BAG OF VISUAL WORDS

A Project Report

submitted by

SOUMYA SARA JOHN

in partial fulfillment of the requirements
for the award of credits in CV and MLSP : AVD864 and AVD863

MASTER OF TECHNOLOGY

DEPT. OF AVIONICS
INDIAN INSTITUTE OF SPACE SCIENCE AND TECHNOLOGY
Thiruvananthapuram - 695547

May 2018

CERTIFICATE

This is to certify that the thesis titled '**Bag of Visual Words**', submitted by **Soumya Sara John**, to the Indian Institute of Space Science and Technology, Thiruvananthapuram, for the award of the degree of **MASTER OF TECHNOLOGY**, is a bona fide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr.Deepak Mishra
Supervisor
Dept. of Avionics
IIST

Place: Thiruvananthapuram
May, 2015

DECLARATION

I declare that this thesis titled '**Bag of Visual Words**' submitted in fulfillment of the Degree of MASTER OF TECHNOLOGY is a record of original work carried out by me under the supervision of **Dr.Deepak Mishra**, and has not formed the basis for the award of any degree, diploma, associateship, fellowship or other titles in this or any other Institution or University of higher learning. In keeping with the ethical practice in reporting scientific information, due acknowledgements have been made wherever the findings of others have been cited.

Soumya Sara John
SC17M048

Place: Thiruvananthapuram
May, 2018

Abstract

An object can be represented as a bag of visual words. These visual words are basically important points in the images. These points are called features, and they are discriminative. This is in contrast to a sharp corner or a unique color combination, where we get a lot of information about the image. We can use the BoVW model for image classification by constructing a large vocabulary of many visual words and represent each image as a histogram of the frequency words that are in the image. BoVW can be used for image classification by extracting these features, generating a codebook, and then generating a histogram.

Contents

List of Figures

List of Tables

1 Introduction

Object categorization is the process of assigning classes or categories to the object in the image.Any categorization system will have to deal with some sort of visual input like color, monochrome, and thermal images, or image sequences.Often, the raw images are processed and features are extracted.But categorization imposes a number of difficult constraints and boundary conditions on established pattern recognition techniques.
There are a number of obvious applications of categorization to image database annotation, image retrieval and video annotation. But potential applications of categorization go far beyond that. Reliable categorization in real-time will open up applications in surveillance, driver assistance, autonomous robots, interactive games, virtual and augmented reality and telecommunications.
The steps involved in Object Categorization are :

- Select keypoints in the image. This could be done by Dense method or Hessian or Harr or Harris corner detection method.
- Extract feature descriptors at these keypoints ; it could be simple colour feature or HOG(Histogram of Oriented Gradients) or SIFT feature. SIFT or Scale Invariant Feature Transform finds keypoints first and then find a feature similar to HOG in the algorithm.
- Give these feature values and label to a classifier,say SVM classifier.

Once the classifier is trained, find the keypoints and features for the test image and test it using the trained classifier. Accuracy of the classifier is defined as the number of examples it classified correctly and is usually defined in terms of the confusion matrix. If tp represents the true positive, tn represents the true negatives, fp represents the false positive, fn represents the false negatives,

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn}$$

True means the example actually belongs to the class that we are trying to identify, and false means that the example doesnt belong to that class.
Or in simple words,
True Positive = Correctly identified
False Positive = Incorrectly identified
True Negative = Correctly rejected
False Negative = Incorrectly rejected

2 Bag of Visual Words

Bag of Visual Words is an extention to the NLP algorithm Bag of Words used for image classification. BOVW was developed by G. Csurka, C. R. Dance, L. Fan , J. Willamowski and C. Bray. lt essentially creates a vocabulary that can best describe the image in terms of extrapolable features.The vocabulary is a way of constructing a feature vector for classification that relates new descriptors in query images to descriptors previously seen in training. Other than CNN, this method is quite widely used.

Steps involved in training the system considering multiple vocabularies:

1. Detection and description of image patches for a set of labeled training images
2. Constructing a set of vocabularies: each is a set of cluster centres, with respect to which descriptors are vector quantized.
3. Extracting bags of keypoints for these vocabularies
4. Training multi-class classifiers using the bags of keypoints as feature vectors
5. Selecting the vocabulary and classifier giving the best overall classification accuracy.

2.1 Algorithm

Training

1. Seperate the training data and testing data
2. Initialize descriptor = [ ], label = [ ]
3. For each image in each class ,do
   (a) find the keypoints using dense method
   (b) append the labels
   (c) for each keypoint , find the hog feature vector and append it to descriptor array
4. Give descriptor array to k-Means clustering method and get kcluster means
5. Find the distance measure of feature descriptors at the points from the k cluster mean values using Eucleidean distance
6. Train an SVM using these distance measure and labels

Testing

1. Initialize descriptor = [ ], label = [ ]
2. For each image in each testing data ,do
   (a) find the keypoints using dense method
   (b) append the labels
   (c) for each keypoint , find the hog feature vector and append it to descriptor array
3. Find the distance measure of feature descriptors at the points from the k cluster mean values using Eucleidean distance
4. Test using the trained SVM using these distance measure and labels
5. Calculate accuracy

For example, consider the three classes : cellphones,faces and elephants. We find the keypoints by dense methd and these points can be viewed as:

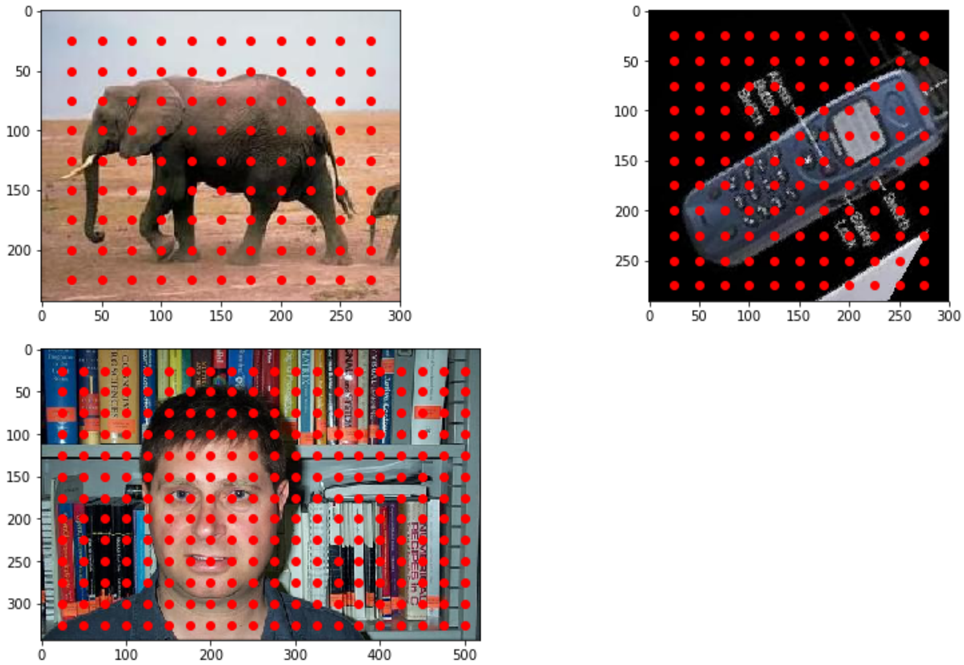

Figure 1: Keypoints

At each of these keypoints, hog features are extracted and then given to the k-Means Classifier, that gives back k mean feature descriptors. Each of this mean value is a visual word. Pictorially, a vocabulary of six visual words can be shown as :

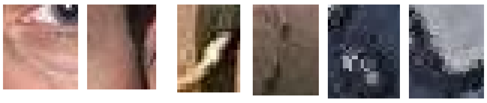

Figure 2: Visual Vocabulary

Once the vocabulary is made, we find the distance of each of the feature points from these six features and this is the simplified new feature descriptor of the points. This along with the label, is given for SVM classifier to create seperating hyperplanes.

3 Experimental Details and Results

3.1 Experimnetal Details

Used Caltech - 101 dataset and tried object categorization on 3 classes : Elephant, face and cellphone and found out keypoints via dense method, i.e. the keypoint was selected one out of every 25 pixel values. The number of keypoints varied in each image as the images were of different sizes. The feature descriptor used was HOG or Histogram of Oriented Gradients over a patch size of (21,21). Each histogram was 9 bins and the patches were divided into 7 x 7 blocks and then concatenated, giving a feature vector of length 81 x 1 at one point. Python libraries skimage,sklearn and scikit-learn were used for hog, k-Means Clustering and svm respectively. k was taken as 15 for k-Means clustering and hence there were 15 visual words. Classifocation was done using SVM classifier. 7 images from each class were used for testing.

3.2 Experimental Results

21 testing images were used. The following two figrures show the classficcation done using BOVW.

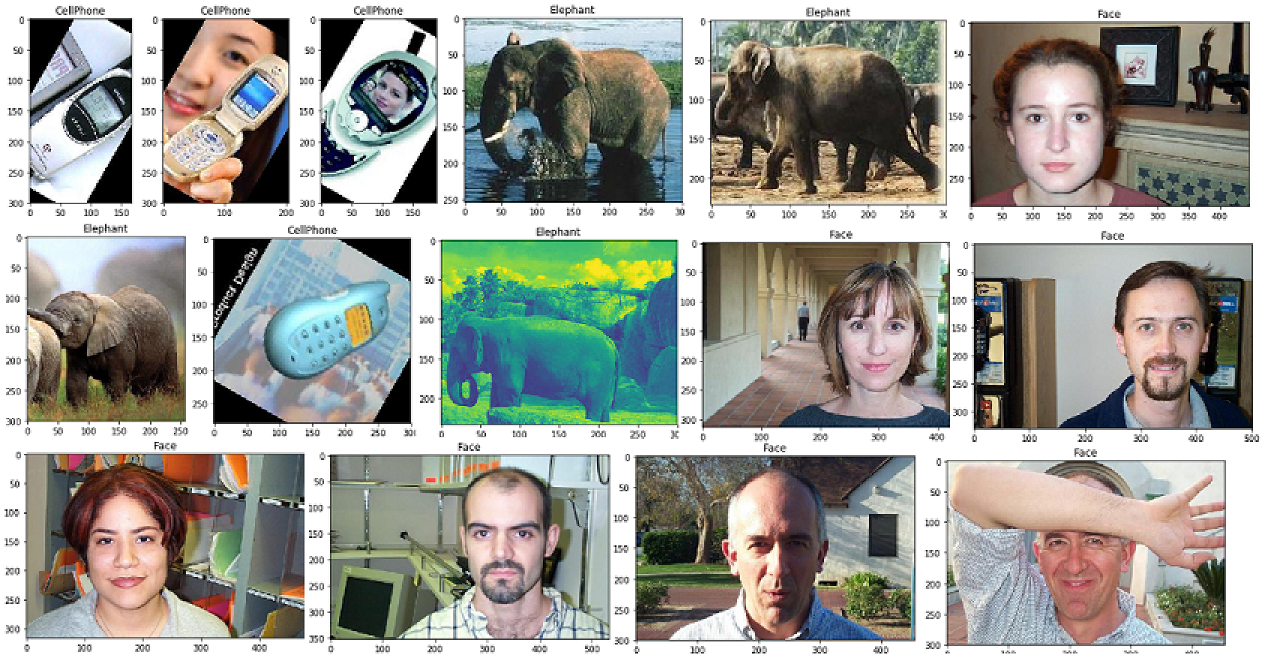

Figure 3: Properly Classified Images

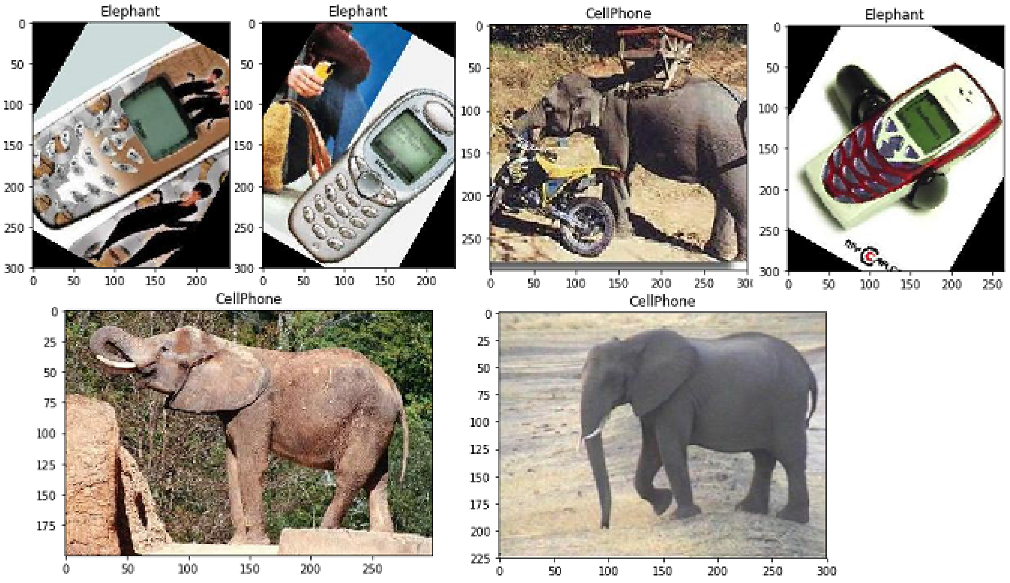

Figure 4: Misclassified Images

Class	No. of Training images	No.of testing images	Accuracy in each class
Elephant	30	7	42.85%
Face	30	7	100%
Cellphone	30	7	71.43%

Table 1: Seperate class accuracies

Class	No. of Training images		
Elephant	10	20	30
Face	10	20	30
Cellphone	10	20	30
Overall Accuracy over testing images	66.6%	71.42%	80%

Table 2: Overall Accuracy over the testing data as the number of training images increases

4 Conclusion

Obtained an accuracy of 80% when 30 training images were used in each class. When more than 30 images where given in each class, the accuracy reduced because of overfitting.The number of testing images remained the same all along. The value of k was also unchanged. Thus BOVW may show overfitting if the number of training and testing images are not selected properly.

Bibliography

1. G. Csurka, C. R. Dance, L. Fan , J. Willamowski and C. Bray ”Visual Categorization with Bags of Keypoints” ,ECCV,2004

2. Bag of Visual Words, Lecture 15,Stanford University