VOCABULARY LENGTH EXPERIMENTS FOR BINARY IMAGE CLASSIFICATION USING BOV APPROACH

S.P.Vimal[1], Eshaan Puri[2] and P.K.Thiruvikiraman[3]

[1,2]Department of Computer Science and Information Systems
Birla Institute of Technology and Science, Pilani, Rajasthan, India
[3]Department of Physics, Birla Institute of Technology and Science,
Hyderabad Campus, Andra Pradesh, India

ABSTRACT

Bag-of-Visual-words (BoV) approach to image classification is popular among computer vision scientists. The visual words come from the visual vocabulary which is constructed using the key points extracted from the image database. Unlike the natural language, the length of such vocabulary for image classification is task dependent. The visual words capture the local invariant features of the image. The region of image over which a visual word is constrained forms the spatial content for the visual word. Spatial pyramid representation of images is an approach to handle spatial information. In this paper, we study the role of vocabulary lengths for the levels of a simple two level spatial pyramid to perform binary classifications. Two binary classification problems namely to detect the presence of persons and cars are studied. Relevant images from PASCAL dataset are being used for the learning activities involved in this work

KEYWORDS

Image classification, Bag-of-Visual-words, Spatial Pyramid

1. INTRODUCTION

Image classification is a supervised learning task of classifying the image to belong to one of the several known semantic categories. Object detection is also a supervised learning task of detecting the presence of one or more objects in the given image and optionally to detect their possible location and scale in the given image. Identifying the appropriate representation of object(s) or the image(s), formulation of the learning problem and evaluation of the learned hypothesis are key activities in any supervised learning problem. The approaches to image classification/detection vary along these lines.

Bag-of-Visual-words (BoV) [1],[4] representation for images is a successful representation for wide range of image analysis applications, which is inspired from the Bag-of-Words (BoW) [1], [4] approach to text mining. The success of BoV approach largely lies in the appropriate definition of visual vocabulary consisting of visual words extracted from the domain of image classes / object classes being under consideration. The visual vocabulary is built from the key point descriptors extracted from images by standard clustering techniques by explicitly choosing the number of visual words required for the vocabulary. Once the vocabulary of visual words is constructed, each image is then represented by a distribution of visual words present in the image. The number of words in the visual vocabulary and the nature of description (color, intensity, gradient information, texture etc.) available with each word in the vocabulary are task dependent and they are generally chosen empirically. Preserving the spatial information, i.e. the spatial region where each visual word in the visual vocabulary is typically found, is important for better performance of the classification/ detection task. There are different approaches in the literature addressing this aspect of representation. Once the representation of visual vocabularies and the

DOI : 10.5121/sipij.2013.4608

distribution of visual words in each image are arrived at, the learning and classification can proceed as if it is for text documents.

In this paper, we study histogram tiling technique used for building the visual vocabulary for two simple binary classification problems. The binary classification problems we considered are (i) to detect if the image consists of person(s) (ii) to detect if the image contains a car. The visual vocabulary for detecting the presence of a car(s) or person(s) in an image consists of visual words from positive and negative examples of all car(s) or person(s) from VOC 2007. We construct four types of histogram tiling inspired from [1] [3] [4]. We considered the maximum of two levels of the spatial pyramid in which the second level represents the image tessellated into four quarters and reported the relative importance of these levels. The relative importance of these levels(scales) is studied with the help of the required vocabulary length for better classification.

The rest of the paper is organized as follows: Section-II briefly reviews the related works, Section-III outlines the Bag-Of-Visual-words (BoV) Approach implemented in this paper, Section-IV explains the histogram tiling and the variations considered in this paper, Section-V explains the experiments conducted along with the explanation on the dataset, Section-VI Discusses the results obtained and the brief discussions and the Section-VII concludes the paper.

2. RELATED WORKS

The analogy of text retrieval problem to the content based image retrieval was discussed in [5] which motivated a lot of fruitful research in this direction. Sivic and Zisserman [6] extended the Bag-of-Words (BoW) approach to text retrieval for the computer vision problem of matching objects in videos. Their work provided systematic formulation of textual retrieval problem for retrieving objects from videos. The visual words forming the visual vocabulary are constructed using descriptors like Histogram of Oriented Gradients [7] , filter banks [8],local image patches [9] and SIFT [10] features. Lowe's SIFT [10] is among the most commonly used local descriptor. The local features extracted are vector-quantized to form a visual vocabulary. The BoV obtained by quantizing local features and representing image documents by the histograms introduces sparcity and sacrifices the locality of local features [6].

The spatial information which is lost in the traditional BoV is vital for scaling the classification/detection process for larger number of categories and Lazebnik et. al [1] proposed a spatial pyramid based approach to address this. The spatial pyramid proposed by Lazebnik et. al. constructs the pyramid by repeatedly subdividing the image to obtain the successive levels in the pyramid instead of gaussian subsampling approach. The concatenated histogram at each level is used for learning purpose. Also, Lazebnik et. al [1] proposed a matching scheme for such concatenated histograms. Similar spatial pyramids have been used for the problem of indoor-outdoor classification in [11] along with local features like color, textures and frequency information. Bosch et. al [2] used a similar spatial pyramid scheme to represent the shape of object categories by extending Histogram of Oriented Gradients [7] and reported considerable improvements in performance when used along with appearance feature [12]. Histogram tiling approach to address the loss of spatial information can be considered as a special case of spatial pyramid approach [1] [2] in which we consider only scale by partitioning an image in a regular fashion and the histograms for each partition are concatenated. Viitaniemi et al. [13] experimented various partitioning techniques for spatial tiling along with the hard and soft tiling approaches and hard and soft histograming approach. They confirm the usefulness of taking the spatial distribution into account. The encoding of spatial information by means of SPM [1] encodes each region on every scale of the pyramid from the vocabulary generated from the entire image. The sky often falls in the top region of the image and similar spatial arrangements can be seen from various semantic categories. Zhang et al. [3][4] proposed to encode the regions individually within the spatial constraints. This requires learning the vocabularies for regions independently, in an attempt to better capture the inherent discriminative features of the individual regions concerned.

The role of various aspects of visual vocabulary building are studied in the literature [4][13][14]. In this paper, we study the relative importance of the vocabulary at level-0 to the vocabulary at level-1 of the spatial pyramid representation for both SPM [1] and SPC [3] based BoV approach for a binary classification problem. Our results reveal that the shorter vocabulary is needed for level-0 relative to level-1 in the spatial pyramid. This implies the level-0 of the spatial pyramid serves as the context for level-1.

3. BAG OF VISUAL WORDS APPROACH

There are many different approaches to the construction of visual vocabulary from the given image database. We follow a simple approach to construct the visual vocabulary in which we quantize the SIFT descriptors into visual words. We use Lowe's SIFT descriptors [10] for this purpose which are densely extracted from the image. Each SIFT descriptor is a 128 dimensional vector. For PASCAL dataset, 1416 key points are typically extracted per image on an average [14]. We use a random subset of key points extracted from all the images and then perform k-means clustering for constructing varied lengths of vocabulary. The length of visual vocabulary is the number of visual words present in the vocabulary. The length of visual vocabulary impacts the learning performance. The vocabulary of length r leads to the r-dimensional histogram representation for each image in the database representing the frequency of visual words in the corresponding image.

4. SPATIAL INFORMATION WITH BOV APPROACH

Spatial extensions to the classic BoV approach yields better classification performance. There are various approaches to incorporate spatial information in the feature vector for images [1] [2] [13] [15] in the literature. Histogram tiling refers to forming the histogram representation for the whole image from the histograms computed on parts of the image. Image is typically partitioned in a specific spatial arrangement [13] and the histograms are computed for individual partitions. Such histograms are concatenated to form the feature vector for the image. Histogram tiling can also be obtained by applying various tiling masks on the image and combining the responses from each such mask. The masks can be overlapping in which the histogram and the tiling process are either hard or soft [13].

The histogram tiling approach employs various ways of partitioning the image and there is no single partitioning method which applies to all the learning problems. For natural scenes, in which the sky, water and the sand appear together, the partition 3x1 is more appropriate. Partitioning the image in 2^kx2^k where k = 0, 1, 2 etc. is more common in the literature. To obtain the best results in the classification, the histograms obtained in many resolutions are combined in systematic way, i.e. combining the histograms obtained for k=0, 1, 2 etc. This approach to incorporate the spatial information, results in the spatial pyramid [1] [2]. The vocabulary is learned for the whole image and then the distribution of visual words for each partition is computed systematically. Zhang et al. [3][4] computed vocabularies for individual partitions and use them to find the distribution of visual words in the corresponding partitions. In this paper, we consider two levels of resolutions namely level-0 and level-1 which is shown in the Figure. 1. The level-0 considers the entire image and level-1 is a partitions level-0 into 2x2.

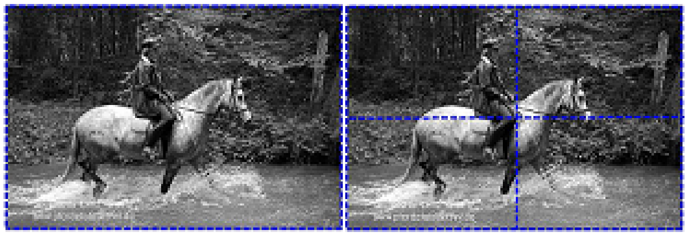

Figure. 1. Spatial pyramid with two levels. The image on the left is level-0 and the partitioned image on the right is level-1.

We consider four cases for our comparison namely case-1, case-2, case-3 and case-4. The case-1 and case-3 considers coding only level-1 and these cases does not code level-0. For case-1, we learned the visual vocabulary for the entire image and used this vocabulary to code the all four partitions of level-1. This case reflects a simple histogram tiling. For case-3, we learned the separate visual vocabularies from all four partitions of level-1 and used them to code the respective partitions. Case-2 and case-4 involves both levels of the pyramid. The case-2 learns a global vocabulary and we use it to code all the partitions of level-0 and level-1. Individual vocabularies are learned for all partitions and the partitions are coded with the respective vocabulary for case-4. The cases 1 and 2 are inspired from works on histogram tiling and SPM [1][2] and the cases 3 and 4 are inspired from SPC [3] [4].

5. EXPERIMENTS AND RESULTS

We used the images from PASCAL-2007 [16] dataset for the classes, person and car. For training purpose, we ensured the positive and negative are equal in numbers and we used the validation images provided in the dataset for our testing purpose. The car class has 713 positive training examples, and we picked 713 negative training examples which make 1426 images for training the car class. The testing images for car class consisting of 1442 images out of which 721 images are positive and 721 images are negative. Similarly, the person class has 4016 images for training out of which 2008 are positive training examples and 2008 are negative training examples. The testing images in person class are 4014 in which 2007 are positive testing images and 2007 are negative testing images.

We experimented with the vocabulary length for all the four cases. We varied the vocabulary length and computed the mean average precision for each case. For case-2 and case-4, we varied the vocabulary length of level-1 for a range of vocabulary lengths of level-0. For case-2, fixing the vocabulary length for level-0 and varying it for level-1 implies learning two vocabularies of different lengths from the entire image and code two levels separately. For cases 1 and 3, the length of the feature vector for each image is the length of the vocabulary multiplied by 4. For cases 2 and 4, the length of the feature vector is the sum of vocabulary length for level-0 and the length of feature vector for level-1. For cases 1 and 3, we varied the vocabulary lengths from 25 to 250, thereby varying the length of the feature vector from 100 to 1000. For cases 2 and 4, we varied the vocabulary length for level-1 from 25 to 250 for the vocabulary length of level-0 varying from 50 to 250. For example, when the level-0 vocabulary length is 50 and level-1 vocabulary length is 25, we have the shortest feature vector of length 50 + (4 x 25) = 150 for case-2 and case-4.

The classification is learned using SVM classifier with a linear kernel. Figure. 2 and Figure. 3 show the performance for detecting person and car respectively for various vocabulary lengths for cases 1 and 3. For the problem of detecting persons, the performance of case-3 dominates the case-1 for the entire range of vocabulary lengths. The mean average precision increases as the vocabulary length increase from 25 for both case-1 and case-3 here. The performance gain is not significant for case-1 beyond the vocabulary length 100. The performance for case-3 is significantly higher for the vocabulary length of 150 than for case-1. This huge difference could be due to the fact that case-3 captures the regional features better in the presence of a person in an environment. The environment in a human being is present hardly contains any visual words representing human face in the bottom partitions and similarly legs cannot be expected in the top two partitions.

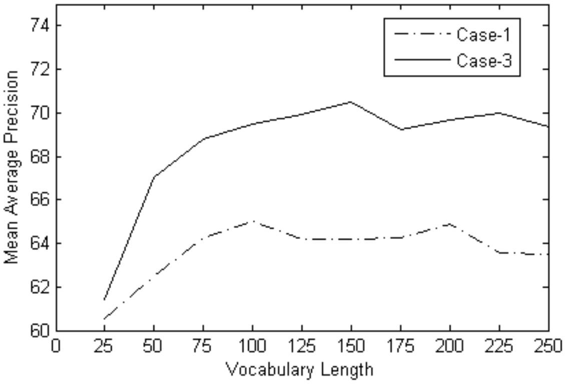

Figure. 2. The Mean Average Precision for various vocabulary lengths for detecting person for case-1 and case-3

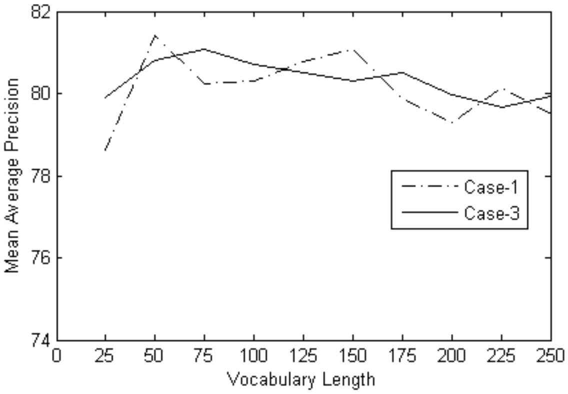

Figure. 3. The Mean Average Precision for various vocabulary lengths for detecting car case-1 and case-3

Figure. 3 indicates that the performance of case-1 and case-3 are almost similar. In contrast to detecting persons, neither case-1 nor case-3 dominates in the range of vocabulary lengths tested for detecting the presence of car. This implies that the vocabularies learned for the individual regions of level-1 in case-3 were not discriminative as compared to a single vocabulary learned from all the images of the training set in case-1. A relatively shorter vocabulary length of around 50 for case-2 has the highest performance on detecting the presence of a car. The performance on case-3 is higher for vocabulary length of 75. In either case, we observed better performances in relatively smaller vocabulary lengths. We also observed that the performance gets lower as we increase the vocabulary lengths. We feel that this phenomenon is due to the fact that, the larger vocabulary need not necessarily increase the detection performances for all the classes, especially where the features are more uniform in nature. That is, the features in an environment in which a car can be found is mostly constructed environment like road, near buildings.

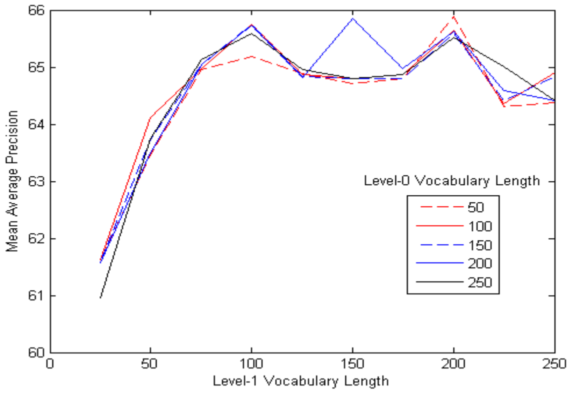

Figure. 4. The Mean Average Precision for various vocabulary lengths of level-1 fixing the vocabulary length of level-0 for detecting person for Case-2 (best viewed in color)

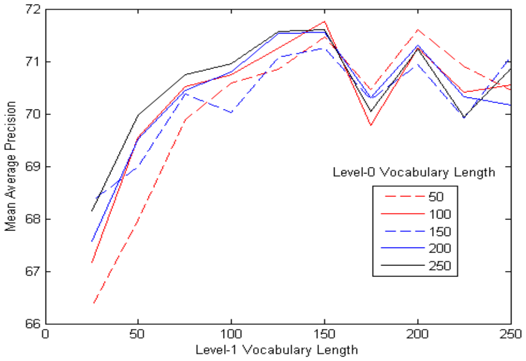

Figure. 5. The Mean Average Precision for various vocabulary lengths of level-1 fixing the vocabulary length of level-0 for detecting person for Case-4 (best viewed in color)

Figure. 4 shows the classification performances obtained for detecting the presence of persons in an image with one vocabulary per level (i.e. case-2). Figure. 5 shows the classification performances obtained for detecting the presence of persons in an image with one vocabulary per region (i.e. case-4). In both cases, the classification performance peaks when the vocabulary length of level-0 is 50 and the level-1 vocabulary length is 200. This implies that when the vocabulary length of level-1 is sufficiently large, the vocabulary for level-0 acts as a context for the words in level-1. Also for case-2, the classification performance for level-0 vocabulary length of 200 (blue continuous line) , dominates others when the level-1 vocabulary lengths are 100 and 200.

The plots for case-4 clearly depict the relationship between the vocabulary lengths of level-0 and 1. Here, the curves for smaller vocabulary lengths of level-0 dominate the later region of the plot and the longer vocabulary lengths of level 0 dominate the initial region. This clearly indicates that the vocabulary for level-0 compensates for the loss of global context due to the locally learned vocabularies for level-1. As we keep increasing the length of the vocabulary at level-1, the required length of level-0 vocabulary comes down where the vocabularies are learned for individual regions. However, the classification performance is better for relatively shorter length of vocabulary at level-0. These observations on case-2 and case-4 are made for detecting the presence of persons in a given image. We can also see that for the problem of detecting persons, case-4 yields better performance comparing to case-2 with shorter overall feature vector length.

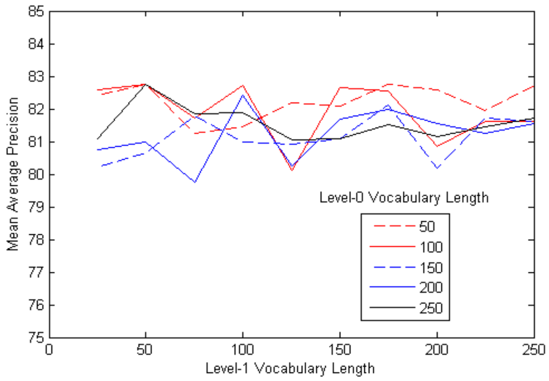

Figure. 6. The Mean Average Precision for various vocabulary lengths of level-1 fixing the vocabulary length of level-0 for detecting cars for Case-2 (best viewed in color)

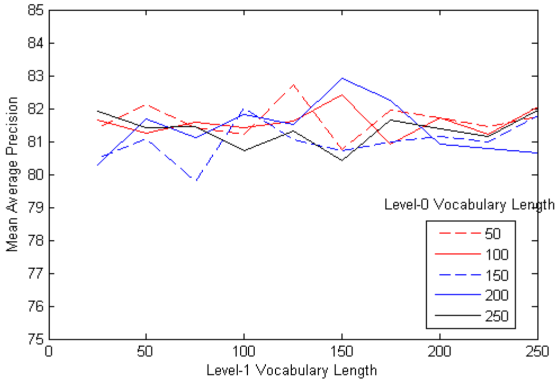

Figure. 7. The Mean Average Precision for various vocabulary lengths of level-1 fixing the vocabulary length of level-0 for detecting cars for Case-4 (best viewed in color)

Figure. 6 and 7 are the classification performances obtained for detecting the presence of cars in an image with one vocabulary for levels (case-2) and one vocabulary for regions (case-4) respectively. In contrast to the person detection problem, learning vocabularies for levels marginally performs better than learning vocabularies from regions for detecting the presence of cars in the image. In case-4 for detecting cars, the peak classification performances are attained for level-0 vocabulary length of 200 and level-1 vocabulary length of 150. Apart from this, for case 2 for detecting cars, the classification performances are higher and better than case-4 when the level-0 vocabulary length is 50 and 100 which are relatively shorter. We believe that the benefit of learning more spatially discriminating features by the way of learning vocabularies for regions separately is absent for detecting cars because the distribution of visual words in the images containing cars is more uniform across the image.

6. Conclusions and Future Work

We performed experiments to understand the role of vocabulary lengths of two levels and their impact on the final classification performance. The experiments for both classification problems suggest that an additional level to simple tiling improves the classification performances. When the classification problem concerned requires learning spatially discriminating features, the relatively shorter vocabulary length for level-0 is found to be more appropriate. For such cases, we believe that the level-0 vocabulary provides the context in which the visual words of level-1 could be found. Also, the choice between learning vocabulary for levels and for regions is category sensitive. We need to extend similar experiments over more classes and more levels in the spatial pyramid to understand better the relationships of the vocabulary lengths of successive levels to the level of information it provides to the classification. This relationship is category sensitive as we uncovered with our experiments. Better understanding of this phenomenon should help to optimize the lengths of feature vector and optimize the learning efforts.

REFERENCES

[1] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proc. of IEEE CVPR, volume 2, pages 2169–2178, 2006.

[2] Bosch, Anna, Andrew Zisserman, and Xavier Munoz. "Representing shape with a spatial pyramid kernel." Proceedings of the 6th ACM international conference on Image and video retrieval. ACM, 2007.

[3] Zhang, Chunjie, et al. "Image classification using spatial pyramid coding and visual word reweighting." Computer Vision–ACCV 2010. Springer Berlin Heidelberg, 2011. 239-249.

[4] Yang, Jun, et al. "Evaluating bag-of-visual-words representations in scene classification." Proceedings of the international workshop on Workshop on multimedia information retrieval. ACM, 2007.

[5] Squire, David McG, et al. "Content-based query of image databases: inspirations from text retrieval." Pattern Recognition Letters 21.13 (2000): 1193-1198.

[6] Sivic, Josef, and Andrew Zisserman. "Video Google: A text retrieval approach to object matching in videos." Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE, 2003.

[7] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[8] Fei-Fei, Li, and Pietro Perona. "A bayesian hierarchical model for learning natural scene categories." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 2. IEEE, 2005.

[9] Varma, Manik, and Andrew Zisserman. "Texture classification: Are filter banks necessary?." Computer vision and pattern recognition, 2003. Proceedings. 2003 IEEE computer society conference on. Vol. 2. IEEE, 2003.

[10] Lowe, David G. "Object recognition from local scale-invariant features." Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Vol. 2. Ieee, 1999.

[11] Szummer, Martin, and Rosalind W. Picard. "Indoor-outdoor image classification." Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on. IEEE, 1998.

[12] Bosch, Anna, Andrew Zisserman, and Xavier Munoz. "Scene classification via pLSA." Computer Vision–ECCV 2006. Springer Berlin Heidelberg, 2006. 517-530.

[13] Viitaniemi, Ville, and Jorma Laaksonen. "Spatial extensions to bag of visual words." Proceedings of the ACM International Conference on Image and Video Retrieval. ACM, 2009.

[14] Yang, Jun, et al. "Evaluating bag-of-visual-words representations in scene classification." Proceedings of the international workshop on Workshop on multimedia information retrieval. ACM, 2007.

[15] Zhang, Chunjie, et al. "Image Classification Using Spatial Pyramid Robust Sparse Coding." Pattern Recognition Letters (2013).

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results