One-pass Keypoint Selection to Construct Codebook for Patch-based Object Classification

Veerapathirapillai Vinoharan
Computer Unit
University of Jaffna
Jaffna, Sri Lanka
vvinoharan@univ.jfn.ac.lk

Amirthalingam Ramanan
Department of Computer Science
University of Jaffna
Jaffna, Sri Lanka
a.ramanan@univ.jfn.ac.lk

Abstract—In a patch-based object classification system, one of the most popular image representation approach is the bag-of-features (BoF) representation. However, the number of patch-based features extracted from images to create the BoF vector is usually very large, which causes two problems: (i) Increased amount of computational needs during the vector quantisation step, and (ii) Ambiguous descriptors from training images increase false positive rate in the classification. To overcome these issues we introduce a one-pass feature selection approach followed by an entropy-based filtering technique to eliminate the ambiguous features from initial large feature set. In this work, a discriminative BoF representation for object recognition is constructed using patch-based descriptors that are informative in distinguishing object categories. Following the construction of a codebook a subset of codewords which is not activated enough in images is eliminated based on statistical measures and visual-bit representation of codewords. The proposed technique is evaluated on (i) Xerox7, (ii) UIUCTex, (iii) PASCAL VOC 2007, and (iv) Caltech101 image datasets. The proposed feature selection step increases the discriminant power of a codebook, while the codeword selection method maintains the codebook to be more compact. The proposed framework would help to optimise the BoF representation to be effective with stable performance.

Index Terms—Bag-of-features, Keypoint selection, Codeword selection, Image representation.

I. INTRODUCTION

Object classification is a process of predicting the presence or absence of a specific object in a digital image or video sequences. There are lot of applications available based on object classification such as pose estimation [9], describing photos [16], driver assistance system [11], intruder detection system [12], etc. Moreover the object classification tasks face lot of challenges owing to different lighting conditions, background, clutter, occlusion, poses, scales, and variety of image sizes. Over the last decade the object classification with the use of bag-of-features (BoF) approach has shown promising results in large scale evaluation such as PASCAL VOC Challenge [7], Caltech256 [5], and ImageNet [17].

The BoF representation of an image conveys the presence or absence of the information of each codeword in the image with respect to a visual codebook constructed using a set of descriptors extracted from training images. The mid-level representations of BoF approach [15], [18]–[20] computed from dense sampled local descriptors of an image have shown state-of-the-art performances in visual object classification. In such BoF approach, the codebook plays an important role in determining the discriminative power and the compactness of the feature vector representation. The discriminative power of a codebook determines the quality of the codebook model, whereas the size of a codebook controls the complexity of the model. We optimize the process of building codebooks with less memory requirement and accelerate the approach while maintaining the discriminative power and compactness in the classification of visual objects.

The main contribution of this paper is summarised below:

- A feature selection technique referred to as one-pass feature selection (OPFS) which is followed by an entropy-based feature selection (EBFS) approach is proposed to filter out ambiguous descriptors from initially extracted large descriptors set.
- Two distinct codeword selection algorithms are presented: (i) Statistical-based measures to compute the inter-category, intra-category, and combined confidences, and (ii) Visual-bit representation of codewords to select informative codewords from an initially constructed large codebook.

In particular, we show that retaining too many similar features per image category and less informative codewords per codebook not only increases the computational cost to generate the BoF vector, but also degrades the classification rate. The proposed method provides an effective way to improve the object categorisation performance when using the BoF model with low-dimensional representation.

The rest of this paper is organised as follows. Section II briefly describes the background needed for our work. Section III summarises the related work that has been used to construct a discriminative and compact codebook for object recognition. Section IV explains the proposed methodology in detail. Section V describes the experimental setup followed by testing results. Finally, section VI concludes this paper with discussion.

978-1-5386-9418-3/18/$31.00 ©2018 IEEE
II. BACKGROUND

In this section, the standard BoF approach is summarised.

A. Bag-of-Features (BoF) Approach

The BoF approach is widely used in image scene classification [10] and object classification tasks [21]. The construction of the BoF vector \( h \) from an image \( \mathbf{I} \) can be summarised in four steps: (i) Descriptors are automatically detected in \( \mathbf{I} \) using local invariant detector, (ii) local descriptors pooled from training set are computed over those region, (iii) all the descriptors are quantised into codewords, and (iv) counted to build the BoF representation of the image \( \mathbf{I} \).

In such BoF representation, the construction of codebook plays a crucial role that affects the model's complexity.

B. Scale-Invariant Feature Transform (SIFT)

SIFT [1] is one of the most widely used patch-based descriptor in object classification. It is a technique that extracts distinctive features from gray-level images, by filtering images at various scales and patches of interest that have sharp intensity changes. SIFT descriptors is a 128 dimensional vector that can be used in the context of recognition and matching of the same scene or object observed under different viewing conditions.

C. K-means Clustering

K-means is one of the popular and simplest clustering technique which partitions \( N \) data points into \( K \) disjoint sets by minimizing the sum of squared distance between all data points. The time complexity of the traditional K-means method is \( O(Ndkm) \), where \( N \) is the number of data points, \( d \) is the dimension of data, \( K \) is the number of centroids, and \( m \) is the number of iterations.

D. Support Vector Machine (SVM)

SVM is a popular statistical learning method that has proven to be successful for several years in image classification tasks. The objective of SVM learning is to find a hyperplane that maximises the inter-class margin of the training data. The data in the input space are projected into a high-dimensional feature space by a kernel function.

III. RELATED WORK

In the literature, several techniques have been proposed on the BoF approach for object recognition, which focused on both discriminative power and compactness of codebooks.

In [14], the authors have proposed a codeword selection method based on BoF representation for pedestrian detection problem to reduce the intensive computational requirement at the classifier. In the first stage, the difference in the total appearance frequency for each visual word of the pedestrian and non-pedestrian images are calculated. The visual codebook that exhibits greater absolute values are considered to be more efficient for pedestrian detection, and are selected. Experiments were used on publicly available Daimler-DB dataset [6]. Randomly selected 3000 pedestrian and 3000 non-pedestrian images are used as the training samples. For the test samples, 3000 pedestrian and 3000 non-pedestrian images are selected from the remainder of the dataset. The dense-SIFT feature descriptors extracted from all training samples are used to construct a codebook using K-means clustering with \( K = 500 \). SVMs were then trained using RBF kernels. In their experiments, 200 efficient visual words in the original codebook have resulted in almost the same performance as with all 500 visual words. Authors’ results show that the miss rate was reduced by 3% than the previous method.

IV. METHODOLOGY

In this work SIFT descriptors were extracted from training images. Before constructing a relatively large initial codebook, unambiguous descriptors are selected using a one-pass feature selection (OPFS) method which is then followed by an entropy-based feature selection (EBFS) method to increase the discriminative power of the codebook. A codebook is informative keypoints using K-means algorithm. Experiments using the Caltech101, Caltech256 and PASCAL VOC 2007 datasets demonstrate that using keypoint selection to generate both BoF and spatial pyramid matching allows the SVM classifier to produce better classification results compared to the previous techniques that are used without the keypoint selection method.
then constructed by means of K-means approach. Finally, indistinctive codewords are eliminated based on statistical measures or visual-bit representation of codeword to obtain a compact codebook. We assign each keypoint in images to the closest codeword and create a histogram representation for each image, which records how many times keypoints corresponding to the codeword occur in the image. We then apply linear SVM classification algorithm to those fixed-length feature vectors. The overall framework of the proposed method is illustrated in Figure 1.

A. Feature selection

Set of keypoints are extracted and described using SIFT algorithm from an image that can be transformed to an intermediate representation. In such set, some of the detected descriptors make the intermediate representation more distinguishable. Selecting these important descriptors increase the discriminative power of the codebook. We propose: (i) A one-pass feature selection approach, and (ii) An entropy-based feature selection approach to eliminate ambiguous SIFT descriptors in order to retain high-quality descriptors. These approaches reduce the computational complexity of the clustering and increase the categorisation precision at the later stage of the BoF representation.

1) OPFS approach: Is a simple and extremely fast technique to select informative keypoints that carves the feature space as fixed-size hyperspheres. OPFS selects unambiguous descriptors to retain high-quality descriptors. The one-pass feature selection algorithm used in this work is summarised in Algorithm 1.

2) OPFS followed by EBFS approach: The descriptors selected using OPFS technique are those with rich internal structure and reduce the false positive rate by orders of magnitude. This process is summarised in Algorithm 2.

B. Codeword selection

A codebook is usually constructed by using a clustering algorithm. In such a codebook, the goal of codeword selection is to remove the redundancy and noise in a codebook. Elimination of indistinctive codewords not only reduces the overall computational complexity but also increases the categorisation precision. A compact codebook has advantages in terms of both computing efficiency and storage requirement. In this work, we compare two types of codeword selection techniques:

Algorithm 1: One-pass feature selection (OPFS)

Input: Training images (trnImgs)

Output: Selected descriptors set (selDesc)

for all \( img_i \in trnImgs \) do
    interestPts \( \leftarrow \) detectPts\((img_i)\)
    descrips\{i\} \( \leftarrow \) describePts(interestPts)
end for

\( r \leftarrow \) predefined radius,
\( K \leftarrow \) #clusters
// Initialise selected feature set
selDesc \( \leftarrow \) descrips\{1\}
i \( \leftarrow \) 2

for all \( desc_i \in descrips \) do
    if \( min\|desc_i - selDesc\|^2 > r^2 \) then
        Create a new hypersphere of \( r \) such that,
        \( selDesc \leftarrow \{selDesc \cup desc_i\} \)
    end if
    i \( \leftarrow i + 1 \)
end for

return selDesc
Algorithm 2: OPFS followed by EBFS

Input: Training images (trnImgs)
Output: Selected features (selFts)

for all $img_i \in trnImgs$ do
    interestPts ← detectPts($img_i$)
    descrips(i) ← describePts(interestPts)
end for

features ← OPFS(descrips)
selFts ← []
thresh ← predefined value
i ← 1
K ← #clusters

for all $fts_i \in features$ do
    entVal ← 0
    j ← 0
    for $j \leq 255$ do
        entVal ← entVal + $p_j(fts_i) \times \log_2 p_j(fts_i)$
        where, $p_j(fts_i) \leftarrow \frac{|k|_{|k=j}}{128}$, $k \leftarrow 0, 1, 2, \ldots, 255$
        j ← j + 1
    end for
    if $entVal \leq thresh$ then
        selFts ← $\{selFts \cup fts_i\}$
    end if
    i ← i + 1
end for

return selFts

(i) Statistical measures, and (ii) Visual-bit representation of codewords.

1) Statistical measures: In-distinctive codewords are eliminated based on categorical confidence measures. Categories having similar histogram distribution may increase ambiguity of the categorisation system and a high variance histogram value of a codeword interrupts categorisation process, i.e., it makes the classifier (e.g., SVM) difficult to classify categories. The inter-category, intra-category and combined confidences of the categorisation system and a high variance histogram value of a codeword interrupts categorisation process, i.e., it makes the classifier (e.g., SVM) difficult to classify categories. The inter-category, intra-category and combined confidences of the categorisation system and a high variance histogram value of a codeword interrupts categorisation process, i.e., it makes the classifier (e.g., SVM) difficult to classify categories. The inter-category, intra-category and combined confidences of the categorisation system and a high variance histogram value of a codeword interrupts categorisation process, i.e., it makes the classifier (e.g., SVM) difficult to classify categories.

2) Visual bit representation of codewords: The best subset of codewords is selected by eliminating inconsistent codewords within each category based on the statistics of the codewords in the initial codebook. Categories having sparse histogram distribution may increase ambiguity of the categorisation performance. The selected codeword subset enlarges the distribution difference. The size of a codebook is reduced by means of visual-bit representations of images as described in [13]. This technique helps to optimise larger codebooks to fewer codewords with stable performance.

V. TESTING RESULTS


A. Dataset

1) Xerox7: It contains 1776 images from seven categories with different resolutions. The object poses are highly variable and there is a significant amount of background clutter, some of which belongs to the other categories making the classification task fairly challenging.

2) UIUCTex: It contains twenty five texture classes with 40 images per class. It has surfaces whose texture is mainly due to albedo variations, 3D shape, as well as a mixture of both. It also has significant viewpoint changes, uncontrolled illumination, arbitrary rotations, and scale differences within each class.

3) PASCAL VOC 2007: It is widely used in large scale evaluation of visual object categorisation task. The dataset consists a total of 9,963 images containing 24,640 annotated objects split into training, validation, and testing sets labelled with twenty object classes.

4) Caltech101: It consists of a total of 9,146 images, split between 101 different object categories, as well as an additional background/clutter category. Each object category contains between 31 and 800 images.

B. Experimental Setup

For the image sets: Xerox7, UIUCTex, and Caltech101 we used 70% for training and 30% for testing from each class. The classification for PASCAL VOC 2007 was performed on each of the twenty classes by training the classifiers on the provided ‘trainval’ set and evaluating on the testing set. We used SIFT descriptors in extracting the features from those image sets. The visual codebook is then constructed by clustering the descriptors that were extracted from the training images using the K-means algorithm with $K = 500$ for all datasets. The linear OVA-SVMs in classification were considered and the reported classification rates are of mean average precision (mAP) [8].

C. Testing Results

1) OPFS: On average about 5%, 7%, 14%, and 7% of training descriptors were found by OPFS with radius $r = 0.65$ for Xerox7, UIUCTex, PASCAL VOC 2007, and $r = 0.70$ for Caltech101 sets that are selected from the initially extracted descriptor set in order to construct codebook for each classification tasks. This selection of reduced number of keypoints enhances the discriminative power of the codebook. We compare the proposed feature selection technique with the traditional BoF approach. It has been noted that the OPFS technique selected around 10% of the descriptors that outperformed traditional BoF approach in all datasets. The performance comparison of BoF approach prior to applying OPFS vs after applying the technique is presented in Table I.
### Table I: Comparison of Mean Average Precision (mAP) with Number of Training Features and Codebook Size (CB): Traditional BoF Approach and Proposed Feature Selection Method without and with Post-processing (i.e., Codeword Selection [CS])

<table>
<thead>
<tr>
<th>Approach</th>
<th>Dataset</th>
<th>#Descriptors</th>
<th>Without CS</th>
<th>CS using Statistical Measures</th>
<th>CS using Visual bits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>inter CB mAP mAP</td>
<td>CB mAP</td>
</tr>
<tr>
<td>Traditional</td>
<td>Xerox7</td>
<td>4,046,578</td>
<td>987</td>
<td>67.63</td>
<td>803</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>70.28</td>
<td>67.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>172,006</td>
<td>500</td>
<td>71.14</td>
</tr>
<tr>
<td>OPFS</td>
<td></td>
<td>212,294</td>
<td></td>
<td></td>
<td>375</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>409</td>
</tr>
<tr>
<td>OPFS+EBFS</td>
<td></td>
<td>172,006</td>
<td></td>
<td></td>
<td>375</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>72.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>406</td>
</tr>
<tr>
<td></td>
<td>UIUCTex</td>
<td>4,543,590</td>
<td>1032</td>
<td>93.40</td>
<td>835</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>96.42</td>
<td>93.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>157,094</td>
<td>500</td>
<td>97.12</td>
</tr>
<tr>
<td>OPFS</td>
<td></td>
<td>314,724</td>
<td></td>
<td></td>
<td>400</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>97.53</td>
</tr>
<tr>
<td>OPFS+EBFS</td>
<td></td>
<td>157,094</td>
<td></td>
<td></td>
<td>400</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>97.12</td>
</tr>
<tr>
<td></td>
<td>PASCAL VOC 2007</td>
<td>1,760,400</td>
<td>1049</td>
<td>67.60</td>
<td>847</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>68.78</td>
<td>68.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>181,248</td>
<td>500</td>
<td>69.58</td>
</tr>
<tr>
<td>OPFS</td>
<td></td>
<td>245,327</td>
<td></td>
<td></td>
<td>400</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>69.50</td>
</tr>
<tr>
<td>OPFS+EBFS</td>
<td></td>
<td>181,248</td>
<td></td>
<td></td>
<td>400</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>69.35</td>
</tr>
<tr>
<td></td>
<td>Caltech101</td>
<td>5,659,137</td>
<td>925</td>
<td>77.52</td>
<td>742</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>79.74</td>
<td>75.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>393,024</td>
<td>500</td>
<td>79.92</td>
</tr>
<tr>
<td>OPFS</td>
<td></td>
<td>393,024</td>
<td></td>
<td></td>
<td>400</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>79.68</td>
</tr>
<tr>
<td>OPFS+EBFS</td>
<td></td>
<td>286,925</td>
<td></td>
<td></td>
<td>400</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>79.11</td>
</tr>
</tbody>
</table>

2) **OPFS followed by EBFS**: The reduced set of keypoints obtained by OPFS in training images are further reduced by eliminating keypoints having entropy value $E(F) \leq 4.0$ for Xerox7, PASCAL VOC 2007, Caltech101, and $E(F) \leq 4.5$ for UIUCTex datasets. This process enhances further the discriminative power when constructing codebook. We compare the proposed feature selection technique with the traditional BoF approach. It has been noted that the filtering technique selected around 6% of the descriptors outperformed the traditional BoF approach in all datasets. The performance comparison of BoF approach prior to applying OPFS followed by EBFS vs after applying the technique is presented in Table I.

3) **Codeword selection using statistical measures**: On average 80% of the codewords were selected using inter-category, intra-category, and combined confidence with $\alpha = 0.4$ and $\beta = 0.6$ for the initially constructed large codebook without the use of our proposed post-processing technique. In Table I, we compare the proposed post-processing technique: codeword selection using statistical measures, with and without the use of preprocessing techniques. This process achieves on average 75% and 60% of reduction in the initially constructed codebook while maintaining comparable performance with the traditional approach, respectively.

4) **Codeword selection using visual-bit representation**: On average 40% of the codewords were selected using the visual-bit representation of codeword approach (with $\lambda = 4$) from the initially constructed large codebook without the use of our proposed post-processing technique. In Table I, we compare the proposed post-processing technique, codeword selection using visual-bit representation of codewords method, with and without the use of preprocessing techniques. This process achieves on average 75% and 60% of reduction in the initially constructed codebook while maintaining comparable performance with the traditional approach, respectively.

VI. DISCUSSION AND CONCLUSION

BoF approach is a standard image representation scheme used in patch-based visual object classification. In such patch-based classification system, the key role of a codebook is to provide a way to map the low-level features into a fixed-length feature vector in histogram domain to which standard classifiers can be directly applied.

Many of the large number of keypoints detected from training images do not contribute for the recognition and the computational cost required for the vector quantisation step for the generation of BoF vectors is very high. A larger sized codebook increases the computational needs in terms of memory requirement for generating the histogram of each image which is proportional to the size of the codebook. The high dimensional image representation could make many machine learning algorithms which become inefficient and unreliable or even a breakdown. The proposed ideas in this paper are to generate a compact and discriminative codebook, that can be obtained by selecting representative keypoints and the elimination of indistinctive codewords. These processes reduce the overall computational complexity while maintaining the BoF model to be efficient with stable performance.

### References


