



Tweet categorization and image retrieval using proposed feature extraction method with subspace clustering algorithms

M. Vadivukarassi^{1*}, N. Puviarasan², P. Aruna³

^{1,3}Department of Computer Science and Engineering,

²Department of Computer and Information Sciences, Annamalai University,
Annamalainagar, Tamilnadu, India

Abstract : The massive amount of user generated content about the real world events is accumulated by social media at every minute. Twitter has gained tremendous popularity with more than 500 million tweets per day for the past few years. The twitter data can be monitored through the Twitter Streaming API. This paper proposed a new feature descriptor method on clustering techniques to improve the performance of the image retrieval system from Twitter dataset. The real time dataset is downloaded from Twitter Streaming API using Python script. From these dataset, the tweets and images are extracted and stored in the separate database. The tweets are analyzed using Naïve Bayes algorithm for tweet categorization. Here these tweets are categorized into different category and the top most categories are detected for further process. Then toppest category tweets are filtered separately and the toppest keyword is detected from these filtered tweets. Based on the keyword detection, the images are retrieved using the proposed Seg_SIFT feature extraction method and subspace clustering algorithms. These algorithms are compared and analyzed based on the elapsed time of image retrieval based on the performance measures such as precision, recall and accuracy. The detailed experimental results show that the elapsed time of k-subspace is lesser than the seq-k-subspace clustering algorithm while retrieving the images from the database. Seq-k-subspace clustering algorithm performs better than the k-subspace clustering algorithm. This proposed work enables us to discover and understand the tweets and images easily.

Keywords : micro blogging; Twitter; Seg_SIFT; k-subspace clustering; seq-k- subspace; social media.

1. Introduction

The recent explosion of social networks is gaining popularity every day and has become a very important platform to discuss about the social events. Using micro-blogging services, the users post messages about their daily life and initiate discussions on different topics by sharing their personal opinions and emotions. These accumulated huge amounts of available behavioural data in online social networks give a chance for knowledge

M. Vadivukarassi et al /International Journal of ChemTech Research, 2018,11(08): 325-346.

DOI= <http://dx.doi.org/10.20902/IJCTR.2018.110841>

discovery. Data mining techniques detects implicit or hidden information's available within social networking sites. Twitter provides micro-blog services and also it is an online social media website. It allows writing messages up to 140 characters at one time in which typically not more than 30 words. It can be possible for the social analysts to understand the public attitude regarding different social issues. Finding the present topmost discussing issues regionwise in the social website, is the need of the day. But it is difficult to search them with better accuracy. Hence, this research paper proposed a new method for tweet categorization and image retrieval using proposed feature extraction method and clustering algorithms. The structure of this paper is as follows: Section 2 reports the related work of keyword detection, feature extraction, image clustering and retrieval; Section 3 describes proposed system of the work; Section 4 describes experimental setup and the results are discussed using tables and graphs. Finally, we provide concluding summaries in Section 5.

2. Related works

The various techniques of CBIR such as k-means clustering, k-nearest neighbours Algorithm(KNN), Colour Structure Descriptor(CSD), Text based image retrieval (TBIR) techniques which increase the effectiveness of fast retrieval are discussed and analysed (K. B. Jayarraman et al., 2016). The keyword search and the list of items are displayed and analyzed. Then, the collected twitter items are stored in the database. Finally, the top results of twitter data are detected and these events are visually located on online Google map. This visualization enables us to discover and understand the events easily (P. Aruna et al., 2016). A novel framework to detect complex social events over streams, which fully exploits the data of social media over multiple extents, is proposed. They discussed a graphical model called location-time constrained topic (LTT) to capture the content, time, and location of social messages (Xiangmin Zhou et al., 2013).

The K-means clustering to accommodate extended clusters in subspaces, such as line shaped, plane-shaped clusters, and ball-shaped clusters are demonstrated. He studied a wide range of subspace clusters in various literatures, carried extensive experiments on both synthetic and real-world datasets, and results are demonstrated the effectiveness of the algorithm. The algorithm retains much of the K-means clustering flavours: easy to implement and fast to converge (Dingding Wang et al.). The two algorithms, TDA (Topic Detection using AGF) and TCTR (Topic Clustering and Tweet Retrieval), are implemented which will help to overcome this problem. From various experimental results and it is observed that the proposed method can maintain good performance irrespective of the size of the data set (Amrutha Benny et al., 2015). A non linear filtering of the image which preserves edges are described and on the use of a color point detector based on the Harris detector are investigated (Parvathy Ram et al., 2016). A method for extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene are presented and also describes an approach to using these features for object recognition (D. Lowe, 2004). The keywords are collected from Twitter using Twitter API and the extracted raw data are pre-processed using Natural Language Toolkit techniques. The sentiments of the online tweets are evaluated based on feature selection of score words (M.Vadivukarassi et al., 2017). The effectiveness of topic-focused trustworthiness estimation method with extensive experiments using real Twitter data are demonstrated (Liang Zhao et al., 2015).

3. Proposed System

This paper proposes a new system for the fast retrieval of efficient images from the database based on the keyword detected from the Twitter dataset. This proposed system uses the real time microblog image search. Fig. 1 shows the block diagram of the proposed system which has the following steps:

1. Tweets for different users are collected manually based on the different keywords between the certain periods of time using Twitter streaming API.
2. This collected twitter data is the input query. The tweets are extracted and categorized using Naïve Bayes classifier to categorize it under a certain label (like sports, politics, entertainment, technology, hospitality, etc.). Then the topmost category of the tweets is detected from the twitter data.
3. The detected category tweets are filtered separately and by using preprocessing and term frequency-inverse document frequency (tf-idf) method, the relevant keywords are detected.

4. Then related to detected keywords, the similar images are retrieved from the image database using proposed Seg_SIFT feature extraction method with different subspace clustering algorithms.
5. Finally, different subspace clustering algorithms are compared using precision, recall and accuracy measures.

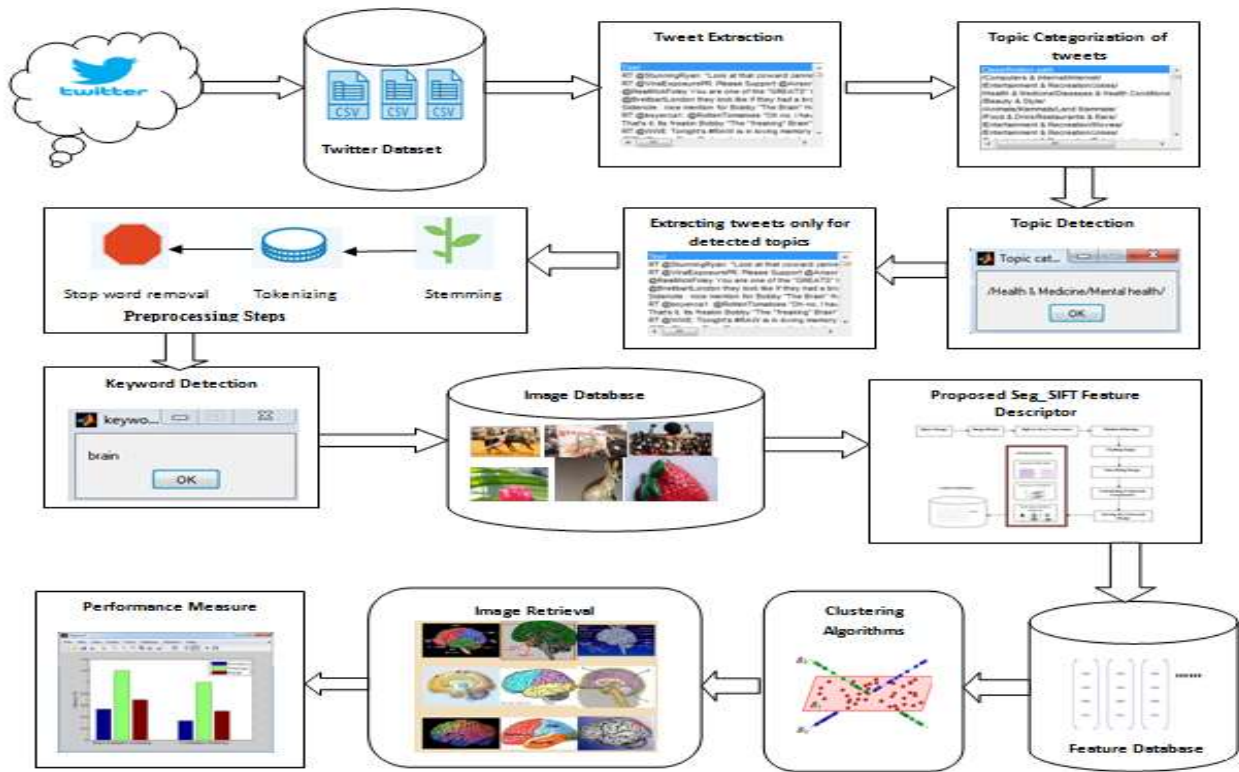


Fig. 1. Block diagram of the proposed system

3.1 Collecting tweets and images

Twitter allows users to download tweets with the help of Twitter streaming API. The Twitter stream is monitored and then tweets are collected based on the particular event. In this paper, Twitter dataset and the relevant images are manually collected and stored in the database. Each tweet contains the URLs of the images but not the images. The URLs are separated for corresponding images from the image storage Twitter websites. Once the tweets are stored in the database, they are preprocessed before applying to keyword detection.

3.2 Tweet Categorization

In this paper, Naive Bayes classifier, a probabilistic learning method is used for tweet categorization. It is the grouping of tweets into a fixed number of predefined classes. Tweet vectors are constructed, and commonly used term frequency-inverse document frequency weights are assigned and Naive Bayes (NB) classifier is used to classify the tweets. These tweets in NB would model as presence and absence of particular tweets. A variation of NB is known as Naive Bayes Multinomial (NBM), considers the frequency of words and can be denoted as:

$$P\left(\frac{c}{d}\right) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k / c) \quad (1)$$

Where $P(\frac{c}{d})$ is the probability of a document d being in class c , $P(c)$ is the prior probability of a document occurring in class c , and $P(t_k/c)$ is the conditional probability of term t_k occurring in a document of class c . A document d in our case is trend definition or tweets related to each trending topic. The input tweets are preprocessed and term-document matrix is applied to produce unigram and bigram features from the tweets and Naive Bayes classifier model is built to classify the label such as computers & internet, entertainment & recreation, health & medicine respectively.

Process flow for the Tweet classification.

Step 1: Reading the Data from .csv file.

Step 2: Divide the dataset into two parts as training dataset and testing dataset.

Step 3: Create a corpus for training dataset and testing dataset.

Step 4: Perform the following data processing transformation on the training dataset

and testing dataset.

a. Transform characters to lower case.

b. Convert into plain text document.

c. Remove punctuation marks.

d. Remove digits from the documents.

e. Remove the words which are redundant in the text mining (e.g. Pronouns,

conjunctions). Then these words as stopwords (“English”) which is a

built-in list for English language.

f. Remove extra white spaces from the tweets.

Step 5: Now create the “Term document matrix”. It describes the frequency of each

term in each tweet in the corpus and perform the transposition of it.

Step 6: Train Naïve Bayes model using transposed “Term document matrix” data

and Target class vector.

Step 7: Apply the prediction on generated model for testing dataset.

In Fig. 2, feature extractor is used to transform each input value into a feature set. These feature sets, which capture the basic information about each input that should be used to categorize it. Pairs of feature sets and labels are fed into the machine learning algorithm to produce a model during training set. Similarly during prediction, the same feature extractor is used to transform unobserved inputs to feature sets. These feature sets are then fed into the model, which produces predicted labels. The topmost category is detected by using predicted labels and related tweets are filtered based on the topmost category.

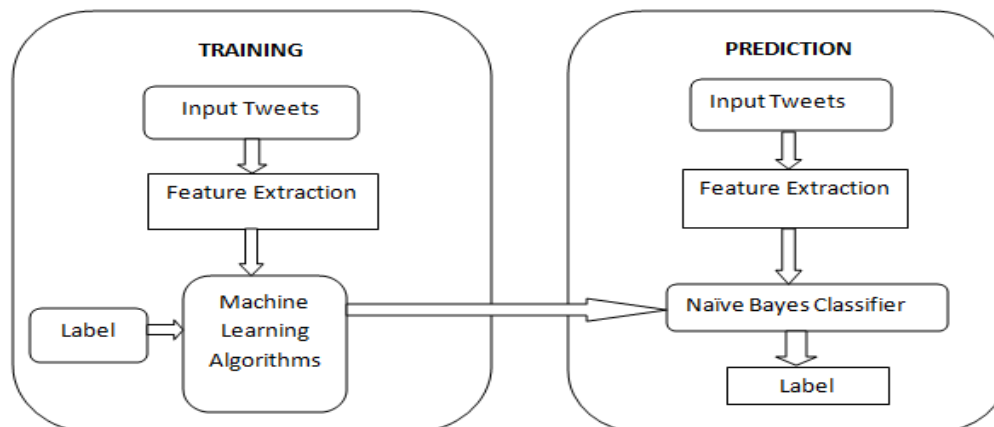


Fig.2. Flow diagram of tweet classification

3.3 Keyword Detection

In general, messages are written as sentences or sets of words in Twitter. At first, the texts are cleaned by removing punctuations, numbers, hyperlinks and stop words, followed by stemming and tokenization (M. Vadivukarassi et al., 2017) which is one of the most basic steps in text analysis. This tokenisation is used to split a flow of text into smaller units, usually called as tokens, words or phrases. Figure 3 presents the preprocessing steps of tweet message. In order to detect the keywords easier, extract the noun words from each tweet messages. Extracting keywords from Twitter data has proved as the most difficult and challenging task. In this paper, to identify frequently repeated keywords in the tweets, word frequency algorithm is used based on co-occurrence of words. The term frequency-inverse document frequency (tf-idf), is a numeric measure that is used to achieve the importance of a word in a tweet. Based on this measure, how often it appears in that text is calculated in a given collection of texts. The intuition for this measure is: If a word appears frequently in a text, then high score is assigned to that word. But if a word appears in too many other tweets, it's probably not a unique identifier; therefore assigned a lower score to that word. The math formula for this measure:

$$tf - id(t, d, D) = tf(t, d) \times idf(t, D) \quad (2)$$

where t denotes the terms; d denotes each text; D denotes the collection of text. The following steps are used to extract the unique keywords (i) Import the words from the text file into a cell array. (ii) All the characters are cleared but that should not be letters or numbers. (iii) Entries in words that have zero characters are removed. Now using the equation (1), appearing number of times of each word is counted. (iv) words without duplicates and the frequency of each word in unique words are counted. (v) Finally the results are printed. Using the above mentioned steps, the keywords are detected from the twitter data and applied to the next feature extraction process.

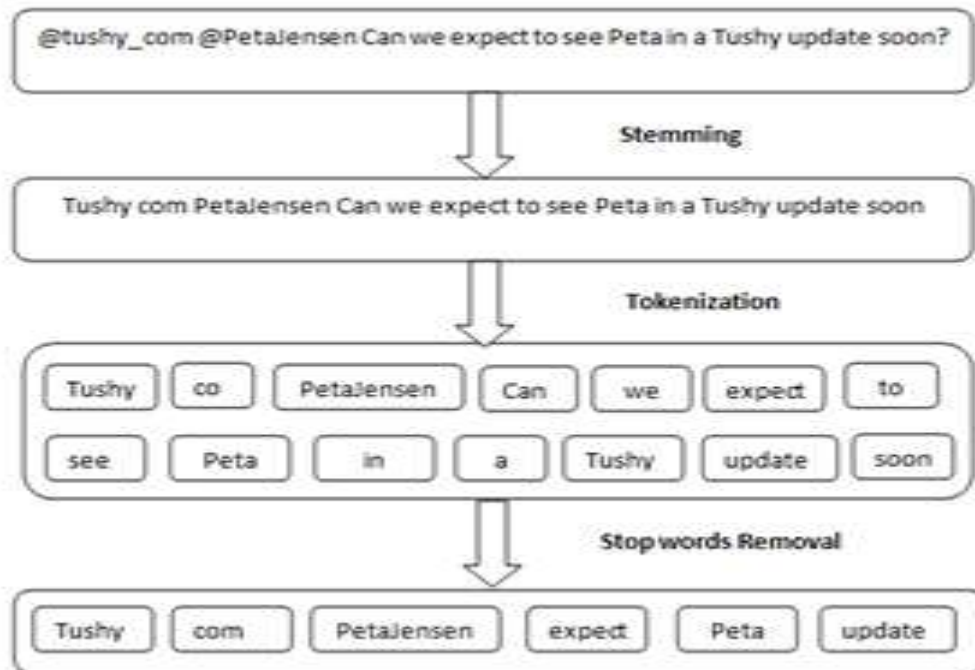


Fig. 3. Steps for preprocessing on tweet message

3.4 Feature extraction

Feature extraction is the process of detection, isolation and extraction of various desired portions or features of a digitized image. These features are used to compute the keypoints of a segmented image. Here, proposed Segmented Scale invariant feature transform (Seg_SIFT) techniques are applied to extract features from image.

3.4.1 Scale Invariant Feature Transform

In SIFT descriptor, difference of Gaussian detector is used to detect interest points and the extracted regions are described by a vector of dimension 128. The descriptor becomes illumination invariant when dividing the normalized descriptor vector by square root of sum of the squared components. A histogram of gradient orientation and location scale is used as a descriptor. A vector is associated with it as a descriptor (Jyoti Joglekar et al., 2010).

3.4.2 Proposed Seg_SIFT Descriptor

The proposed Seg_SIFT descriptor algorithm is shown as block diagram in Fig 4. The following are the steps of proposed Seg_SIFT algorithm:

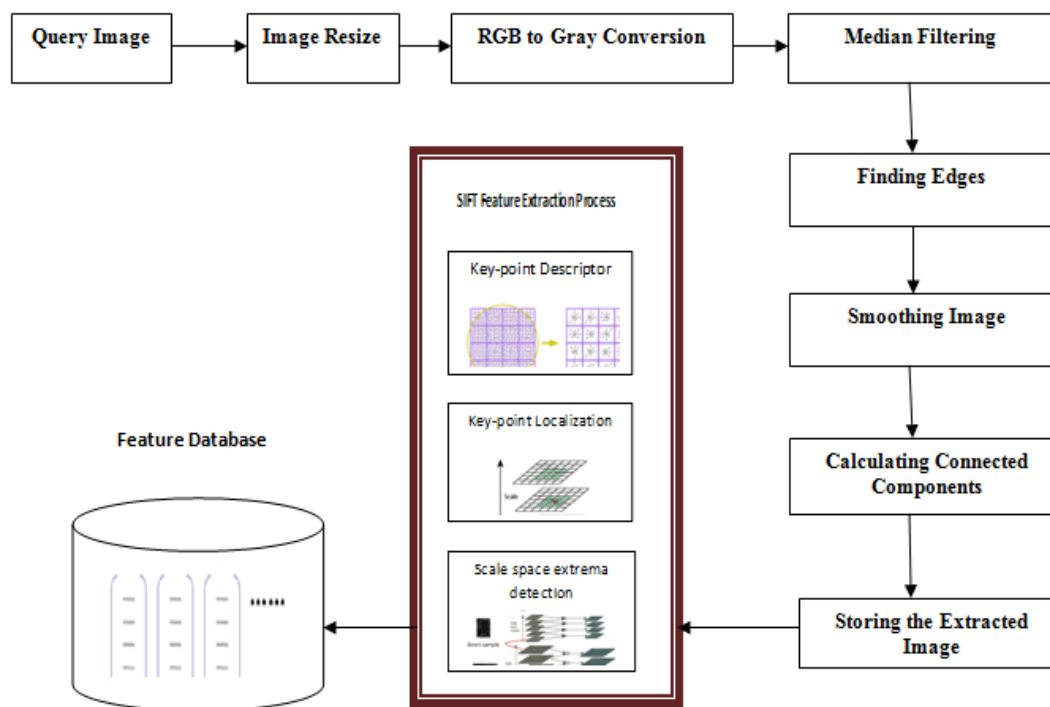


Fig. 4. Block Diagram of the proposed Seg_SIFT feature extraction

1. Segmented image Computation

In this proposed work, the color image is resized and converted into grayscale image. Then median filter is used to remove noise in the image and edges are found in the intensity image. The edge takes intensity or a binary image I as its input, and returns a binary image BW of the same size as I , with 1's where the function finds edges in I and 0's elsewhere. Here, Sobel method is used for finding edges using the Sobel approximation to the derivative. It returns edges at those points where the gradient of I is maximum. Smoothing image is done to reduce the number of connected components and maximum number of connected components are obtained. The computed extracted image is the input to the SIFT extraction process to extract the features.



Fig. 5. Sample Twitter images

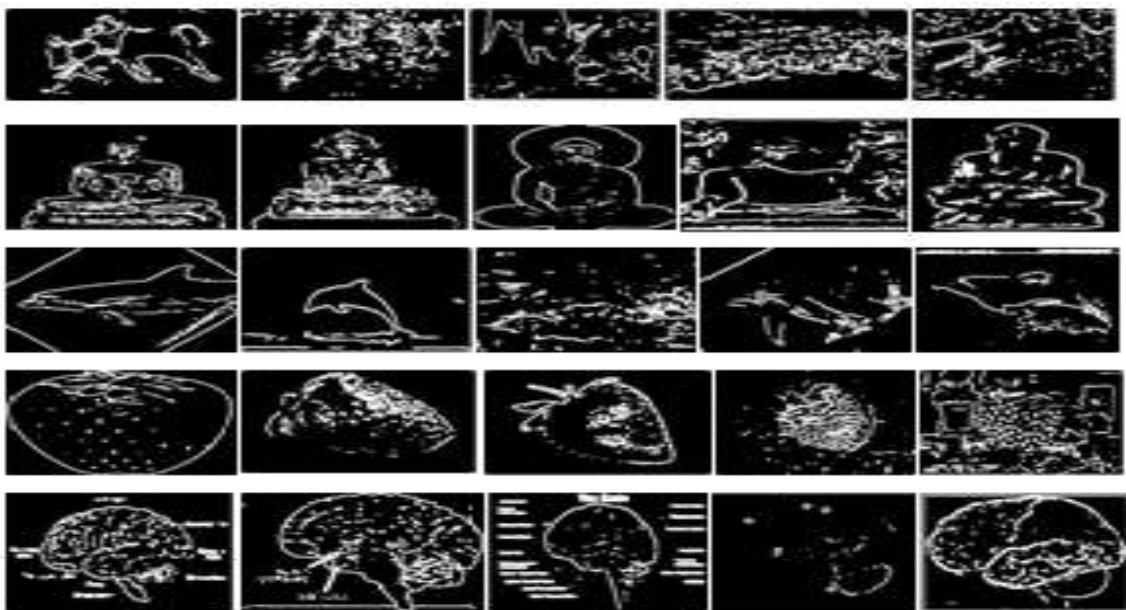


Fig. 6. Sample segmented images

Fig. 5 shows the sample query images used for testing the proposed work. The sample images of segmented images of the corresponding query images are shown in Fig. 6.

2. Constructing a scale space using SIFT descriptor

In SIFT feature extraction method, the scale space can be constructed by obtaining the original image to the half size and the number of blurred out images are produced again to form an octave (vertical images of the same size). In the proposed Seg_SIFT, first the original image is processed using median filter and smoothing of image is done to produce segmented image. Then segmented image is used as the input image to construct the four octaves and the individual image by increasing “scale” (the amount of blur). The number of octaves and scale depends on the size of the original image (Raval Vidhi et al., 2014). Mathematically, “blurring” is referred to as the convolution of the Gaussian operator and the image. Gaussian blur has a particular expression or “operator” that is applied to each pixel is given in equation (3).

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (3)$$

where

- L is a blurred image
- G is the Gaussian Blur operator
- I is an image
- x, y are the location coordinates
- σ is the “scale” parameter. Greater the value, greater the blur.
- The $*$ is the convolution operation in x and y . It “applies” Gaussian blur G onto the image I .

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (4)$$

The above equation 4 represents Gaussian Blur operator. The Laplacian of Gaussian technique is calculated by the difference between two consecutive scales for the image. This is computed using the given equation (5)

$$G(x, y, k\sigma) - G(x, y, \sigma) \approx (k - 1)\sigma^2 \nabla^2 G \quad (5)$$

The difference of Gaussians is a feature improvement algorithm that is computed from the subtraction of one blurred version of the magnitude image from another version of the same image which is less blurred (Dingding Wang et al., 2009). In the proposed feature extraction, after obtaining magnitude from HOG descriptor, SIFT feature extraction method is used to extract the salient features of the image.

3. Finding Key-points

The major key-points for matching purpose are recognized and retrieved the image. These keypoints are obtained by finding the edges, corners, blobs or even ridges. The first step of finding the image key-points is to find the maximum and minimum pixels from all its neighbours. The successive stage in the application of the SIFT algorithm is key-point localization. In order to compute the location and scale information, a model is fitted for each key-point candidate location. These key-points are chosen on the basis of the measurement of their stability and the interpolation process is carried out by utilizing the quadratic Taylor expansion of the Difference-of-Gaussian scale-space function (Alitappeh et al., 2012). The Taylor’s expansion is represented as follows:

$$D(x) = D + \frac{\delta D^T}{\delta x} x + \frac{1}{2} x^T \frac{\delta^2 D}{\delta x^2} x \quad (6)$$

where $x=(x, y)$ denotes the offset from this point. Once the key point is localized and refined, the low contrast points are rejected as depicted in Equations (7) and (8).

$$x = -\frac{\delta^2 D^{-1} \delta D}{\delta x^2 \delta x} \quad (7)$$

$$D(x) = D + \frac{1\delta D^T}{2\delta x} x \quad (8)$$

Then the next step is to determine and delete bad key-points which are edges and low contrast regions via fitting three dimensional quadratic functions, meanwhile this step will make algorithm robust and efficient. For this reason, a concept of a Harris corner detector is used. If the magnitude of the intensity at the current pixel in the difference of gaussian image is less than a certain value, then the key-point is rejected.

Algorithm 1: Proposed Seg_SIFT Descriptor

Input: Query image in the database

Output: Generating key points descriptor for the image and storing in feature database

Method:

Step 1: Read the query image u as the input

Step 2: Convert u_{RGB} to $u_{grayscale}$

Step 3: Compute median filter of $u_{grayscale}$ to remove the noise.

Step 4: Find the edges of the image u_{edge} using Sobel operator.

Step 5: Smoothing image is done to reduce the number of connected components and calculate the maximum number of components.

Step 6: Then extracted segmented image $u_{segment}$ is obtained.

Step 7: Initialize the empty octave matrix as $octave []$ and different octaves are formed using different sigma σ values such as $k^{2\sigma}, k^{4\sigma}$ using SIFT feature extraction method.

Step 8: Compute the Gaussian scale space $v_{scale space}$ of $u_{segment}$.

Step 9: Compute the Difference of Gaussian w_{DOG} of the $v_{scale space}$

Step 10: Find keypoints from w_{DOG} to produce list of discrete extrema $\{(x_d, y_d, \sigma_d)\}$

Step 11: Filter the unstable keypoints from w_{DOG} and $\{(x, y, \sigma)\}$ due to noise and laying on edges

Step 12: Assign a reference orientation to each keypoints $\{(x, y, \sigma, \theta)\}$ from the scale space gradient $(\partial_m V, \partial_n V)$ and list of keypoints $[(x, y, \sigma)]$

Step 13: Build the keypoint descriptor $\{(x, y, \sigma, \theta, f)\}$

4. Keypoint Descriptor

The keypoint localization based on Taylor series method is used to fit every key-point location and scale in SIFT feature extraction method. For each image, gradient magnitude and orientation is computed using pixel difference. But in the proposed Seg_SIFT method, segmented image is computed in the first step using HOG descriptor method. This is used as to find the key-point descriptor in SIFT process. At every key point, image gradient with one or more orientations are assigned. These orientations and magnitudes are used to construct a histogram at selected scales for the region around the key-point. These gradient magnitude and orientation are used at each image sample point is weighted by window to compute the descriptor. The algorithm of the proposed Seg_SIFT descriptor is described as shown below in algorithm 1. Here, 16×16 window is broken into sixteen 4×4 windows to generate the key-point. Within each 4×4 window N key-points for the image I , described by $f = \{x, y, angle, orientation\}$ and descriptor $d = f \times 128$. These descriptors are used further for matching the key-points (Iyad Abu Doush et al., 2016). At this step, Seg_SIFT feature vector is generated.

3.5 Clustering algorithms

In several areas of research, clustering techniques have been used in finding appropriate method for solving complex problems. These techniques involve the classification of data set into groups and has widely been accepted as a time saving method which has huge acceptability. The data in the equal subclass or group will have peculiar characteristic features that connect them and it makes out them from data items in other subclasses (Enikuomohin et al., 2016). Every data point is a vector of measurements in a multidimensional space. Clustering is an unsupervised data mining process of grouping similar data points into clusters without any prior knowledge of the underlying data distribution. A variety of distance measures can be used to quantify the similarity or density of the data points. Traditional clustering algorithms are designed to generate clusters in the full-dimensional space by measuring the nearness between the data points using all dimensions of a dataset (Zhang T et al., 1996). These classical clustering algorithms are ineffective as well as inefficient for the high-dimensional data. The high-dimensional data suffers from the curse of dimensionality which has two main implications:

(1) The relative contrast among similar and dissimilar points becomes less, when the dimensionality of data grows (Beyer K et al., 1999).

2) Under the different sets of dimensions, the data tend to group together differently (Parsons L et al., 2004).

Therefore, it becomes essential to find the clusters in the relevant subsets of dimensions of the data.

3.5.1 Subspace clustering

The process of finding clusters in the subspaces of the dataset is called subspace clustering. The k -dimensional data can have up to $2^k - 1$ possible subspaces. The exponential number of these subspaces creates single computational test for mining data with large number of magnitude. Assume $DB = \{P_1, P_2, \dots, P_n\}$ is a database of n points in a k -dimensional space, where each point is a k -dimensional vector, $P_i = P_{1i}, P_{2i}, \dots, P_{ki}$. A subspace S is a subset of the original attribute-set D , such that, an m -dimensional subspace is denoted as $S = \{d_1, d_2, \dots, d_m\}$ where, $d_i \in D$ and $1 \leq m \leq k$. A subspace S' is the projection of a high dimensional subspace S , if $S' \subset S$. The dimensionality of a subspace refers to the total number of dimensions participating to construct this subspace. Figure 7 shows the structure of subspace clustering.

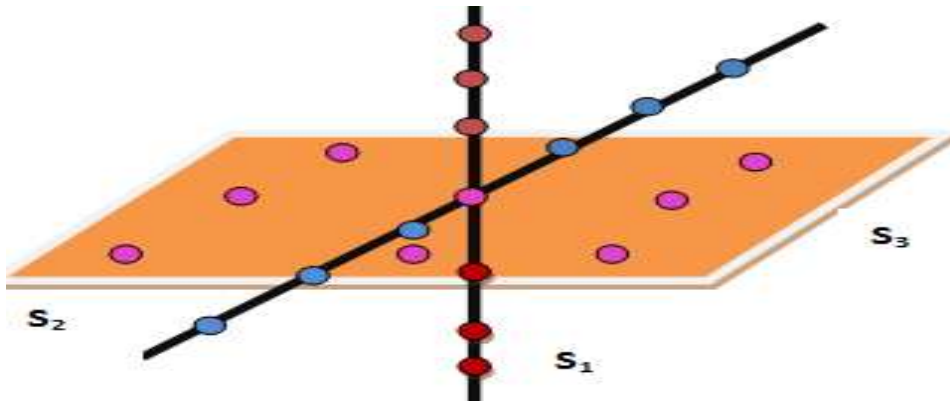


Fig. 7. Structure of subspace clustering

In this paper, subspace clustering techniques are used for effective retrieval of relevant images. These techniques are classified into two iterative clustering based methods called k-subspace clustering and seq-k-subspace clustering methods and they are similar to the k-means algorithm. However, instead of using cluster centres to model the data as in k-means they use subspaces.

3.5.1.1 k-subspace clustering algorithm

The k-subspace is developed as an extension to the well known as k-Means algorithm. The main function of k-subspace clustering algorithm is to form randomly oriented subspaces and to assign data points clustering that reside in them. The main basic steps of k- subspace clustering are

1. Form initial subspaces based on the number of clusters (k).
2. Assign each object taken from the data set to the nearest subspaces based on the data distance from the subspaces to complete the initial grouping process.
3. By the average of all data points that are assigned to the clusters are recalculated to form the new subspaces.
4. Repeat the steps 2 and 3 until there is no need for the subspaces to be adjusted.
5. Thus, the final clusters are formed.

The main advantage of k-subspaces is that it is easy to implement as each iteration corresponds to essentially assigning points to different clusters and estimating the subspaces. The clusters are generated by computing IDX (maximal subspace cluster). It indicates the cluster of each data and represented using formula in equation (9).

$$IDX = \max \sum_{i=1}^{iter} \sqrt{\sum_{k=1}^{SS(size)} (imp * imp)} \quad (9)$$

where *iter* indicates iteration, *SS(size)* indicates the size of the subspace and *imp* indicates inner product of each base of each subspaces. Algorithm for k-subspace clustering for the proposed work is given in algorithm 2.

Algorithm 2: K-Subspace clustering**Input:** X: data**Output:** IDX: cluster indexes of each data**Method:**

Step 1: **if** n_arg = 3 **then** th \leftarrow 0.75 //th-threshold, n_arg: number of arguments

Step 2: **if** n_arg = 4 **then** iter \leftarrow 3 // iter- maximum number of the iteration , //X-data

Step 3: th = th*th

Step 4: Find IDX by computing N by N matrix of zero X

Step 5: Find X_n by calculating the square root of sum of X // IDX-cluster indexes of each data

Step 6: Let SS \leftarrow [], k \leftarrow 1, IDX \leftarrow 0 // SS-initial subspaces, k: the number of the clusters

Step 7: **while** (sum[IDX]>0) **do** // IDX: cluster indexes of each data
 pos = IDX; tIDX = IDX(pos);
 XX = Xn(pos,:); XXX = X(pos,:);
 if (size_subspace \leq dim) **then** //dim- dimension of the subspaces
 Compute ipn by calculating maximum norm in projected subspaces of SS and XX
 return ipn //ipn-inner product of each base of each subspace
 Find ss by computing orthogonal basis of subspaces XX with dimension

Step 8: **for** i=1 **to** iter **do**
 Compute $P_{xx} \leftarrow XX * ss^T * ss$ // SS: basis vectors which represent subspaces

Step 9: $in_1 = (\text{sum}(P_{xx} .* XX, 2))$

Step 10: **if** ($in_1 \geq th$) **then**, assign $V \leftarrow XXX(in_1)$
 Compute V by finding eigen value of X
 return V, ss = V(1:dim,:);

Step 11: tIDX(in_1) = k; IDX(pos) = tIDX;

Step 12: SS \leftarrow concatenate SS and ss

Step 13: k \leftarrow k+1 // k-number of clusters,

Step 14: end

3.5.1.2 Seq-k-subspace clustering algorithm

Seq-k-subspace is an algorithm that attempts to find groups in the data. It is a vector space whose elements are infinite sequences of real or complex numbers. This algorithm works similar as k-subspace clustering. The main difference between these clustering algorithms is threshold value. In k-subspace, the k^{th} subspace is obtained without threshold value. But in seq-k-subspace clustering is done by assigning threshold value. It takes lesser number of iteration to generate the maximum norm in projected subspaces. At last, concatenation process is repeated by incrementing the k^{th} clusters. Algorithm for seq-k-subspace clustering for the proposed work is given in algorithm 3.

Algorithm 3: Seq-k-subspace clustering

Input: X: data

Output: IDX: cluster indexes of each data

Method:

Step 1: **if** n_arg =4 **then** assign SS←[] //n_arg-number of arguments, SS-subspaces

Step 2: **if** isempty(SS) **then** //SS-initial subspace

Generate SS₁ from random values of normal distribution using dim, size(X, 2), and k

for i=1to k **do** //k-number of clusters,

Calculate SS from orthonormal basis for the range of SS₁.

Step 3: Assigning SS as N // N-basis vectors which represent subspaces

Step 4: **if** n_arg =5 **then** assign iter←128

Step 5: **for** i=1to iter **do** // iter- max no. of iteration

Compute inpnorm in each subspaces of SS, X

//inpnorm-: max. norm in projected subspaces

Find maximum of inpnorm in subspaces and assign it to IDX

for j=1 to k **do**

if (IDX==j) **then** p = IDX // IDX- cluster indexes of each data

if (sum of elements in p > dim) **then** XX = X(p,:);

Calculate V[^] by finding eigen values of a square matrix XX

Assigning V[^] as V and **return** V

Assign SS(:,j) ←V(1:dim,:); **else**

Calculate SS by orthonormal basis for the random range of dim, size(X,2)

for m=1to size(SS,1) **do**

if (sum(SS(m,:)) < 0) **then** SS(m,:)= -SS(m,:);

Step 6: Calculate 'd' by finding absolute difference between SS and N

Step 7: d= max(d(:));

Step 8: **if** n_arg = 6 **then** assign ep ← 1E-3;

Step 9: **if** (d < ep) //ep - stoping criteria of the iteration

Step 10: end

The efficiency of the image retrieval is measured based on the performance of the feature extraction and clustering algorithms. Here, precision, recall and accuracy are the factors used to measure the performance of the retrieved images. The precision (P) in image retrieval is the ratio of the retrieved relevant images to the total retrieved images.

$$Precision = \frac{No\ of\ relevant\ image\ retrieved}{Total\ number\ of\ images\ retrieved} \times 100 \tag{10}$$

Recall (R) in image retrieval is the ratio of the retrieved relevant images to total relevant images in the database.

$$Recall = \frac{Number\ of\ relevant\ images\ retrieved}{Total\ number\ of\ relevant\ image\ in\ the\ database} \times 100 \tag{11}$$

After calculating precision and recall, Accuracy (A) is calculated which is the average value of precision and recall.

$$A = \frac{Precision + Recall}{2} \tag{12}$$

4. Experimental Results and Discussion

This section discusses more details of the effectiveness and performance of the proposed system. The discussion begins with experiment setup and design, and follows to cover data collection and implementation. Relevant results are also discussed and presented for a better understanding. 2500 tweets are collected with the help of Twitter streaming API. The tweets for the keywords such as Jallikattu, Strawberry, Brain, Buddha, and Dolphin are collected and are saved as excel file containing 44 attributes respectively. The attributes include id, text, source, lang, media_url etc. The tweets are collected during the month of January 2017 for six days on 26.01.2017 to 31.01.2017 respectively. A sample files for tweets are shown in Figure 8. All these tweets are pre-processed and top 10 keywords are detected from the tweets. The URL images in the attribute of each tweet are downloaded and saved as the image database. Table 1 shows the summary of the sample collection of the keywords. Table 2 indicates the summary of the keyword detection. Table 3 shows the tweet categorization using Naive Bayes classifier model. Using this model, each tweets is classified into different categories such as Computers & Internet, Health & Medicine, Entertainment & Recreation respectively. Finally, Health & Medicine category has the highest number of topics followed by other category.

	A	B	C	D	E	F	G
	id	Text	Source	Lang	CreatedAt	worltsCoulsRetw	
1	826278691298755000	RT @kalakkalcinema: கலாக்கலினிமா LIL தலை இலக்கலிபுத்தலை	<a href="http://twittv	ta	1/31/2017 3:59	0	TRU
2	826278669500957000	Inkhabar: #jallikattu சே சம்பலி சமீ யாபிகாசி பர #SupremeCou	<a href="http://ittt.c	hi	1/31/2017 3:59	0	FALS
3	826278547673252000	RT @lam_K_A: Pics speaks Truth!! #Thala #NadigarSangam #jallikatt	<a href="https://mob	en	1/31/2017 3:58	0	TRU
4	826278537438958000	RT @Cheatedbuyer: Animation on how forming groups can help rea	<a href="http://twittv	en	1/31/2017 3:58	0	TRU
5	826278442337522000	Jallikattu, like Shah Bano https://t.co/LnSgPqRW7N via @IndianExp	<a href="http://twittv	in	1/31/2017 3:58	0	FALS
6	826278392660111000	RT @Telugu360: To @WhackedOutMedia: we stand by article on @p	<a href="http://twittv	en	1/31/2017 3:58	0	TRU
7	826277911585911000	RT @Veeran_p: 6+ lakhs ppl were at marina beach for 90hrs #jallikat	<a href="http://dfg.p	en	1/31/2017 3:56	0	TRU
8	826277881651212000	RT @Veeran_p: 6+ lakhs ppl were at marina beach for 90 hrs #jallika	<a href="http://dfg.p	en	1/31/2017 3:56	0	TRU
9	826277870829920000	RT @Veeran_p: 6+ lakhs ppl were at marina beach for 90 hrs #jallika	<a href="http://dfg.p	en	1/31/2017 3:56	0	TRU
10	826277860063130000	RT @Veeran_p: 6+ lakhs ppl were at marina beach for 90 hrs #jallika	<a href="http://dfg.p	en	1/31/2017 3:56	0	TRU
11	826277826236137000	RT @CMOTamilNadu: Now #jallikattu is permanent!!!Hon President	<a href="http://twittv	en	1/31/2017 3:55	0	TRU
12	826277735848931000	சுடீல ஈகலபு... ஈலலகலகல குல ஈா கு ஈலகலகல ஈல	<a href="http://www	ta	1/31/2017 3:55	0	FALS
13	826277732816282000	#jallikattu சே சம்பலி சமீ யாபிகாசி பர #SupremeCourt ஈ சூல	<a href="http://twittv	hi	1/31/2017 3:55	0	FALS

Fig. 8. Collection of tweets

Table 1. Summary of the sample tweet

Keyword Name	Number of tweets collected	Number of images collected
Jallikattu	500	35
Buddha	500	85
Dolphin	500	65
Strawberry	500	35
Brain	500	40

Table 2. Summary of the keyword detection

Keyword Name	Number of words detected	Number of unique words detected
Jallikattu	5593	1589
Buddha	4549	1800
Dolphin	5321	1347
Strawberry	3703	1299
Brain	5347	1898

Table 3: Some categorization of Tweets

Classification path	Level 1 label	Level 1 probability
/Computers & Internet/Internet/	Computers & Internet	0.162
/Entertainment & Recreation/Jokes/	Entertainment & Recreation	0.297
/Health & Medicine/Diseases & Health Conditions/	Health & Medicine	0.442
/Beauty & Style/	Beauty & Style	0.436
/Animals/Mammals/Land Mammals/	Animals	0.281
/Food & Drink/Restaurants & Bars/	Food & Drink	0.230
/Entertainment & Recreation/Television/	Entertainment & Recreation	0.647
/Animals/Mammals/Land Mammals/	Animals	0.197
/Health & Medicine/Mental health/	Health & Medicine	0.333
/Entertainment & Recreation/Television/	Entertainment & Recreation	0.131
/Humanities/Social Science/	Humanities	0.350

The category of health and medicine related tweets are filtered separately and those tweets are used for keyword detection. The collected tweets are read from a plain text document and displays the most frequently used words in the tweets. Figure 9 illustrates the top 10 detected keywords from the twitter dataset. The first column consists of the most frequently used words in this text. The second column consists of the frequency of the word (i.e. the number of times that word appeared in the tweets). The last column contains the relative frequency of the word which is simply the frequency of the word divided by the total number of words in the tweets. This might be useful for statistical purposes and the words are case-sensitive, which means 'Great' and 'great' are treated as two different words. The keyword "Brain" is detected from the top most classified tweets. Figure 10 represents the graphical representation of detected keywords with relative frequency.

'WORD'	'FREQ'	'REL. FREQ'
'brain'	[143]	'3.025%'
'strawberri'	[67]	'1.417%'
'jallikattu'	[60]	'1.269%'
'buddha'	[55]	'1.164%'
'dolphin'	[45]	'0.952%'

Fig. 9. Top 5 detected keywords

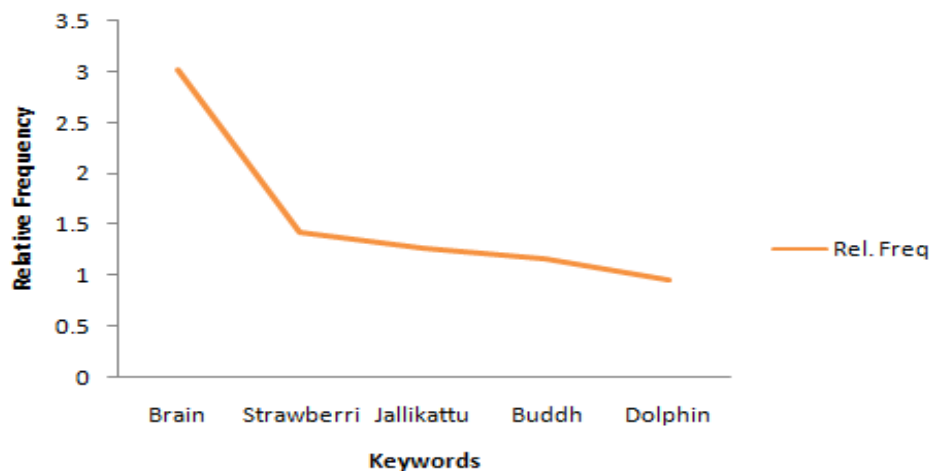


Fig.10. Frequencies of the detected keywords

Table 4 shows the comparisons of the different images using SIFT and proposed Seg_SIFT. In the detected Brain image, the time for Gaussian scale space construction, differential scale space construction, time for finding keypoints, total number of keypoints extracted and descriptors calculation for SIFT feature extraction method are 7.631sec, 0.013 sec, 0.050 sec, 976 keypoints and 0.943 sec respectively. Similarly for proposed Seg_SIFT feature extraction methods are 7.761 sec, 0.016 sec, 0.045sec, 787 keypoints and 0.738 sec respectively. It is experimentally shown that the total number of keypoint extraction of proposed Seg_SIFT feature extraction method takes lesser time compared to the SIFT feature extraction method. Similarly the time for Gaussian scale space construction, differential scale construction, for finding keypoints and for calculating descriptor take lesser time in Seg_SIFT feature extraction than SIFT feature extraction method.

Table 4. Comparison of different images using SIFT and proposed Seg_SIFT





Images								
	Buddha		Brain		Strawberry		Dolphin	
Methods	SIFT	Proposed SEG_SIFT	SIFT	Proposed SEG_SIFT	SIFT	Proposed SEG_SIFT	SIFT	Proposed SEG_SIFT
Feature Extraction methods	SIFT	Proposed SEG_SIFT	SIFT	Proposed SEG_SIFT	SIFT	Proposed SEG_SIFT	SIFT	Proposed SEG_SIFT
Time for Gaussian scale space construction (sec)	7.495	7.694	7.631	7.761	7.687	7.719	7.476	7.632
Time for Diff. scale space construction (sec)	0.012	0.012	0.013	0.016	0.012	0.017	0.012	0.013
Time for finding key points(sec)	0.057	0.045	0.050	0.045	0.062	0.046	0.056	0.045
Total number of key points extracted	1513	775	976	787	2254	814	1421	743
Time for calculating descriptor (sec)	1.450	0.699	0.943	0.738	2.374	0.766	1.355	0.657

Figure 11 shows the GUI (Graphical User Interface) model of the proposed system. The images are retrieved using the clustering algorithms. From these, a simple yet effective retrieval method based on exploring the keyword to image similarities is proposed. In this work, SIFT and proposed Seg_SIFT feature descriptor are compared to analyze the results for different images in the database. The image retrieval process is represented with the help of the clustering techniques. The comparison of clustering algorithms is done based on the performance measures. Table 5 gives the comparison of retrieval time of the different clustering algorithms based on SIFT and proposed Seg_SIFT extraction method. In SIFT feature extraction method, the elapsed time of k-subspace clustering and seq-k-subspace clustering are 0.5925 sec and 0.1396 sec respectively. Similarly for the proposed Seg_SIFT feature extraction method, the elapsed time of k-subspace clustering and seq-k-subspace clustering are 0.5975 sec and 0.0499 sec respectively. It can be observed that the elapsed time of k-subspace clustering algorithm is lesser than the seq-k-subspace clustering algorithm for both different feature

descriptors. The elapsed time is the amount of time that passes from the beginning to its end of image retrieval. So the retrieval of the images using seq-k- subspace clustering algorithm is faster than k-subspace clustering algorithm.

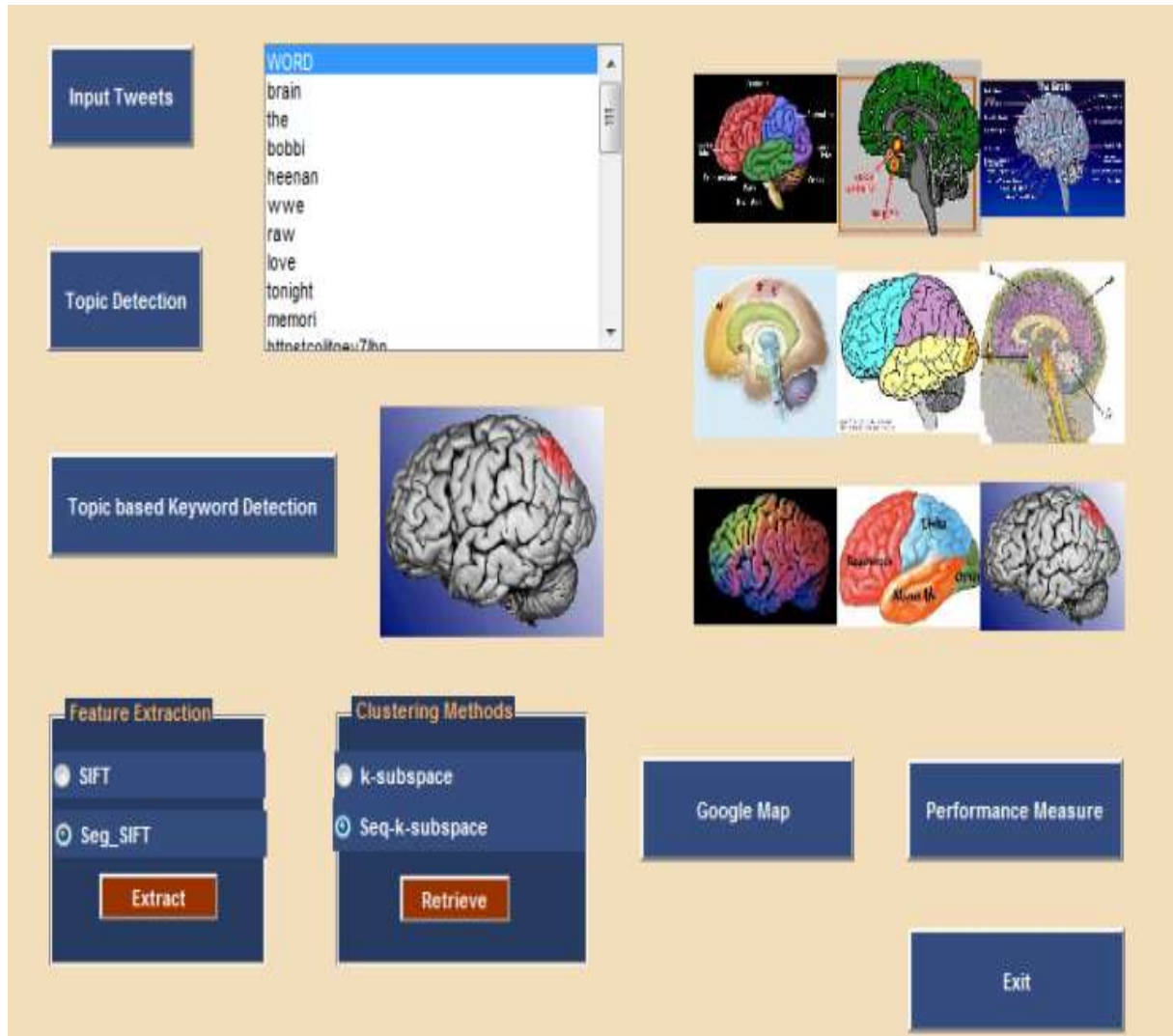


Fig. 11. Twitter image retrieval based on clustering algorithms

Table 5. Elapsed Time (sec) for clustering algorithms

Images	Clustering Algorithms	Feature Extraction Methods			
		SIFT		Proposed Seg_SIFT	
		k-subspace	Seq-k-subspace	k-subspace	Seq-k-subspace
Brain		0.5925	0.1396	0.5975	0.0499
Strawberry		1.3930	0.1517	0.5992	0.0759
Dolphin		0.6289	0.1521	0.6153	0.0616

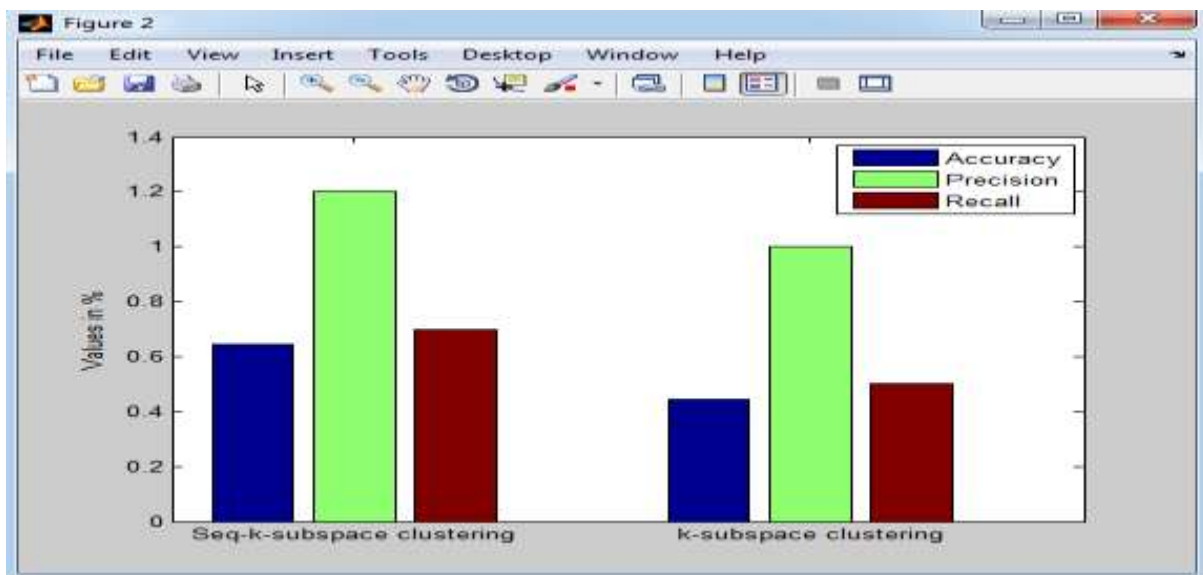
**Fig. 12. Performance measure of the clustering algorithms**

Fig. 12 shows the average performance measure of the compared model of clustering algorithms. The average of accuracy, precision and recall values of the k-subspace clustering algorithms obtained are 80.15%, 84.76% and 79.43% respectively. Similarly for seq-k-subspace clustering algorithm, average accuracy, average precision and recall are 84.00%, 87.35% and 86.32% respectively. Therefore, in this paper Naive Bayes classifier is used to classify the tweets from the twitter data the this performance analysis shows that the seq-k-subspace is better compared to the k-subspace clustering algorithms.

5. Conclusion

The images are collected and extracted from the image database based on the keyword detection. Then, clustering algorithms such as k-subspace and seq-k-subspace algorithms are implemented for retrieving the similar images. These algorithms are compared and analyzed based on the elapsed time of image retrieval. The detailed experimental results show that elapsed time of k-subspace is lesser than the seq-k-subspace clustering algorithm while retrieving the images from the database. So seq-k-subspace clustering algorithm is chosen better than the k-subspace clustering algorithm. In this paper, proposed feature descriptors that can improve the performance of the image retrieval system are suggested. The keyword based image retrieval based on proposed Seg_SIFT and subspace clustering algorithms are implemented and they are very fast and efficient image retrieval. The proposed work is experimented using keywords and images from tweets in Twitter. This work enables us to discover and understand the tweets easily. The proposed Seg_SIFT descriptor is obtained for

extracting the key points from the image database. In future this research aims to develop a framework and integrate all the popular social networking sites like Facebook, Google+, hi5 and YouTube and also plan to extend the keyword-based image retrieval method by investigating with or without different clustering algorithms of image retrieval.

References

1. Aggarwal CC, Reddy CK (2013) Data clustering: algorithms and applications. Data Mining Knowledge and Discovery Series 1st. CRC Press.
2. Ali Vashae , Reza Jafari , Djemel Ziou , Mohammad Mehdi Rashidi," Rotation invariant HOG for object localization in web images",Signal processing,2016.
3. Amrutha Benny, Mintu Philip: Keyword Based Tweet Extraction and Detection of Related Topics. In: Procedia Computer Science 46 (2015) 364 – 371
4. Berkant Basa," Implementation of hog edge detection algorithm", Procedia - Social and Behavioral Sciences 174 (2015) 1567 – 1575.
5. Beyer K, Goldstein J (1999) When is nearest neighbor meaningful? Proc 7th Int Conf Database Theory. In: Database Theory –ICDT'99. Lecture Notes in Computer Science. Springer, Berlin Heidelberg Vol. 1540. pp 217-235
6. D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. Int. Journal of Computer Vision, 2004
7. Dingding Wang, Chris Ding, and Tao Li: K-Subspace Clustering. In: Springer, pp. 506–521,(2009)
8. Enikuomelin A. O, Rahman M. A, Zubair A. F," Fuzzy K-means Application to Semantic Clustering for Image Retrieval" Advances in Computing 2016, 6(1): 1-5
9. Günnemann S, Boden B, Seidl T (2012) Finding density-based subspace clusters in graphs with feature vectors. In:Data mining and knowledge discovery. Springer, US Vol. 25. pp 243–269
10. Iyad Abu Doush , Sahar AL-Btoush," Currency recognition using a smartphone: Comparison between color SIFT and gray scale SIFT algorithms", Journal of King Saud University – Computer and Information Sciences (2016).
11. Jyoti Joglekar , Shirish S. Gedam ," Image Matching With Sift Features – A Probabilistic Approach ",Vol. Xxxviii, Part 3b – Saint-Mandé, France, September 1-3, 2010.
12. K. B. Jayaraman, Salomia Brigitha.J: Survey on Content Based Image Retrieval Technique. In: IJARCSSE, Volume 6, Issue 3 (2016)
13. Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary," Twitter Trending Topic Classification", 11th IEEE International Conference on Data Mining Workshops(2011).
14. Li T, Ma S, Ogihara M (2004) Document clustering via adaptive subspace iteration. In: Proceedings of the 27th annualinternational ACM SIGIR conference on research and development in information retrieval. ACM, USA. pp 218–225
15. Liang Zhao, Ting Hua, Chang-Tien Lu, Ing-Ray Chen: A topic-focused trust model for Twitter. In: Computer Communications , Elsevier (2015) 1–11
16. Nagaraja S. and Prabhakar C.J.," Low-Level Features For Image Retrieval Based On Extraction Of Directional Binary Patterns And Its Oriented Gradients Histogram", Computer Applications: An International Journal (CAIJ), Vol.2, No.1, February 2015
17. K. Nalini, L. Jaba Sheela," Survey on Text Classification", International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Volume 1 Issue 6 (July 2014)
18. Parsons L, Haque E, Liu H (2004) Subspace clustering for high dimensional data: a review. ACM SIGKDD Explor Newsl 6(1):90–105
19. Parvathy Ram, S.Padmavathi : Analysis of Harris Corner Detection For Color Images. In: International conference on Signal Processing, Communication, Power and Embedded System (SCOPEs)(2016)
20. Raval Vidhi, Shah Khushbu," An Integrated Technique Dct & Sift For Detection Of Image Forgery", International Journal For Technological Research In Engineering , Volume 1, Issue 10, June-2014.

21. Reza Javanmard Alitappeh, Kossar Jeddi Saravi, Fariborz Mahmoudi, “A New Illumination Invariant Feature Based on SIFT Descriptor in Color Space”, *Procedia Engineering* 41 (2012) 305 – 311
 22. Ryfial Azhar, Desmin Tuwohingide, Dasrit Kamudi, Sarimuddin, Nanik Suciati,” Batik Image Classification Using SIFT Feature Extraction, Bag of Features and Support Vector Machine”, *Procedia Computer Science* 72 (2015) 24 – 30
 23. Takamu Kaneko, Keiji Yanai: Event photo mining from Twitter using keyword bursts and image clustering. In: *Neurocomputing*, Elsevier, (2015)
 24. Vadivukarassi.M, Aruna.P and Puviarasan.N: Real Time Prediction of Twitter users location on Google map using Python. In: *Australian Journal of Basic and Applied Sciences*, 10(12) July 2016, Pages: 91-97
 25. Vadivukarassi.M, Puviarasan.N and Aruna.P: Sentimental Analysis of Tweets Using Naive Bayes Algorithm. In: *World Applied Sciences Journal* 35 (1): 54-59, 2017
 26. Vidal R, Tron R, Hartley R (2008) Multiframe motion segmentation with missing data using PowerFactorization and GPCA. *Int J Comput Vis* 79(1):85–105
 27. Xiangmin Zhou · Lei Chen: Event detection over twitter social media streams. In: *The VLDB Journal*, Springer, (2013)
 28. Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: an efficient data clustering method for very large databases. In: *Proc. of the ACM SIGMOD international conference on management of data*, vol. 1. ACM Press, USA. pp 103–114.
- Zhegao Piao · Hyung - Geun Ahn · Seong Joon Yoo Yeong Hyeon Gu · Helin Yin · DaWoon Jeong · Zhiyan Jiang · Won Hee Chung,” Performance analysis of combined descriptors for similar crop disease image retrieval”, *Cluster Comput* DOI 10.1007/s10586-017-1145-4.
