A Mini-project Report

on

Latent Topic Encoding for Content-Based Retrieval

carried out as part of the course Information Retrieval (IT362)

Submitted by

Aditi Singh 14IT104

Parvathi M.H 14IT130

Aparna P.L 14IT132

VI Sem B.Tech (IT)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

In

INFORMATION TECHNOLOGY

At

Department of Information Technology

National Institute of Technology Karnataka, Surathkal.

Jan - May 2017

CERTIFICATE

This is to certify that the project entitled **"Latent Topic Encoding for Content-Based Retrieval"** is a bonafide work carried out as part of the course Information Retrieval (IT362), under my guidance by

1. Aditi Singh (14IT104)
2. Parvathi M.H (14IT130)
3. Aparna P.L (14IT132)

students of VI Sem B.Tech (IT) at the Department of Information Technology, National Institute of Technology Karnataka, Surathkal, during the academic semester of Jan - May 2017 in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Information Technology, at NITK Surathkal.

Place: ______________________________

(Signature of the Guide)

Date:

DECLARATION

We hereby declare that the project entitled **"Latent Topic Encoding for Content-Based Retrieval"** submitted as part of the partial course requirements for the course Information Retrieval(IT362) for the award of the degree of Bachelor of Technology in Information Technology at NITK Surathkal during the Jan - May 2017 semester has been carried out by us. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Further, we declare that we will not share, re-submit or publish the code, idea, framework and/or any publication that may arise out of this work for academic or profit purposes without obtaining the prior written consent of the Course Instructor.

Name and signature of the Students:

1. Aditi Singh (14IT104)
2. Parvathi M.H (14IT130)
3. Aparna P.L (14IT132)

Place:

Date:

Abstract

This report offers a novel approach for encoding centred on latent topics which is particularly intended for Content-Based Retrievals (CBR). The vast compilations of videos and images have increased the semantic gap for CBR. Hence, innovative and effective approaches to reduce the semantic gap are needed. The uniqueness of the suggested Latent Topic Encoding (LTE) technique can be seen in: (1) describing the vocabulary of visual words matching to the unseen patterns which are revealed from the local features; (2) each content encoding is obtained by collecting the amount of its local descriptors over the topics. Some retrieval activities using the obtained datasets have been done to test the effectiveness of the Latent Topic Encoding and compared with the regular visual Bag of Words (BoW). End result shows that LTE technique can perform better than the visual BoW when the task of retrieval is done in the space of latent topic.

Contents

List of Figures

1 Introduction

Recently, the increase in video and image compilations has made video and image media as the largest form of information available on the Internet. In this situation, one of the most significant tasks is how to retrieve data which is useful to the user from this large collection of information in an efficient way. Traditional search engines retrieve the video or images using only textual information. Thus, the capability of the retrieval process as media content is not understood. Content-Based Retrieval (CBR) has become a very crucial area of research and CBR systems have been developed.

CBR is related to delivering videos or images which satisfy the queries of the users, that is, semantic concepts that is present in the mind of the users and what they are searching for. In the recent years, CBR has been extensively focused by the research scholars and many methods have been developed [1,3]. In spite of all this study, the semantic gap [2] that exists between assess-able features of lower level and theories of higher level makes the CBR a challenging field mainly for big datasets. A CBR system, in overall, has three important parts involved in the process of retrieval: (1) a query, whatever is in the mind of the user; (2) a data compilation, which is used to obtain related content similar to the query of the user; (3) a function for ranking, which keeps the samples of the datasets depending on the relevance of the query of the user. These parts are often combined along with the relevance feedback given by the user [4] to provide the most useful content after many rounds of searching.

1.1 Motivation

Present methods use classification techniques to try to fill the semantic gap. Hence, samples are represented with those features which have no semantic meaning to represent semantic concepts. So, the present techniques in retrieval are only dependable under a certain criteria and these also need huge calculation time. The standard encoding technique used is the visual Bag of Words (BoW) [6].

In a CBR problem, we deal with complex classes with very slight information about the target. A noticeable form of the information, as generated by a vast majority of encoding methods, is not sufficient to distinguish among unconstrained classes. Thus, in this area, techniques based on topic models can be used for characterization of the contents in a semantic meaning of higher level. The offered work is fixated on delivering co-operating content based on latent problems based on this retrieval technique to deal with the challenge of the semantic gap by given by the topic based systems. Thus, using of latent topic representation can be used to show the knowledge in a space of characterization that is semantically closer to the concepts of the user. The probabilistic ranking approach can be used for these representations of topic model which are supposed to be effectually calculated for determining tasks of class in huge retrieval systems.

2 Literature Survey

2.1 Background

Sivic, J., Zisserman [6] et al suggest that in computer vision, the usual method of encoding is the visual Bag of Words (BoW). This encoding method begins with attaining the knowledge of visual language formed from the gathering of the local characteristics of the training dataset. Thereafter, each content of visual words in a histogram are represented by collecting the number of local descriptions in their neighbour clusters. The disadvantage of this technique is the fixed allocation of words which picks the best visual word disregarding the importance and relationships with rest of the clustering centres and effective semantic loss.

Of late developments, substitute the quantization functions which are difficult with substitute encodings that are capable of maintaining more knowledge about the primary descriptions. Two of the main developments in this context are (1) expression of descriptors as pattern of the visual words and (2) noticing the changes between visual words and the corresponding descriptors.

Soft quantization, Co-occurrence models, and local linear encoding are examples of such alternative encoding techniques, where features are expressed as a combination of visual words. On the other hand, Super-vector encoding and Fisher encoding are techniques which record the differences between visual words and the features.

2.2 Outcome of Literature Review

Although, some of the above stated techniques have appeared to provide decent results in the challenge of grouping of data points, the content based retrieval problem is of totally distinct nature. In a general problem of classification, one is provided with sufficient or at least enough information about the data set, class labels and how we would want to classify the dataset into those classes. However, in a problem of retrieval which class to be got is not known in prior, which poses as a challenge on its own. Thus, it becomes dependent up on the user and the way he queries, as to how the model goes around with classification. This often results in dealing with complex classes with very little knowledge about the objective.

Topic models are known to be able to obtain hidden formats from the dataset and the whole information distribution and successfully represent this data according to the extracted patterns. Typically, based on their usage, they have been noted to reduce the dimensionality of the original depiction of the space as extracted by the BOW.

2.3 Problem Statement

The aim of this project is to perform Content-Based retrieval (CBR) of similar items from a database to satisfy the user input query by using a novel feature encoding technique called Latent Topic Encoding (LTE).This technique for Content based retrieval performed in latent space is compared with the BOW technique.

2.4 Objectives

- To find a method that offers a more appropriate encoding method for Content Based Retrieval rather than the usual Bag of Words approach.
- To implement a novel encoding method called Latent Topic Encoding totally based on the concept of latent topics, which describes the vocabulary of visual words by methods of unseen patterns and performs the encoding by considering the effect of every topic to every local descriptor.

3 Methodology

3.1 System Architecture

The work was carried on in a machine with 4 GB RAM and Intel Core i5 processor on a Python base. The overall flow of the proposed approach is shown in Fig.1

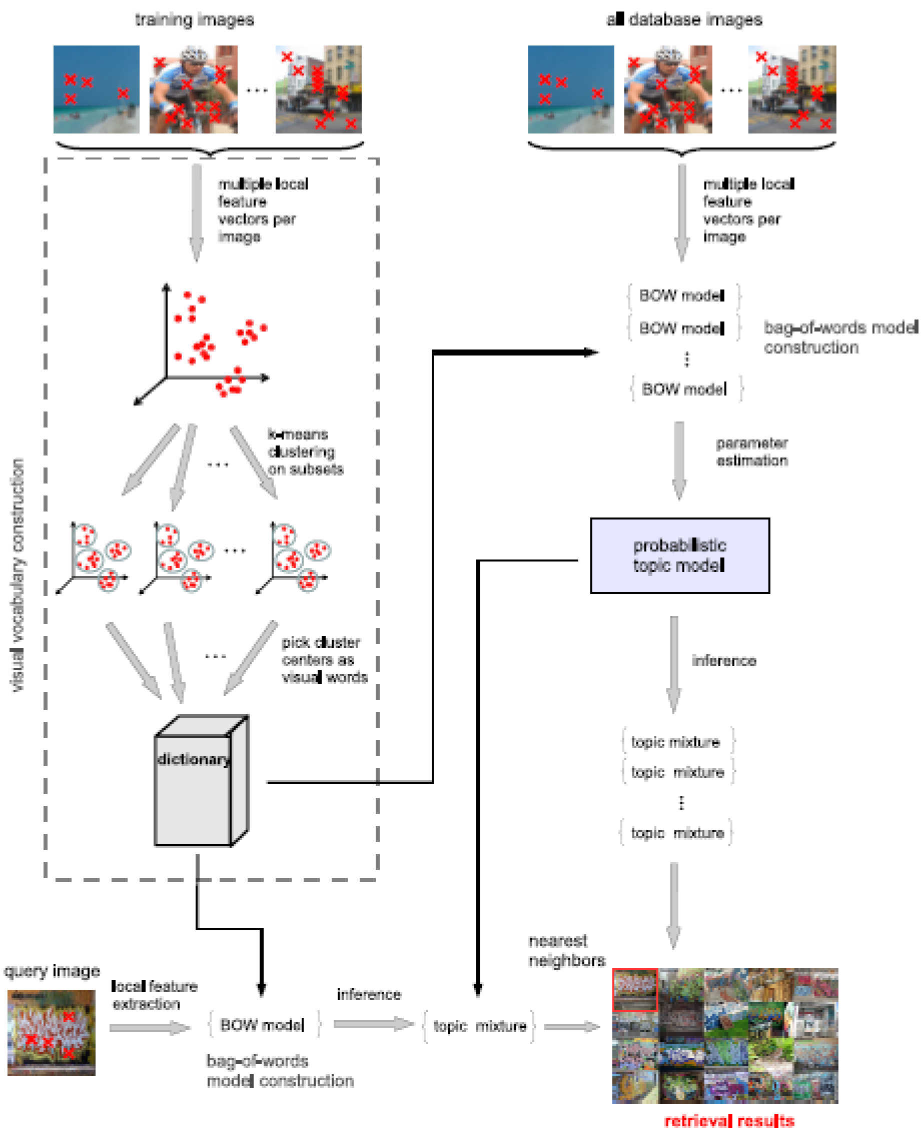


Figure 1: System Architecture

3.2 Algorithms

- SIFT (Scale Invariant Feature Transform): SIFT is an algorithm to detect and describe local features in images. It detects "interesting points" called key points.

  - Building a scale space: This is the initial preparation. Take the original image, and generate progressively blurred out images. This is called an octave. Then, resize the original image to half size and generate blurred out images again. Number of blurred images in each octave is called a scale. The creator of SIFT suggests that 4 octaves and 5 blur levels are ideal for the algorithm. Blurring is obtained through convolution operation of the Gaussian operator and the image.
  - Log Estimation: The aim is to generate another set of images for locating edges and corners in an image through LoG(Laplacian of Gaussian) operator. However, LoG is computationally expensive and it is thus approximated with the DoG(Difference of Gaussian) operator where the set of consecutive images obtained in each octave the previous step are differenced.
  - Obtaining key points: Finding key points is a two part process 1. Locate maxima/minima in DoG images 2. Find subpixel maxima/minima Iterate through each pixel and check all it's neighbours to determine if it's a maxima or a minima. The check is done within the current image, and also the one above and below it in the scale space. These keypoints are the aproximate maxima/minima. The subpixel maxima/minima is obtained using Taylor expansion of the image around the approximate key point.
  - Removing of bad key points: If the magnitude of the intensity (i.e., without sign) at the current pixel in the DoG image (that is being checked for minima/maxima) is less than a certain value, it is rejected.
  - Allotting an orientation to the keypoints:An orienation is assigned to each keypoint. The orientation provides rotation invariance. The magnitude of direction and the orientation are calculated by finding the difference in x and y directions around the pixels for all pixels near the keypoint(by choosing a window around a keypoint). Then a histogram is created for this and any peaks above 80
  - Generating SIFT features:A unique fingerprint isto be generate for each keypoint. A 16x16 window around the keypoint. This 16x16 window is broken into sixteen 4x4

windows. Within each 4x4 window, gradient magnitudes and orientations are calculated. These orientations are put into an 8 bin histogram with range 0-44 degrees add to the first bin, 45-89 add to the next bin, And so on. The amount added to the bin depends on the magnitude of the gradient. The amount added also depends on the distance from the keypoint. Doing this for all 16 pixels, 16 random orientations are compiled into 8 predetermined bins. This is done for all sixteen 4x4 regions. So 4X4X8 = 128 numbers for each keypoint.

- **Visual Bag of Words**

  To represent an image using the BoW model, three steps are followed: feature detection, feature description, and then codebook generation. Feature detection helps finds patches in the image. Feature representation methods deal with representing the patches as numerical vectors. Scale-invariant feature transform (SIFT) technique is used to convert each patch to 128-dimensional vector called feature descriptors. Mini Batch K-means is used to cluster a set of feature descriptors. A codeword represents several similar patches. They are analogous to words in text document. Codewords are then defined as the centers of the learned clusters. Several codewords form CodeBook which is analogous to word dictionary. Finally, the mapping of each patch present in an image to a particular codeword is done by finding the closest codeword with Eucledian distance and then the image can be represented by the histogram of the codewords.

- **Latent Dirichlet Allocation**

  - This technique of encoding tries to deal with the problems of existing encoding techniques by using "latent topics". A topic model is a type of statistical model for discovering the latent topics that occur in a collection of documents.

    Beginning from a particular data matrix $P(W\|C)$ which states a corpus of documents (C) in a specific space $W \subset N^n$ topic algorithms are capable of getting two matrices:

    1. the depiction of topics in words $P(W\|Z)$
    2. the depiction of documents in topics $P(Z\|C)$.

    The topics which are obtained(Z) is a parameter which has to be decided beforehand for this algorithm.

– Generative process: Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA assumes the following generative process for a corpus D consisting of M documents each of length N_i

1. Choose $\theta_i = Dir(\alpha)$, where i ε {1,2,..,M} and Dir(α) is the Dirichlet distribution for parameter α.
2. Choose $\phi_k = Dir(\beta)$, where k ε {1,2,...,K}
3. For each of the word positions i,j where j ε {1,2,...,N_i} and i ε {1,...M}: (a) Choose a topic $z_{i,j}$ = Multinomial(θ_i) (b) Choose a word $w_{i,j}$ = Multinomial($\phi_{z_{i,j}}$)

- **LDA Statistical Inference:** LDA statistical inference is the reverse process of generative process i.e given a collection of documents, we find the document-topic distributions for all documents and topic-word distribution for all topics. This is achieved through a sampling process called Gibbs sampling which is shown in Figure 2.

```
Input: words w ∈ documents d
Output: topic assignments z and counts n_{d,k}, n_{k,w}, and n_k
begin
    randomly initialize z and increment counters
    foreach iteration do
        for i = 0 → N − 1 do
            word ← w[i]
            topic ← z[i]
            n_{d,topic}-=1; n_{word,topic}-=1; n_{topic}-=1
            for k = 0 → K − 1 do
                p(z = k|·) = (n_{d,k} + α_k) (n_{k,w} + β_w)/(n_k + β × W)
            end
            topic ← sample from p(z|·)
            z[i] ← topic
            n_{d,topic}+=1; n_{word,topic}+=1; n_{topic}+=1
        end
    end
    return z, n_{d,k}, n_{k,w}, n_k
end
```

Algorithm 1: LDA Gibbs Sampling

Figure 2: LDA Gibbs Sampling

- **LDA topic modeling for image retrieval**
  – Extract SIFT features for all images in the dataset.

- Apply K-means clustering to obtain k(user-defined) clusters. The centers of the dictionary are considered to be the words of the vocabulary.
- Go through each SIFT feature in obtained in step 1 and assign it to the closest cluster center obtained in step 2. Closest cluster center to the SIFT feature can be found using Eucledian distance measurement. FLANN based matcher was used for finding the closest cluster.
- Count the number of SIFT features belonging to each word in the vocabulary and build a histogram for each image.
- Using the document-word distribution obtained in step 4, apply LDA topic model to obtain document-topic and topic-word distribution.
- When a user inputs a query, find topic distribution for this query and compare it with the topic distributions of all documents in step 5 using Eucledian distance(ED) and Cosine Similarity (CS) and display the top 10 documents to the user.

4 Implementation and Work Done

4.1 Experimental Framework

The simulations were performed on a Linux system with 8 GB RAM, with CPU speed 2.60Hz and x64-based processor. The language of implementation used was Python. Following Python libraries were used:

- Scikit-image: Image processing methods
- Scikit-learn: K-Means, LDA
- Numpy: Matrix manipulation
- OpenCV: Image processing methods

Dataset used:

1. Abnormal Objects Dataset consisting of 623 images belonging to 6 classes.
2. Caltech 101: 9145 images belonging to 102 folders
3. IRMA medical image dataset: 12000 images belonging to 193 categories[11]

4.2 Step by step implementation

- Collect descriptors(128 dimensional) of each image using SIFT feture extraction.Concatenate all of the descriptors obtained for each image
- Quantization: Use k-means to cluster the descriptors obtained and define the centers of the clusters obtained as the vocabulary(dictionary)
- Encoding: Go through every image. Assign every keypoint in the image to a word in the vocabulary by finding the closest word in the dictionary
- Pooling: Build a histogram for every image consisting of frequency of each word of the dictionary in that image using the assigned words in the previous step and build a document-word matrix consisting of frequency of each word in every document.

- Retrieval: Apply any topic model (say LDA) which will take as input the document-word distribution and extracts z(user-defined) topics and also gives Document-Topic distribution for each document and Topic-Word distribution for each topic.

- A user inputs query image and the keypoints are extracted, the keypoints are assigned to vocabulary, and a histogram of vocabulary is built for this image. It is then represented in topic space using LDA. The topic distribution is then compared with topic distribution of each document (using Eucledian distance). The top 10(or 20) documents with least distance(Eucledian and Cosine similarity measures) are displayed to the user.

- BoW: The Document word distribution of the user input query is found and compared with the document word distributions of the images in the dataset. The top 10(or 20) documents with least distance(Eucledian and Cosine similarity measures) are displayed to the user.

4.3 Challenges faced

- K-means clustering is an unsupervised learning algorithm. It partitions the dataset into clusters through minimizing the space between each of the data points and the centers of the cluster. When the numbers of descriptors increase, k-means is known to take a longer time. Our dataset consists of a large numbers of descriptors, owing to which k-means took approximately 20 minutes for a descriptor of size 105. Thus, the time and space complexity was fairly high with k-means clustering, which is why we used mini-batch k-means instead. Mini-batch k-means works similar to k-means, the only difference being that the data points are taken in batches for clustering. Although, the accuracy might get affected due to the usage of k-means, the time spent for training reduces greatly.

- The time consumed to assign each descriptor to a word in a vocabulary is high. Encoding for a descriptor of size 10^5, took about 5 hours, which is a huge load computationally. To overcome this, FLANN matching was used as an alternative to brute force matching to search the closest neighbor. FLANN is short for "fast library for approximate nearest neighbor", and is found to work great for searching the nearest neighbor and finding a system for automatically choosing the best suited algorithm and most optimal parameters depending on the dataset.

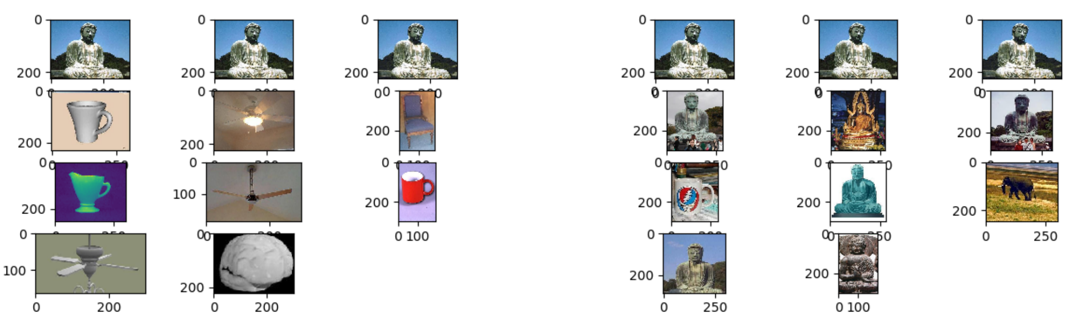
- Computing histograms for each retrieval was a waste of time and resources. Thus, histograms of all documents were stored to files to avoid computing them for each query. Similarly, Document-Topic and Topic-word distribution were stored to a file to avoid computing them for every query.

4.4 Results and discussion

The image rerieval was carried out on 3 databases- Abnormal Objects Dataset: Abnormal objects consisting of 623 images belonging to 6 categories. Subset of IRMA medical image dataset: Radiographs consisting of 180 images belonging to 5 categories. Subset of Normal objects Dataset: Caltech dataset consisting of 550 images of objects like aeroplanes, elephants etc belonging to 9 categories. The results obtained for each of the datasets are shown in Fig.[3-8] where EC stands for Eucledian Distance , CS for Cosine Similarity, BOW stands for Bag of Words model and LDA stands Latent Dirichlet Allocation model(Latent topic Model).

4.4.1 Results for Caltech dataset

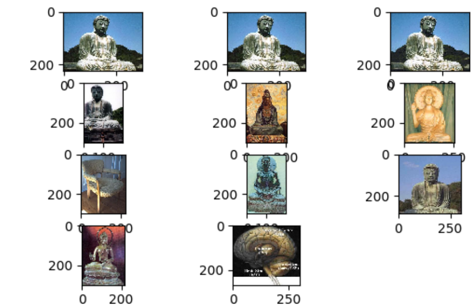
Figure 3 and Figure 4 show the results for BOW and LDA approaches with EC and CS measures.



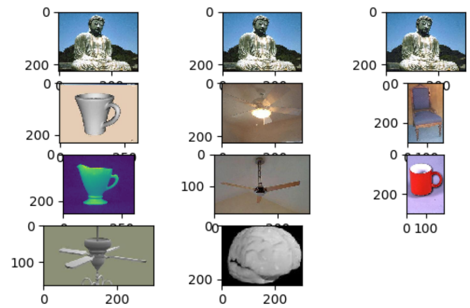
(a) With EC, precision=0.2

(b) With CS, precision=0.8

Figure 3: Results of EC and CS with BOW for Buddha as input image



(a) With EC, precision=0.2

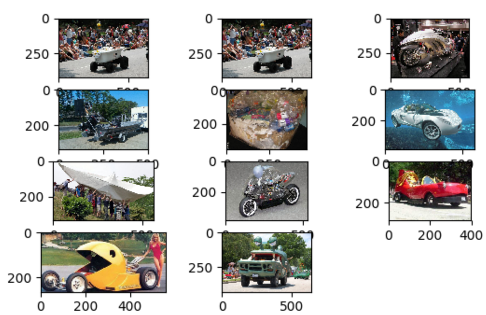


(b) With CS, precision=0.8

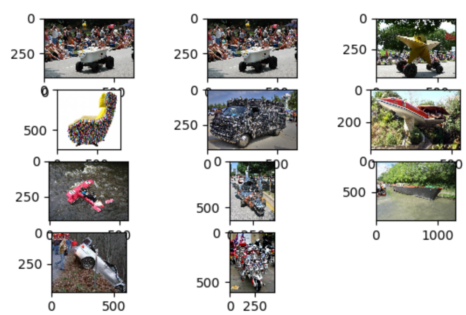
Figure 4: Results of EC and CS with LDA for Buddha as input image

4.4.2 Results for Abnormal object dataset

Figure 5 and Figure 6 show the results for BOW and LDA approaches with EC and CS measures.

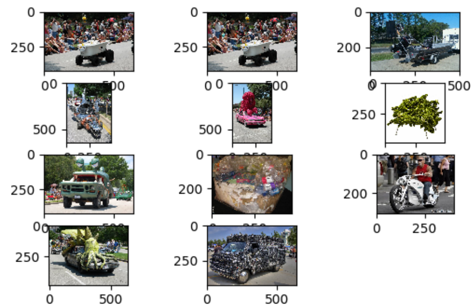


(a) With EC, precision=0.4

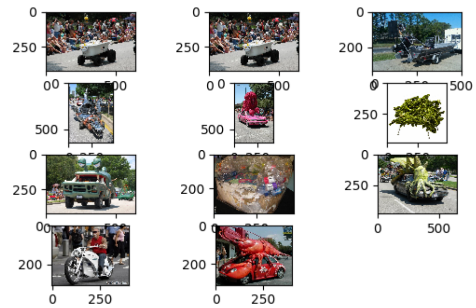


(b) With CS, precision=0.5

Figure 5: Results of EC and CS with BOW for car as input image



(a) With EC, precision=0.6

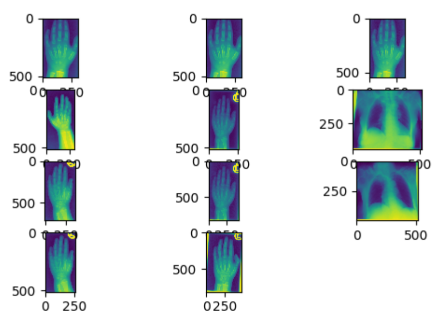


(b) With CS, precision=0.6

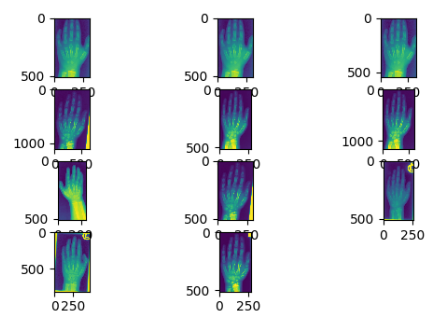
Figure 6: Results of EC and CS with LDA for car as input image

4.4.3 Results for IRMA dataset

Figure 7 and Figure 8 show the results for BOW and LDA approaches with EC and CS measures.

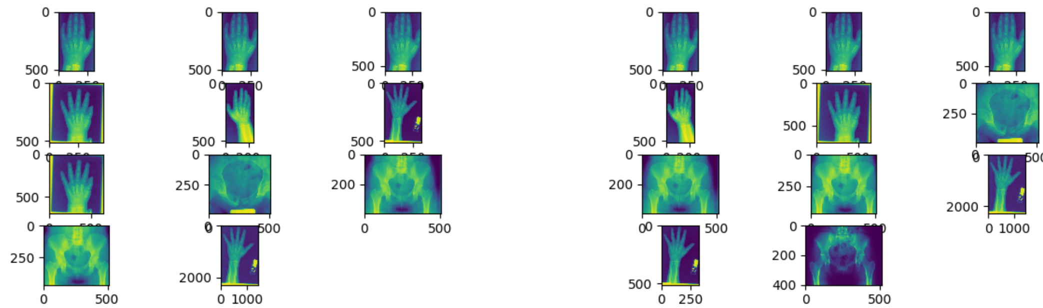


(a) With EC, precision=0.8



(b) With CS, precision=1.0

Figure 7: Results of EC and CS with BOW for hand as input image



(a) With EC, precision=0.7

(b) With CS, precision=0.6

Figure 8: Results of EC and CS with LDA for hand as input image

4.4.4 Analysis of results

CS is found to outperform EC in case of abnormal objects and Caltech dataset and not in medical images because in objects contain have more hidden/latent nature. Since CS is a better metric to measure semantic similarity between items, it is found to perform better. Medical images retrieval is based on visual characteristics and don't possess hidden characteristics. So, EC and CS gave almost the same amount of precision in case of medical images.

Summary of results obtained is shown in Figure 9 The average precision @ 10 and precision @ 20 is computed for each dataset with both BoW and LDA and with both EC and CS similarity measures. CS clearly performs better than EC in case of BoW. In case of LDA, both perform equally well, with EC preforming slightly better. In case of Abnormal objects, both precision at 10 and precision at 20 are better for LDA. For medical images, BoW with CS performs much better than LDA. For normal objects dataset, LDA outperforms BoW.

4.5 Innovative Work

- Compared BOW and LDA model for 2 datasets other than the one mentioned in the chosen paper.
- Used Flann matcher for encoding instead of brute force.

		LDA			BoW	
		EC	CS		EC	CS
	Abnormal	0.5	0.43		0.31	0.38
Precision @ 10	Medical	0.58	0.54		0.4	0.84
	Normal	0.63	0.63		0.16	0.53
	Abnormal	0.34	0.35		0.325	0.29
Precision @ 20	Medical	0.6	0.57		0.36	0.73
	Normal	0.5	0.5		0.12	0.51
	Abnormal	0.64	0.69		0.44	0.67
MAP	Medical	0.57	0.64		0.47	0.87
	Normal	0.65	0.7		0.47	0.67

Figure 9: Summary of Results

4.6 Individual Contribution

Individual contribution of members

- Dataset collection and pre-processing: Parvathi M.H., Aparna PL
- Bag of words: Aditi Singh, Parvathi M.H
- Latent Dirichlet allocation: Aparna PL
- Results and analysis: Aditi Singh, Parvathi M.H, Aparna PL

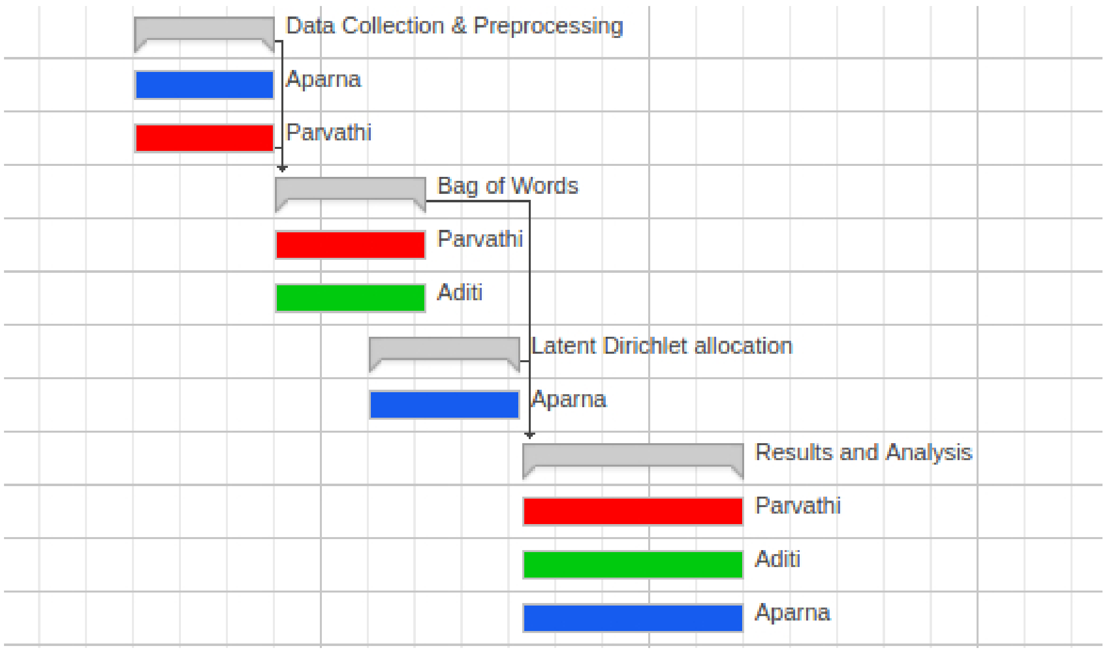


Figure 10: Gantt chart of division of work

5 Conclusion and Future work

The topic models use hidden patterns of the locally available descriptors form the dataset and each content is then shown as the collection of the local descriptors of the content shown in these topics. LDA(Latent Topic Model) was chosen as the topic model and its results were compared with those of BoW(Bag of Words) retrieval for 3 different datasets. CS (Cosine Similarity) was found to outperform EC(Eucledian Distance) for BoW. For LDA, EC was found to perform slightly better than EC.

The future works could be focused on defining a tactic to select the vocabulary size, automatically. For the experiments, SIFT descriptors and LDA model have been used but any other topic model or feature extraction technique could also be utilized.

References

[1] Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: state of the art and challenges. ACM TMCCA 2(1), 1–19 (2006)

[2] Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE TPAMI 22(12), 1349–1380 (2000)

[3] Ren, W., Singh, S., Singh, M., Zhu, Y.S.: State-of-the-art on spatio-temporal information-based video retrieval. Pattern Recogn. 42(2), 267–282 (2009)

[4] Zhou, X., Huang, T.: Relevance feedback in image retrieval: a comprehensive review. Multimedia Syst. 8(6), 536–544 (2003)

[5] Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60(2), 91–110 (2004)

[6] Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: ICCV, vol. 2, pp. 1470–1477 (2003)

[7] Li, F., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: CVPR, pp. 524–531 (2005)

[8] Philbin, J., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: improving particular object retrieval in large scale image databases. In: CVPR (2008)

[9] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong. Y.: Locality-constrained linear coding for image classification. In: CVPR (2010)

[10] Perronnin, F., Sanchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)

[11] T.M Deserno,"12000 images belonging to 193 categories", https://www.irma-project.org, 2008