Australian Journal of Basic and Applied Sciences, 9(16) Special 2015, Pages: 118-125

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

An Experimental Study of Bag of Visual Words (BoVW) for Image Classification with different Vocabulary sizes

[1]P.Arulmozhi, [2]S.Abirami, [3]S.Murugappan

[1]*Department of Information Science and Technology, Anna University, Chennai, India*
[2]*Department of Information Science and Technology, Anna University, Chennai, India*
[3]*School of Computer Science, Tamilnadu Open University, Chennai, India.*

ARTICLE INFO

Article history:
Received 20 January 2015
Accepted 02 April 2015
Published 20 May 2015

Keywords:
Bag of Visual Words Vocabulary Codebook Feature Encoding

ABSTRACT

In today's scenario, Digital Technology advancement in field like digital photography and surplus access to Social Network Media, make the availability of images in enormous quantity. Because of this, the need to manage, categorize and retrieve images as per user needs increases day by day and this lay the foundation to make more and more research works on digital images particularly on Content Based Information Retrieval (CBIR). In this paper, inorder to classify images based on their categories, BoVW representation has been preferred which is a popular technique in Text Mining .The images are representated in BoVW representation and using this representation they are trained to classify based on their categories them by changing the parameters like vocabulary sizes, number of training images, cluster sizes and optimized parameters are selected. Then testing is performed on this trained model and their results are examined using CalTech 256 dataset. Thus it helps us to decide on choosing proper parameters which lead to the improvement in accuracy.

To Cite This Article: P.Arulmozhi, S.Abirami, S.Murugappan An Experimental Study of Bag of Visual Words for Image Classification with different Vocabulary sizes. **Adv. in Nat. Appl. Sci.,** *9(16): 118-125, 2015*

INTRODUCTION

The growth of Internet and advancement in digital images resulted in excessive availability of images. As these large quantities of images are often accessed, their storage in the databases is becoming mandatory. Rather than concentrating on retrieving relevant image alone, now focus has been drifted to retrieve them efficiently and accurately also. One way to achieve this is to perform Image Classification where images are grouped into different categories based on their similarities or dissimilarities. Here, in Image Classification when a query image is given, the performance of relevant Image Retrieval is better. As a result of this Image Classification is now becoming a key Technology for many applications like web content analysis, location recognition, image compression, medical diagnosis, scene classification, near duplicate detection and so on.

Success of any image related operations depend solely on image features and on their representation. Any further operation is done using these feature representation only. As a thumb rule, if feature representation is not appropriate, even best model cannot provide us best result. So this paper is about a feature representation called Bag of Visual words. As Bag of Visual Words(Sivic and Zisserman, 2003) are invariant to transformations, occlusions lighting, simple and compact representation of image content, we have chosen BoVW representation for image classification. Actually Bag of Visual Word is a common representation technique for Text Retrieval Processing and this has been extended to Content Based Image Retrieval (CBIR). More research work has been concentrated in improving this Bag of Visual Words (BoVW) representation and there by efficiency and accuracy of Image Retrieval is improved.

Here in this paper, Image Classification is performed to specify a relevant label(s) to the given test image from the stored database images using BoVW representation. BoVW representation is performed by the following steps. 1. Feature Extraction 2.Feature Preprocessing 3.Codebook generation 4.Feature Encoding 5. Pooling (Piotr Koniusz *et al*. 2013).

The rest of the paper is organized follows. Section 2 provides related research work done using BoVW. Section 3 provides the steps to be performed for creation of BoVW representation of given input images and use this representation for Image Classification. Various metrics used for evaluating the performance of Image classification are specified

Corresponding Author: P. Arulmozhi, Department of Information Science and Technology, Anna University, Chennai, India
E-mail: arulmozhikec@gmail.com

in Section 4.The experimental study of the performance of Image Classification for different vocabulary size are shown in Section 5.Finally the conclusion and further modifications to be concentrated in near future is given in Section 6.

2. Related Works:

We here review relevant research works that are being carried out with Bag of Visual Words and other relevant issues. BoW which is a popular Technique in Text Processing, is extended this representation in Video Object Matching (Sivic and A. Zisserman, 2003). There are many techniques to describe local features like SIFT, GIST (Zhicheng Li and Laurent Itti 2011), but due to invariance property towards translation, rotation and viewpoint changes SIFT has gained popularity (D.G. Lowe, 2004).Always there is a debate about which kind of representation of images is better- local or global and this is elaborately discussed in (Yongjin Lee *et al*., 2005). In this paper, BoVW being a global representation is considered which in turn uses SIFT as local representation. So it possesses both the merits of local and global representation.

The discrimination power can be increased if spatial information is included, which is lacking in BoVW. So lots of research works are performed to improve the discrimination power of BoVW by exploring each and every steps of it. The survey papers (Ken Chatfield *et al*., 2011), (Xiaojiang Peng *et al*., 2014), (Yongzhen Huang *et al*., 2014) discussed numerous ways of Feature Encoding and Pooling Techniques. These papers give tremendous inputs on Feature Encoding Techniques like Fisher Vector (FV), the VLAD, the Super Vector (SV) and the Efficient Match Kernel (EMK) which are basically grouped under Hard and Soft Assignments(Mohammad Mehdi Farhangi *et al*.,2014). In (Jiang Hao and Xu Jie, 2010),GMM ,a Soft Assignment Feature Encoding Technique is applied and used histogram to show improvement in classification performance.

In another direction, to avoid useful information loss, learning codebook has been given importance and many papers has come with modifications (Hongping Cai *et al*., 2010) (Xinmei Tian and YijuanLu, 2013).In(Chunjie Zhang *et al*., 2014), encoding was done by including visual word's governing region which was generated by weighted local features by considering visual words similarities and based on this they proposed a new algorithm called weighted feature sign search algorithm (Chunjie Zhang *et al*., 2014)(Hongping Cai *et al*., 2010)

Pooling, being the last step in BoVW representation got attention now-a-days and this was focused as a major step for improving the accuracy. Here, various histogram representations are tried like Sum Pooling, Average Pooling and Max Pooling (Naila Murray and Florent Perronnin, 2014) (Piotr *et al*., 2013).Each method has their own pros and cons. In (Naila Murray and Florent Perronnin, 2014), a new pooling method called Generalized Max Pooling has effect similar to that of Max Pooling and has been extremely used in Fishers vector Encoding Technique. In (Piotr *et al*.,2013), authors had compared many Feature Encoding methods and Pooling methods of BoVW and investigated descriptor interdependence by applying Pooling methods and proposed a new improved Pooling Algorithm.

To improve the retrieval efficiency and accuracy of BoVW representation, PCA was proposed which reduced feature vector space and applied it to Leader clustering algorithm and showed reduction in Quantization loss (Shusheng *et al*., 2013).In (Yu-BinYang *et al*., 2013), to overcome the weak discrimination power and strong ambiguity of the low level features in BoVW, they proposed to use both the nearest and furthest visual words together for codebook writing and by this way they were able to create more semantic content accurately. Another group has been working on BoVW to provide spatial information (Xiaohui Shen *et al*., 2014), providing fuzzy clustering as in (K.S.Sujatha *et al*., 2012)

In (Yu-BinYang *et al*., 2013), studied visual vocabularies from different datasets are very similar to each other, and concluded that built visual vocabularies for different data set and exchange them without performance loss called as universality of visual vocabularies. To find semantically similar visual words, using semantic local adaptive clustering(SLAC) algorithm with SIFT descriptor, reduced similar keywords and thereby maintain semantic relations between key points and objects(Kraisak Kesorn *et al*., 2011).

In order to avoid the information loss in conventional BoVW representation, an optimized codebook called semantics preserving codebook was used and correspondingly proposed a semantic preserving Bag of words algorithm to reduce the semantic gap (Lei Wu *et al*., 2010). The performance of fuzzy clustering Multiple Dictionary Bag of Words model using Separate dictionary increases the performance (K.S.Sujatha *et al*.,2012).To improve the bow retrieval efficiency and accuracy, they proposed PCA to reduce the feature vector space and then applied Leader clustering algorithm and showed reduction in quantization loss (Shusheng *et al*.,2013)

Thus for doing a lot of improvements in BoVW, the knowledge about the initial working principle of BoVW are needed and as a result this paper talks about how the changes made in vocabulary sizes affect the performance of Image Classification which uses BoVW representation.

3. Image Classification:

Image search plays an increasingly important role in our daily lives due to the explosive growth of Web images. Classification is the problem of

identifying a category for a new observation, on the basis of a training set of data containing observations whose category membership is known. Here, Image classification has its own role to analyze the properties of various image features and it help to organize data into categories.

A. Image Classification:

Classification algorithms typically employ two phases of processing: Training and Testing. In the initial training phase, characteristic properties of typical image features are isolated and, based on these, a unique description (representation) of each classification category, i.e. training class, is created. In the subsequent testing phase, these feature-space partitions are used to classify image features.

Steps performed in Image Classification:

Image Classification is done by the following steps

Step 1.**Feature Representation**: Represent each training and testing images in a vector form .Here it is done using Bag of Visual Words (BoVW) representation.

Step 2.**Training:** Create a model to classify the images based on their similarity/dissimilarity. In this paper, multi-class SVM (Support Vector Machine) is preferred for training.

Step 3.**Testing:** Use the model trained done in step 2 to categories the test images correctly. Again multi-classSVM is for testing purpose.

B. Bag of Visual Words:

The visual representation of image is the fundamental factor to the quality of Content-Based Image Search. Recently, Bag-of-Visual Word model has been widely used for image representation and has demonstrated promising performance in many applications.

It is an order less document representation where the occurrence of each word is used as a feature for training a classifier. It quantizes feature space to make discrete set of visual words.

The following five steps are performed to have BoVW Representation (Piotr Koniusz *et al*. 2013). They are

1. Feature Extraction
2. Feature Pre-processing
3. Generate codebook
4. Feature Encoding
5. Pooling

1. Feature Extraction:

It extracts local patches to identify the key points from images. This step can be done in one of the two ways. The first way is to detect features for all images and then describe the features; and the other way is to directly apply feature descriptors without using any Feature Detection methods. Feature Detection is used to detect various features found in the images using methods like Corner Detector, Blob Detector and Affine Invariant Feature Detectors, Canny Edge Detector, Harris Corner Detector, Hessian, Difference Of Gaussian(DOG).Feature Descriptors are used to represent the key points by converting key points(local patches) into vectors like SIFT (D.G. Lowe, 2004), SURF, HOG,ORB,BRIEF (Muja and David, 2012).

2. Feature Pre-processing:

High dimensional functions tend to have more complex features than low-dimensional functions, and hence harder to estimate and this is called as curse of dimensionality. So local features which are usually high dimensional and strongly correlated are transformed to set of linearly uncorrelated variables.

3. Generate Codeword:

After extracting the feature vectors, learning Visual Vocabulary is to be accomplished. This is done usually in Unsupervised mode which uses clustering algorithms like k-means, hierarchical k means. Here using clustering algorithms, the similar features are grouped into same cluster and for each cluster, a cluster centre called codeword is calculated. These cluster centers together form a Vocabulary or Dictionary for these training set. But to do this visual vocabulary apart from Unsupervised learning, Supervised/Semi-supervised Learning methods are used now-a-days.

4. Feature Encoding:

After finding vocabularies, each feature descriptors are to be mapped to a codeword belonging to a cluster and generate a coding vector with length equal to number of codewords. There are different feature encoding algorithms which alters in the process of selecting codewords for each feature; they are broadly classified as Hard and Soft Quantization. Hard Quantization refers to assigning descriptors to the most nearest visual word; Another category called soft Quantization assign features to the combination of visual words by finding the difference between the features and the visual words. (Eg: Fishers encoding, super vector encoding) In this feature encoding, only hard quantization is tried (Piotr Koniusz *et al*., 2013).

5. Pooling:

Pooling is the operation which involves aggregating several local descriptor encoding into a single representation. This step gives the final global representation of an image. Average (Avg), Maximum Pooling (Max), Mix-order Max-pooling (MixOrd), weighted pooling and an Lp-norm based trade-off (lp-norm) are some of the pooling techniques normally adopted. (Naila Murray and Florent Perronnin, 2014).

Here each and every step has their role to contribute for efficient Image classification.

4. Evaluation Metrics:

Any system should have quantitative metrics to analyze its performance. In this paper, the evaluation metrics considered for Image Classification are Precision, Recall, Accuracy and Average Precision.

1. **Precision:** It is a measure of result relevancy. Precision is the fraction of retrieved images that are relevant.

$$\text{Precision} = TP / (TP + FP) \quad (1)$$

where TP and FP are the numbers of true positive and false positive predictions for the considered class.

2. **Recall:** Recall is a measure of how many truly relevant results are returned. Recall is the fraction of the images that is relevant to the query images that are successfully retrieved.

$$\text{Recall} = \text{Sensitivity} = TP / (TP + FN) \quad (2)$$

where TP and FN are the numbers of true positive and false negative predictions for the considered class.

TP + FN is the total number of test examples of the considered class.

3. **Accuracy:** Accuracy is how close a measured value is to the actual (true) value. The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN) \quad (3)$$

4. **Average Precision:** It is an average of the precision values of the classes.

RESULTS AND DISCUSSION

The experiment was performed in Matlab 2013a and to do training and testing for Image Classification CalTech-256 dataset is considered. It consists of approximately 29,700 images with 256 object categories. Each category has at least 85 elements. Among 256 classes only 4 classes are taken to perform the experiments namely motorbikes, t-shirts, aeroplanes and faces classes. Some sample images of afore mentioned four classes are shown in Fig 1.For training 200, 100, 50 images per class are taken and 10 images are considered for testing. To get BoVW representation from raw images, Interest points are extracted first by applying Harris Corner Detector and from that local patches are obtained using 128 bit SIFT Descriptor. K-means Clustering has been used for Codebook Generation and for Feature Encoding, Hard Quantization Technique called Vector Quantization is performed. The final step of BoVW is pooling operation, where average pooling is applied to the feature encoded histogram. Then to perform Image Classification using BoVW representation, multi-class SVM is applied for both training and testing Caltech images.

Fig. 1: Sample images of four classes

Initially Harris corner detector and SIFT descriptor are used to collect local patches for different Vocabulary sizes namely 250,400, 500, 600, 750, 1000 and 2000. For each vocabulary size, the training images considered are 200,100 and 50.For testing, 10 images are considered for all the four classes. The accuracy obtained for these different parameters are shown in Table I. These accuracy values are plotted as a graph as shown in Fig 2.

Table I: Accuracy for various Vocabulary Sizes and Training Images

Vocabulary size	Training images	Accuracy
250	200	82.5
	100	87.5
	50	77.5
400	200	87.5
	100	85
	50	82.5
500	200	87.5
	100	85
	50	82.5
600	200	87.5
	100	87.5
	50	82.5
750	200	87.5
	100	87.5
	50	90
1000	200	77.5
	100	82.5
	50	85
2000	200	77.5
	100	62.6
	50	60

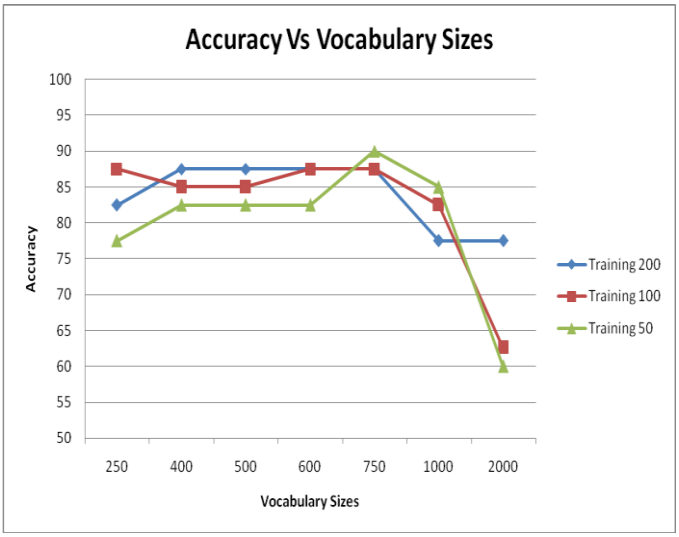

Fig. 2: Accuracy graph for different Vocabulary sizes and training image sizes

From Fig 2, for training set of 200 images, as vocabulary size increases the accuracy increases from 82.50% to 87.5 % and it remains constant till the vocabulary size reaches 750 and after that the accuracy drops down as vocabulary size increases. For 100 training image set it starts with 87.5 % accuracy and the accuracy is changing (increase/decrease) around 87.5 % till vocabulary size comes to 750. For the training set of 50 images, and for the vocabulary size 250, it shows the lowest accuracy value as 77.5 % but when the vocabulary size reaches 750 it gives highest accuracy value as 90% among all the training sets; and after that for both training set 100 and 50 the accuracy deeply decreases when vocabulary size increases. This show that even with smaller training set (50) with proper selection of vocabulary sizes (750), the maximum accuracy can be obtained. Thus the assumption stating that as vocabulary size increases the accuracy also increases is not true for all the cases and it could be hold partially only; and this has been proved from this graph as well.

A Confusion Matrix is resulted for each and every Vocabulary sizes and Training Image size combinations. Infact the accuracy shown in previous table are calculated using this Confusion Matrix only. So as a sample, only one Confusion Matrix of vocabulary size 250 and training image size 100 is shown in Fig. 3.

	Motor bikes	T-shirts	Aero planes	Faces
Motorbikes	8	0	1	1
T-shirts	0	10	0	0
Aero planes	0	1	8	1
Faces	0	0	1	9

Fig. 3: A sample Confusion matrix for vocabulary size of 250 and training Image size of 100

The same Confusion Matrix is used to calculate Precision, Recall and the results are shown in Table II. Here Class A,B,C,D refers to dataset motorbikes, t-shirts, aeroplanes and faces classes respectively. The Precision and Recall calculation done for individual classes are performed using the formulae given as in Evaluation Metrics. Along with this, Average Precision is also calculated for different Vocabulary Sizes and Training Image Sizes. A graph showing the Average Precision for different vocabulary sizes are given in Fig. 4.The Average Precision graph also depicts that for vocabulary size of 750, almost for all the training image sets of 200,100,50 gives high value compared to all other vocabulary sizes which is the same conclusion stated from the accuracy graph.

Table II: Precision, Recall for each class and Average Precision calculated for different Vocabulary Sizes

Vocabulary Size	Training Image size	Class A		Class B		Class C		Class D		Avg. Precision
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	
250	200	0.9	0.9	1	0.9	0.67	0.8	0.78	0.7	0.84
	100	1	0.8	1	1	0.8	0.8	0.8	0.9	0.9
	50	0.9	1	0.88	0.7	0.7	0.7	0.64	0.7	0.78
400	200	1	0.9	0.89	0.8	0.89	0.8	0.77	1	0.88
	100	0.9	0.9	0.9	0.9	0.86	0.6	0.77	1	0.86
	50	0.75	0.9	0.89	0.8	0.81	0.9	0.81	0.9	0.83
500	200	0.75	0.9	1	1	1	1	0.8	0.8	0.89
	100	0.77	1	1	0.9	0.89	0.8	0.77	0.7	0.85
	50	0.69	0.9	1	0.8	1	0.8	0.72	0.8	0.85
600	200	0.75	0.9	1	0.8	0.81	0.9	1	0.9	0.89
	100	0.83	1	0.9	0.9	1	0.7	0.81	0.9	0.89
	50	0.77	1	1	0.6	1	0.8	0.69	0.9	0.86
750	200	0.83	1	0.9	0.9	0.89	0.8	0.89	0.8	0.88
	100	0.83	1	1	0.8	0.89	0.8	0.82	0.9	0.88
	50	0.83	1	1	0.9	0.89	0.8	0.9	0.9	0.9
1000	200	0.76	1	1	0.8	0.8	0.8	0.5	0.62	0.77
	100	0.78	0.7	0.89	0.8	0.83	1	0.8	0.8	0.825
	50	0.88	0.7	0.9	0.9	0.9	0.9	0.75	0.9	0.85
2000	200	0.71	1	0.85	0.6	0.88	0.7	0.72	0.8	0.79
	100	0.64	0.9	0.8	0.4	0.5	0.4	0.61	0.8	0.64
	50	0.64	0.9	0.66	0.4	0.5	0.4	0.58	0.7	0.59

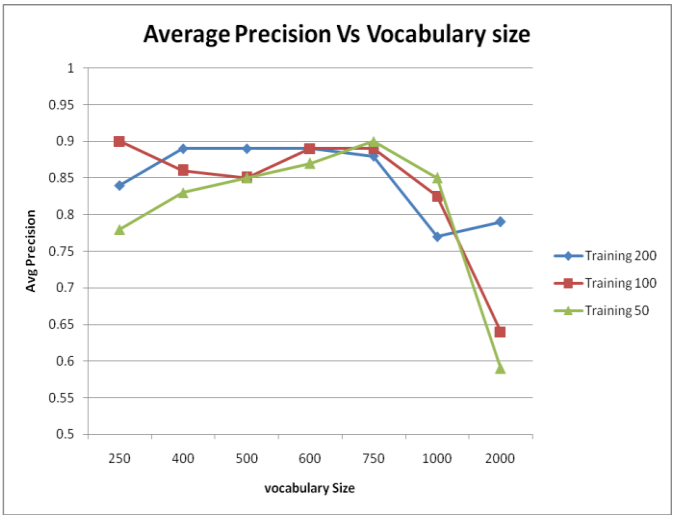

Fig. 4: Average Precision for different Vocabulary sizes

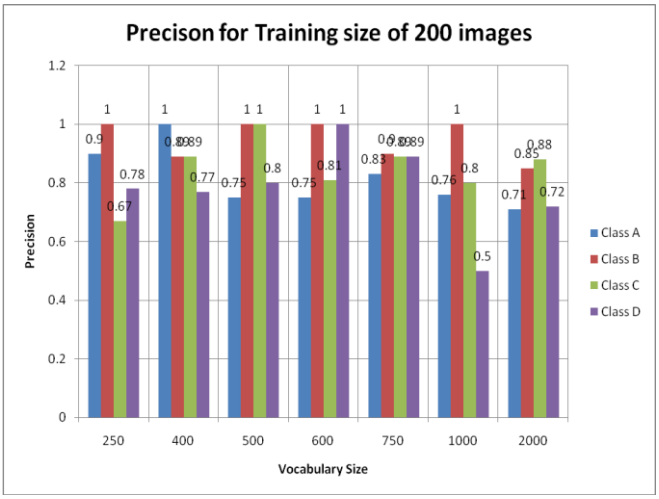

Fig. 5: Precision for different Vocabulary size of Training Image size of 200

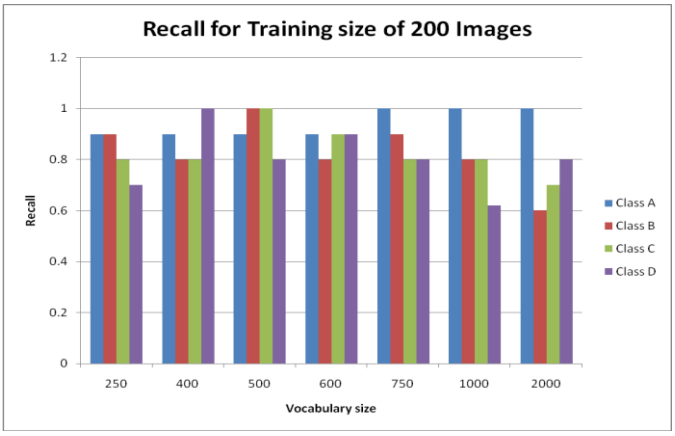

Fig. 6: Recall for different Vocabulary size of Training Image size of 200

Similarly, Fig. 5 and Fig. 6 shows the precision and recall values respectively for different vocabulary sizes and for 200 training image size. For a vocabulary size the precision and recall values of each class is not the same. But for the vocabulary size of 750 both the graphs give better values compared to all other vocabulary sizes which in turn ensure that Image classification rate is higher for this vocabulary size.

Conclusion:

As Bag of Visual Words are simple, compact representation of image content and invariant to transformations, occlusions, lighting, we have chosen BoVW representation for Image Classification .In this paper by changing the parameters like vocabulary size and training image sizes we explored the working behavior of BoVW representation to perform Image Classification. The accuracy of Image Classification depends on these parameters, therefore selection of parameters should be done appropriately. If the taken Vocabulary size is either too low or too high, generalization and overfitting problem occurs respectively. Therefore considered vocabulary size should be in such a way that the Image Classification is not affected by both the problems. Here to do training and testing of Image Classification is examined using CalTech 256 data set and found that high accuracy is obtained for vocabulary size of 750. Similarly, Precision, Recall and Average Precision are used as other evaluation metrics for Image Classification. For further exploration, need to utilize spatial contextual information, apply supervised/semi supervised techniques for learning codewords, apply different pooling techniques to visualize their impact on accuracy improvement for Image Classification and Image retrieval

REFERENCES

Chunjie Zhang, Xian Xiao, Junbiao Pang, Chao Liang, Yifan Zhang, Qingming Huang, 2014. Beyond visual word ambiguity: Weighted local feature encoding with governing region. in Journal of Visual Communication. Image R. Elsevier, 25(6): 1387-1398.

Hongping Cai, Fei Yan, Krystian Mikolajczyk, 2010. Learning weights for codebook in image classification and retrieval. in IEEE International Conference on Computer Vision and, Pattern Recognition., pp: 2320-2327.

Mohammad Mehdi Farhangi, Mohsen Soryani, Mahmood Fathy, 2014. Informative visual words construction to improve bag of words image representation.IET Journals & Magazines, 8(5) DOI: 10.1049/iet-ipr.2013.0449: 310-318.

Hao Lei, Kuizhi Mei, Nanning Zheng, Peixiang Dong, Ning Zhou, Jianping Fan, February, 2014. Learning group-based dictionaries for discriminative image representation. Pattern Recognition, 47(2): 899-913.

Yan Ke and Rahul Sukthankar, 2004. PCA-SIFT: A more distinctive representation for local image descriptors. in proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, CVPR 2004 2: 513-506.

Ken Chatfield, Victor Lempitsky, Andrea Vedaldi and Andrew Zisserman, September 2011. The devil is in the details an evaluation of recent feature encoding methods. in BMVC 2011. proceedings: BMVC.

Kraisak Kesorn, Sutasinee Chimlek, Stefan Poslad, Punpiti Piamsa-nga, September 2011.Visual content representation using semantically similar visual words.Expert Systems with Applications, Elsevier, 38(9): 11472-11481.

Yongjin Lee, Kyunghee Lee, Sungbum Pan, 2005. Local and global feature extraction for face recognition. in proceedings of the 5th international conference on audio- and video-based biometric person authentication, 219-228.

Lei Wu, Steven C. H. Hoi and Nenghai Yu, 2010. Semantics-Preserving Bag-of-Words Models and Applications. IEEE Transactions on Image Processing, 19: 1908-1920.

Lican Daiy, Xiaoyan Sunz, Feng Wuz and Nenghai Yuy, 2013. Large Scale Image Retrieval with visual groups. Image Processing (ICIP), 20th IEEE International Conference on DOI: 10.1109/ICIP.2013.6738532: 2582-2586.

Jiang Hao, Xu Jie, 2010. Improved Bags-of-Words Algorithm for Scene Recognition. Signal Processing Systems (ICSPS), 2010 2nd International Conference on Vol:2 DOI:10.1109/ICSPS.2010.5555494: V2-279 - V2-282.

D.G. Lowe, 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision., 60: 91-110.

Marius Muja and David G. Lowe, 2012. Fast Matching of Binary Features. IEEE Conference Publications: 404-410.

Naila Murray and Florent Perronnin, 2014. Generalized Max Pooling. CVPR 2014 - IEEE Conference on Computer Vision & Pattern Recognition.

Piotr Koniusz, Fei Yan, Krystian Mikolajczyk, 2013. Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection. Computer Vision and Image Understanding, Elsevier, 117(5): 479-492.

Shusheng CEN, Yuan DONG, Hongliang BAI, Chong HUA, 2013. Fast and compact visual codebook for large scale object retrieval. Broadband Network & Multimedia Technology (IC-BNMT), 5th IEEE International Conference on, 17-19: 35-38.

Sivic, J. and A. Zisserman, 2003. Video Google: a text retrieval approach to object matching in videos'. In ICCV, 2003. 809, 810 Computer Vision, Proceedings. Ninth IEEE International Conference on 13-16(2): 1470 -1477 .

Sujatha, K.S., P. Keerthana, S. Suga Priya, E. Kaavya, B. Vinod, 2012. Fuzzy based Multiple Dictionary Bag of Words for Image Classification. International Conference on Modeling Optimisation and Computing., 38: 2196-2206.

Xiaojiang Peng, Limin Wang, Xingxing Wang, Yu Qiao, 2014. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice. arXiv:1405.4506v1 [cs.CV].

Xinmei Tian, YijuanLu. Discriminative codebook learning for Web image search Signal Processing. Elsevier, 93(8): 2284-2292.

Xiaohui Shen, Zhe Lin, Jonathan Brandt and Ying Wu, 2014. Spatially-Constrained Similarity Measure for Large-Scale Object Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(6): 1229-1241.

Yang, J., Y. Jiang, A. Hauptmann, C. Ngo, 2007. Evaluating bag-of-visual-words representations in scene classification. in: International workshop on Multimedia, Information Retrieval, pp: 197-206.

Yongzhen Huang, Zifeng Wu, Liang Wang, Tieniu Tan, 2014. Feature Coding in Image Classification: A Comprehensive Study. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(3) DOI: 10.1109/TPAMI.2013.113: 493-506.

Yu-BinYang, Ling-YanPan, YangGao, Guang-NanHe, YaoZhang, 2013.Visual word coding based on difference maximization.Neurocomputing, Elsevier, 120: 277-286.

Zhicheng Li and Laurent Itti, 2011. Saliency and Gist Features for Target Detection in Satellite Images. IEEE Transactions on Image Processing, 20(7): 2017-2029.

Jian Hou,Wei-Xue Liu,Xu E,Qi Xia, Nai-Ming Qi, 2013.An experimental study on the universality of visual vocabularies. Journal of Visual Communication and Image Representation, Elsevier, 24(7): 1204-1211.