Are Large Scale Training Images or Discriminative Features Important for Codebook Construction?

Veerapathirapillai Vinoharan[1] and Amirthalingam Ramanan[2]
[1]*Computer Centre, University of Jaffna, Jaffna, Sri Lanka*
[2]*Department of Computer Science, Faculty of Science, University of Jaffna, Jaffna, Sri Lanka*

Keywords: Bag-of-Features, Object Recognition, Incremental Learning, Discriminative Features, Visual Codebook.

Abstract: Advances in machine learning and image feature representations have led to great progress in pattern recognition approaches in recognising up to 1000 visual object categories. However, the human brain solves this problem effortlessly as it can recognise about 10000 to 100000 objects with a small number of examples. In recent years bag-of-features approach has proved to yield state-of-the-art performance in large scale evaluations. In such systems a visual codebook plays a crucial role. For constructing a codebook researchers cover a large-scale of training image set. But this brings up the issue of scalability. A large volume of training data becomes difficult to process whereas the high dimensional image representation could make many machine learning algorithms become inefficient or even a breakdown. In this work we investigate whether the dominant bag-of-features approach used in object recognition will continue significantly to improve with large training image set or not. We have validated a one-pass clustering algorithm to construct visual codebooks for object classification tasks on the PASCAL VOC Challenge image set. Our testing results show that adding more training images do not contribute significantly to increase the performance of classification but it increases the overall model complexity in terms of increased storage requirement and greater computational time. This study further suggests an alternative view to the community working with the patch-based object recognition to enforce retaining more discriminative descriptors rather than the reminiscent of the BIG data hypothesis.

1 INTRODUCTION

The bag-of-features approach (Csurka et al., 2004), (Karmakar et al., 2015) is a popular technique for representing image content. In such a system a visual codebook plays a crucial role. An important issue with the visual codebook representation is its discriminative power and dimensionality. Most of the visual codebooks that are used in larger evaluations consist of 10,000 codewords. This higher dimensionality curses the subsequent classifier training procedure. Thus, most of the object recognition systems expect the histogram representation of a bag-of-feature approach to be more compact while maintaining the discriminative power.

The long-term goal of computer vision in object recognition is to achieve near human levels of recognition. Changes in pose, lighting, occlusion, clutter, intra-class differences, inner-class variances, deformations, background that varies relative to the viewer, large number of images and several object categories make the problem of recognition highly challenging. Humans develop the invariance of an object so as to easily recognise different sized objects, orientation, illumination, and perspective objects. Whenever an object is seen, the human brain extracts the features without considering the size, orientation, illumination, perspective, and the object is remembered by its shape and inherent features (Kim, 2011). Moreover, an incremental learning method is adapted by the human visual processing system. When a new instance of an existing object category is seen the previous knowledge base is updated using new invariants. It has been proposed by (Ullman et al., 2002), the human visual system encoded features of intermediate complexity that are class-specific is selected for encoding images within a class of related images. The popular approach in artificial visual object recognition is to use local information extracted at several points or patches in the image. In such a system the construction of a visual codebook is often performed from thousands of images and each image averagely contains hundreds or even one thousand patch-based interest points described in a higher dimensional space of at least one thousand codewords, in order to capture sufficient information for efficient

Vinoharan, V. and Ramanan, A.
Are Large Scale Training Images or Discriminative Features Important for Codebook Construction?.
DOI: 10.5220/0005676201930198
In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016)*, pages 193-198
ISBN: 978-989-758-173-1

classification. A major bottleneck lies in handling such massive scale of datasets.

The PASCAL VOC challenge imageset (Everingham et al., 2010) has become a benchmark dataset in many computer vision tasks. In such challenges, participants normally request to increase the number of training images to train their model in a better way in order to achieve higher recognition rate. It can be observed that there is a steady increase in the training set of this image set over the years 2007 to 2012 consisting nearly 2000 to 6000 images amidst the fact that the number of object categories remains the same as 20. This is an important issue that we are focusing on this paper whether such object recognition system will continue to improve with increasing large number of training images for achieving slightly increased classification rate or is it worth to focus on the selection of discriminative features and the development of better object models.

We optimise the process of constructing codebooks with less memory requirement and speeding up the approach while maintaining compactness and discriminative power in recognition. Our technique constructs a codebook by acquiring information about objects in a sequential way. The strategy that we use to design discriminant codebook is by updating an initially constructed codebook over sequentially arriving training images, and the output classifier accounts for the class-specific discriminant features. At the arrival of each of the training images belonging to an interest or non-interest object category, only the novel information in the codebook will be absorbed as additional entries. The construction of a codebook in this context is achieved by extending the resource allocating codebook (RAC) technique proposed by (Ramanan and Niranjan, 2010). The proposed approach in this paper constructs a codebook for a large-scale object recognition task without the favour of machines that have become fast enough in constructing a codebook on relatively large-scale descriptors.

The rest of this paper is structured as follows. Section 2 briefly describes the background needed for our work. Section 3 briefly describes the objectives of our research. Section 4 summarises different methods used to construct a codebook with compactness and discriminative power that have been carried out in recent years. Section 5 explains the proposed methodology in achieving incremental learning method and constructing codebook with compactness and discriminative power. Section 6 describes the experimental setup and testing results which support our claim. Finally, section 7 concludes this paper.

2 BACKGROUND

2.1 Bag-of-Feature Approach

The bag-of-features (BoF) approach is widely used in image scene classification and object recognition tasks in computer vision (Ramanan and Niranjan, 2011). The pseudocode of bag-of-features approach is given in Algorithm 1

Algorithm 1: BoF representation for images.

```
for all image do
    interestPts ← detectPts(image)
    descriptors ← describePts(interestPts)
end for
codebook ← quantizePts(descriptors(trainingImages))
for all image do
    BOF ← computeHist(codebook, descriptors(image))
end for
```

In such approach, visual codebooks are created as follows. After extracting a large number of local patch descriptors (e.g., SIFT descriptors (Lowe, 2004) from a set of training images, a clustering method (e.g., K-means) is often used to group these descriptors into K clusters, where K is a predefined parameter. The center of each cluster is called the "visual word" or "codeword", and a set of codewords forms a codebook. Each image descriptor is then labeled with the most similar codeword according to the Euclidean distance, and the image is characterised by a K-dimensional histogram of the number of occurrences of each codeword. In fact, the size and effectiveness of the codebook has a critical impact on recognition performance.

2.2 Resource-Allocating Codebook

The Resource-Allocating Codebook (RAC) (Ramanan and Niranjan, 2010) is a simple and extremely fast way to construct a codebook by using a one-pass process, which simultaneously achieves increased discrimination and a drastic reduction in the computational needs.

RAC starts by arbitrarily assigning the first data item as an entry in the codebook. When a subsequent data item is processed, its minimum distance to all entries in the current codebook is computed using an appropriate distance metric. If this distance is smaller than the predefined threshold r (radius of the hypersphere), the current codebook is retained and no action is taken with respect to the processed data item. If the smallest distance to codewords exceeds the threshold, including the current data item as the

additional entry, creates a new entry in the codebook. This process is continued until all data items are seen only once.

2.3 Support Vector Machine (SVM)

SVM is a well-known statistical learning method (Cortes and Vapnik, 1995). In particular, it is effective when the training data consists of a small number of samples in high-dimensional spaces. The objective of SVM learning is to find a hyperplane that maximises the inter-class margin of the training samples. Feature vectors are projected into a high-dimensional space by the kernel function.

3 OBJECTIVES

Discriminative power and compactness of a codebook are important to control the model complexity. In this regard we formulate the following:

- Not all training images contribute to the discriminative power of a codebook. That is, a unique or different image of the same class will contribute to the construction of a codebook.
- The incremental construction of a codebook using the training images is more appropriate to retain discriminative features similar to the human visual perceptual system than processing all the required images.

4 RELATED WORK

There is an extensive body of literature in the area of visual object recognition systems. (Zhu et al., 2012) have investigated the question of whether existing feature detectors will continue to improve as data grows or the development of better object detection models is needed. The authors have found that additional data does help, but only with correct regularisations and treatment of noisy examples in the training data and compositional mixtures (implemented via composed part) that give a much better performance in recognition. However, a straightforward but effective approach lies in the use of a codebook model. A compact codebook can also be achieved by carefully selecting the codewords from an initially constructed large codebook (Kirishanthy and Ramanan, 2015).

(Yang et al., 2008) have proposed a unified codebook generation that is integrated with classifier training. Unlike clustering approaches that associate each image feature with a single codeword in their approach, images are represented by means of visual bits associated with different categories, i.e., an image which can contain objects from multiple categories is represented using aggregates of visual bits for each category that constitutes the semantic vocabulary. These visual bits are augmented iteratively to refine visual words based on the learning performance of the classifier. The iterative process is carried out until a desired performance is achieved. Harris Laplace corner detectors are used in detecting interest points and are described by SIFT descriptors. The proposed framework is mainly evaluated on the PASCAL VOC Challenge 2006 dataset which contains 10 visual challenges. Their training set consists of 100 randomly-selected images. Their framework outperforms the baseline K-means and SVM on every category and demonstrates significant improvements over extremely-random classification forest algorithm on 8 out of 10 classes.

(Li et al., 2008) have proposed an approach for learning optimal compact codebook by selecting a subset of discriminative codeword from a large codebook. An initial codebook was constructed using K-means clustering algorithm. Each codeword in this codebook is then modeled by a spherical Gaussian function through which an intermediate representation for each training image is obtained. A Gaussian model for every object category is learned based on this intermediate representation. Following this step, an optimal codebook is constructed by selecting discriminant codewords according to the learned Gaussian model. The discriminative capability is measured either by likelihood ratio or by Fisher score. Interest points in their experiments were detected by the DoG detector and are described by SIFT descriptors. Classification is performed using SVM classifiers with RBF kernel. The proposed framework is mainly evaluated on the Caltech-4 dataset consisting four object categories. In their experiment 100 images were randomly choosen for training and the rest was used for testing. They report superior performance of object categorisation compared with traditional K-means method with same size of codebooks.

(Winn et al., 2005) have proposed to optimise codebooks by hierarchically merging visual words in a pair-wise manner using the information bottleneck principle from an initially constructed large codebook. The final visual words were represented by the Gaussian Mixture Models of pixel appearance. Training images are convolved with different filter-banks made of Gaussians and Gabor kernels to generate a set of filter responses. The resulting filter responses are clustered by K-means method with a large value of

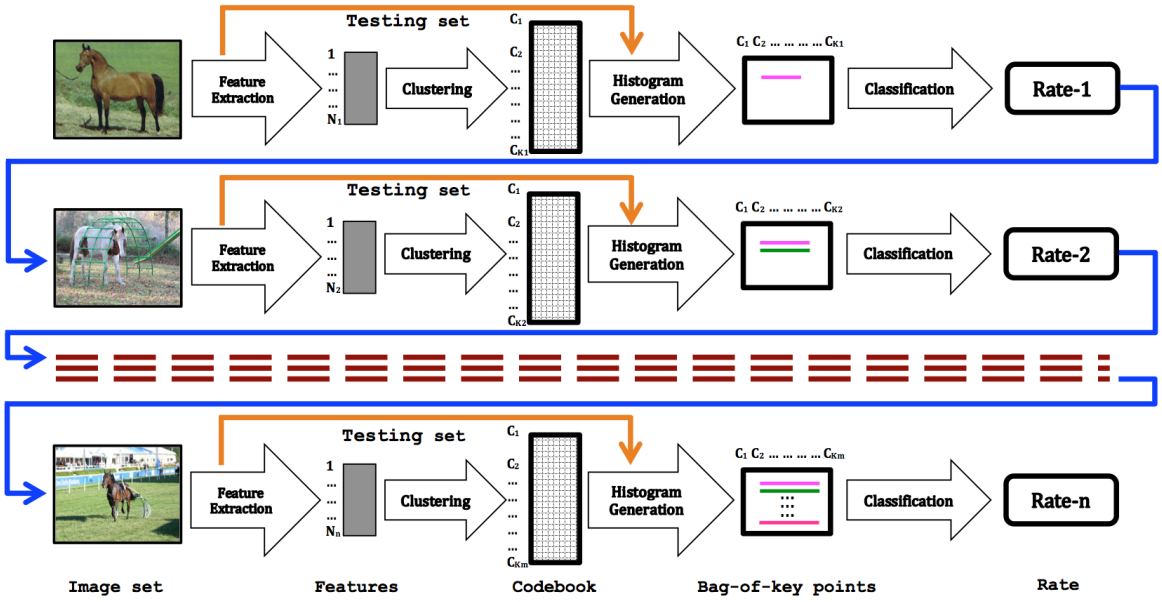

Figure 1: The overall framework of the proposed technique to sequentially constructing visual codebook for an object-specific category (e.g. horse).

K in the order of thousands. The learned cluster centre and their associated covariance define a universal visual codebook. Following the construction of this large codebook, a histogram is constructed over the initial codebook which is processed by each region of training images. A mapping between corresponding histogram and pair-wise merging operation is used to produce a much more compact visual codebook. The proposed framework is mainly evaluated on the in-house database with faces from Caltech dataset. Class models are learned from a set of 240 manually segmented and annotated images belonging to the nine object categories. In order to measure the classification accuracy the image set is split into 50% training and 50% test sets. Their framework yields an accuracy of 93%.

5 METHODOLOGY

The proposed sequential learning method to constructing codebook is extremely fast and efficient when compared to the approaches. The proposed technique shows a better way to construct a compact codebook while maintaining its discriminative power. The following steps describe the process of sequentially constructing a codebook:

Step 1: A randomly chosen image from a training set is processed to extract the features and then cluster those features using RAC technique in a one-pass manner to construct an initial codebook. Based on this codebook, images are then represented as histograms for all training and testing image sets. Thereafter, the classification rate is computed using a standard classifier.

Step 2: The subsequent image in the training set is processed to extract features. If the smallest distance to a codeword in the codebook obtained in step 1 exceeds the radius of RAC, the current feature is recorded as an additional informative codeword that creates a new entry in the codebook by updating the obtained codebook in the step 1; otherwise no action is taken with respect to the processed feature. Based on the updated codebook images of training and testing sets are represented as histograms. The classification rate is computed using the same standard classifier. This process will be continued until all images have been considered.

Figure 1 shows the overall framework of the proposed technique to sequentially construct a visual codebook for an object-specific category and the pseudocode of this approach is given in Algorithm 2. The stopping criteria for Algorithm 2 can be implemented either by processing sequentially all the images in the training set or when achieving a desired classification rate.

Algorithm 2: Sequentially constructing codebook.

Input: Training images (trImgs), Testing images (teImgs)
Output: Visual codebook (CB), Classification accuracy (rate)
for all $img_i \in \{trImgs, teImgs\}$ **do**
  interestPts ← detectPts(img_i)
  descrips ← describePts(interestPts)
end for
r ← predefined value
// Initialise the codebook CB
D ← descrips(img_1) //where $img_1 \in trImgs$
CB ← D_1
i ← 1
for all $img_i \in trImgs$ **do**
  D ← descrips(img_i)
  j ← 1
  while $(j \leq size(D))$ **do**
    if $min\| D_j - CB \|^2 > r^2$ **then**
      Create a new hypersphere of r such that,
      $CB \leftarrow \{CB \cup D_j\}$
    end if
    j ← j + 1
  end while
  trainHist ←computeHist(CB, descrips(trImgs))
  testHist ←computeHist(CB, descrips(teImgs))
  $rate_i$ ← classify(trainHist, testHist)
  i ← i + 1
end for

6 TESTING RESULTS

We test our approach on PASCAL VOC 2007 Challenge dataset. It consists of 9963 images from 20 categories. For constructing visual codebooks, SIFT features were clustered independently using K-means with K = 250 (2K = 500) and RAC with r = 0.89. Experiments in this work were mainly carried out to validate our objectives. Table 1 shows the performance comparison of BoF approach with K-means and RAC whereas Table 2 details the classification rate of RAC with the proposed sequential learning technique.

Based on the statistical t-tests performed on the results of Table 1 and 2, we conclude that RAC and K-means are of near performance, whereas the proposed sequential learning technique outperforms the RAC method, at the level of significance 0.05. Moreover, on average about 13-22% of the training images provided in PASCAL VOC 2007 is only needed to construct a discriminative codebook for each binary classification tasks listed in Table 2. This proves that not all images are required for constructing a discriminative codebook in a similar way as the human visual processing system. Furthermore, the proposed technique constructs a compact codebook which is around 60% size of the codebooks constructed either by K-means or RAC method.

Limited experimentation with reordering of the images was carried out to check the evolution of codebook size and the classification rates during the exe-

Table 1: Comparison of two codebook generation methods tested on a selected binary classification tasks from the PASCAL VOC 2007 dataset. K-means (KM) with K = 250 (2K = 500) and RAC codebook sizes nearly equal to 500. The r of RAC was 0.89.

Object	KM+SVM	RAC+SVM
Aeroplane vs Bird	**0.84**	0.83
Aeroplane vs Boat	0.77	**0.80**
Aeroplane vs Horse	**0.88**	0.87
Aeroplane vs Sofa	**0.88**	0.87
Bicycle vs Motorbike	**0.69**	0.67
Bird vs Cat	**0.77**	0.73
Boat vs Bus	**0.86**	0.82
Boat vs TVmonitor	**0.91**	0.88
Bottle vs Pottedplant	**0.68**	0.64
Bus vs Train	**0.71**	0.70
Cat vs Dog	0.64	**0.65**
Chair vs Dog	0.**83**	0.81
Cow vs Sheep	0.**66**	0.63
Diningtable vs Pottedplant	0.58	**0.61**
Pottedplant vs TVmonitor	0.68	**0.68**
Train vs TVmonitor	**0.89**	0.85

Table 2: Classification rate with codebook (CB) size and number of training images for the proposed sequential learning method with r = 0.89.

Object	#imgs avail	#imgs used	CB size	Ours
Aeroplane vs Bird	568	34	279	**0.87**
Aeroplane vs Boat	419	22	236	**0.80**
Aeroplane vs Horse	525	58	341	**0.90**
Aeroplane vs Sofa	467	90	356	**0.88**
Bicycle vs Motorbike	488	22	274	**0.68**
Bird vs Cat	667	72	315	**0.75**
Boat vs Bus	367	46	303	**0.84**
Boat vs TVmonitor	437	14	165	**0.89**
Bottle vs Pottedplant	489	110	348	**0.65**
Bus vs Train	447	86	392	**0.71**
Cat vs Dog	758	30	247	**0.65**
Chair vs Dog	866	32	255	**0.81**
Cow vs Sheep	237	42	233	**0.65**
Diningtable vs Pottedplant	445	34	263	**0.64**
Pottedplant vs TVmonitor	501	132	373	**0.70**
Train vs TVmonitor	517	46	293	**0.86**

cution. We report the mean classification rates of ten independent runs where each run is carried out by fixing the same total number of images considered in Table 2 and by randomly shuffling the order of presence of the images in the process of constructing a codebook. In the Aeroplane vs Horse example the average size of codebook was 338 ± 17 with a classification rate of 0.88 ± 0.01, whereas for the Diningtable vs Pottedplant example the size of the codebook and classification rate were 259 ± 25 and 0.61 ± 0.02, respectively. While we have included the standard deviation for completeness, we noted that these are the estimates of uncertainty for a very few trials. The construction of a codebook using K-means algorithm was performed in an average time of 16536 seconds, while the proposed method required an average time of 42 seconds only on a desktop computer with an Intel Core i5 running at 3.2GHz and 8GB of RAM.

7 DISCUSSION AND CONCLUSION

This paper addresses the problem of object classification of images together with a sequential learning technique. Our system starts to progress in extracting features from the training images using SIFT algorithm. These features are converted into a codebook using an extended RAC method. The codewords then serve to construct a histogram for representing an image. These histograms are then fed into a binary SVM classifier to classify the objects. We construct the codebook by sequentially processing images to retain only the discriminative or rare features by allocating new codewords using the extended RAC technique. Our test results show that it is worth to select discriminative features, instead of increasing the number of training images, to yield better classification rate by means of a compact codebook.

In the literature of BoF approach, the codeword size is manually selected by the user and is commonly defined up to tens of thousands for ensuring enough information encoding. However, such a huge size of codewords causes an enormous computational cost. To create a discriminative BoF representation, we present a technique that well approximates the distribution of visual words in an image and the output classifier accounting for class-specific discriminant features. Thus, this paper suggests an alternative view to the research community working with the patch-based object recognition to emphasize the retaining of more discriminative descriptors rather than the reminiscent of the BIG data hypothesis.

REFERENCES

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision*, volume 1, pages 1–2.

Everingham, M., Eslami, S. M. A., Gool, L. V., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The PASCAL Visual Object Classes VOC Challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338.

Karmakar, P., Teng, S. W., Lu, G., and Zhang, D. (2015). Rotation invariant spatial pyramid matching for image classification. In *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 653–660.

Kim, S. (2011). Robust object categorization and segmentation motivated by visual contexts in the human visual system. *EURASIP Journal on Advances in Signal Processing*.

Kirishanthy, T. and Ramanan, A. (2015). Creating compact and discriminative visual vocabularies using visual bits. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 258–263.

Li, T., Mei, T., and Kweon, I. S. (2008). Learning optimal compact codebook for efficient object categorization. In *IEEE Workshop on Applications of Computer Vision*, pages 1–6.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Ramanan, A. and Niranjan, M. (2010). A one-pass resource-allocating codebook for patch-based visual object recognition. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 35–40.

Ramanan, A. and Niranjan, M. (2011). A review of codebook models in patch-based visual object recognition. *Journal of Signal Processing Systems, Springer*, 68(3):333–352.

Ullman, S., Vidal-Naquet, M., and sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature neuroscience*, 5(7):682–687.

Winn, J., Criminisi, A., and Minka, T. (2005). Object categorization by learned universal visual dictionary. In *IEEE International Conference on Computer Vision*, volume 2, pages 1800–1807.

Yang, L., Jin, R., Sukthankar, H., and Jurie, F. (2008). Unifying discriminative visual codebook generation with classifier training for object category recognition. In *proceeding of IEEE conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pages 1–8.

Zhu, X., Vondrick, C., Ramanan, D., and Fowlkes, C. (2012). Do we need more training data or better models for object detection? In *British Machine Vision Conference (BMVC)*.