Image Classification Using Bag Of Visual Words Model With FAST And FREAK

Neetika Singhal
Department of Information Technology
Indira Gandhi Delhi Technical
University For Women, New Delhi
Email: neetikasinghal16@gmail.com

Nishank Singhal
Department of Computer Science and Engineering
BITS Pilani, Dubai Campus
Email:nishanksinghal20nov@gmail.com

V.Kalaichelvi
Department of Electrical and Electronics Engineering
BITS Pilani, Dubai Campus
Email: kalaichelvi@dubai.bits-pilani.ac.in

***Abstract* - This paper presents a novel technique of image classification using BOVW model. The entire process first involves feature detection of images using FAST, the choice made in order to speed up the process of detection. Then comes the stage of feature extraction for which FREAK, a binary feature descriptor is employed. K-means clustering is then applied in order to make the bag of visual words. Every image, expressed as a histogram of visual words is fed to a supervised learning model, SVM for training. SVM is then tested for classification of images into respective classes. The maximum accuracy obtained by the method proposed is 90.8%.**

Keywords –BOVW, K-means clustering, FAST, FREAK, SVM

I. INTRODUCTION

The main aim of the paper is to perform bi-linear classification of images, deciding between car and non-car images. Looking at the various techniques used previously for the same, this method is unique in its methodology and implementation. The use of FAST and FREAK as detectors and extractors along with a bag of visual model makes the technique unique, fast and efficient.

There is a wide variety of feature detectors that vary in the kinds of keypoints detected, repeatability, time complexity and space complexity. Examples of a few image the detectors are as follows: SIFT (Scale-invariant feature transform), FAST[1], SURF[2], BRIEF[3], ORB[4], BRISK[5]. FAST has been used in the process and discussed in more detail in the paper. Similarly, there is a wide variety of features extractors such as BRISK, ORB, BRIEF, FREAK[6] that can be used, but FREAK has been used in the process and discussed in more detail in the paper.

K-means clustering provides a base to form the dictionary of the visual words known as the bag of visual words, which is the most important framework used in the entire process.

Supervised learning model, SVM is an important classifier and used in many linear and non-linear classification problems. This problem, being a linear classification problem, SVM has been used to train the positive and negative images as car and non-car images respectively and hence perform the classification of test images.

II. BACKGROUND

A. Feature Detection using FAST

Feature Detection is the earliest operation performed on an image in image processing and often plays the role of an important deciding factor for the proficiency of an algorithm. The unique and distinct interest points detected by the detector are called keypoints.

A Feature Detector performs a low level function by abstracting the features (keypoints). It processes the image pixel by pixel to make a local decision whether a feature is present at that pixel. A detector has an efficient property of 'repeatability' by virtue of which it detects identical features in images containing identical scenes.

There are different types of image features which can be utilized as the basis of an algorithm for feature detection. The different types of image features include edges, corner points and blob points. In edge detection, the basis of detection is based on the properties of the edges of the images, examples include Canny, Sobbel, Kayyali. In corner/interest points detection, the corner points defined as the intersection of two edges are detected along with the interest points defined as the points which are not at the corner of image and have maximum local intensity. SUSAN, FAST are the examples of corner image detection. In blob detection, an area is considered as a basis for detection rather than individual points, as in the case of interest points detection, examples include Difference of Gaussians, MSER.

The feature detector used in this paper is FAST. FAST is an algorithm proposed originally by Rosten and Drummond for identifying interest points in an image. Other Feature detectors such as SIFT (DoG), Harris and SUSAN are good methods which yield high quality features, however they are too computationally intensive for use in real-time applications of any complexity [1].

The most promising advantage of the FAST corner detector is its computational efficiency. Adhering to its name, it is fast and indeed it is faster than many other well-known feature detection methods stated above.

B. Feature Extraction using FREAK

Feature extraction, form of reduced dimensionality, is the extraction of particular set of features from full size input to form a feature vector. Extracted features contain only relevant information about the image and describe the input data accurately.

978-1-5090-3239-6/17/$31.00©2017IEEE

In other words, Feature extraction is making of a feature vector extracting only a particular set of features that contain relevant information and can describe the data accurately. FREAK [6], a binary descriptor has been used for the process of feature extraction.

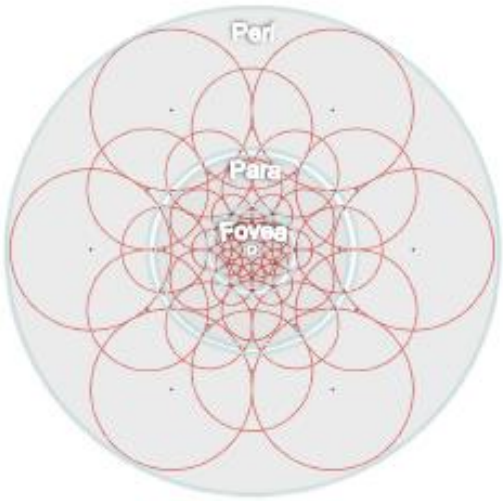

Figure 1. FREAK sampling pattern

A binary descriptor usually comprises of three parts, which are sampling pattern, orientation compensation and sampling pairs. The sampling pattern used in the case of FREAK is retinal[7], which depicts a circular grid with higher density of points near the center. The density of the points decreases exponentially as we move away from the center. Each sampling point is smoothed with a Gaussian kernel where the radius of the circle depicts the size of the standard deviation of the kernel. Figure 1 represents the sampling pattern of FREAK.

For the orientation assignment, in order to compensate for rotation changes, FREAK measures the orientation of the keypoint and rotates the sampling pairs by measure angle. FREAK's mechanism for measuring the orientation differs from BRISK in that instead of using long distance pairs, FREAK uses a predefined set of 45 symmetric sampling pairs.

The sampling pairs are learned by choosing pairs having maximum variance and minimum correlation. A coarse-to-fine approach of the pairs matches with the model of the human retina. The sampling pairs in the outer rings are selected and the pairs in the inner rings are selected at the last. Hence by following this approach along with the cascade approach, the matching is fastened. Cascade approach involves the compare of only first 128 bits and moving to the other 128 bits only in the case of distance being smaller than the threshold. As a result, a cascade of comparisons is performed accelerating the matching even further since more than 90% of the candidates are discarded with the comparison of the first 128 bits of the descriptor.

C. *Bag of visual words*

The dictionary of visual words which can be used to define each image in terms of the frequency of each word present in an image is known as the bag of visual words. The generation of visual words or the codebook generation happens after the keypoints have been detected and extracted to form a descriptor. The detected keypoints that have similar descriptors are grouped together to form a cluster. The most widely used algorithm for clustering in unsupervised learning is K-means Clustering [8] which is a quite efficient technique and follows a greedy approach. In this, firstly all the objects/entities are considered to be the canters of the cluster. Next, to each cluster, the objects that are very much similar to the center of the cluster are assigned to it. For this, the distance of each object from the center of the cluster is computed. After this, the centers of the clusters are recomputed and with the newly computed centers, the second and third steps are repeated until no more objects change their groups. Therefore, after K-means clustering, desired clusters are obtained each having a particular cluster center that represents a visual word. The number of visual words in a dictionary corresponds to the size of the dictionary. The size of the dictionary is decided on the basis of the number and classes of images present in the database. Once the size of the dictionary is decided, each image in the training set is expressed as the histogram of the visual words in the dictionary. Figure 2 demonstrates the bag of visual words showing the images as histogram of visual words. Now, this gives a fixed length vector of descriptors irrespective of the number of keypoints detected. This is one of the advantages of using a bag of visual words model. The fixed length feature vector thus obtained is fed to Support Vector Machine or any other supervised learning model for training.

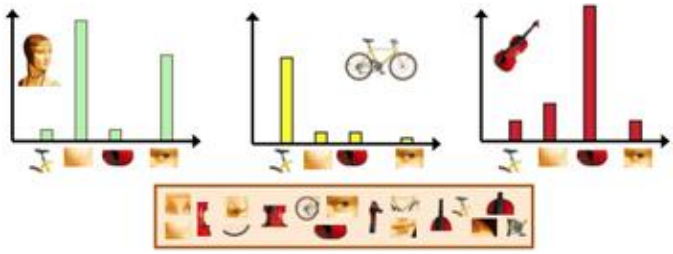

Figure 2. Bag of Visual Words Model

D. *Supervised Vector Machine Classifier*

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [9]. SVM can be defined as a discriminative classifier, formed of an optimal hyperplane. The hyperplane separates the various classes of examples which helps in classification. SVM runs an algorithm which helps in finding the hyperplane which is optimal based on the training data that is fed to it. The optimal hyperplane is chosen such that the distance of the hyperplane from the nearest data point on either side is maximum. Twice, this

distance is known as the margin. The optimal separating hyperplane maximizes the margin of the training data.

Therefore, given a labelled training data, SVM outputs an optimal hyperplane that predicts the response of new examples and classifies them into a category. Figure 3 shows the SVM classifier along with its optimal hyperplane.

SVM for this project has been used for the task of binary classification, that is it classifies into any two categories. However, SVM can also be used for performing non-linear classification.

SVM is thus an important model used in machine learning and finds its application in object classification.

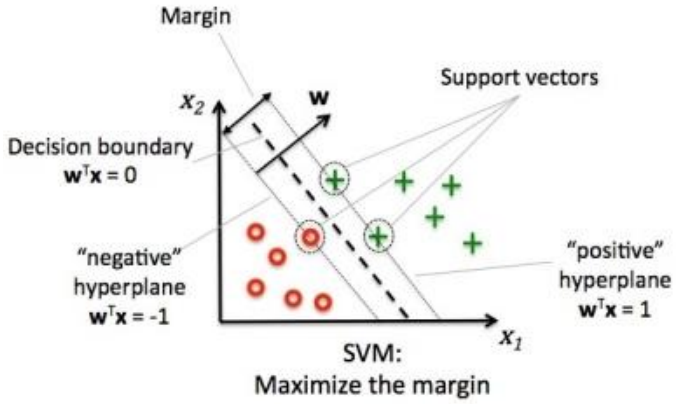

Figure 3. SVM classifier

III. Datasets, Implementation and Results

A. *Dataset for evaluation*

The images used in the project are divided into two categories, car images and non-car images.

The total number of images used correspond to 1000, where the number of training images is equal to the number of testing images which is equal to 500. Stanford car dataset [10] has been used for the training and testing of car images.

Dataset has 600 640*480 images. The data is split into 300 training images and 300 testing images, where data has been split roughly in a 50-50 split. Figure 4 shows a glimpse of car dataset.

Figure 4. Car Dataset

INRIA dataset [11] has been used for the training and testing of non-car images. It includes photos with objects of high complexity and high intra-class variability on cluttered backgrounds. It has 400 640x480 or 480x640 pixels images. The data is split into 200 training images and 200 testing images. Figure 5 shows a glimpse of non-car images.

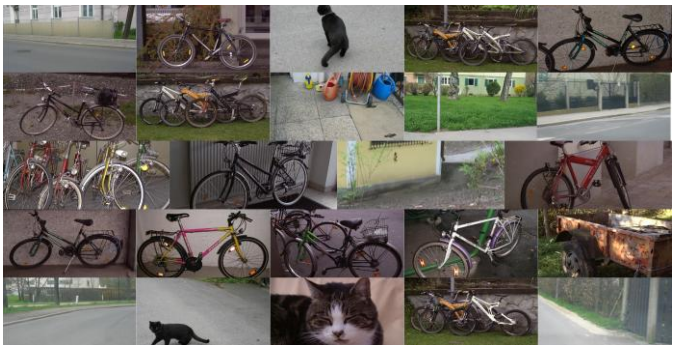

Figure 5. Non-Car Dataset

B. *Implementation*

Figure 6 demonstrates the steps of implementation. To explain, first of all, the images are read from the dataset and keypoints are detected using the FAST detector. The features are then extracted using FREAK to form descriptors. FREAK, being a binary descriptor extracts the keypoints in uchar. Since BOW trainer works only with float keypoints, it first converts uchar to float i.e. CV32F. All the descriptors are then pushed back into a single Mat object. The unclustered descriptors so formed are clustered using k-means clustering, hence forming the vocabulary. The value of k is varied from 2 to 150. The dictionary is built in float and again converted back to uchar i.e. CV8U. Here, the size of dictionary is varied from 50 to 300. For matching in the case of binary features like FREAK, hamming distance is used. Hence, the matcher used in FREAK is Brute Force-Hamming matcher.

Brute Force-Hamming Matcher is used for descriptor matching and finally the bag of visual words is made.

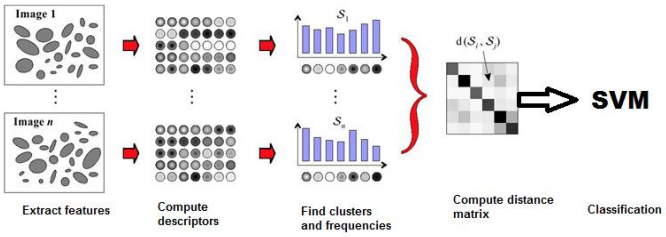

Figure 6. Steps of Implementation

The training data so obtained from Bag of Visual Words model in the process is passed as a training vector along with labels to the Support Vector Machine for Linear Classification. SVM is now trained with positive and negative examples and hence ready for bi-linear classification of images. The results of the test images so passed after the training of SVM are predicted through SVM::predict() function with the help of labels. The accuracy is then calculated based on the labels predicted by SVM.

C. *Results and Discussion*

To strengthen the analysis of simulation parts, training and testing data sets are varied in different proportions.

Figure 7 helps in selecting the optimal dictionary size and thus shows the simulation results of accuracy vs dictionary

size. The optimal dictionary size is determined to achieve the maximum accuracy of the binary classification. Accuracy, here is defined as number of correct classifications over the number of validation images multiplied by 100. If there are too many words in the dictionary, the training data gets over-fit and quantization effects are seen on building of histogram (as seen by the rapid drop in accuracy as we approach to 250 words). If there are too few words in the dictionary, the data gets under-fit and is not descriptive enough to distinguish between car and non-car images with only a few visual words. Notice that the number of vocabulary words needed to describe all the images can be partly dependent on the resolution of images (average number of keypoints). During testing, it was noticed that running on the dataset, having roughly 240 words in the vocabulary was optimal. This can be attributed to the fact that in higher resolution images, the number of keypoints per image are more, so there should be a few more bins in the histogram without worrying about quantization effects. The ratio of average number of keypoints to number of bins is important in the histogram process in order to avoid overfitting. Thus, it is found that with the optimal dictionary size, which is 240, the maximum accuracy of 90.8% is obtained.

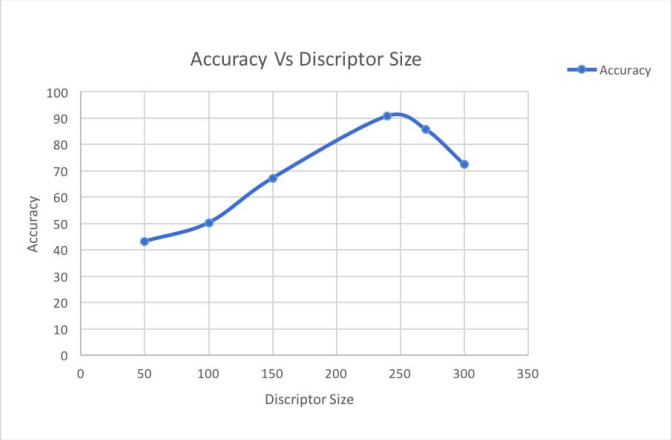

Figure 7. Selecting the Optimal Dictionary Size

Further, simulation was carried out by varying K-means clustering value, which is shown in Figure 8. The accuracy is highly dependent on the K-means clustering value. Accuracy increases at a very fast rate on increasing the K-means value, which can be easily seen in the figure. However, after a particular value the accuracy starts to decrease. It is therefore observed that the maximum value of accuracy is obtained at a K-means value of 150.

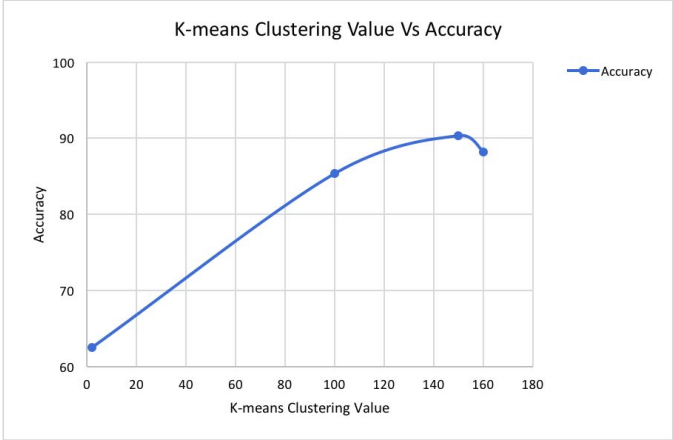

Figure 8. Accuracy

Figure 9 shows the simulation results of Accuracy Vs Training-Testing proportion by keeping descriptor size and K-means clustering value constant. After taking different proportions of training to testing images, it is observed that maximum accuracy is obtained in the case of equal number of training and testing images, that is a 50-50 split. The accuracy obtained in the case of 40-60 proportion is less owing to the fact that the number of training images is less than required and also less than the number of testing images. For a 60-40 split, it is observed that the performance is lesser than that with a 50-50 split, owing to the fact the performance starts to degrade after increasing the number of training images with respect to testing images beyond a certain number.

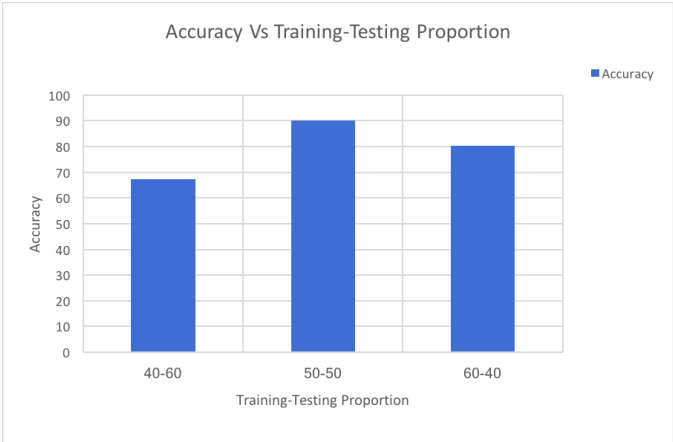

Figure 9. Accuracy

IV. Conclusions and Further Works

The binary classification of images, between car and non-car images is hence successfully performed. The method above proposed produces an error rate of 9.2%, signifying 90.8% of success rate. Hence, it is experimentally proven that

the proposed method is an efficient method of performing bilinear classification of car images using the given dataset.

Future work involves performing the classification using higher number of images and hence a larger dataset. Correspondingly, it is required to validate that the proposed method works in the same efficient manner as in the case of smaller dataset. Also, it is aimed to perform the classification using neural networks and comparing the same with the already performed classification using SVM.

REFERENCES

[1] E. Rosten and T. Drummond, "Machine learning for high speed corner detection", in 9th Euproean Conference on Computer Vision, vol. 1, 2006, pp. 430–443.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features," Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346–359, June 2008.

[3] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in European Conference on Computer Vision, vol. 6314, September 2010, pp. 778–792.

[4] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," in IEEE International Conference on Computer Vision, November 2011, pp. 2564–2571.

[5] M. C. Stephan Leutenegger and R. Siegwart, "BRISK: Binary Robust Invariant Scalable Keypoints," in IEEE International Conference on Computer Vision, November 2011, pp. 2548 –2555.

[6] Alahi, A., Ortiz, R., Vandergheynst, P.: FREAK: fast retina keypoint. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 510–517 (2012)

[7] Wohrer, A.: Model and large-scale simulator of a biological retina with contrast gain control. Ph.D. thesis, University of Nice Sophia-Antipolis (2008)

[8] K. Alsabti, S. Ranka, and V. Singh, An Efficient k-means Clustering Algorithm, Proc. First Workshop High Performance Data Mining, Mar. 1998.

[9] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.

[10] http://ai.stanford.edu/~jkrause/cars/car_dataset.html

[11] http://www.emt.tugraz.at/~pinz/data/GRAZ_02/