International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

Performance Evaluation and Study of Bag of Visual Words and Cascade Classifiers in Object Recognition

Rahul Vijay Soans[1], Pradyumna G.R.[2]

[1]*Student, Dept. of E&C Engineering, N.M.A.M Institute of Technology, Karnataka, India*
[2] *Assistant Professor, Dept. of E&C Engineering, N.M.A.M Institute of Technology, Karnataka, India*

Abstract - *Object Recognition has many applications in the field of automation, robotics, medical, defence, surveillance and security. Two approaches for recognizing an object are discussed and implemented in this paper. First method of recognizing and classifying an object is based on Bag of Visual Words and the second method is using Cascade Classifiers. The two methods of object recognition are compared for various performance evaluation metrics like accuracy, precision, miss rate and hit rate by using a Confusion Matrix table.*

Key Words: Object Recognition, SIFT, SURF, Cascade Classifier, Haar, Confusion Matrix, Performance Evaluation.

1. INTRODUCTION

Object Recognition is one of the most important task in the computer vision field. The increasing demand for real time image processing, has led to a great deal of research in Object Recognition algorithms [1]. Object Recognition is widely used in various applications like automation, surveillance, medical field, etc. The goal of Object Recognition is to automatically detect the objects in the screen and classify them according to their properties. This process has to be repeated for all the frames of the captured images. The region of interest is determined by training a model based on samples features.

Recognition algorithms are used to identify the object of interest in video or real-time web camera. There are various recognition techniques like feature extraction and boosted learning algorithms, Bag of Visual words models with features such as SURF and SIFT, gradient-based matching approaches, image segmentation, template matching and blob analysis, etc. The main focus is on Bag of visual words and Cascade classifiers.

The Bag of features or Bag of words is a well-known classification method for object recognition. An image feature identification algorithm has key point detection and descriptor extraction using SIFT or SURF method [2] because of their invariance to the scale, the orientation and almost to the illumination. Key feature points are extracted from an image and based on its feature, a descriptor vector is created. Bag of Words is a representation based on visual histogram. By clustering all extracted features from the training set, a visual words are created. Then, each SIFT feature is mapped into a visual word and is represented by the histogram. The Support Vector Machine (SVM) is a supervised classification algorithm used while testing an unknown image.

The Cascade Classifiers based on Adaboost is robust against false detections. The cascade classifiers consists of a number of stages, where each stage has weak learners. The weak learners are used in several stages to build a strong classifier. The objects are detected by moving a window over the image to find the features like Haar. Each stage labels the region of the window as either positive or negative. If the object of interest is present in the image, it is passed on to the next stage for further processing, else the image is discarded. If the image has enough matched features and passes all the training stages, the recognition will be successful.

Bag of features has shown many applications, like object recognition [3] 4], mapping for mobile robots [5] and text classification [6].

Cascade classification is one more method of recognizing objects which was motivated by face recognition algorithm developed by Viola and Jones as in paper [7]. Haar features are calculated with “integral image” to speed up and AdaBoost to select a few of thousands of Haar features, and trained one cascaded classifier. The training of cascade classifier involves positive and negative set of images which must be optimized for better accuracy and less false positives. Cascade classifiers are used to detect face [8], people [9], and many other objects which has distinct features.

2. SYSTEM DESIGN

The visual categorization is an important task in the computer vision, and it is widely used for the object recognition. Most research in this field has focused on recognizing faces, objects, scenes, and characters. Out of many recognition algorithms there is trade-off between the accuracy and speed of classifiers. There is a need of algorithm with better accuracy and sometimes with better speed. The aim of this work is to build a learning model based on different object samples to classify objects and then compare two different methods of object recognition. By evaluating the performance metrics, the robustness of an algorithm can be known.

2.1 Object Recognition using Bag of Features

Bag of Visual Words or Bag of Features is one of the popular technique used for visual data classification. A bag of words is a vector of number of words which is a histogram over the vocabulary. Image classification is done based on its

information. It extracts the features of the training images and forms a codebook which is compared with the features of unknown test images. The various steps involved in Bag of Features method is shown in Fig -1.

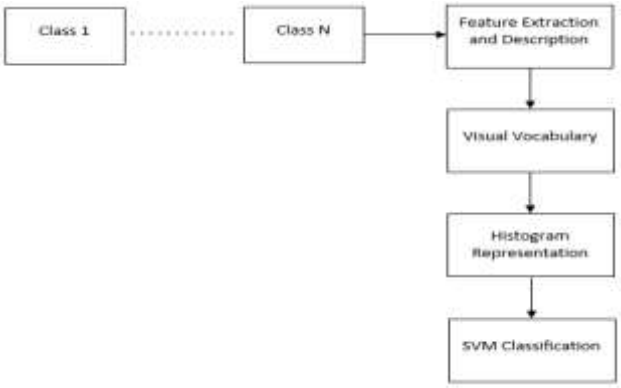

Fig -1: Steps involved in Bag of Features method

The database consists of N classes to be trained for classification. The first step is the feature extraction or keypoint extraction. Meaningful blob like feature points on the object can be extracted by using SIFT method. These are extracted from training image sets, and stored as numpy array. Vector quantization technique is used to cluster descriptors in their feature space. Each cluster is a "vocabulary" which represents the pattern that exist in the key points. Then these clusters are represented as the histogram of occurrences to define a Bag of Words containing meaningful visual words. Finally the classifier stage classifies the vocabulary based on the prediction on the clusters of feature space.

1) Feature Extraction and Description: Key points are important features that become more and more utilized in the image analysis. Local feature extraction is based on local geometrical features such as Scale Invariant Feature Transform (SIFT) which is the successful algorithm for feature detection. To perform a good recognition, the features extracted from the training image must be recognised even under various image scale, noise and illumination. Important points usually exist in object edges and corners. A point feature can be edge or a corner. SIFT extracts blob like feature points which are large binary objects. Which is then stored as a descriptor with is invariant to scale, illumination, and rotation. Fig -2 depicts feature extracted from various objects.

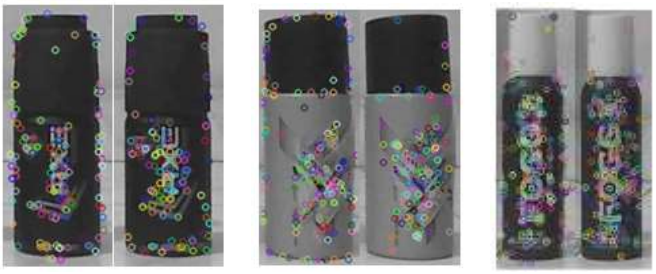

Fig -2: SIFT and SURF feature points extracted

2) Visual Vocabulary: After the extraction of descriptors, the next step is to cluster descriptors using vector quantization technique. After describing each of the images inside a class with the feature transform, a model is constructed that represents all the images which are not of the same object but within the same class of a particular object. K-means clustering is used for quantizing the feature points in feature space before providing the data to the SVM classifier.

K-means clustering is a vector quantization technique in signal processing. K-means clustering will partition 'n' observations into 'K' clusters where each point is assigned to the cluster with the nearest mean. The K-means clustering algorithm is done by selecting initial centroids at random at the beginning and assign each key point to the cluster with the nearest centroid. For the next iteration, a new centroid is calculated which is the mean of new cluster points. This process is repeated until the centroid won't change further. While implementing in Python using OpenCV function, the value of 'K' is controlled depending on the number of clusters required.

3) Histogram Representation: Histogram representation is a vector which contains number of occurrence of visual words in the image. Each of the clusters' centre is represented as a word by quantization. The number of occurrences of each word in the model database is calculated which gives the probability of the number of words in each class of objects. Subsequently, it is a step towards computing a codebook or a dictionary of several classes of object.

4) SVM Classification: While testing a new set of images, classification method is used to predict which class a new data point will be in. SVM is a linear classifier which builds a model that assigns new examples to one of the classes, from the training sets. An SVM model predicts and maps new points with clear separation gap. Data points are dimensional vector which is given to the SVM classifier to train a model. There are many hyperplanes that separate the two different classes. Best hyperplane with maximum separation is considered which is also called as 'Optimal Hyperplane'.

2.2 Object Recognition using Cascade Classifiers

The word "cascade" means that the classifier consists of several stages of classifiers. It is applied on the image until it is passed or rejected by all stages. Classifiers are trained with hundreds of "positive" samples of a particular object and few "negative" images. After the classifier is trained it can be applied to a region of an image and detect the object under test. While searching the object in a frame, the window is moved across the image and checked for the location of classifier. This can be used for object detection and tracking. Viola and Jones proposed the first cascading classifier for face detection.

The various steps involved in training a Cascade classifiers is shown in Fig -3. Each class of images consists of positive and negative samples. Feature extraction is performed by using Haar features on every positive samples collected. Thousands

of features are extracted from an image using a sub window. Using boosting techniques only the relevant features can be saved and others can be discarded. Boosting is a technique where a strong classifier is constructed from several weak ones. These features are hard to evaluate in a single stage. So cascade of classifier stages are used to quickly discard windows which are not of interest and only pass those windows which are positively classified for the next stage.

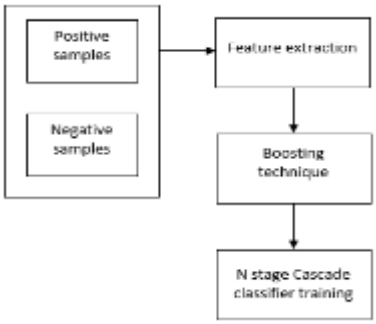

Fig -3: Steps involved in training a Cascade classifier

1) Feature Extraction: Image features are the important factor for object recognition. Haar wavelets were first used in real-time face detector. A Haar-like features has a detection window which consist of an adjacent rectangular region. It sums up the pixel values in the region and finds the difference between them. The rectangle window acts as a bounding box to the target object. The kernel window is moved over the input image, and Haar features are calculated. The difference of the pixels is then compared with a threshold value that determines the objects. A large number of similar features are used to find an object with good accuracy since they are weak learners.

2) Boosting Technique: AdaBoost is a Boosting algorithm. Strong classifier is constructed from many weak classifiers. AdaBoost is a statistical learning technique which also reduces training errors. In detecting objects, AdaBoost based approaches has two main steps. In the first step, strong classifiers will be constructed from weak classifiers. In the second step, the strong classifiers will be combined sequentially to create a cascade of boosted classifier. Final classifier is a weighted sum of weak classifiers.

3) Cascade Classifier stages: The cascade classifier consists of many stages, where each stage consists of many weak learners. Object is detected by moving a window over the image. Each stage of the classifier labels the specific region of the window as either positive or negative. Positive meaning that an object was found or negative means that the specified object was not found in the image. If the labelling yields a negative result, then the classification of this specific region is complete and the location of the window is moved to the next location. If the labelling gives a positive result, then the region moves of to the next stage of classification. The classifier yields a final decision of positive, when all the stages gives a positive, indicating the object is found in the image.

3. RESULTS AND DISCUSSIONS

The two object recognition techniques, Bag of Visual words and Cascade classifiers are evaluated by training them with the same set of database images. It uses different feature extraction and classification methods. The training process is implemented using Python 2.7 and OpenCV libraries in Intel Core i7, 8 GB RAM, CPU at 2.20 GHz. The training process results the final classifiers in different formats which is later used for testing. The classifiers are tested for a set of test images and the various evaluation metrics are obtained and compared.

The Bag of Visual word uses SIFT and SURF feature extraction. The feature extraction from the learning process is performed on different classes of objects. For the evaluation purpose various deodorant bottles are used to increase the complexity of recognition. The keypoints are extracted by using sift and surf detection functions from OpenCV. Fig -4 shows the feature extraction performed on various objects.

Later, the descriptors are computed and then used in further processing in creating a visual vocabulary. After the finding the histogram and codebook, a set of test images are checked for the correct classification. If they are classified correctly, the classification is said to be true positive, or else if it is classified as some other object, then it is a false positive. The objects can be recognized in real time using a web camera. The frames are processed continuously and the feature points are compared with the database. When enough number of feature points are obtained, more than a threshold value, the object is said to be present and the bounding box is drawn to recognize the object. Fig -4 shows the recognized objects using SIFT BoF descriptors.

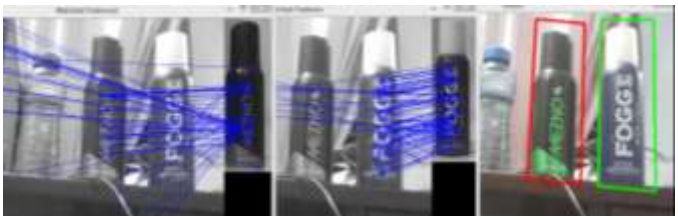

Fig -4: Object Recognition using SIFT feature extraction

The Cascade classifier uses Haar feature to extract features from the given positive dataset. Later a sliding window approach is used to detect the target object. The training for each objects must be done separately from the positive and negative samples. After cascade stages of training, xml file is generated which can be used to recognize objects in real time. Fig -5 shows the Object Recognition using trained xml files.

Fig -5: Object Recognition using Cascade Classifiers

In the field of machine learning and specifically the problem of statistical classification is a Confusion Matrix as or Error matrix. It is a table that allows visualization of the performance of an algorithm. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class. It gives the number of false positives, false negatives, true positives, and true negatives. This allows more detailed analysis of correct classifications like accuracy, precision, hit rate, miss rate, etc.

The Confusion Matrix for object recognition using Bag of features method is shown in Table -1. The Confusion Matrix for object recognition using Cascade Classifiers method is shown in Table -2. Form these information, performance evaluation metrics are calculated which is shown in Table -3.

Table -1: Confusion Matrix for Bag of Features

		Predicted class						
		Yardley	Engage	Fogg	Mezno	Axe	Ajile 1	Ajile 2
Actual class	Yardley	18	2	0	0	0	0	0
	Engage	1	19	0	0	0	0	0
	Fogg	0	0	20	0	0	0	0
	Mezno	0	0	0	20	0	0	0
	Axe	0	0	0	0	18	2	0
	Ajile1	0	0	0	0	0	20	0
	Ajile2	0	0	0	0	4	0	16

Table -2: Confusion Matrix for Cascade Classifiers

		Predicted class						
		Yardley	Engage	Fogg	Mezno	Axe	Ajile 1	Ajile 2
Actual class	Yardley	20	0	0	0	0	0	0
	Engage	0	20	0	0	0	0	0
	Fogg	0	0	20	0	0	0	0
	Mezno	0	0	0	20	0	0	0
	Axe	0	0	0	0	20	2	0
	Ajile1	0	0	0	0	0	20	0
	Ajile2	0	0	0	0	0	0	20

Table -3: Performance evaluation metrics

Bag of Words	Cascade Classifier
$Precision = \frac{TP}{TP+FP} = 0.935$ $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = 0.97$ $Hit\,rate = \frac{TP}{TP+FN} = 0.94$ $Miss\,rate = 1-TPR = 0.05$	$Precision = \frac{TP}{TP+FP} = 1.00$ $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = 1.00$ $Hit\,rate = \frac{TP}{TP+FN} = 1.00$ $Miss\,rate = 1-TPR = 0.00$

3. CONCLUSION

A visual system is implemented to detect and recognize various objects in the scene and also a study is made between two different object recognition techniques. Also the performance evaluation is done for various object classes and metrics like accuracy, precision, hit rate and miss rate are calculated. Advantages and disadvantages of two different object recognition methods are observed. The Bag of Visual word can train images quickly by extracting the features and generating the codebook whereas the Cascade classifier has a slow training phase which takes more number of samples. But the accuracy and hit rate of this is more and can be used in rapid real time object detection. The accuracy of Bag of Words approach was 97.7% and that of Cascade Classifier was found to be 100% for the same test data. The precision of Bag of Words approach was 93.5% and that of Cascade Classifier was found to be 100%.

REFERENCES

[1] Jacinto C. Nascimento and and Jorge S. Marques, "Performance Evaluation of Object Detection Algorithms for Video Surveillance," IEEE Transactions on Multimedia, vol. 8, no. 4, 2006.

[2] Yuki Sakai, Tetsuya Oda, Makoto Ikeda, Leonard Barolli, "An Object Tracking System Based on SIFT and SURF Feature Extraction Methods," International Conference on Network-Based Information Systems, 2015.

[3] Jacinto C. Nascimento and Jorge S. Marques, "Performance Evaluation of Object Detection Algorithms for Video Surveillance," IEEE Transactions on Multimedia, vol. 8, 2006.

[4] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, "Contextual bag-of-words for visual categorization," IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 4, pp. 381–392, 2011.

[5] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," IEEE International Conference on Robotics and Automation, pp. 3921–3926, 2007.

[6] J. Bai, J.-Y. Nie, and F. Paradis, "Using language models for text classification," in Proceedings of the Asia Information Retrieval Symposium, Beijing, China, 2004.

[7] Paul Viola, Michael Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.

[8] Jie Zhu, Zhiqian Chen, "Real Time Face detection System Using Adaboost and Haar-like Features," International Conference on Information Science and Control Engineering, 2015.

[9] M. Sialat, N. Khlifat, F. Bremond, K. Hamrouni, "People detection in complex scene using a cascade of Boosted classifiers based on Haar-like-features," IEEE Intelligent Vehicles Symposium, 2009.