensity values, or gray levels.

- Gray Level Co-Occurrence Matrix: A statistical method for examining the texture that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. The GLCM functions characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix. Many Statistics such as Contrast, Homogeneity, Correlation and Energy are derived from the GLCM. These characteristics provide the details about the Texture of an image. Contrast measures the local variations in the gray-level co-occurrence matrix. Correlation measures the joint probability occurrence of the specified pixel pairs. Energy provides the sum of squared elements in the GLCM also known as uniformity or the angular second moment. Homogeneity measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. Comparing to Texture analysis statistics and Texture filter characteristics, texture analysis method provides the shape feature characteristics. Texture filter method provides the Range filter which calculates the local range of an image, Standard deviation filter which calculates the local deviation of an image; Entropy filter calculates the local entropy of a gray scale image. Entropy is a statistical measure of randomness.
### Table 1: Texture Analysis

<table>
<thead>
<tr>
<th>Image</th>
<th>Contrast</th>
<th>Homogeneity</th>
<th>Correlation</th>
<th>Energy</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Img01</td>
<td>0.1414</td>
<td>0.9423</td>
<td>0.9819</td>
<td>0.192</td>
<td>5.702</td>
</tr>
<tr>
<td>Img02</td>
<td>0.1181</td>
<td>0.9450</td>
<td>0.9506</td>
<td>0.264</td>
<td>6.939</td>
</tr>
<tr>
<td>Img03</td>
<td>0.1406</td>
<td>0.9350</td>
<td>0.9321</td>
<td>0.277</td>
<td>6.854</td>
</tr>
<tr>
<td>Img04</td>
<td>0.1504</td>
<td>0.9385</td>
<td>0.9773</td>
<td>0.150</td>
<td>7.414</td>
</tr>
<tr>
<td>Img05</td>
<td>0.1652</td>
<td>0.9217</td>
<td>0.9478</td>
<td>0.169</td>
<td>7.281</td>
</tr>
<tr>
<td>Average</td>
<td>0.1</td>
<td>0.9</td>
<td>0.9</td>
<td>0.1</td>
<td>7.5</td>
</tr>
</tbody>
</table>

### 2.3 Shape Feature Extraction

Shape is also an important low-level feature in image retrieval system; since an object, in most case, can form by a set of shape (e.g. a car is consisted of a few rectangles and a few circles), most similar objects have a high correlation in the set of shapes. Shape-based image retrieval should extract the shapes from images by segmentation, and classify the shape, where each shape should have their own representation and should variant to scaling, rotation, and transition [18, 19]. In shape-based image retrieval the user need to choose an reference image or sketch a desired shape, since the user may not only want the shape that exact matched, so shape based image retrieval should be able to identify similar shapes.

- **Moments**: In general, the moment’s technique provides a better recognition. In this case, the moments [20] of gray levels at each pixel in the image is calculated in the same way as mechanical moment [18]. The shape moments are calculated by canny edge detection method of gray image.

### Table 2: Shape Moments

<table>
<thead>
<tr>
<th>Image name</th>
<th>Centroid</th>
<th>Eccentricity</th>
<th>EquivDiameter</th>
<th>Perimeter</th>
</tr>
</thead>
<tbody>
<tr>
<td>smlImages01</td>
<td>18.1132</td>
<td>0.940882</td>
<td>14.2283</td>
<td>306.225</td>
</tr>
<tr>
<td>smlImages02</td>
<td>2.8</td>
<td>0.901281</td>
<td>2.52313</td>
<td>6.24264</td>
</tr>
<tr>
<td>smlImages03</td>
<td>2.8</td>
<td>0.901281</td>
<td>2.52313</td>
<td>6.24264</td>
</tr>
<tr>
<td>smlImages04</td>
<td>2.3333</td>
<td>0.730297</td>
<td>1.95441</td>
<td>3.41421</td>
</tr>
<tr>
<td>smlImages05</td>
<td>2.8</td>
<td>0.901281</td>
<td>2.52313</td>
<td>6.24264</td>
</tr>
</tbody>
</table>

### 2.4 Similarity Measures

CBIR systems are working with some measures for calculating distance between similarities of two images during image retrieval time. The measures are sometimes similar or otherwise dissimilar between two images. CBIR system frequently used similarity measures for an accuracy of image retrieval from large...
image database. CBIR has overviewed with various similarities measures. These distance measures are mainly used for identify, sum, calculate of similarities distance between images, regions, features vectors, color histograms etc. Content Based Image Retrieval includes the standard measures. Those measures are the initial measures of the image retrieval process. The measures are Recall, Precision, and Fallout. These measures are used to improve the image retrieval process with various measures. The similarity measures include a benchmark measures that is Euclidean distance between two similar images. This measure is used to improve the image retrieval process in an efficient way with various types of measures with different combinations of two object of an image. Similarity measures are very useful to calculate the different object of an image for accuracy and fast image retrieving.

2.5 Relevance Feedback
Relevance feedback (RF) is a query modification technique which attempts to capture the user’s precise needs through iterative feedback [4] and query refinement. It can be thought of as an alternative search paradigm, complementing other paradigms such as keyword-based search.

- **Learning-Based Advancements** are used to modify the feature set or similarity measure.
- **Feedback Specification Advancements** each round consisting of one set each of positive and negative examples in relation to the intended query.
- **User-Driven Methods** the user’s point-of-view in providing the feedback has largely been taken for granted.
- **Probabilistic Methods** used where uncertainty about the user’s goal is represented by a distribution over the potential goals.

3. EXISTING METHODS
The classification is used to group with homogeneous characteristics, with the multiple objects from each other within the image. Classification divides the feature space into several classes based on a decision rule. Supervised classification is the sampling [5] of training data from clearly identified training areas. Unsupervised classification is working with two types they are multiple groups from randomly selected sampling data and clustering classes.

3.1 Semi-supervised learning
Semi-supervised learning deals with methods for automatically exploiting unlabeled data in addition to labeled data to improve learning performance. That is, the exploitation of unlabeled data does not need human intervene. Here the key is to use the unlabeled data to help estimate the data distribution. Transductive learning [6] is a cousin of semi-supervised learning, which also tries to exploit unlabeled data automatically.

3.2 The Graph Mincut Learning Algorithm
This algorithm is [7] working with weighted graph \( G = (V,E) \) by using edge. The classification of vertices is connected by edges of infinite weight to the labelled examples. The weights are assigned based on relationships between example vertices. The minimum cut for weights of the graph the minimum total weights of the edges whose removal disconnects from \((v^+, v^-)\). Finally assign the positive and negative labels to all unlabelled examples of \((v^+, v^-)\).

3.3 k-Means
K-means clustering is an algorithm to classify or to group your objects [9] based on attributes/features into \( K \) number of group. \( K \) is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus the purpose of K-mean clustering is to classify the data.

3.4 K Nearest Neighbors Algorithm
K-nearest neighbor is a supervised learning algorithm where the result of new instance query is classified based on majority [10] of K-nearest neighbor category. The purpose of this algorithm is to classify a new object based on attributes and training samples. The classifiers do not use any model to fit and only based on memory. Given a query point, we find \( K \) number of objects or (training points) closest to the query point. The classification is using majority vote among the classification of the \( K \) objects. Any ties can be broken at random. K Nearest neighbor algorithm used neighborhood classification as the prediction value of the new query instance.
An existing CBIR system working with various type of image classification methods, they are D-EM (Discriminant EM) Algorithm [11] using the approach Expectation-Maximization (EM) with unlabeled image as data set. Support Vector Machines using the approach structural risk minimization with unlabeled image as two classes relevant and irrelevant. Relevance Feedback using the approach high confidence degree with unlabeled image as rank. BMMA AND SEMIBMMA FOR SVM RF [12] using the approach Laplacian regularizer with unlabeled image as sample. Semi-Supervised/Active learning using the approach co-training paradigm, co-testing paradigm with unlabeled image as training set. SSAIR (Semi-Supervised Active Image Retrieval) using the approach Shrunk with unlabeled image as data set. TRANSDUCTIVE LEARNING METHOD [13] using the approach Probability with unlabeled image as ranking scores. ACTIVE LEARNING METHOD using the approach Convergence with unlabeled image as two examples (positive for relevant and negative for irrelevant). Bootstrapping SVM active learning using the approach images are ranked in descending order with unlabeled image as initial SVM classifier. Local Weakly Similar Patches using the approach Semi-supervised learning algorithm with unlabeled image as labeled log image. SEMANTIC KERNEL LEARNING[10] using the approach Factorization with unlabeled image as kernel matrix. Semi-Supervised Multiple Instance Learning (SSMIL) [14] using the approach Optimization with unlabeled image as instance of mapped bag. New Analysis of the Value of Unlabeled Data using the approach Distributions with unlabeled image as positive and negative examples. Label propagation using the approach Binary approach with unlabeled image as fuzzy relevance. The Graph Mincut Learning Algorithm using the approach Weighted graph with unlabelled data as example vertices. Self-taught Learning Algorithm using the approach Correlations between rows of pixels with unlabelled data as example data. Anchoring Edge Fragments to Local Patches using the approach SIFT descriptor with unlabeled image as a set of semi-local region features.

4. PROPOSED METHOD
In proposed system using Modified K-Nearest Neighbor Algorithm for image classification in Content Based Image Retrieval system. Algorithm consists of two parts:

- Validity of the train samples that takes every training sample must be validated at the first step. The validity of each point is computed according to its neighbors.
- Applying weighted KNN means weighted KNN is one of the variations of KNN method which uses the K nearest neighbors, regardless of their classes, but then uses weighted votes from each sample rather than a simple majority or plurality voting rule.

4.1 Modified K-Nearest Neighbor Algorithm
Inspired the traditional KNN algorithm, the main idea is classifying the test samples [15] according to their neighbor tags. This method is a kind of weighted KNN[17] so that these weights are determined using a different procedure. The procedure computes the fraction of the same labeled neighbors to the total number of neighbors.

The main idea of the presented method is assigning the class label of the data according to K validated data points of the train set. In other hand, first, the validity of all data samples in the train set is computed.

```plaintext
Output_label := MKNN ( train_set , test_sample )
Begin
For each process x in the train data do
Validity(x) := Compute Validity of x-th sample;
End for;
Output_label:=Weighted_KNN(Validity,test_sample);
Return Output_label ;
End.
```

Figure 4: Pseudo-code of the MKNN Algorithm

4.2 Validity of the Train Samples
The validation process is performed for all train samples once. After assigning the validity of each train sample, it is used as more information about the points. To validate a sample point in the train set, the H nearest neighbors of the point is considered. Among the H nearest neighbors of a train sample x, validity(x) counts
the number of points with the same label to the label of \(x\).

\[
Validity(x) = \frac{1}{H} \sum_{i=1}^{H} S(\text{lbl}(x), \text{lbl}(N_i(x)))
\]

Where \(H\) is the number of considered neighbors and \(\text{lbl}(x)\) returns the true class label of the sample \(x\). Also, \(N_i(x)\) stands for the \(i\)th nearest neighbor of the point \(x\). The function \(S\) takes into account the similarity between the point \(x\) and the \(i\)th nearest neighbor.

### 4.3 Applying Weighted KNN

Each of the \(K\) samples is given a weighted vote that is usually equal to some decreasing function of its distance from the unknown sample. For example, the vote might set be equal to \(1/(d+1)\), where \(d\) is Euclidian distance. These weighted votes are then summed for each class, and the class with the largest total vote is chosen. This distance weighted KNN technique is very similar to the window technique for estimating density functions. For example, using a weighted of \(1/(d+1)\) is equivalent to the window technique with a window function of \(1/(d+1)\) if \(K\) is chosen equal to the total number of training samples. In the MKNN method, first the weight of each neighbor is computed using the \(1/(d+0.5)\).

Then, the validity of that training sample is multiplied on its raw weight which is based on the Euclidian distance. In the MKNN method, the weight of each neighbor sample is derived according to

\[
W(i) = \text{Validity}(i) \times \frac{1}{d_i + 0.5}
\]

Where \(W(i)\) and \(\text{Validity}(i)\) stand for the weight and the validity of the \(i\)th nearest sample in the train set.

![Proposed offline model of CBIR system](image)

**Figure 5: Proposed offline model of CBIR system**

### 5. CONCLUSION AND FUTURE WORK

In this paper we presented CBIR system using feature classification with Modified K-Nearest Neighbor Algorithm. CBIR system based on visual feature of image that color of image enables to classify the image database as labeled and unlabeled. The offline model contains two major parts, namely Validity of the Train Samples, Applying Weighted KNN. MKNN algorithm adds a new value named — Validity|| to train samples which cause to more information about the situation of training data samples in the feature space. The validity takes into accounts the value of stability and robustness of the any train samples regarding with its

### REFERENCES


[19] T. Dharani, I. Laurence Aroquiaraj, —Content Based Image Retrieval System Using Feature Classification with Modified KNN Algorithm[12], International Journal of Computer Trends and


[22] John Eakins, Margaret Graham, University of Northumbria at Newcastle.