Content based Retrieval Management Systems in Web Engineering

Prameela Devi. Chillakuru, T. Kumanan, CH. Sarada Devi

Abstract—The tremendous upsurge and abundant availability of graphic documents on the internet directed to the high concern in research on Content-Based Image Retrieval (CBIR). It has cemented the attitude for a massive sum of innovative procedures and schemes, and growing curiosity in respective fields to preserve such research. Existing technologies were discussed about various CBIR techniques such as k-means, content-based image, and hybrid clustering, etc. in combination with the feature vector of information images and texture features. In similar cases, it is construed in stating the user’s semantic knowledge to permit information distribution and reuse. Hence models ought to be managed within repositories, where they might be retrieved upon users’ queries. Due to the lack of adequate tools on incisive/handling for visual content, this research work proposes an Efficient Density-based Clustering Algorithm (EDBC) for CBIR technique that will enhance scalability, reduce the user search time, and lower the maintenance cost. Using the existing CBIR, the color, texture, and shape features of an image is identified by integrating with the proposed EDBC algorithm and thereby, it improves the scalability and user search time. When comparing the feature between the color histograms of RGB and CMYK, it shows better color characteristics in CMYK by using the proposed technique. Also, the grouping of color, shape and texture features based image retrieval improves the accuracy when compared with existing methods.

Keywords---CBIR, EDBC, Clustering, Support Vector Machine (SVM), RGB, CMYK, Colour Histograms, Texture and Shape, Feature Extraction.

I. INTRODUCTION

Recently, there is incredible attention in classifying the problem of search and retrieval of images by relevance feedback. In an image database, the image features are extracted to store it for future reference. An enormous calculation is needed in this step which is very challenging for image feature extraction. The most commonly used image features for representation are shape, image, and color, which is considered as a low-level feature to represent the image. The stored image feature in the database matches with the user query image to provide the best possible image [1, 2].

A traditional CBIR system is as follows: At first, the feature extraction is done in which each image is associated with the quantitative vector. This vector is known as image’s feature vector and features of each image present in the database are also calculated. Secondly, an input image which is also called as a query image calculates the distance measure of the feature vector with the imagemeries present in the database. At last, the user gets the closest images having the smallest distance measure, and this well-known CBIR system is specified in [3]. Based on the parameters such as color, shape, and texture, the CBIR is developed to analyze both grayscale and color images, and the user can execute the query to acquire multiple results which arise to match the different parameters [4].

Various CBIR algorithms were developed, but still, the problematic retrieval of images based on the pixel gratified remains unsolved. So, the EDBC algorithm is suggested to extract the images depending on the density function and gives accurate results and scalability in color histograms.

II. LITERATURE REVIEW

For the past three decades, CBIR has been a dynamic area of research in consequence of an extensive range of presentations in techniques related to retrieval of images. The verse “content-based” denotes the exploration procedure that assesses the exact contents of pictures rather than utilizing outdated techniques of image explanation for recovery. The word “content” in this context states the consistency, color, figure, or any additional data that might be obtained from the source itself [5].

2.1 Review of CBIR and Its Algorithms

According to the CBIR data mentioned in [6], there are numerous categories of image recovery methods. Dissimilar attention based indicators and descriptors have been projected or extraction of features in a retrieval procedure [7].CBIR employs low-level attributes such as dimension, surface, shape, etc. which portrays the content of an image. However, finding a single best illustration of material is incredible which segregates the various classes with characterized boundaries to a great extent. Additionally, CBIR which measures the distance is not so effective while describing low-level attributes. To overcome this, a genetic algorithm is used which elevates the loads extended from feature extractions of a picture and incorporating it with machine learning techniques [8]. However, [9] suggested low-level features (line, dots, edges, etc..) to enhance the accuracy of CBIR by attaining more images similar to the original image. It increases the average precision (accuracy) to 72 %. However, the CBIR method improves the accuracy, [10] used K means-CBIR algorithm for retrieving images in both local and global features. It gave better accuracy in search results than link-based and content-based retrieval techniques.
Moreover, also provides visually and semantically similar output images. A histogram-based color feature extraction technique is simple, fast and efficient. The retrieved results in [11] were based upon texture by using Gary-Level Co-occurrence Matrix (GLCM) which provided better results. Also, the combined techniques perform well when compared with individual techniques. Integration of HSV color histogram with GLCM and Discrete Wavelet Transform (DWT) enhances the retrieval performance. Moreover, for the combination of the local features of color, edge, and texture of an image, the SVM classifier provided better optimization results in [12, 13].

Corel image database indicated that the Memetic Algorithm (MA) is very useful to categorize color, shape and texture features. This method is used for the identical images from the images database. It provides better results in recall rates and average precision when compared with other techniques. Temporarily, the MA retrieved images correlated to the inquiry image effectively based on the similarity measure. An investigation of this method was accompanied as per the core image. It shows outstanding performance in differentiating color, shape and color texture features which are ascertained in [14, 15].

Meanwhile, the relevant features are selected from the original feature set for facilitating image classification, using the Self-Adaptive Harmony Search (SAHS) algorithm. Then, two methods “one-against-one” and “one-against-all” were planned to build Support Vector Machine (SVM) classifiers trained using the relevant features selected by SAHS [16]. In this scheme of CBIR, the visual features were take out by evaluating the statistical parameters of quantized histograms of Hue (H) and Saturation (S) constituents of HSV color image through the texture determined from Gray-Level Co-Occurrence Matrices (GLCM) of the multi-resolution images of Value (V)-component color space of HSV. A clustering technique used to create a CBIR system was described in [17, 18] in which the groups decreased the images without any presented number of data parameters or clusters. It showed the like images being appropriate to the similar cluster and the shaped image data are healthily grouped.

The problem involved in CBIR of remote sensing images was reduced by applying deep features extraction from the last layers of a trained convolutional neural network from deep learning approaches which gave higher performance than feature extraction using shallow methods [19]. Then, the six features of the image such as RGB Color moments, RGB Color histogram, HSV Color histogram, Co-occurrence Matrices, Local Binary Pattern (LBP) and HSV Color moments were selected by RICE algorithm selection model [20, 21] which gave high performance. Its effectiveness was improved by the Scale Invariant Feature Transform (SIFT) descriptor which represents the visual gratified of the images and then utilized the distance ratio as the threshold to control the number of matched feature points. The visual structures sideways with textual features were extracted and fused them composed by the Kernel technique to retrieve top rank same images. This kernel method supported three manners of finding such as Image query, Keywords, and Mixture of both. It revealed that the textual features are efficient compared to the visual features for CBIR applications as mentioned in [22] and it is suitable only for the non-textual images. So, searching for the images in content-based took more time, and it also fetches some irrelevant images.

From the existing studies, CBIR has semantic gap problems and also lacks the effectiveness during image indexing and retrieval. The main task of the CBIR system is to compare the similarities while extracting a feature of every image based on its pixel values. These features are suitable for measuring the similarity between the images present in the other database. Depending on the similarity value, images are graded for the retrieval process. So, regaining the desired images from the large database needs an efficient and effective algorithm for content image retrieval. An evolutionary algorithm called EDBC algorithm is proposed for comparing the similarities present in the CBIR system.

### 2.2. The Contribution of the Study

The contribution of this study is as follows,
- To enhance the scalability, accuracy and user search time using the proposed EDBC algorithm based on texture-based image recovery technique.
- To extract the local features by the dense matrix and global features by the color histogram.
- To validate the results with the existing methods

### III. PROPOSED METHODOLOGY

A CBIR system is used to search the same kind of images to a query image from an extensive image database using EDBC for efficient picture retrieval system is proposed. It is used for enhancing the scalability, accuracy and user search time. The experiments are implemented in Java and system configuration as Intel Core 2.93GHz, 4GB RAM with Windows 10. The flow of data of the proposed technique is exemplified in figure 1.

Figure 1 demonstrates the data flow diagram of the proposed method; it collects the unstructured dataset like images for content with quality based retrieval images using a data mining approach. The smoothing filter is employed to filter the unwanted information from an image.
It is also utilized to filter or remove many types of noises from the images. Generally, this feature is collaborating aspire-processing involves alteration of an image by altering the pixel brightness values to progress its visual impact structures like shape, color, texture features, etc. Extract features with low within-class variability and high between class variability and discards redundant information. The procedure of using samples of recognized informational classes (training sets) to categorize pixels of unknown individuality and this type of classification requires a huge number of unidentified pixel that divisions into some classes based on natural groupings current in the copy values. Finally, it provides more accuracy, scalability and reduces the cost and user search time.

3.1 EDBC Algorithm

Consider the dataset as D, and let the algorithm’s clustering radius be, and the lowest number of objects in the neighbourhood were. The basic ideas of the algorithm [23] are presented in Table 1.

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eps-neighbouring area</td>
<td>Let ( P ) be the sphere’s center in the dataset ( D ). For data within the radius ( Eps ) of the object’s area, a gathering of points enclosed in the sphere ( { y \in D : dist({ y }, P) \leq Eps } ). The neighboring area is exposed in figure 2a.</td>
</tr>
<tr>
<td>Density</td>
<td>At the location of the data point ( P ) in the dataset ( D ), the number of points, Num. confined in the neighborhood ( Eps ) its density.</td>
</tr>
<tr>
<td>Core point</td>
<td>At the position of data point ( p ) in the dataset ( D ), if the density ( \text{Num}(p) ) in the neighborhood ( Eps ) satisfies the condition ( \text{Num} \geq \text{MinPts} ), then it is known as a core point.</td>
</tr>
<tr>
<td>Border point</td>
<td>At the location of the data point ( P ) in the dataset ( D ), if the density in the neighborhood ( Eps ) satisfies the condition ( \text{Num} \leq \text{MinPts} ), it is present inside the sphere called as a border point.</td>
</tr>
<tr>
<td>Noise point</td>
<td>Every object except the point ( P ) present in the core and border.</td>
</tr>
<tr>
<td>Reachable-Direct density</td>
<td>An object ( P ), ( Q \in D ), where there exists a core point, and exclusively the ( Eps ) neighborhood of ( Q ) is said that from ( p ) to ( q ) is straight density-reachable, i.e., ( q \in N_{Eps}(P), N_{Eps}(P) \geq \text{MinPts} ), as presented in figure 2b.</td>
</tr>
<tr>
<td>Reachable-Density</td>
<td>Given objects ( P_1, P_2, P_3, P_4, \ldots, P_n \in D ), where ( P_1 = \emptyset, P_2 = Q ), if ( P_1 ) is direct density-reachable from ( P_2 ), ( P_2 ) is density-reachable from ( P_3 ), this illustrated in figure 2c.</td>
</tr>
<tr>
<td>Density-connected</td>
<td>Assumed objects ( P ), ( Q \in D ), if there is a point on ( Q ) that is density-reachable from ( P ), then ( P ) and ( Q ) are density-connected. It is represented in figure 2d.</td>
</tr>
</tbody>
</table>

The various concepts were represented in the form of figures as,
From the above-stated descriptions, the knowledge of the EDBC algorithm is as follows: The search can begin from the neighbouring area of any compelling data point. When enough data points are given in the neighbourhood (≥), the cluster enlarges or else, a data point is noticeable as noise for some time. This point further originates in other neighbourhoods and characterized a portion of a cluster. When a data object is marked as a core, then the neighbourhood is too part of that group. Therefore, altogether the points present in the core neighbourhood and neighbourhood, are included in the group. This procedure is frequent until density-connected groups are entirely originated. Lastly, noise or deeper clusters are found by the retrieval of new and untreated points. After this step, objects present in the D dataset were verified and came to an end.

According to previous clarification of simple ideas and notions of the EDBC algorithm, its procedure is summarized as follows: Assume the spatial dataset D, a radius of clustering as, the negligible amount of neighbouring objects is, and the present assortment of entities is N1.

1. Every entity in dataset Dis unconstrained. Initial from some abandoned data position, mark it as 'tested,' then check neighbourhood and compute the sum of substances in the vicinityn. If m satisfiesm ≥, then generate a new gatheringC1, and add to C1, in the meantime add every point in the region to the gathering of objects N1.
2. Aimed at the gathering of substances, N1, if object was not patterned, then mark as ‘checked,’ and chequered its neighbourhood and compute the number of entities in the vicinity. If n satisfies n ≥, then it is included in the collection of object. It does not fit into some cluster, then include to C1.
3. Repeat step 2, and remain to a chequered object set N1 awaiting its vacancy.
4. Repeat steps 1 to 3 when every entity of information is denoted as 'checked,' and it completes the algorithm.

Algorithm 1 explains the detailed implementation which is as follows:

Algorithm 1. EDBC implementation of algorithm steps

**Input Data:** Process the dataSet D

**Output Data:** Satisfies the Cluster requests.

**Parameters:** Minimal number of neighboring points MinPts and Clustering Radius Eps

**Step 1:** Start

**Step 2:** DensityCluster (D, Eps, MinPts )

\{ ClusterNum = 0

**Step 3:** for each unchecked point M as noise in D

**Step 4:** set M as checked then

NeighbourResult = NeighbourSearching(M, Eps )

if sizeof(NeighbourResult) ≥ MinPts

ClusterNum = UpdateClusterNum

enlarge NeighbourPart (M, Neighbour Outcome, ClusterNum, Eps, MinPts )

else

set M

**Step 5:** set the Sub function:

expandNeighbourPart (M, Neighbour Outcome, ClusterNum, Eps, MinPts )

NeighbourSearching(M, Eps )

**Step 6:** End the process.

3.2 Proposed Architecture

The image retrieval are utilized in a specific combination. The various modules that correspond to the system are,

- Dataset collection module
- Pre-Processing
- Testing and Training Dataset
- Feature Extraction
- Classification

3.2.1 Module Description

The modules mentioned aboverefer to various subsets of architecture system which aids in retrieving the images through this recovery process. These modules are clarified as,

a. Dataset Collection

This module is utilized in collecting the unstructured datasets like images are reused for content with quality based retrieval process through data mining methods.

b. Pre-Processing

In this module, filtering based on smoothing spatial algorithm is used to strainer the needless noises present in an image. The image is filtered to accentuate certain features or remove other features. In most of the cases, this feature is interactive. In general, pre-processing is the alteration of an image by altering the brightness values of pixels to enhance the visual influence of images.

c. Testing and Training Dataset

This module deals with the testing of the contents that are being processed in the pre-processor unit. These datasets are then trained to possess the attributes that are to be utilized for retrieving the image.

d. Feature Extraction

To describe the content of an image, the various features like shape, texture, color, etc. are generally utilized where image’s feature can be categorized into primitives. Here, the information from the raw data which is more relevant for discrimination amid the classes is extracted.

Extraction of features has to be made over low within-class patchiness and high between-class changeability. Also, the redundant information has to be discarded during extraction.

e. Classification

This deals with the procedure of utilizing samples of recognized informational classes, i.e., training sets to categorize pixels of unknown characteristics.
This type of classification deals with examining a huge number of unidentified pixels and divides it into some classes founded on natural groupings existing in image values.

By employing all these modules, the accuracy of image recovery is in the course of development and ultimately provides an output with enhanced. Figure 6 depicts the structural design of this technique.

**Figure 3: Architecture diagram of the proposed method**

Figure 3 explains the proposed method. When an input image is given, pre-processing is done by smoothing spatial algorithm to reduce the unwanted distortions or improves some features in the given image. A low-level feature is extracted by an improved EDBC algorithm which helps in attaining the images from the large database rapidly. In this method, a construct indexing is applied to an image features database based on Haar wavelet transform [24] technique. During this process, the matching image concept uses features to find similarity among the images. Depending on the values of similarity, the images are partitioned into clusters. The proposed CBIR technique is estimated by visualization and query formulation method. This interface permits a user to identify a query using a query design and to visualize the recovered like images. Then, the query images determine the precision values of the retrieved results. The query-processing module excerpts a feature vector from a query design and apply a metric such as the Euclidean distance to calculate the similarity between the database images and query images. It ranks the database images in descending instruction of match to the query image and onwards the most like images to the boundary module. Finally, that database images are often indexed rendering to their feature vectors to improve recovery, resemblance computation, and image identification. In this type, classification is a technique which inspects a large number of unidentified pixels and divisions it into many classes founded on natural groups existent in the image values. Ultimately, our approach provides more accuracy.

**IV. RESULTS AND DISCUSSIONS**

The outcomes obtained revealed that content-based image and data related to repossession of text enhanced the scalability lowered the maintenance costs significantly and progressed the efficacy of software development. In this study, an image dataset is considered which includes 347 images in jpg, gif, tiff, bmp formats. Various interpretations achieved through the tests conducted were expressed as,
Figure 4 shows the login page of the content image retrieval processing. Username and Password are entered for further process.

Figure 5: Upload image dataset in Jpg format

Figure 5 displays the uploaded input image dataset in Jpg format which consists of both RGB and CMYK conversion. JPEG images include metadata which describes the content of the file.

Figure 6: Upload image dataset in GIF format

Figure 6 shows the uploaded input image dataset in GIF format which consists of both RGB and CMYK conversion. GIF format contains more than 256 colors. It is performed by isolating the image into multiple blocks which contains unique 256 colors.

Figure 7: Upload image dataset in BMP format
Figure 7 shows the uploaded input image dataset in GIF format which consists of both RGB and CMYK conversion. In this image, each scan line contains consecutive bytes representing the pixels present in the scan line in the order left to right.

Figure 8 shows the uploaded input image dataset in PNG format which consists of both RGB and CMYK conversion. It prevents the transparency of the uploaded files.

Figure 9: Indexing all image dataset based on cluster
Figure 9 displays the indexing set of all input images from training dataset.

Figure 10 depicts the group of pictures obtained in the dataset. This might be achieved over the suggested practice of image retrieval employing EDBC procedure. The clustered image present in the dataset is shown in Figure 9. This collection has to be made from the designated group only. It can be noted that an image should be chosen by the user which acts as the input for the extraction procedure. This section seems like the configuration window in which, the operator has to inject his query or might search an image through this algorithm. In this study, 329 features are extracted from an image dataset.

Figure 11 (a): Illustration for Content and shape based image extraction
Figure 11 (b): Illustration for Content and shape based image extraction

Figure 11 (c): Illustration for Content and shape based image extraction
Figure 12: Input image RGB histogram

Figure 12 shows the histogram of the input image, and it has RGB values to point out the variations present in the uploaded image.

Figure 13: Input image CMYK histogram

Figure 13 depicts the color variation with image quality. Here, the image quality ratio is high in CMYK when compared with RGB.

Figure 14 characterizes the graphical demonstration of content-based image extraction, which displays the assessment amid the entire sum of descriptions and the amount of extracted images. Here, the local features and global features are obtained by dense matrix and color histograms respectively.

Figure 14: Comparison graphs of total and search datasets
From the obtained results, it was evident that, from this system of image recovery, the operator can acquire a selective array of search data retrieved through the extraction process from the complete dataset.

Firstly, to evaluate the developed process and compare it with advanced techniques, three datasets clusters are selected and verified. In these experiments, the suggested scheme is associated with every approach utilizing the similar dataset where every source was recognized as an inquiry image. If a recovered source is the same class as of the inquiry image, then, the method is identified as successful; otherwise, it fails. Here, the precision is utilized as a performance measure.

\[
\text{Precision} = \frac{\text{Number of Relevant Images}}{\text{Number of Relevant Images} + \text{Number of Irrelevant Images}}
\]  

The investigational outcomes and comparisons are presented in Table 2. Here, the table displays the comparison of fallouts for three image extraction techniques. Figure 15 represents the evaluation of the proposed EDBC method with existing methods [25].

**Table 2: Comparison of the accuracy with the other methods**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Image 1 (%)</th>
<th>Image 2 (%)</th>
<th>Image 3 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed EDBC Method</td>
<td>87.37</td>
<td>92.6</td>
<td>95.3</td>
</tr>
<tr>
<td>Color-Based Search (HSV)</td>
<td>55</td>
<td>67.16</td>
<td>75</td>
</tr>
<tr>
<td>Color-Based Search (YCbCr)</td>
<td>65</td>
<td>68.1</td>
<td>70.5</td>
</tr>
<tr>
<td>Color and wavelet Based Search (Symtal)</td>
<td>68.2</td>
<td>70.2</td>
<td>76.62</td>
</tr>
<tr>
<td>Color and wavelet Based Search (Daubechies)</td>
<td>70</td>
<td>71.2</td>
<td>78.25</td>
</tr>
<tr>
<td>Fuzzy Partition Based Search</td>
<td>71</td>
<td>72</td>
<td>75.24</td>
</tr>
<tr>
<td>Curvelet Based Search</td>
<td>72</td>
<td>75</td>
<td>81.165</td>
</tr>
</tbody>
</table>

![Figure 15: The accuracy measure of the EDBC method with existing methods](image-url)
Figure 15 illustrates the experimental outcomes regarding accuracy for image 1, image 2 and image 3. It shows that the EDBC method outperforms all the existing techniques and hence it is appropriate for retrieving images similar to the query images. The proposed technique shows 95.3% accuracy which suits well for the recovery of imageries from the database.

Comparison of proposed EDBC method with existing methods is presented in Table 3.

Table 3: Comparison of proposed EDBC method with existing methods

<table>
<thead>
<tr>
<th>Existing and proposed methods</th>
<th>Accuracy (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Neural Network (DNN) &amp; Bacteria foraging optimization algorithm [1]</td>
<td>88</td>
</tr>
<tr>
<td>Convolutional Neural Network (CNN) [22]</td>
<td>92</td>
</tr>
<tr>
<td>CNN-based image retrieval [16]</td>
<td>83.4</td>
</tr>
<tr>
<td>Proposed EDBC method</td>
<td>95.6</td>
</tr>
</tbody>
</table>

Figure 16 illustrates the proposed results of the CBIR system with image feature descriptors which provides efficient accuracy of 95.6% using EDBC method. Clustering is performed to group the similar features of the images into a single group. The proposed method provides better processing time and image retrieval accuracy for query image.

V. CONCLUSION AND FUTURE WORK

In this paper, an improved and proper content with texture-based image retrieval method through Efficient Density-based Clustering Algorithm (EDBC) mechanism is introduced for enhancing scalability. Moreover, it gave a useful image feature for image retrieval techniques. The extensive features are taken out from the entire image and the extracted features are used to measure the similarities between the images. Then, the entire image performs color histogram like CMYK and RGB. Extracting the features with low within-class variability and high between class variability discards the redundant information. The procedure of using samples of recognized informational classes (working out sets) is performed to classify pixels of unknown identity. In this type of classification, this method inspects a large number of unknown pixels and splits it into many classes grounded on natural groupings existent in the image values. The experiments have validated the efficiency and effectiveness of the proposed methodology. Ultimately, the proposed approach provides accuracy to 95%. The proposed method is compared with existing techniques and proves that the suggested algorithm provides better accuracy and scalability.

In the future, researchers can use different kinds of features along with existing features and machine learning techniques like neural network classification in the successful recovery of media. One such direction for further investigation of this method is to customize this in handling other media types like video and audio through a texture-based retrieval technique.

REFERENCES

9. Saad S, Kumar H & Tewari T K (2017 August) Efficient content-based image retrieval using SVM and color histogram. In Contemporary Computing (IC3) 2017 Tenth International Conference on (pp 1-3) IEEE.