Feature Selection using Bag-Of-Visual-Words Representation

Faheema AG and Subrata Rakshit
Centre for AI and Robotics(CAIR),
DRDO Complex,
C V Raman Nagar,
Bangalore - 560093, India
faheema,subrata@cair.drdo.in

Abstract

In this paper, we introduce an efficient method to substantially increase the recognition performance of object recognition by employing feature selection method using bag-of-visual-word representation. The proposed method generates visual vocabulary from a large set of images using visual vocabulary tree. Images are represented by a vector of weighted word frequencies. We have introduced on-line feature selection method, which for a given query image selects the relevant features from a large weighted word vector. The learned database image vectors are also reduced using the selected features. This will improve the classification accuracy and also reduce the overall computational complexity by dimensionality reduction of the classification problem. In addition, it will help us in discarding the irrelevant features, which if selected will deteriorate the classification results. We have demonstrated the efficiency our method on the Caltech dataset.

Index terms: **PCA-SIFT, Feature extraction,Visual Words, Vocabulary tree, Feature Selection**

1 INTRODUCTION

Object recognition is an active area of research in computer vision. Several object recognition algorithms based on the bag-of-visual-words have been reported in the literature[5,6,7,9]. Bag-of-visual-word model is getting more attention due to its simplicity and good performance on object, scene and activity recognition problems. Due to huge success of bag-of-words(BoW) approach in information retrieval, it has been extensively applied on images/videos in computer vision. In the BoW model, a document is represented by histogram of words. In order to employ the BoW model to an image, we need to first find the words in the images. Recently, there has been a lot of work on extraction of local descriptors[1,2,3,4] from images, and these have demonstrated very good recognition results. These local descriptors are analogous to the words in the text documents. To represent an image using BoW model, a large number of local descriptors or interest points are extracted. These are quantized into visual words by employing clustering algorithms. Visual words are typically found by clustering features extracted from the trainig dataset. A large number of data clustering algorithms have been reported in literature, such as k-means, hierarchical clustering(single, average, max and min linkages) and k-medoid etc, but mostly k-means algorithm is employed due to its simplicity, linear time and space complexity. The problem with k-means is that, we need to specify the number of clusters a priori, which leads to undesirable partitioning of data. Intuitively, when the number of specified cluster is too small, the resulting visual words are non-discriminative thereby generating many false positive matches. On the other hand, when the number of cluster is too large, local descriptors from the same scene/object region in different images can be assigned to different clusters thus generating false negative matches. Nister and Stewenious[6] proposed a visual vocabulary organized in tree data structure. Visual vocabulary tree seems to overcome the difficulty of specifying the number of clusters a priori. A large set of representative images are used to extract local descriptors vectors[1,2,4]. These descriptor vectors are used for unsupervised training of the tree. The vocabulary tree defines the hierarchical quantization that is built by calling hierarchical k-means clustering algorithm. A tree with L levels and B branch factor is created first. The descriptors are fed to the root. An initial k-means clustering is carried out on the entire training data, defining B cluster centers. The training data in each node of the tree is further partitioned by running k-means clustering into B groups. In the process each node gets the descriptor vectors closest to its cluster center. The same process is recursively carried till the maximum specified level is reached. As per Nister and Stewenious[6], it has been demonstrated that $L = 6$ and $B = 10$ gives a very good

978-1-4244-4791-6/10/$25.00 ©2010 IEEE

recognition results. In this paper, we have used the visual vocabulary organized in a tree data structure to generate our visual codebook. Once the visual vocabulary is determined, the database images are represented by a vector of visual word frequencies. For the query image also such a vector is computed. The query image vector is then matched with the database image vectors and set of similar images containing the target object present in the query image are retrieved. Sivic and Zisserman[5] developed video google, which is an image search method based on the idea of text retrieval in large document collection. They have demonstrated image search algorithm based on inverted file structure and $\mathbf{tf} - \mathbf{idf}$[11] feature weighting. Mostly, large dimension feature vector is employed for matching. This is due to the fact that visual vocabularies have to be large to be effective. Thousands of visual words have to be used to get a good recognition performance. This will lead to a huge sparse bag-of-visual-word vector representation. The dimensionality of feature space is very high and finding nearest neighbor is hard. In this paper, we have introduced a new feature selection algorithm. This algorithm selects the relevant features from the large weighted word frequency query vector. It assigns a feature relevance measure to each feature in the feature vector. Based on the feature relevance measure, we select the top K features for classification purpose. This will substantially reduce the dimensionality of the vector and also throws away spurious features, which if selected will decrease the classification accuracy. The paper is organized into five sections. In Section 2, we will briefly review the feature extraction and learning using vocabulary tree. The proposed method of feature selection from visual word vector is discussed in Section 3. In Section 4, the proposed method for object recognition is described. The experimental results and conclusion are presented in Section 5 and Section 6 respectively.

2 FEATURE EXTRACTION

Feature extraction is the most important module of any object recognition system. The local descriptors are most widely employed in a number of real-world applications such as object recognition, image and video retrieval systems, due to their resistance to partial occlusion and background clutter. The image key points or interest points are the descriptors detected[1,3,4] from the local image patches containing rich local information about an image. The local descriptors can be computed efficiently and are relatively insensitive to changes in viewpoint. The SIFT[1] descriptors have shown to be very effective for object recognition problems. The dimensionality of SIFT descriptors were further reduced from 128 to 32 by PCA-SIFT[4]. In this paper, we have extracted PCA-SIFT[4] descriptors from a large set training images. These descriptors are then fed to the hierarchical k-means algorithm to generate the visual vocabulary. The leaf nodes of the vocabulary tree represents the visual words. The set of visual words are commonly known as codebook or visual vocabulary. We have trained our tree with $L = 6$ and $B = 6$ yielding us 15891 visual words. During training some of the nodes does not get enough data point for clustering. In such cases all the data points are assigned to the first child and remaining children of the nodes will not get any data points. The nodes with no data points are not considered for further clustering. Due to this, the number of visual words are less than B^L. The database images descriptors are then matched with visual words in the codebook. These descriptors are than assigned to the nearest visual word. The image is represented by a histogram of visual word frequencies. This will provide us with a more compact vector representation as compared to individual PCA-SIFT[4] vectors.

3 FEATURE REPRESENTATION

Once the images are represented by bag-of-visual-words vector, we can employ text classification algorithms. There are many ways in which the visual word features can be represented. The most commonly employed methods in text retrieval systems are term weighting and stop word removal. But in the case of images, size of the vocabulary plays an important role. The size of the vocabulary is unique to images. Empirical studies have shown that the vocabulary has to be big enough to be effective. The size of vocabulary is an important feature which really affect the classification/recognition result.

Term weighting is a key technique used in information retrieval. In this section, we briefly review the standard weighting that is employed. We apply the popular term weighting scheme of information retrieval. The $\mathbf{tf} - \mathbf{idf}$[11] is a standard weighting technique often used in information retrieval. Essentially, $\mathbf{tf} - \mathbf{idf}$[11] works by determining the relative frequency of words in a specific document compared to the inverse proportion of that word over the entire collection of documents. It is a statistical measure used to evaluate how important a word is to a document in a given collection or corpus. It is computed as follows. Suppose we have a vocabulary of W words, then each document is represented by a vector

$$V_d = (term_1, term_2, ..., term_i, .., term_W)^T \quad (1)$$

of weighted word frequencies with component

$$term_i = \frac{n_{ij}}{n_d} \log \frac{D}{D_i} \quad (2)$$

where n_{ij} is the number of occurrence of word i in document j, n_d is the total number of words in the document d,

D_i is the number of documents containing the term i, and D is the number of documents in the whole database. This weighting is product of two terms: the word frequency, $\frac{n_{ij}}{n_d}$, and the inverse document frequency, $\log \frac{D}{D_i}$. A high weight value is obtained by high term frequency in the given document and a low document frequency of the term in the whole corpus or collection of document, the weights tend to filter out common terms. The weight value will always be greater than or equal to zero. Intuitively, this calculation determines how relevant a given word is in a particular document. Word frequency assigns higher weights to words occurring more often in a particular document and thus describes it well, while inverse document frequency lower the weights of the word that appear often in the collection or corpus and therefore do not help in discriminating between different documents.

3.1 FEATURE SELECTION

The objective of feature selection is to find a subset of features with classification/recognition performance comparable to the full set of features. Given a set of features M, it deals with the problem of choosing appropriate features m, where $m << M$ for a given query. The feature selection module in a recognition system should meet to the following requirements.

1. It must select a subset of features that provide comparable classification/recognition results as with full set.

2. It must reduce the overall computational complexity by reducing the dimensionality of the classification problem.

3. The feature selection is applied to each query, it has to be efficient. It should have linear time complexity w.r.t to number of features.

The need for feature selection method arises from the fact that dimensionality of the weighted frequency vector is huge. Moreover, these vector are sparse in nature. Consider a visual vocabulary of W visual words. If each image contains R regions and M distinct visual words such that $M < R$, then some regions of the image are represented by the same visual word. Each image is represented by a vector with M non-zero entries and remaining entries will be zero. At run-time, the task is to compute the matching score between the query vector and each feature vector of database image. Our feature selection method evaluate the relevance of the features based on the estimation of statistical parameters like mean and variance. These quantities are well defined for any distribution. In this paper, we are utilizing these statistical parameters for characterizing distribution and estimating the relevance of features. Our method

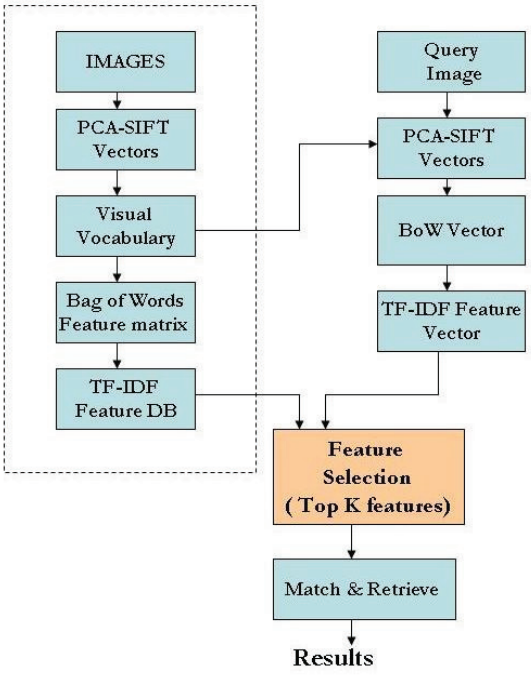

Figure 1. Overview of algorithm

of defining the relevance measure for a feature is as follows:

$$F_i = \frac{|x_i - \mu_i|}{\sigma_i} \tag{3}$$

where x_i is the i^{th} feature of the query feature vector and μ_i is the mean of the i^{th} features of the database images and σ_i is the variance of the i^{th} feature of the database images. The relevance of a feature is measured as how far it is from the mean of the i^{th} feature of the database images(normalized by the variance of the i^{th} feature of database images). If the feature relevance measure is high, then interest in the feature increases; it has potential and good discriminative power. If feature relevance measure is low, then that feature is not useful. The feature selection method is applied to the query feature vector. The features are sorted in the ascending order of their relevance measure and top K feature are selected for image matching.

4 OVERVIEW OF PROPOSED ALGORITHM

The Figure 1. depicts the overall proposed algorithm. The dashed rectangle indicate one time off-line processing. The remaining parts of the flow diagram indicate on-line processing. The off-line processing includes set of representative images, extraction of PCA-SIFT[4] descriptors, generating visual vocabulary using hierarchical k-means,

representing images using bag-of-visual-words model, and finally assigning $\mathbf{tf-idf}$[11] weights to the features. To generate good vocabulary a vast set of representative images are chosen. From these selected images PCA-SIFT[4] descriptors are extracted. The set of descriptors from all the images constitutes the training data for learning. The training data is fed to the vocabulary tree for hierarchical quantization specified by the branch factor of the tree. Same process is recursively carried out until the max specified level is reached. The cluster means of the leaf nodes of the tree gives the visual words. The cluster means of set of all the leaf nodes forms visual vocabulary or visual codebook. Using the visual vocabulary the training images are represented by histogram of visual words. The standard $\mathbf{tf-idf}$[11] weights are assigned to each feature of the feature vector. This constitutes the feature matrix of training database images. This entire process is carried out off-line.

To query the system, user has to submit a query image. For the query image the PCA-SIFT[4] descriptors are extracted. Using visual vocabulary generated during off-line processing, the query image descriptors are assigned to their nearest visual word. The query images is represented by the bag-of-visual word vector. Each feature in the query feature vector is assigned the standard $\mathbf{tf-idf}$[11] weights. The feature selection method described in previous section is applied to the query vector. The feature relevance measure is determined for each feature of the query feature vector. The feature are ranked in ascending order of their relevance measure. The top K features are selected for image matching. The remaining features are discarded as their inclusion will deteriorate the classification results. The training database vectors are also reduced using the output of the feature selection method. This will help in reducing the huge dimensionality of the training database. The query feature vector is matched with database feature vector using dot product. The top matching images are retrieved from the database and presented to the user.

5 EXPERIMENTS AND RESULTS

We study the performance of proposed method by choosing 5 different categories from Caltech101 dataset. They are Airplane, Face, Motorbike, Leopard and Wrist Watches. We have used holdout method for preparing the training and test data, i.e., half the data is used for training and the remaining half data is used for testing. We have extracted PCA-SIFT[4] descriptors from the training images. The extracted set of descriptor are used to generate visual vocabulary(codebook). We generated codebook of size 15881 by setting L=6 and B=6. The proposed method was applied to retrieve relevant images. The retrieval result is shown in Figure 3. In this figure, the top left is the query image and the remaining images are top ranking retrieved images on application of our feature selection method on query image. We have selected the top 800 features from the query feature vector, and reduced the training database with selected feature indices. In the Figure 2, we have plotted Feature Relevance measure v/s number of features for the Face query. It is clearly evident from the plot that a large number of features are irrelevant for the Face image with our feature relevance measure and selecting only the features whose relevance measure is high is enough. Hence, this idea was tried by varying the top features whose relevance measure is high. During this experiment, we found that if increase the top relevant features from 800 to 1000, we started retrieving irrelevant images. Figure 4, shows the retrieval result by varying the feature vector size to 800,1000,3000,and 5000. For the Face query, top 800 features gave very good recognition result, and the result start deteriorating as we increase the features. We tried our experiment by varying the top features from 200,800,1000,5000 features. The outcome of this experiment was that, the top 800 features are enough to get good recognition performance. The recognition accuracy is shown in Table 1. The first column gives the category name, and remaining columns describes the recognition accuracy for varied feature vector size. The reason for getting low recognition accuracy for the leopard and the wrist watches is due to the fact that the number of sample images were very less for these categories. Moreover, our dataset is relatively small. The vocabulary should be generated from a large set of representative image set. We wanted to demonstrate the fact that it is worth applying feature selection method instead of using the entire feature vector for further processing. The advantage of our method is that, we do not require any specialized sparse coding technique, as these features will be thrown away in the feature selection.

Objects	800	1000	3000	5000
Face	70%	66%	59%	57%
Motor Bike	78%	72%	75%	78%
Airplane	79%	73%	75%	77%
Leopard	42%	40%	44%	42%
Watch	68%	58%	65%	68%

Table 1. Recognition Accuracy

6 Conclusions and Future Work

In this paper, we have presented a new feature selection method using bag-of-visual word representation. The feature selection algorithm suggested in this paper exploits database mean anad variance to determine the relevance feature measure. Using this relevance measure, we are able to reduce the dimensionality of feature vector substantially

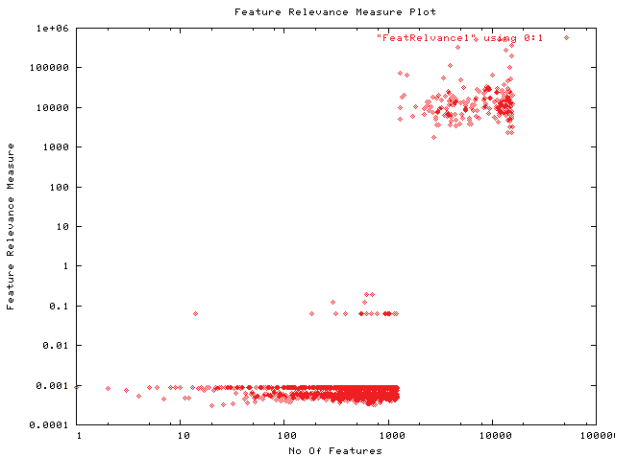

Figure 2. Feature Relevance

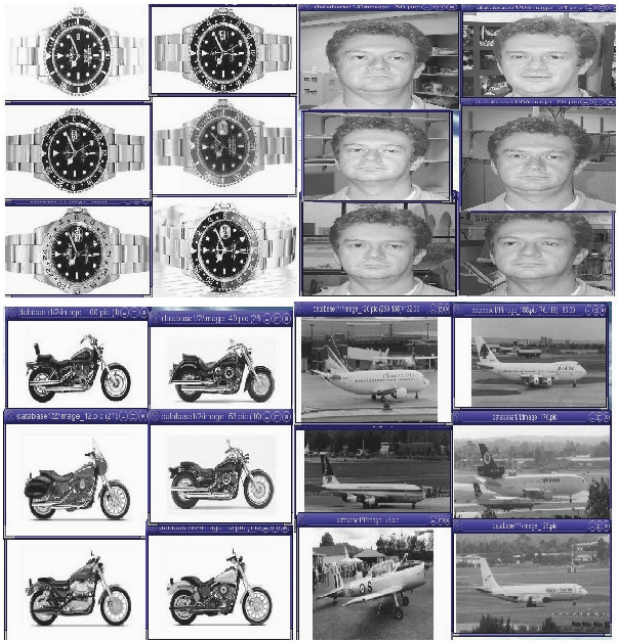

Figure 3. Retrieval Results with Feature Vector size=800: Top left is query image for each category and remaining images are retrieved images

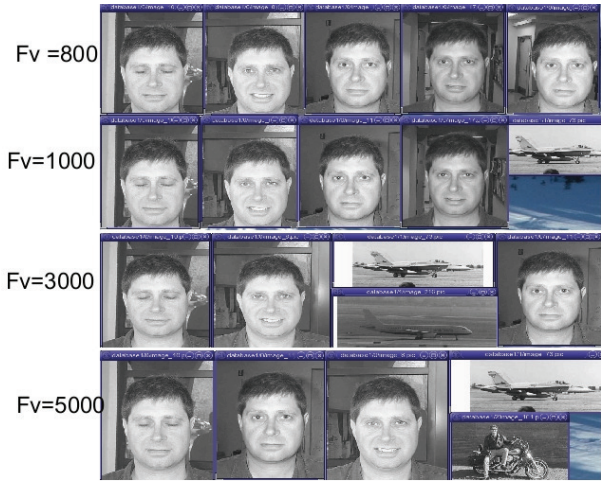

Figure 4. Retrieval Results with feature Selection, First image of each row is query image

and thereby increasing the classification performance by throwing out the irrelevant features. Results shown in the previous section suggest that it is worth doing feature reduction before matching. We are extending the current approach by employing other feature selection algorithms and using color and texture features.

Acknowledgment

This work is sponsored by CAIR, Bangalore. The authors wish to thank Director, CAIR, for his support.

References

1. D. G. Lowe, Distinctive image feature from scale invariant keypoints. In International Conference on Computer Vision,pages 1150-1157,1999
2. D. G. Lowe, Local feature view clustering for 3D object recognition. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, Dec 2001, pp. 682-688
3. Scott Helmer and D. G. Lowe, Object recognition with many local features, Workshop on Generative Model Based Vision 2004, Washington, D.C., July 2004
4. Y.Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. CVPR 2004, 02:506-513,2004
5. J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In Proceedings of the International Conference on Computer Vision, Vol 2,pages 1470-1477, Oct.2003.

6. D.Nister and H.Stewnius. Scalable recognition with a vocabulary tree. In CVPR '06: Proceedings of the IEEE Computer Society Conference on Computer vision and Pattern Recognition, pages 1261-1268, Washington DC, USA, 2006.

7. R. Fergus, L. Fei-Fei, P. Perona R., and A.Zisserman. Learning object categories from googles image search. In ICCV 2005.

8. V.Shiv Naga Prasad, A.G Faheema and Subrata Rakshit. Feature Selection in Example-Based Image Retrieval Systems. In the Proceedings Of ICVGIP 2002, Dec.2002.

9. Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann and Chong-Wah Ngo. Evaluating Bag-Of-Visual-Words Representation in Scene Classification. In the Proceedings of the International Workshop on Multimedia Information Retrieval,pages 197-206, MIR07,2007

10. Caltech101 dataset http://www.vision.caltech.edu/html-files/archive.html

11. http://en.wikipedia.org/wiki/Tf-idf