International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 2 148 – 153

Clustering of Images from Social Media Websites using Combination of Histogram and SIFT Features

M.Vadivukarassi[1], N. Nanthini[2], N. Puviarasan[3], P. Aruna[4]
[1]Research Scholar, [2] M E Student, [3] Associate Professor, [4]Professor
[1, 2, 4] Department of Computer Science and Engineering, [3]Department of Computer and Information Science
Annamalai University
Chidambaram, India
[1] *vadivume28@gmail.com*, [2]*nanthini0294@gmail.com*, [3]*npuvi2410@yahoo.co.in*, [4]*arunapuvi@yahoo.co.in*

Abstract— In recent years, the rapid growth of high dimensional datasets has created an emergent need to extract the knowledge. With the tremendous growth of social network, there has been a development in the amount of new data that is being created every minute on the networking sites. This work presents an efficient analysis of SIFT and color histogram features with spectral clustering algorithm. In this work the images from social media websites are downloaded. The downloaded images are stored in the database. The proposed feature extraction technique is based on combination of both SIFT descriptor and color Histogram to increase the efficiency. The extracted features are then clustered using spectral clustering algorithm. The spectral clustering method is a clustering area which achieves the clustering goal in high dimension by allowing clusters to be formed with their own correlated dimension.

Keywords- *social media; SIFT; histogram; spectral clustering.*

*****

I. INTRODUCTION

With millions of people spending countless minutes on social media to share, communicate, connect and create user-generated data at an unparalleled rate. Social media has become unique source of big data. This novel source of rich data provides unmatched opportunities and great potential for research and development. Social media sites have grown increasingly popular in the last few decades. Initially, they were created as tools utilized in communication and connecting people, but in the recent years they have evolved extremely. Social media have strongly changed our lives and how we interact with one another and the world around us.

Image retrieval is the fast growing and challenging research area with regard to both still and moving images. Retrieval focuses at developing new techniques that support effective searching and browsing of large digital image libraries based on automatically derived imagery features. It is a rapidly expanding research area situated at the intersection of databases, information retrieval, and computer vision. Meanwhile, the next important phase today is focused on clustering techniques. Clustering algorithms can offer superior organization of multidimensional data for effective retrieval. Clustering algorithms allow a nearest neighbor search to be efficiently performed. Hence, the image mining is rapidly gaining more attention among the researchers in the field of data mining, information retrieval and multimedia databases. The main goal is to mine the images from the social media by extracting the features of the images using SIFT descriptor from the image database and then display the results according to the human expectations based on spectral clustering algorithm.

This paper is made further as: Section II discusses related work examined till now. Section III describes overall system design. Section IV presents result and analysis of the work using the graphical analysis. Section V closes with the conclusions and presents future work.

II. RELATED WORKS

Hsin-Chien Huang, et al. (2012) proposed an affinity aggregation spectral clustering algorithm ring, SIFT, spectral clustering. For aggregating affinity matrices for spectral clustering, it was more immune to ineffective affinities and irrelevant features. Also, it enables the construction of similarity measures for clustering less crucial. Hanqiang Liua, et al. (2012) proposed a simple algorithm called non-local spatial fuzzy spectral clustering which incorporates both the spatial and neighborhood relationships of the pixel. in this method, the detail of the images are well preserved. Hasan Mahmud, et al. (2016) presented a hand gesture recognition system using SIFT features, they applied the SIFT features on binary images and keypoints from the images are used in k-means clustering to reduce the feature dimensions. Osameh Biglari, et al. (2014) studied the Scale invariant feature transform and speeded up robust features to compare them in different color spaces for human detection. Jerrin Varghese (2015) studied image search based on scale invariant feature transform descriptors using k-means clustering algorithm.

III. PROPOSED SYSTEM

The block diagram of the proposed image retrieval method is shown in Fig.1. There are three modules employed in the proposed work. They are image database collection, feature extraction and image clustering.

A. Image Database collection

Image database is a collection of image data, typically associated with the activities of one or more related

organizations. It focuses on the organization of images and its metadata in an efficient manner. The first part of our proposed work deals with collecting images from social media website. By giving the URL of the social media websites, the images are downloaded based on the keyword. In this work, "Vardah" cyclone is considered as a keyword. The images related with "Vardah" cyclone are downloaded from social media websites like Facebook, Twitter, etc. Here, 100 images are taken for the proposed work.

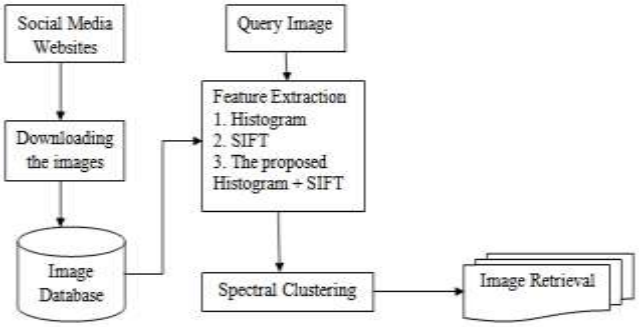

Fig.1. Block diagram of the proposed work

B. Feature Extraction method

The second module of the work deals with feature extraction. Feature extraction involves reducing the amount of resources required to describe a large set of data. Analysis with a large number of variables generally requires a large amount of memory and computation power. Also, it may cause a classification algorithm to over fit to training samples and generalize poorly to new samples. This is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. SIFT and color histograms are the feature extraction methods used to extract the salient features of the images.

SIFT:

Scale-invariant feature transform (SIFT) is an algorithm in computer vision to detect and describe local features in images. SIFT detects and uses a much larger number of features from the images, which reduces the contribution of errors caused by these local variations in the average error of all feature matching errors.

Step 1: SIFT octaves are generated by constructing a scale space using,

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (1)$$

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (2)$$

Step 2: From Gaussian values, difference of Gaussians which are equivalent to laplacian of Gaussians and it can be calculated using the formula,

$$G(x, y, k\sigma) - G(x, y, \sigma) \approx (k-1)\sigma^2\nabla^2 G \quad (3)$$

Step 3: Then, key points of objects are first extracted with Taylor expansion from a set of reference images and stored in a database.

An object is recognized in a new image by individually comparing each feature from the new image to this database and finding candidate matching features based on Euclidean distance of their feature vectors. From the full set of matches, subsets of key points that agree on the object and its location, scale and orientation in the new image are identified to filter out good matches.

Step 4: Finally the probability that a particular set of features indicates the presence of an object is computed, given the accuracy of fit and number of probable false matches.
Fig. 2 shows the extraction of SIFT features from "Vardah" cyclone images downloaded from the social media websites.

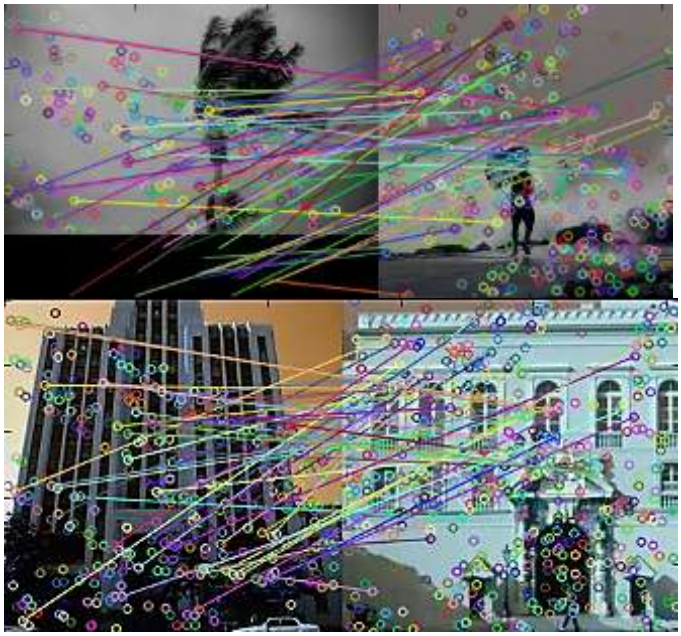

Fig. 2 Extraction of SIFT features from images

Color histogram:

Color histogram is a representation of the number of pixels of every color present in the image taken. While separating the color pixels, if any of the color space is huge, then it is broken down into smaller intervals. Those intervals are called bins. In our experiments, the number of bins considered is 2. This process is called color quantization. The final step is to obtain the color histogram using bin values.

C. Image Clustering algorithm

Image clustering is an application of cluster analysis to images which is nothing but it divides a set of images into cluster, so that image within each cluster is similar in content. Clustering algorithms provides a useful tool to explore data structures.

In this proposed work, spectral clustering algorithm is used for image retrieval. The spectral clustering methods are easy to implement and reasonably fast especially for sparse data sets up to several thousands. It treats the image clustering as a graph partitioning problem without making any assumption on the form of the image clusters. It makes use of spectral-graph structure of an affinity matrix to partition data into disjoint meaningful groups. It requires robust and

appropriate affinity graphs as input in order to form clusters with desired structures. Constructing such affinity graphs is an on trivial task due to the ambiguity and uncertainty inherent in the raw data. Most existing spectral clustering methods typically adopt Gaussian kernel as the similarity measure, and employ all available features to construct affinity matrices with the Euclidean distance, which is often not an accurate representation of the underlying data structures, especially when the number of features is large.

Spectral Clustering is a three stage process. They are as follows

1. *Pre-processing*

In this step, the graphs are constructed and the dataset represents the similarity matrix.

2. *Spectral representation*

In this step, the associated Laplacian matrices are formed and the Eigen values and Eigenvectors of the Laplacian matrix are computed. Then, each point to a lower-dimensional representation based on one or more Eigenvectors are mapped respectively.

3. *Clustering*

In this step, the points to two or more classes, based on the new representation are assigned. Data matrixes from feature extraction are given as input to cluster the images using spectral clustering algorithm. By using, this output we are retrieving the images similar to the query image.

Algorithm 1 shows the steps involved in spectral clustering.

Algorithm 1: Spectral clustering Algorithm

Step 1: Form the affinity matrix $A \in R^{nXn}$

Step 2: Define $A_{ij} = e^{-\|s_i - s_j\|^2/2\sigma^2}$; If i ≠ j, $A_{ii} = 0$

Step 3: Define D a diagonal matrix whose (i,i) element is the sum of A's row i

Step 4: Form the matrix $L = D^{-1/2}AD^{-1/2}$

Step 5: Find $x_1, x_2, \ldots \ldots, x_k$ the k largest eigenvectors of L.

Step 6: These form the columns of the new matrix X.

Step 7: Form the matrix Y, Renormalize each of X's rows to have unit length, $Y_{ij} = X_{ij}/(\sum_j {X_{ij}}^2)^2$; $Y \in R^{nxk}$ .Treat each row of Y as a point in R^k

Step 8: Cluster into k clusters via K-means

Step 9: Final Cluster Assign point to cluster j if row i of Y was assigned to cluster j.

In this proposed work, above SIFT descriptor and color histogram are combined to form feature vector. The image retrieval process is represented with the help of the hierarchal index tree as shown in Fig 3. In hierarchal index tree, non-leaf nodes have variable number of children within some pre-defined limit. When data is altered from a node then, its child gets changes. To maintain the pre-defined tree depth, non-leaf nodes may join or split. This tree is kept balanced by keeping all leaf nodes at same depth. The root node's children has upper limit as same as the non-leaf nodes.

For example, consider 35 images in the database; it takes user query as the root node. Related images are taken as children in the first level; this process iteratively goes on until the similar images are obtained.

IV. RESULTS AND DISCUSSION

The proposed work is implemented in MATLAB. In this work, 100 images are downloaded from social media websites like Facebook, Twitter, etc., related to "Vardah" cyclone which distracted Chennai, Tamilnadu, India in the month of December 2016. Here, various feature extraction methods were used and implemented. Color histogram feature is used to extract only the color feature of the image. Then, SIFT feature is used to extract the salient feature of an images. In this proposed work, the combinations of color and SIFT features were used to extract the features of the images.

Experiments are conducted on exclusive histogram and SIFT features separately and on the proposed combination of histogram and SIFT features.

Performance measures:

Performance is measured with three factors: Precision, Recall and Accuracy. Precision is ratio of the number of relevant image retrieved and total number of image retrieved. Precision is denoted by P.

$$Precision = \frac{Number\ of\ Relevant\ image\ retrieved}{Total\ number\ of\ image\ retrieved} X100$$

Second factor, recall is ratio of number of relevant image retrieved and total number of image retrieved.

$$Recall = \frac{Number\ of\ Relevant\ image\ retrieved}{Total\ number\ of\ relevant\ images\ in\ the\ database} X100$$

Third factor, Accuracy is a weighted arithmetic mean of precision and its inverse as well as a weighted arithmetic mean of recall and its inverse.

$$Accuracy(A) = \frac{Precision + Recall}{2}$$

It is found from the Table 1 that the proposed combination of histogram + SIFT features with spectral clustering gives better performances with precision of 92%, recall of 83% and accuracy of 90%.

TABLE 1 Comparison of existing and proposed feature extraction methods

TECHNIQUES	PRECISION (%)	RECALL (%)	ACCURACY (%)
Histogram	86	58	72

SIFT	72	56	64
Proposed Histogram+SIFT	92	83	90

From Fig. 4, it is observed that the accuracy of the proposed combination of both histogram and SIFT feature extraction is better than the exclusive histogram and SIFT features separately. Fig. 5(a & b) shows the screenshot of the image retrieval of the system using histogram and SIFT separately. Fig. 6 shows the image retrieval of the system using the proposed combination of histogram and SIFT feature extraction.

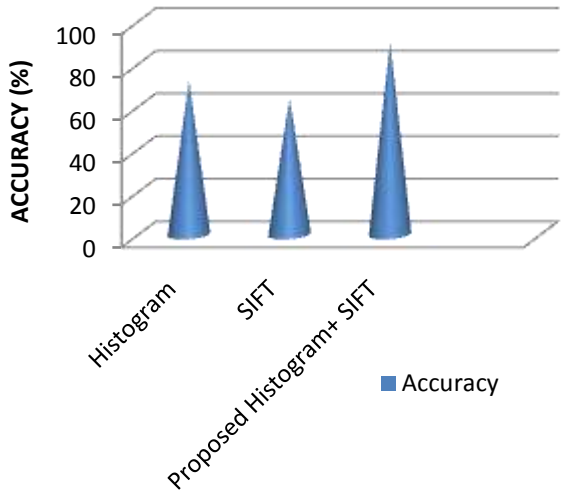

Fig 4: Accuracy of the various feature extraction methods

V. CONCLUSION

In this work, various feature extraction methods were implemented with spectral clustering algorithm in MATLAB. The experiment results shows the accuracy of feature extraction with histogram as 72%, SIFT as 64% and the proposed combination of histogram and SIFT as 90%. From the results, it is observed that the proposed technique of feature extraction which is built on both combination of SIFT and histogram shows higher performance than extraction with either SIFT or color histogram. Hence, the combination of histogram and SIFT feature extraction can be used to retrieve relevant images than exclusively other techniques.

REFERENCES

[1] Hanqiang Liua, Feng Zhaob, Licheng Jiao," Fuzzy spectral clustering with robust spatial information for image segmentation," Proceedings of applied soft computing, Elsevier, vol.12, pp. 3636-3647,2012.

[2] Hasan Mahmud, Md. Kamrul Hasan, Abdullah-Al-Tariq, M. A. Mottalib," Hand Gesture Recognition Using SIFT Features on Depth Image," Ninth International Conference on Advances in Computer-Human Interactions, ISBN: 978-1-61208-468-8,2016.

[3] Mingling Zheng, Zhenlong Song, Ke Xu, and Hengzhu Liu, "Parallelization and Optimization of SIFT Feature Extraction on Cluster System," International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol.6, 2012.

[4] Osameh Biglari1, Reza Ahsan, Majid Rahi, "Human Detection Using SURF and SIFT Feature Extraction Methods in Different Color Spaces," Journal of mathematics and computer Science, vol.11, pp. 111-122, 2014.

[5] Jerrin Varghese , "GUI Based Large Scale Image Search with SIFT Features," International Journal of Science and Research,Vol.4, Issue 9, pp.2319-7064, September 2015.

[6] Shan Zeng, Rui Huang, Zhen Kang, Nong Sang, "Image segmentation using spectral clustering of Gaussian mixture models," Proceedings of Neurocomputing, vol.14, pp. 0925-2312, 2014.

[7] Frederick Tung, Alexander Wong, David A. Clausi, Enabling scalable spectral clustering for image segmentation:in proceedings of Pattern Recognition, Elsevier, vol.43, pp. 4069-4076,2010.

[8] Yifang Yanga,c,, Yuping Wangb, Xingsi Xueb ,A novel spectral clustering method with superpixels for image segmentation: in proceedings of Optik, Elsevier, vol.127,pp.161–167, 2016.

[9] Dipesh Patel, Darshan Patel, "Improvement in Performance of Image Retrieval using Various Features in CBIR System," International Journal of Computer Applications,Vol.138, pp. 0975 – 8887, 2016.

[10] Aboli W.Hole, Prabhakar L.Ramteke," Design and Implementation of Content Based Image Retrieval Using Data Mining and Image Processing Techniques,"International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 3, March 2015.

[11] Annkur S.Mahalle, Snehal H. Kuche," Image Mining and Clustering Based Image Segmentation," International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 3, March 2015.

[12] Jerrin Varghese," GUI Based Large Scale Image Search with SIFT Features," International Journal of Science and Research (IJSR), Volume 4 Issue 9, September 2015.

 ISSN: 2321-8169
148 – 153

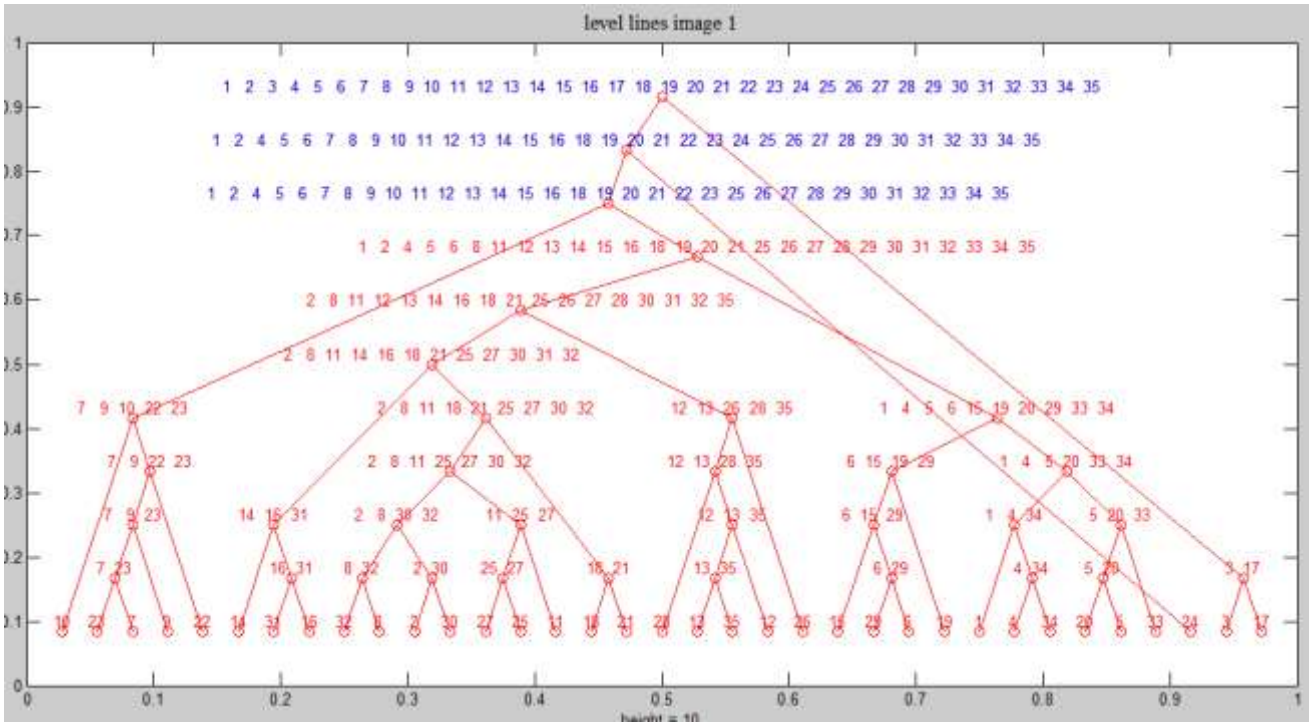

Fig. 3 Hierarchical index tree

Fig. 5(a) shows the retrieval of images using SIFT

 ISSN: 2321-8169
148 – 153

Fig. 5(b) shows the retrieval of images using Color histogram.

Fig. 6 Retrieval of images using the proposed combination SIFT and color histogram