

An Improved Bag-of-Features Approach for Object Recognition from Natural Images

Saikat Basak

School of Education Technology
Jadavpur University
Kolkata, India

Ranjan Parekh

School of Education Technology
Jadavpur University
Kolkata, India

ABSTRACT

This paper proposes a novel approach for adding spatial information with local appearance features for improved classification accuracy using the Bag-of-Features approach. Spatial information can describe the probability of finding local appearance features within a sub-region of an image. Speeded-Up Robust Features (SURF), describing the appearance of small regions within an image, are extracted from sets of images used for training and testing. Extracted local image features are extended using quantized xy-coordinates that serve as spatial features. The classification is done using a Support Vector Machine (SVM) and comparisons with previous approaches have been drawn. It is observed that the proposed approach produces a significant increase in classification accuracy.

General Terms

Object Recognition, Scene Recognition.

Keywords

Bag-of-Features, SURF, Spatial Visual Vocabulary.

1. INTRODUCTION

The ability to classify visual objects is a necessary trait desired for systems that aim to be situationally aware. Systems such as robot-assisted search and rescue, unmanned aerial vehicle (UAV), unsupervised video surveillance make critical use of this capability. Search engines on the Internet provide functionality which lets users search for any image that either matches a keyword or a similar image. Digital photo managers can classify and group images that have similar content. These are all examples showing how important it is to have a framework or method that can classify visual objects accurately. The task of visual object recognition from natural images involves many challenges. Visual objects vary in appearance. There can be a ton of features that distinguish one object from another of a different class or category. And there can be dissimilarities between two objects that belong to the same class. It is essential to identify features that can be used to separate out two different objects and at the same time not get confused by dissimilarities within the same class of objects. The visual appearance of an object in an image may also depend on illumination conditions, viewing angle, focal length of the lens used and many other aspects of photography.

The Bag-of-Features (BOF) [1] model has become immensely popular in the last decade as it is an effective model for classification of visual objects. Local image features that describe the visual appearance of small interest points are extracted from a training dataset. Extracted features are clustered and quantized to get a fixed set of representatives or “visual words” that describe similar features. The frequency

of occurrence or histogram of these visual words represents an image. The similarity between histograms of two images indicates a match.

Although the classical Bag-of-Features model exhibits moderate classification accuracy, it has major shortcomings. Images are viewed and analyzed holistically, and the spatial arrangement of features is not contemplated. Succeeding research [2][3][4] addressed this problem by introducing a technique known as “spatial pyramid matching”. Images are partitioned into increasingly fine sub-regions in a quadtree-like fashion, and histograms of local features are computed for each sub-region. Local histograms are concatenated as a single “feature vector” to describe the whole image. These feature vectors often tend to be extremely high-dimensional, even up to 8,000 visual words, gathered over 21 sub-regions [5]. Such representation requires huge amounts of computational resources. This definitely is a major disadvantage for practical use cases.

An approximation of spatial pyramid matching has been suggested by Grzeszick et. al. [6] using “spatial visual words”. Quantized xy-coordinates are added to the feature vectors as spatial features. This method dramatically reduces the computational complexity but preserves the classification accuracy seen in spatial pyramid matching. One disadvantage of both the techniques that aim to preserve spatial information is the implicit assumption that an object resides, roughly, in the center of an image.

The motivation behind this paper was to overcome these challenges and come up with a framework that allows accurate identification regardless of the positioning of objects within an image while keeping the use of computational resources at a minimum. Images partitioned in $n \times n$ sub-regions, keeping the object in the center, retain the probability of finding local features that are similar in the same sub-regions for all images, in the dataset. Which means, when objects of a particular category appear in the center, there is a probability that the local feature which was extracted from the ij^{th} sub-region of an image will be found in the same ij^{th} sub-region throughout the dataset. Images, where the objects are not in the center, do not retain this advantage and this is where previous approaches fall short. The proposed method also ensures that no more than 4 sub-regions are considered assuring less usage of computational resources.

Section 2 of this paper outlines the proposed method and explains in brief about the process of feature extraction, clustering and quantization of extracted features, and finally the classification method involved with it. Section 3 summarizes the experimental results and provides the proof of concept for the proposed method. Section 4 presents an in-depth analysis of the reported experimental results and

illustrates why the proposed approach is better than its predecessors. Section 5 includes the conclusions and future scopes.

2. PROPOSED METHOD

The Bag-of-Features strategy, in scene recognition, is inspired by the Bag-of-words model used to match documents. A histogram, describing the frequency of occurrence of particular keywords, is used to match or categorize documents. In the case of images, there are no keywords. Instead, local appearance features, extracted from a set of training images, are clustered in a finite number of groups. This gives a “visual vocabulary” where every cluster represents one “visual word”. The frequency of occurrence of visual words represents an image. Fig 1 shows a graphical illustration of the classical Bag-of-Features model.

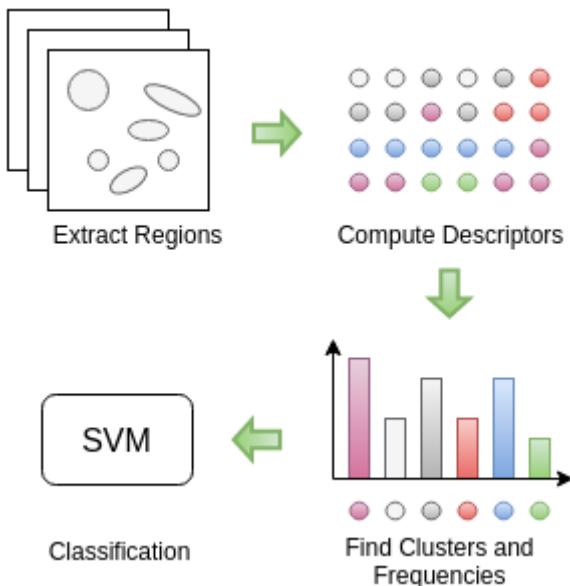


Fig 1: Graphical illustration of the classical Bag-of-Features model.

The basic idea behind the proposed approach is to extend the extracted feature vectors by adding spatial information. Using spatial feature extraction technique described in section 2.2, spatial features are extracted and concatenated with appearance features (described in section 2.1) before they are clustered and quantized. The new feature vector v is a concatenation of appearance features a , and spatial features s .

$$v = (a_0, a_1, \dots, a_n, s_0, s_1, \dots, s_n) \quad (1)$$

These feature vectors, after clustering, yields in “spatial visual words” representing both appearance and the spatial information. Once clustered, the entire set of spatial visual words will be referred to as the “spatial visual vocabulary”. In order to create the vocabulary, the combined feature descriptors (also referred to as the “bag” of features) are clustered using the K-means algorithm [7]. Once clustered, histograms representing the frequency of visual word occurrences are generated for the images in the training and test dataset. A Support Vector Machine (SVM) is used to classify a test image into one of the training classes. Fig 2 illustrates the proposed method of creating the spatial visual vocabulary and classification of visual objects using visual word histograms.

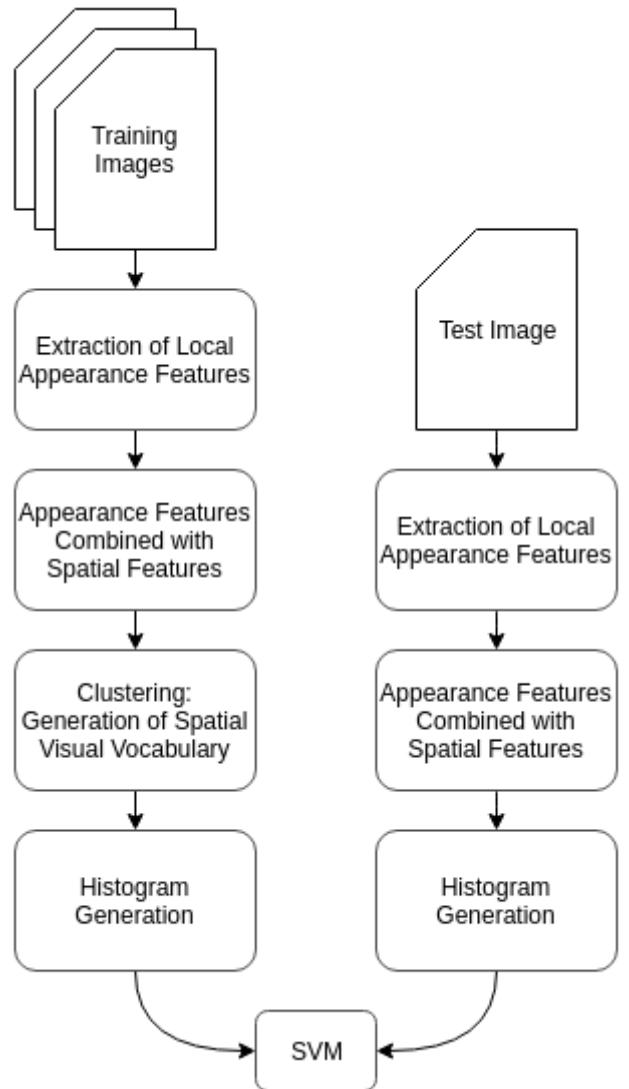


Fig 2: Proposed classification method based on spatial visual vocabulary.

Following sub-sections briefly illustrate the proposed method.

2.1 Local Appearance Features

Speeded-Up Robust Features (SURF) [8] is a local feature detector and descriptor that uses an integer approximation of the determinant of Hessian blob detector which can be computed with 3 integer operations using an integral image. The SURF feature descriptor is based on the sum of Haar Wavelet responses around a point of interest. SURF provides feature descriptors that usually have either 64 or 128 dimensions. Implementation of the proposed method uses 128-dimensional SURF feature vectors as these are more robust compared to the 64-dimensional ones. Extracted SURF keypoints are drawn on sample images and shown in Fig 3.

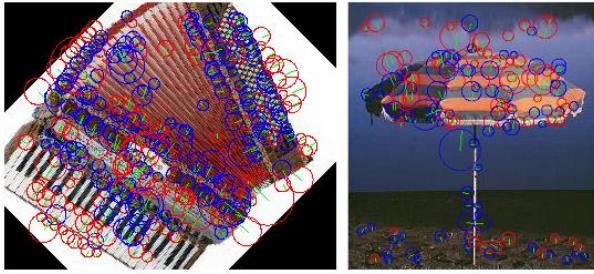


Fig 3: SURF keypoints are extracted and drawn on sample images.

2.2 Spatial Features

As previously mentioned, spatial visual words suggested by Grzeszick et. al. approximate the spatial pyramid matching technique by adding quantized xy-coordinates $q(x)$ and $q(y)$ with the feature vectors. Thus the combined feature vector becomes,

$$v = (a_0, a_1, \dots, a_n, q(x), q(y)) \quad (2)$$

In order to achieve this, the image is subdivided into 2×2 sub-regions. Quantized xy-coordinates representing the top-left, top-right, bottom-left, and bottom-right sub-regions are used as spatial features. For example, the 2×2 sub-regions can be represented by $[(0, 0), (0, 1), (1, 0), (1, 1)]$. This method can also be applied for 4×4 sub-regions for better classification accuracy. The biggest disadvantage of Grzeszick's method is the assumption that objects are, roughly, centered within an image. This might not be the case for most natural scenes. The following approach helps to overcome this limitation for improved classification accuracy.

Quantized center coordinates (c_x, c_y) are calculated taking the quantized mean of xy-coordinates of extracted local feature points.

$$\begin{aligned} c_x &= \frac{\text{mean}(x_0, x_1, \dots, x_n)}{w} \\ c_y &= \frac{\text{mean}(y_0, y_1, \dots, y_n)}{h} \end{aligned} \quad (3)$$

Where (x_i, y_i) suggests spatial coordinates of the i^{th} feature point. w and h are the width and height of the image, respectively.

The Euclidean distance d_i between (x_i, y_i) and (c_x, c_y) serves as one of the two spatial features.

$$d_i = \sqrt{(c_x - x_i)^2 + (c_y - y_i)^2} \quad (4)$$

The image is then subdivided into 2×2 sub-regions. These sub-regions reside in the northwest, northeast, southwest, and southeast side of the quantized center coordinates, (c_x, c_y) , as shown in Fig 4.

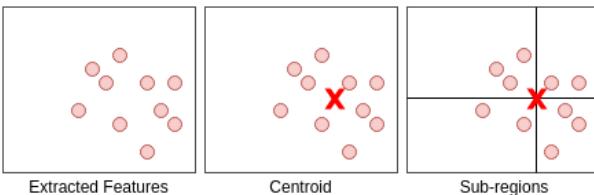


Fig 4: Images are subdivided into 2×2 sub-regions around the centroid.

The sub-regions are represented by decimal values 0, 1, 2, and 3. Thus, the combined feature vector becomes,

$$v = (a_0, a_1, \dots, a_n, d_i, l) \quad (5)$$

Where, $l = 0, \dots, 3$.

This partitioning scheme works even when the object is not in the center within an image.

2.3 Clustering and Quantization

In order to create a vocabulary of size $|V|$ the feature vectors are clustered using the K-means algorithm. Once the vocabulary is created, histograms are generated for each image in the training dataset. For convenience in classification, an average histogram for each class of objects can be generated. Fig 5 shows histograms generated during training for some of the object classes, namely, accordion (a), umbrella (b), and wheelchair (c).

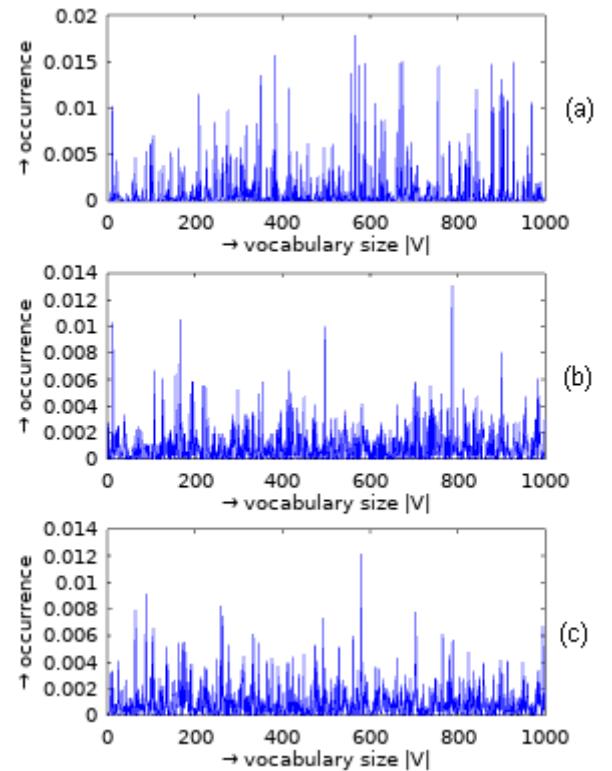


Fig 5: Histograms representing object classes, accordion (a), umbrella (b), and wheelchair (c), have been plotted.

During the tests, histograms have to be generated for individual images. Fig 6 shows histograms for sample test images from the same classes, i.e. accordion (a), umbrella (b), and wheelchair (c). Vocabulary size $|V|$ is kept at 1000 for consistency.

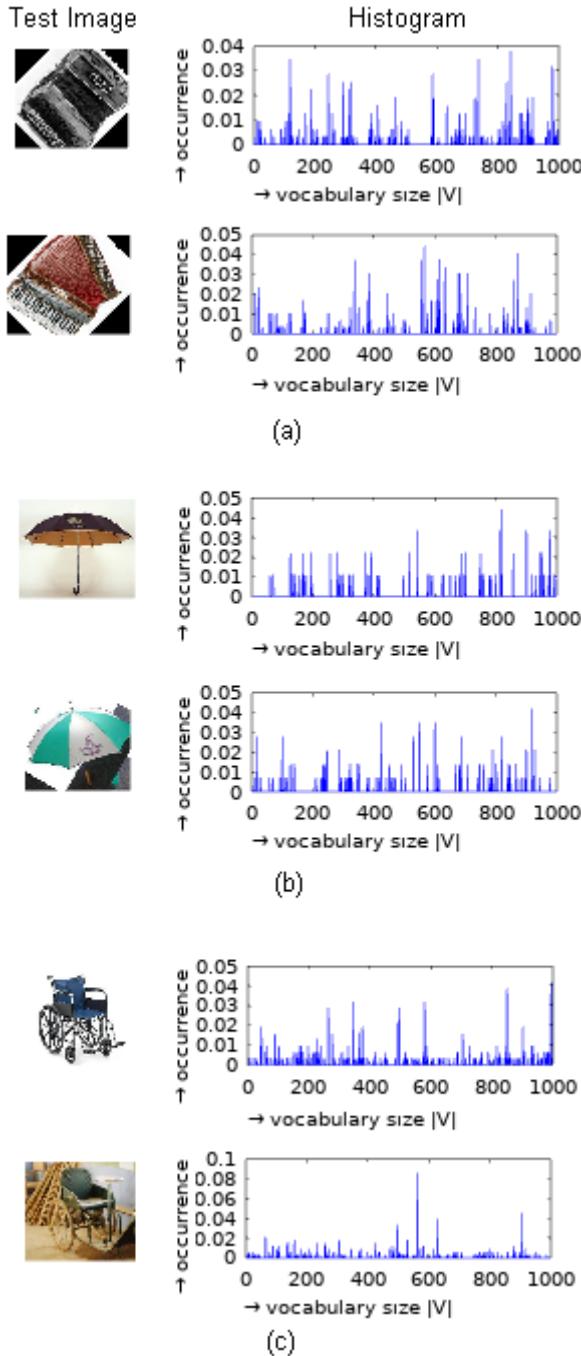


Fig 6: Histograms representing individual test images from the classes, accordion (a), umbrella (b), and wheelchair (c), have been plotted. And the actual test image is also shown for reference.

At this moment, comparisons between histograms generated from test images and the average histogram representing each of the classes can be made. For (a), it is visible that test image histograms contain frequent high peaks. Test image histograms for (b) contain lesser number of high peaks and a decent concentration of bars having almost the same height at the bottom area. Histograms obtained from test images that

belong to class (c) contain even lesser number of high peaks and a good amount of smaller bars having almost the same height at the bottom area. These traits are also traceable to the average histograms representing the classes (a), (b), and (c).

3. IMPLEMENTATION AND RESULTS

The dataset used to test the accuracy of the proposed method is the Caltech101 [9] image dataset which consists of 101 image categories. 30 images per category, divided in 8:2 ratios for training and testing purposes, have been used to train and test the system. The size of each image is, roughly, 300 × 200 pixels.

The proposed method has been implemented using Octave and tested with vocabulary sizes ($|V|$) 200, 500, and 1000. The vocabulary size clearly dictates the possible variations among spatial visual words. For classification a Support Vector Machine (SVM) has been used. Fig 7(a) shows a confusion matrix for the proposed method showing the classes with the highest classification accuracy with a vocabulary size of 1000. Fig 7(b), on the other hand, shows the confusion matrix for the proposed method showing the classes with the lowest classification accuracy with the same vocabulary size.

	accordion	schooner	dollar_bill	stapler	Faces	snoopy	airplanes	Leopards	minaret	scissors
accordion	.83									
schooner		.83								
dollar_bill			.83							
stapler				.83						
Faces					.83					
snoopy						.67				
airplanes							1.00			
Leopards								1.00		
minaret									.83	
scissors										.67

(a)

	lamp	dolphin	lobster	Motorbikes	headphone	wrench	Inline_skate	saxophone	laptop	scoccer_ball
lamp	.17									
dolphin		.17								
lobster			.33							
Motorbikes				.33						
headphone					.50					
wrench						.17				
Inline_skate							.33			
saxophone								.50		
laptop									.50	
scoccer_ball										.50

(b)

Fig 7: (a) Confusion matrix showing classes with the highest classification accuracy. (b) Confusion matrix showing classes with the lowest classification accuracy.

Fig 8(a) shows sample images from the test dataset that have been classified accurately using the proposed method. Fig 8(b), on the other hand, shows some of the misclassified images from the test dataset.

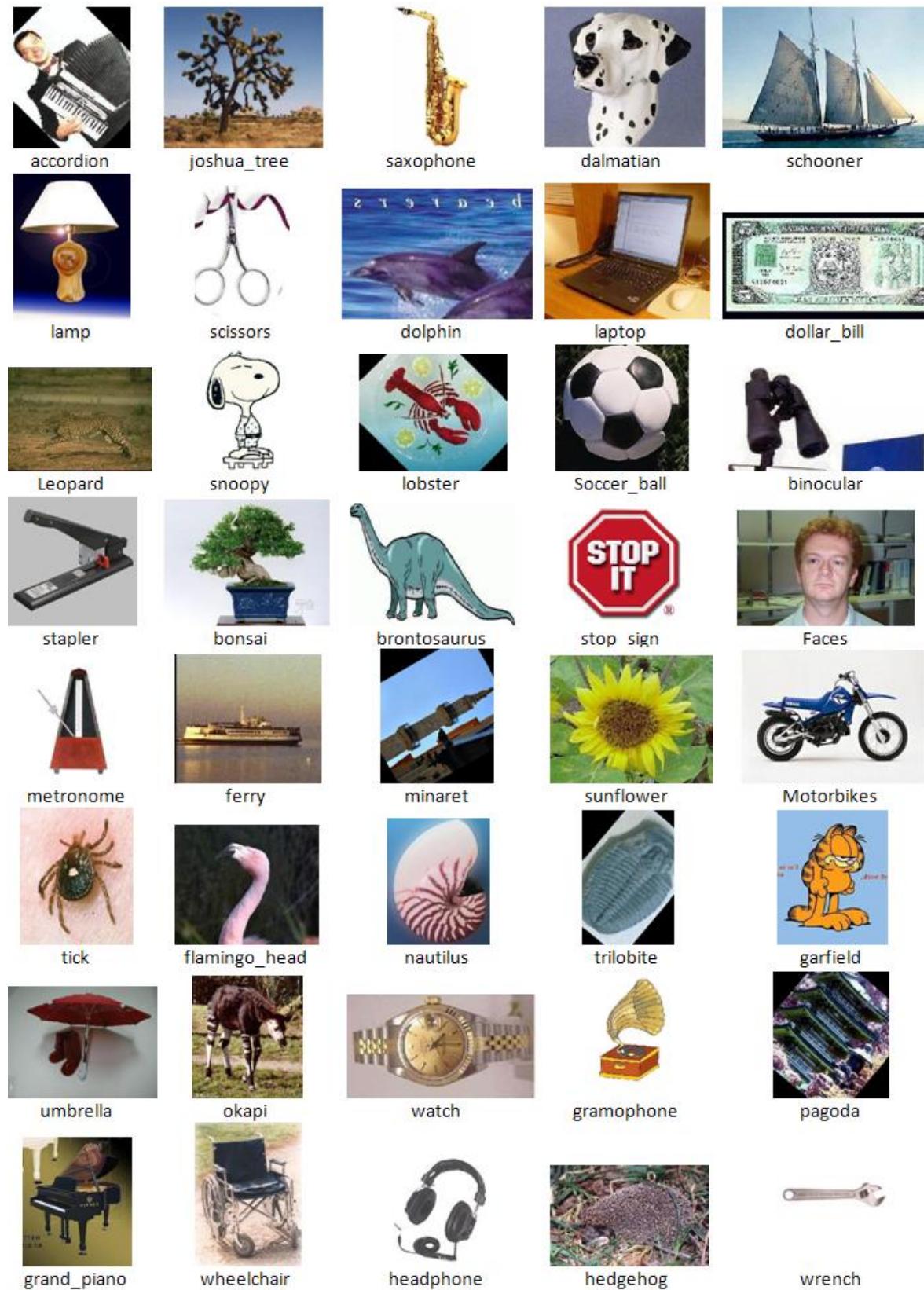


Fig 8: (a) Sample images along with respective labels from the test dataset that have been classified accurately.

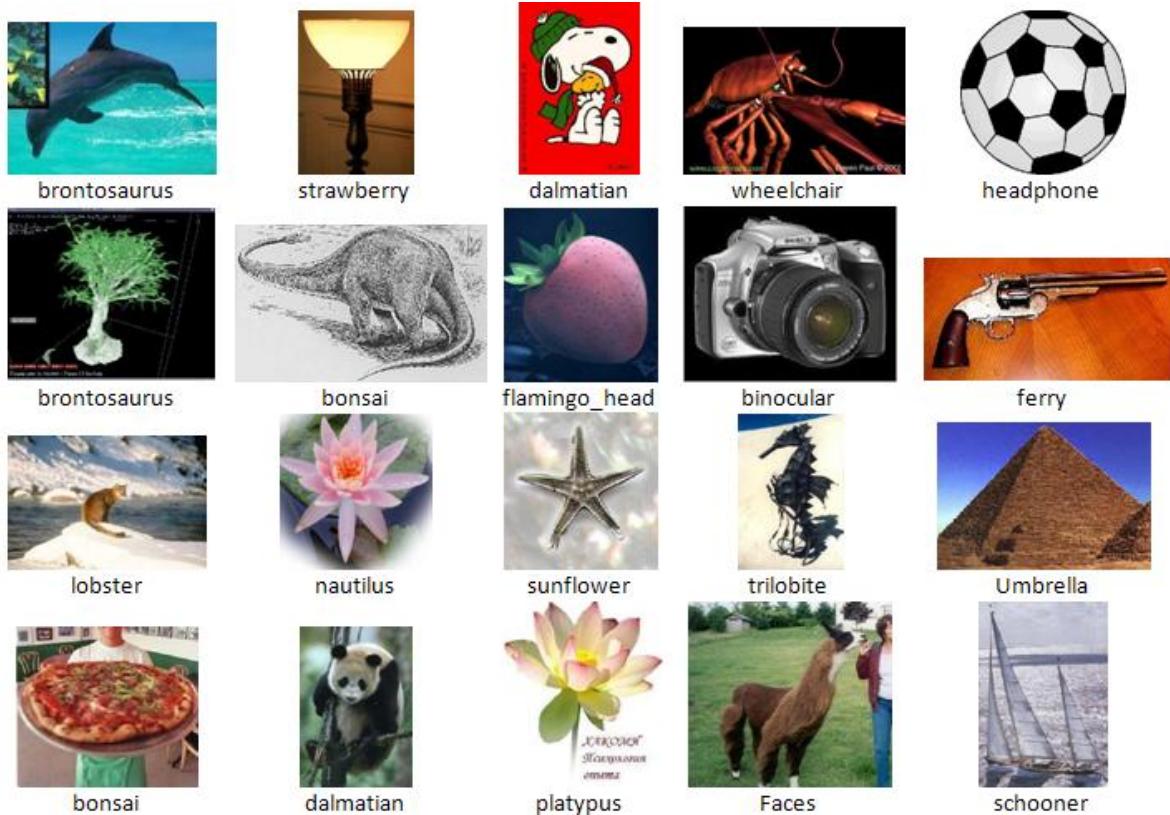


Fig. 8: (b) Sample images along with predicted labels from the test dataset that have been misclassified.

4. ANALYSIS

Three of the models that have been discussed, namely, the classical Bag-of-Features (BOF), Grzeszick's model with spatial visual vocabulary for 2×2 and 4×4 sub-regions, and the proposed approach with 2×2 sub-regions have been tested for the purpose of drawing a comparison. Table 1 shows the classification rates for each of these methods.

Table 1: Classification accuracies on Caltech101 dataset

Method	Vocabulary Size	Average Classification Accuracy (%)
Bag-of-Features	200	49.00
	500	54.33
	1000	56.67
Grzeszick's method (2×2 sub-regions)	200	49.33
	500	57.33
	1000	62.33
Grzeszick's method (4×4 sub-regions)	200	54.33
	500	60.22
	1000	64.67
Proposed Method	200	55.67
	500	63.33
	1000	65.00

Fig. 9 shows a comparison of classification rates between the methods listed above for each of the vocabulary sizes.

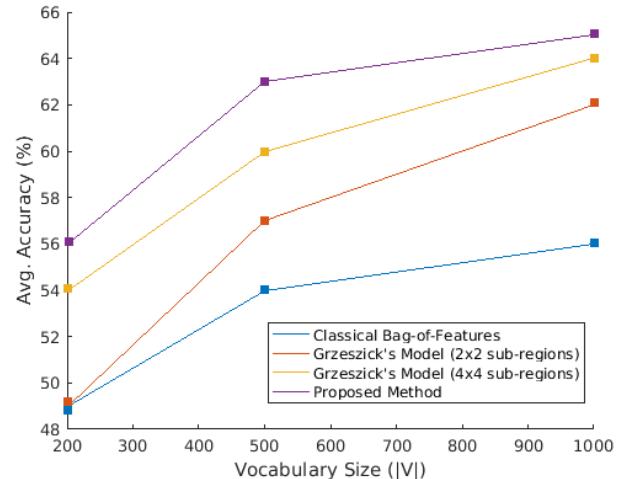


Fig 9: Comparison between the classical BOF, Grzeszick's method with 2×2 and 4×4 sub-regions, and the proposed method vocabulary size 200, 500, and 1000 have been drawn.

The above graph clearly suggests a significant increase in classification accuracy using the proposed method.

It is observed that previous partitioning schemes, proposed by Grzeszick, misclassify images when objects in the image are not properly centered. With the help of Table 2, it is shown how the proposed partitioning scheme works better than

Grzeszick's partitioning schemes as the proposed method accurately classifies images even when the objects are not properly centered.

Table 2: Sample images where objects are not centered within the image are tested using previously mentioned approaches

Test Image	Method	Prediction
 sunflower	Grzeszick's (2 × 2 sub-regions)	Faces_easy
	Grzeszick's (4 × 4 sub-regions)	minaret
	Proposed Method	sunflower
 stop_sign	Grzeszick's (2 × 2 sub-regions)	bonsai
	Grzeszick's (4 × 4 sub-regions)	okapi
	Proposed Method	stop_sign
 saxophone	Grzeszick's (2 × 2 sub-regions)	accordion
	Grzeszick's (4 × 4 sub-regions)	accordion
	Proposed Method	saxophone

5. CONCLUSION AND FUTURE SCOPE

In this paper a novel approach for packing spatial information with appearance features, for Bag-of-Features representation of visual objects, have been proposed. The proposed method, along with other methods of adding spatial information with appearance features and the classical Bag-of-Features method have been implemented, and tested with the Caltech101 image dataset. In the tests, the proposed method shows better classification accuracy than the rest.

The issue with this approach, however, is the background of an image. For example, an image of a car might also have a street sign in it. An image of an airplane on the runway might also have trees, or grass in it. These background objects also contribute to the “bag” of features and hamper the classification accuracy. An attention based Bag-of-Features model [10] can overcome these limitations. Combining the

proposed approach with attention based BoF will be a future extension to this research.

6. REFERENCES

- [1] C. Schmid, G. Dorko, S. Lazebnik, K. Mikolajczyk, and J. Ponce, "Pattern Recognition with Local Invariant Features", in *Handbook of Pattern Recognition and Computer Vision*, 2005.
- [2] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, Volume 2, pp. 2169-2178.
- [3] A. Bosch, A. Zisserman, and X. Munoz, "Representing Shape with a Spatial Pyramid Kernel", in *6th ACM International Conference on Image and Video Retrieval*, 2007, pp. 401–408.
- [4] K. Grauman, and T. Darrell, "The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features", in *IEEE International Conference on Computer Vision (ICCV)*, 2005, Volume 2, pp. 1458–1465.
- [5] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The Devil is in the Details: An Evaluation of Recent Feature Encoding Methods", in *British Machine Vision Conference*, 2011, pp. 76.1-76.12.
- [6] R. Grzeszick, L. Rothacker, and G. A. Fink, "Bag-of-Features Representations using Spatial Visual Vocabularies for Object Classification", in *IEEE International Conference on Image Processing (ICIP)*, 2013.
- [7] S. Lloyd, "Least Squares Quantization in PCM", *IEEE Transactions on Information Theory*, Volume 28, No. 2, pp. 129–137, 1982.
- [8] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-Up Robust Features (SURF)", *Computer Vision and Image Understanding*, Elsevier Science Inc., 2008, Volume 110, Issue 3, pp. 346-359.
- [9] L. Fei-Fei, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories", in *Computer Vision and Pattern Recognition*, IEEE, 2004.
- [10] Q. Wang, S. Wan, L. Yue, and C. Wang, "Visual Attention Based Bag-of-Words Model for Image Classification", *SPIE, Sixth International Conference on Digital Image Processing (ICDIP)*, Volume 9159, 2014.