Alike Content Detection in Image and Video

Mr. Shibi Thambi k
PG Student (M. Tech)
Department of Applied Electronics & Communication
Thejus Engg. College

Ms. Vidya R Menon
PG Student (M. Tech)
Department of Applied Electronics & Communication
Thejus Engg. College

Mr. Sreerag S
PG Student (M. Tech)
Department of Applied Electronics & Communication
Thejus Engg. College

Abstract

This paper recognizes the problem of detecting alike content from images and video for content based retrieval applications. With the development of digital multimedia data types and available bandwidth there increase a demand of image retrieval systems and the users shift from text based retrieval systems to content based retrieval systems. Selection of extracted features play an important role in content based retrieval systems. These features are used for indexing, selecting and ranking according to the user. Good features selection also reduces the time and space costs of the retrieval process. Content based image retrieval (CBIR) is a technique in which it uses visual contents to search images from the large image or video databases. Alike features can be derived from a set of feature extraction methods like SIFT and MSER. Here the features generated are then combined together to form feature code book. Both mid-level feature description techniques and multi class support vector machine is used as classifiers for detecting purposes. Analysis is done using four different image classes' viz. cars, flowers, buildings and human faces.

Keywords: CBIR, CBVR, MSER, Retrieval System, SIFT, SVM

I. INTRODUCTION

Techniques for content based image retrieval (CBIR) are mainly based on low level features like color, texture and shape. Detection of such contents has been extensively studied over the last couple of years. However researches are more in image retrieval based on alike or high level content i.e. images or video frames are need to be retrieved which are almost similar to some given image, rather than visually factor. Systems which can automatically detect, analyses, and search image databases have been developed both in researches and in commercial concerns. Most of the early research system performs retrieval based on low level image features, such as color and texture. They expect the user to give a query image which is an example of images that is to be retrieved. The query image features are then extracted and database images with similar feature values are returned, mainly in basis of similarity. The commercial versions of these systems are not proven to be useful because human users, who are looking for images for illustrations, think in terms of high level concepts that should appear in an image and they usually do not have an example of the image to show. Commercial systems such as those provided by Corbis, Inc. and Getty Images still use keyword indexing. There is already a wealth of literature in the area of content based image retrieval related to global color and texture features. One approach is to define a structural feature that captures the structure of a class of images [1] [2]. Another is to segment the images into different regions, which have characteristic color [3] and texture or characteristic shape [4]. Queries can also request images that have regions with certain properties in spatial relationships [6]. Another approach is to employ user relevance feedback [5] to refine the query results, but this is mainly paired with the query by example approach.

For alike content detection therefore systems are first designed to extract a set of low level feature values and then at a higher level a mapping is done to associate a set of low level features with high level concept. This paper addresses the problem of detecting alike content concepts from images and video by using a number of low level feature extraction methods. The paper is organized as follows: section 2 provides an overview of related work, section 3 outlines the proposed methodology, section 4 provides details of the dataset and experimental results obtained, and section 5 provides the overall conclusions and the scope for future research.

II. RELATED WORKS

Lin et al. [7] proposed a color texture and color histogram based image retrieval system (CTCHIR). They are (1) three image feature values based on color, texture and color distribution as color co-occurrence matrix (CCM), difference between pixels of scan pattern (DBPSP) and color histogram for Kmean (CHKM) respectively and (2) a method for image retrieval by integrating CCM, DBPSP and CHKM to improve image recognition rate and simplify the computation of image retrieval

process. From the experimental result analysis they found that, their proposed method outperforms the Jhanwar et al. [8] and Hung and Dai [9] methods. Raghupathi et al. [10] have done a comparative study between image retrieval techniques having different feature extraction methods like color histogram, color histogram+gabour transform, Gabor Transform, Contourlet Transform and color histogram+contourlet transform. Hiremath and Pujari [11] proposed CBIR system based on color, texture and shape features by dividing the image into tiles. The features analyzed on tiles serve as the local descriptor values of color and texture keypoints. The color and texture methods are analyzed by using two level grid frameworks and the shape feature is analyzed by Gradient Vector Flow. The comparison of the experimental result of proposed method with other system [12] [13] found that, their retrieval system gives better performance than that of others [14].

III. PROPOSED APPROACH

The proposed approach mainly makes use of two stages for the CBIR based detection 1) Feature extraction and representation, 2) Classification of the content. Figure 1 shows the architecture of the proposed scheme. It shows the training and testing sections by using SVM classifier.

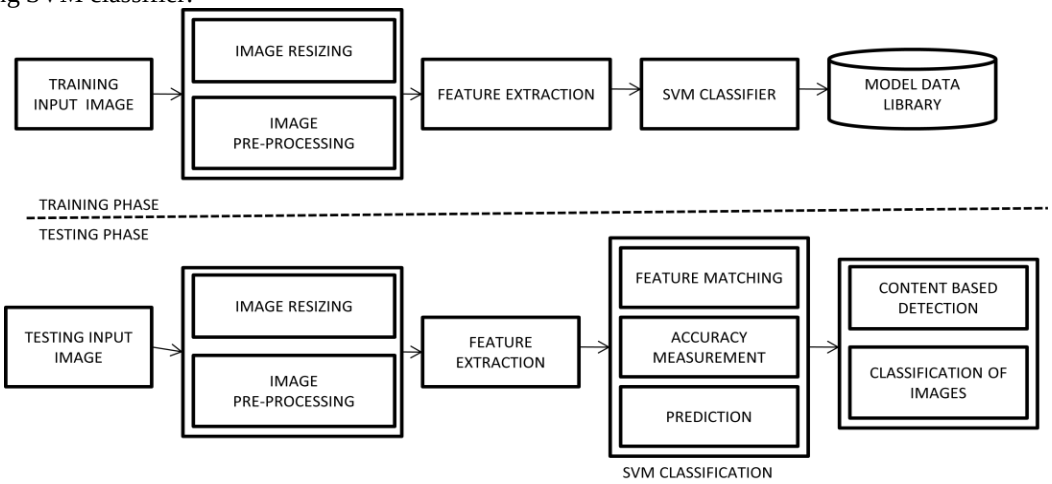

Fig. 1: proposed system architecture

A. Feature Extraction and Representation:

Feature extraction explains the relevant shape information contained in an image so that the task of arranging the image is made easy by a simple procedure [15]. In video processing and in image processing, feature extraction is a special form of dimensionality reduction method. The main aim of feature extraction is to extract the most useful information from the data and to represent that information in a lower dimensional space. When the input data to an extraction algorithm is very large to process then the input data will be transformed into a reduced representation of features or features vector. Transforming these input data into set of feature values is called feature extraction. If the features obtained are carefully chosen it is expected that the reference features set will extract the information from the input data in order to perform the required task using this reduced representation instead of the entire size input. Feature extraction is done after the preprocessing stage in character recognition systems. The primary stage of pattern recognition is to take an input pattern and correctly arrange in the possible output classes. This process can be further divided into two general stages: Feature descriptor selection and Classification. Feature selection is critical to the entire process since classifiers will not be able to detect from selected features.

1) *Maximally Stable Extremal Region (MSER):*

In this paper we are using Maximally Stable Extremal Region (MSER) features [17]. In computer vision techniques, maximally stable extremely regions (MSER) are used as a method of detection. This technique is mainly used to find out correspondences between image elements from two images with different viewpoints. This method of extracting a number of corresponding image elements contributes to wide baseline matching. The algorithm is based on the number of pixels. It starts by first sorting the pixels by intensity. After sorting, pixels are then marked in the image. In practice these steps are very fast. During this process, the area of every connected component represents as a function of intensity and it is stored to produce a data structure. A merge of two components is analyzed as termination of existence of smaller component and insertion of all pixels of the smaller component into larger one. MSERs are region that are either darker, or brighter than their surroundings and that are stable across a wide range of thresholds of the intensity. MSERs have also been defined on other scalar functions [18]. In the extremal regions, the 'maximally stable' ones are those corresponding to the thresholds where the relative area changes as a function of relative changes of threshold at a local minimum, i.e. the MSER are the part of the image where local binarization is stable over a wide range of thresholds.

According to MSER algorithm [16], the maximally stable extremal region R is defined as:

$$\rho(R;\Delta) = \frac{|R_{+\Delta}|-|R_{-\Delta}|}{R} \qquad (1)$$

Where R is a maximally stable extremal region when $\rho(R;\Delta)$ is minimum. The extremal region also called as α-connectivity and it is a maximal connected component of a level set $S(i)$ in which the pixel intensity is not greater than i. $R_{+\Delta}$ is the smallest extremal region containing R and has intensity which exceeds of at least Δ intensity of R . $R_{-\Delta}$ is the biggest extremal region contained by R, and has intensity which is exceeded at least Δ intensity by R. The MSER procedure is carried out from $i = 0$ to L_m for minimum MSERs, and from $i = L_m$ to 0 for maximum MSERs, L_m is the maximum gray value of the input image.

2) *Scale-Invariant Feature Transform (SIFT):*

Scale-invariant feature transform is an algorithm in computer vision to detect and extract local features in images. The algorithm was published by David Lowe in 1999. SIFT can easily identify objects even among high clutter and under partial occlusion, because the SIFT feature descriptor is invariant to uniform orientation, scaling and partially invariant to affine distortion and illumination variations. SIFT key points of objects are first extracted from a collection of reference images and stored in a database. An object is detected in a new image by individually comparing every feature from the new image to this database and finding matching features based on Euclidean distance of their feature vectors [19]. From the full group of matches, subgroup of key points that agree on the object and its scale, location, and orientation in the new image are analyzed to filter out good matches. The determination of consistent clusters is then performed rapidly by utilizing an efficient hash table implementation of the Hough transform. Each cluster of three or more features that agree on an object and its pose is then subject to detailed model verification and subsequently outliers are avoided. Finally the probability that a particular set of features indicates the presence of an object is find out, given the accuracy of match and number of probable false matches.

3) *There Are Mainly Four Steps Involved In SIFT Algorithm:*

a) Scale-Space Extrema Detection:

In order to detect larger corners we have to use larger windows. For this, scale space filtering is used. Here Laplacian of Gaussian is found out for the image with different σ values. LoG acts as a blob detection which detects blobs in various sizes due to change in σ. In short, σ acts as a scaling parameter. But this LoG is a little costly, so SIFT algorithm uses Difference of Gaussians which is an approximation of LoG. Difference of Gaussian is obtained as the difference of Gaussian blurring of images with two different σ.

b) Keypoint Localization:

Once potential keypoints locations are found, they have to be filter to get more accurate results. DoG has higher response for image edges, so edges are need to be removed. For this purpose, a algorithm similar to Harris corner detector is used. A 2x2 Hessian matrix (H) to compute the principal curvature is used. We know from Harris corner detector that for edges, one eigen value is larger than the other. So it eliminates any low contrast keypoints and edge keypoints and what remain will be strong interest points.

c) Orientation Assignment:

Now an orientation is given to each keypoint to obtain invariance to image rotation. A surrounding region is taken around the keypoint location depending on the scale and the gradient and direction is calculated in that region. An orientation histogram with 36 bins covering 360 degrees is created. The largest peak in the histogram is taken and any peak above 80% of it is considered to calculate the orientation. It creates keypoints with same location and scale, but with different directions.

d) Key point Descriptor:

Now keypoint descriptor is formed. A 16x16 neighbourhood around the keypoint is taken. It is divided into 16 sub blocks of 4x4 sizes. For each sub block, 8 bin orientation histograms are created. So a total of 128 bin values are obtained. It is represented as a vector to get keypoint descriptor. In addition to this, several procedures are taken to achieve immunity against illumination changes.

e) Keypoint Matching:

Keypoints between two images are then matched by identifying their nearest neighbors. But in some of the cases, the second closest match may be very near to the first one. It may happen due to noise or some other reasons. In that case, ratio of closest distance to the second closest distance is taken. If it is greater than 0.8, they are rejected. It eliminates around 90% of false matches while discards only 5% correct matches.

B. SVM Classification:

SVM is the useful technique for data classification. A classification task usually involves with training and testing data which consist of data instances. Each instance in the training set contains one target value and several attribute values. The goal of SVM is to produce a model which predicts target value of data instances in the testing phase which are given only the attributes. Classification in the SVM is an example of Supervised Learning. If the feature descriptors in the original feature space are not linearly separable, SVMs preprocess and represents them in a space of higher dimension where they become linearly separable. The dimension of the transformed space may vary. With a suitable nonlinear mapping to a high dimension, data from two different classes can always be made in to linearly separable, and separated by a hyper plane. The choice of the nonlinear mapping depends on the information available to the designer. If such information is not permitted, one might choose to use Gaussians, polynomials, or other types of basis functions.

Here LIBSVM package [20] is used to solve the standard SVM problem in the learning framework of different video activities. For multi-class classification, we apply the one- vs -all training scheme.

IV. EXPERIMENTAL RESULTS

The experiment was conducted by using a dataset having images collected from internet, 80 low resolution images with 4 different subjects. Four classes such as car, flower, human face and buildings were chosen for conducting the experiment. First at all the images are preprocessed then training and testing is done. 10 images from each image class were taken and used for training.ie; the number of samples used for training is 40 and for testing all four separate images for each classes were used i.e., 40 testing samples were used. The feature extractions on each image are done by MSER features and SIFT features. The descriptors like orientation, axes, location of interest points were analyzed. The classification is then done by using multi class SVM classifier. Here we were used LIBSVM package for one vs-all classification and for their accuracy measurement.

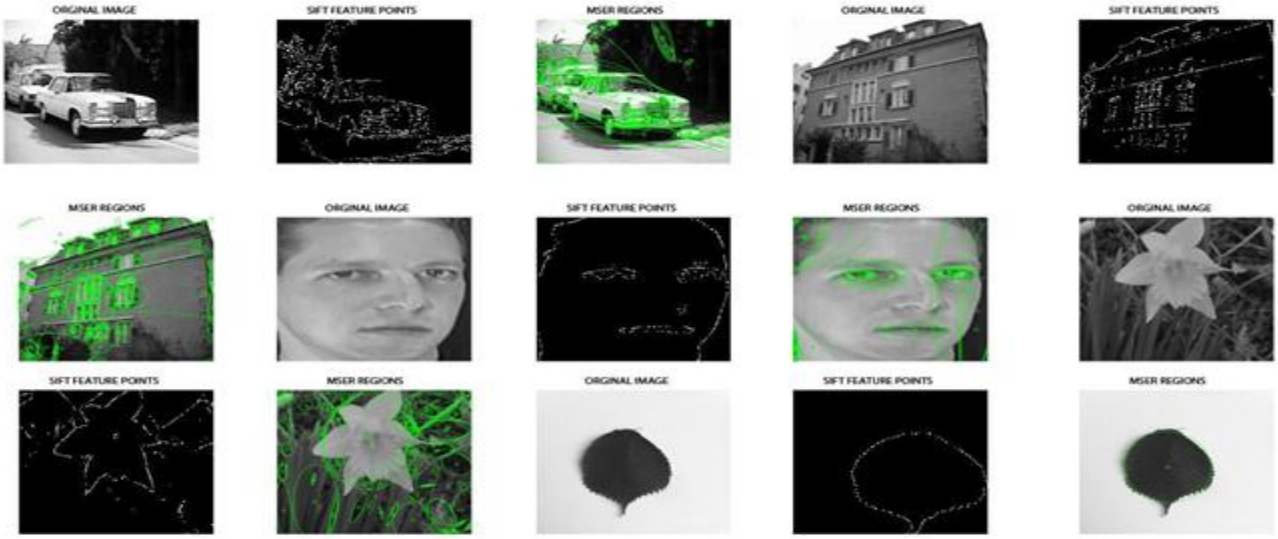

Fig. 2: Feature extraction (MSER and SIFT)

Table-1:
Percentage of accuracy for MSER

TEST NO	*CAR*	*FLOWER*	*FACE*	*BUILDING*
1	*44.2*	*32.1*	*39.8*	*39.9*
2	*44.2*	*31.7*	*39.1*	*39.5*
3	*37.3*	*31.6*	*41.6*	*38.5*
4	*37.3*	*31.5*	*26.8*	*41.6*
5	*44.2*	*33.4*	*36.4*	*34.6*
6	*44.2*	*18*	*23*	*40*
7	*20.6*	*30.9*	*42.8*	*37.5*
8	*20.6*	*33*	*23.7*	*34.9*
9	*44.2*	*21.3*	*25.2*	*37.1*
10	*44.2*	*31.6*	*33.6*	*27.2*
MEAN	*38.1*	*29.51*	*33.2*	*37.08*

Table-2:
Percentage of accuracy for SIFT

TEST NO	*CAR*	*FLOWER*	*FACE*	*BUILDING*
1	*45.9*	*39.9*	*71.9*	*57*
2	*45.9*	*40.6*	*61.2*	*81.2*
3	*45.9*	*39.5*	*69.8*	*64.4*
4	*48.8*	*17.8*	*73.6*	*77.5*
5	*48.8*	*32.5*	*73.2*	*63.5*
6	*47.2*	*55.5*	*67.9*	*74.7*
7	*47.2*	*45.8*	*67.2*	*69.2*
8	*37*	*45.1*	*69.9*	*69*
9	*37*	*31.8*	*68.2*	*83*
10	*70.6*	*40*	*66.9*	*72.2*
MEAN	*49.9*	*38.85*	*68.98*	*71.17*

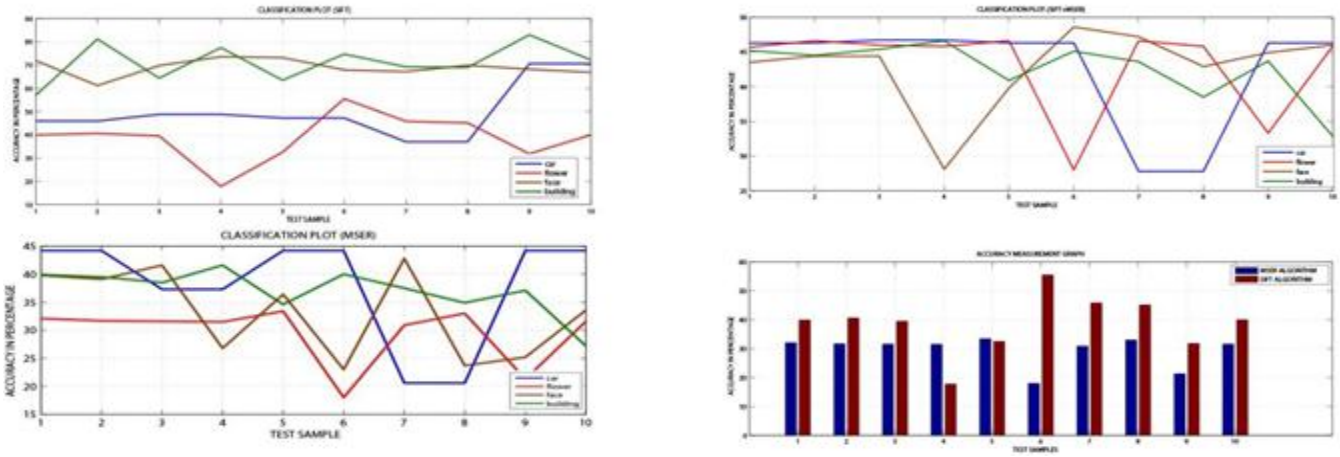

Fig. 3: Classification and accuracy plot

V. CONCLUSION

This paper proposes a system for detecting alike contents from images and video frames using multiple features extraction methods, SIFT and MSER algorithms. Multi class SVM is used as classifier for testing the efficiency of the system. From the analysis it is found that the SIFT approach shows higher accuracy and but have a higher complexity thereby resulting in slower retrieval. Individually the MSER produces varying accuracy but have a greater speed. Thus the accuracy can be increased by combining these two feature extraction methods. The SVM approach only gives good classification accuracy result with multiple image classes. And for probabilistic approach and complex procedure it's a fastest classifying approach. This work can be extended further by incorporating an advanced learning system wherein the retrieved videos can be reused to teach the system further thereby providing fine tuning of the process. Here only image based retrieval concept is taken in account and only four classes have been included. In future, more image classes and data samples can be included. Also the system can be further improved into audio and video retrieval by using high level feature extraction methods.

REFERENCES

[1] A. Vailaya, A. K. Jain and H.-J. Zhang, "On Image Classification: City Images vs. Landscapes", Pattern Recognition, Vol. 31, pp 1921-1936, 1998.
[2] S. X. Zhou, Y. Rui, and T. S. Huang, "Water-filling Algorithm: A Novel Way for Image Feature Extraction Based on Edge Maps," in Proc. IEEE Int. Conf. on Image Proc., 1999.
[3] C. Carson, S. Belongie, H. Greenspan, J. Malik, "Regionbased Image Querying," Proceedings of the 1997 IEEEWorkshop on Content-Based Accesss of Image and Video Librayies, pp. 42-49, June 1997.
[4] A. Del Bimbo, P. Pala, S. Santini, "Visual image retrieval by elastic deformation of object sketches," IEEE Symposium on Visual Languages, pp. 216-223, 1994
[5] Y. Rui, T. Huang, and S. Mehrotra, "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval," Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries, pp. 67–74, 1997.
[6] J.R. Smith and S.-F. Chang, "VisualSEEK: A Fully Automated Content-Based Image Query System," ACM Multimedia, pp. 87-98, November, 1996.
[7] C.H. Lin, R.T. Chen and Y.K. Chan, "A smart content-based image retrieval system based on color and texture feature", Image and Vision Computing vol.27, pp.658–665, 2009.

[8] N. Jhanwar, S. Chaudhurib, G. Seetharamanc and B. Zavidovique, "Content based image retrieval using motif co-occurrence matrix", Image and Vision Computing, Vol.22, pp-1211–1220, 2004.
[9] P.W. Huang and S.K. Dai, "Image retrieval by texture similarity", Pattern Recognition, Vol. 36, pp- 665–679, 2003.
[10] G. Raghupathi, R.S. Anand, and M.L Dewal, "Color and Texture Features for content Based image retrieval", Second International conference on multimedia and content based image retrieval, July-21- 23, 2010.
[11] P. S. Hiremath and J. Pujari, "Content Based Image Retrieval based on Color, Texture and Shape features using Image and its complement", 15th International Conference on Advance Computing and Communications. IEEE. 2007.
[12] Y. Chen and J. Z. Wang, "A Region-Based Fuzzy Feature Matching Approach to Content Based Image Retrieval", IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 24, No.9, pp. 1252-1267, 2002.
[13] Y. Rubner, L. J. Guibas and C. Tomasi, "The earth mover's distance, multidimensional scaling, and color-based image retrieval", Proceedings of DARPA Image understanding Workshop. Pp- 661-668, 1997.
[14] Manimala Singha* and K.Hemachandran$ "Content Based Image Retrieval using Color and Texture" Signal & Image Processing: An International Journal (SIPIJ) Vol.3, No.1, February 2012
[15] Gaurav Kumar, Pradeep Kumar Bhatia "A Detailed Review of Feature Extraction in Image Processing Systems" 2014 Fourth International Conference on Advanced Computing & Communication Technologies
[16] J. Matas, et al., "Robust wide-baseline stereo from maximally stable extremal regions," In Proc. BMVC, pp.384-393, 2002
[17] J. Matas, O. Chum, M. Urban, and T. Pajdla. "Robust wide baseline stereo from maximally stable extremal regions". In 13th BMVC, pages 384–393, September 2002.
[18] S. Obdrz´alek. "Object Recognition Using Local Affine Frames" PhD thesis, Czech Technical University, 2007.
[19] David G. Lowe "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, 2004.
[20] C.-C. C h a n g and C.-J. Lin, "LIBSVM: A library f o r support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, p. 27, 2011