iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** V **Month of publication:** May 2018

DOI: http://doi.org/10.22214/ijraset.2018.5097

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue V, May 2018- Available at www.ijraset.com

Scene Recognition & Classification Using Computer Vision Techniques

P.Harshith Bhargav[1], Prachi Kaushik[2]
[1, 2]*Electronics and Communication Engineering Bhagwan Parshuram Institute of Technology, Bhagwan Parshuram Institute of Technology New Delhi*

Abstract: This paper tries to explain the uses and ways of recognizing a scene and classifying into different data sets using computer vision. A scence is required to be recognized by a computer since we are using them in several intelligent systems such as Self driving car, Robots, Upcoming Vision Aide systems, Augmented Reality e.t.c .The need to classify increases as there are several overlapping categories of scenes the data of which could be further reused to make the system more efficient at recognizing the surroundings .We are going to use Several Image processing and Computer vision algorithms to test and compare the results of Final outputs.
Keywords: bag of words (BOW), confusion matrix, speed up robust features (SURF)

I. INTRODUCTION

A scene is mainly characterized as a place in which we can move around. In terms of this project a scene could be a Beach or a Kitchen. A scene could be either indoor or outdoor. Scene recognition is an important upcoming real world problem which finds application in the fields of robotics (autonomous systems), surveillance (wearable camera footage, evidence photographs, etc), and personal assistance devices (Google Glass).There has been a steady progress in the field and this paper explains the developments made in this field leading to the current state-of-the-art approach, documenting the results obtained from our implementation with standard benchmarks and classifiers. Our motivation towards taking up this project was its multiple applications in a various domains. Understanding the world in a single glance is one of the most accomplished feats of the human brain it takes only a few tens of milliseconds to recognize the category of an object or environment, emphasizing an important role of feed forward processing in visual recognition. The applications are endless and as the field is new more applications are bound to emerge.
The paper is organized as: Section II is Literature Review. Section III describes the proposed work with explanation of the method used to get the results. The effectiveness of our method is evaluated by performing tests using Scene database. Evaluation is done, results and confusion matrix illustrated in Section IV. Finally, Section V concludes the proposed work and also discusses the future scope of this work.

II. LITERATURE REVIEW

The basic and earlier approach [1] was of each image resized in to tiny photos of 1 pixel each for training set and then nearest neighbour a unsupervised clustering algorithm was run to compare the L2 pixel-wise distance between the training set and the test set images. With the introduction of SIFT feature next, Bag of words was introduced as an idea to capture the information about particular scene which uses scale invariant feature points which are unique. In the last couple of years there have been subtle advances which include higher order vector representation [3] and spatial pyramid matching [2]. There are some other state of the art approaches which have recently made their way based on above approaches- in one of them trained object detection SVM's as filter bank is used distinctively and uses the response from these filters to create the histogram. Another approach goes even further by detecting these objects for the filter bank directly from the training set.
The bag-of-word (BOW) model, whose main idea is to count the occurrences of all the words in a language library and represent a document with histogram model, is derived from text categorization. Fei-fei proposed a method [4] which represents an image by using the BOW model. The image is regarded as a document, therefore the features in the image can be defined as words. Then the histogram model is obtained from the features of all the image blocks [5]. Recently, the BOW model has been widely used in the field of computer vision such as video / image retrieval, image automatic labelling, object recognition and scene classification. The BOVW model represents an image as a set of unordered visual-word features, and extracts local features from the sample images. Some clustering methods (e.g. K-means clustering, mean shift clustering, fuzzy clustering) are used to get the cluster centres as

visual words. A set of visual words constitutes the visual dictionary, and the histogram model will be input to classifier for learning and recognition.

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. 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[6]

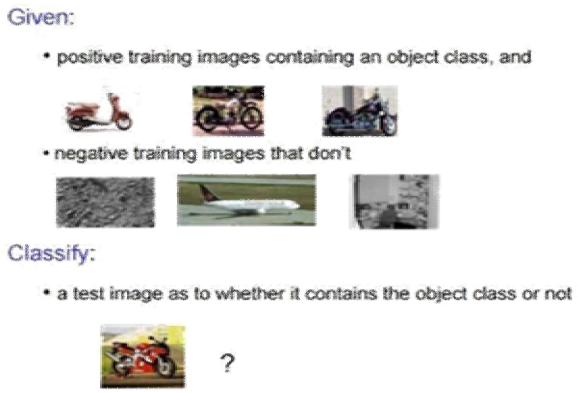


Fig. 1 Classification of an image

III.PROPOSED WORK

In This section the proposed work will be described in a step-by-step manner.

A. Dataset Used

For Classification of scenes using Computer Vision large Dataset of Images of each category is required. Acquiring these many by self is not possible in the scope of the project. Therefore we are going to use one of the most widely cited and used data sets i.e. Ponce Research Group's Fifteen Scene Categories. It contains 15 scene categories but to reduce the complexity and burden on the according to hardware resources available we are going to use just 7 scene categories having 200 images each.

B. Feature Extraction in Images

Feature extraction of an image is the most crucial part of classification of an image into several categories of images. For this reason, the extraction of different texture and edge descriptions is performed. Image features descriptors are visualized in the form of a vector or a histogram based on the feature extraction used. By matching the feature representation, similarity between two images could be found.

C. SURF Descriptors

In Computer Vision Speeded up Robust features (SURF) is used for local feature description and detection in Computer Vision. Matching, interest point detection and local neighbourhood description are the main three parts of SURF. For approximation of Gaussian smoothing squared-shaped filters are used. To detect interest points, Determinant of Hesssian blob detector is taken and it's integer approximation is used by the SURF. Feature description is based on Sum of the Haar wavelet response is obtained which generally is around the point of interest.

D. Bag-of-Words (BOW) model in Computer Vision

The bag-of-visual-words model takes image feature and interprets them as words which are then used for classification. BOW can also be seen as an independent image features which are represented as a histogram [4]. In terms of Computer Vision, Vector of the count of occurrences of image feature vocabulary is considered to be the bag of visual words. The Flow Chart of The Proposed Work is shown in Figure 1.

Image categorization problem can be carried out by the BOW technique. Confusion matrix is used as an evaluation metric for categorization of multiple labels. For BOW feature detection, feature description and codebook generation are the three steps taken into consideration.

1) *Feature detection:* The image features present are considered while computing image information for which a number of methods could be used. Edges, corners, blobs, ridges are the various image features. SURF, MSER feature detectors could be used to detect these
2) *Feature representation:* Number of local image patches represent each image .The patches are encoded into numerical vectors know as feature descriptors in this method. SIFT is considered to be a good descriptor and gives a 128- dimensional vector for each patch present in the image [7]. Moreover, each image is considered to be a collection of 128 - dimensional descriptor *vectors*
3) *Codebook generation:* "Codewords" are made by conversion of vector-represented patches which is the last steps and it is equivalent to words in a essay. A codebook is formed by combining all the codewords which is analogous to word dictionary. For this purpose K-means clustering is used. Codewords are a centers of clusters in this method [8]. Therefore, each image mapping is carried out through clustering to a certain present codeword and the image can then be represented in form of the codewords using the histogram.

E. Classification In Images

Classification holds an important part in image recognition and computer vision techniques. It has an utmost essential role. In our techniques we have taken into consideration the following classification techniques which are Surf Feature with Bag of visual words.

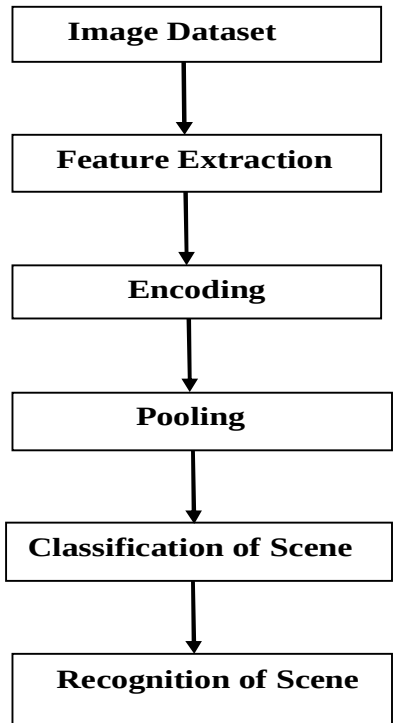


Fig. 2 Flow Chart of Proposed Work

IV.RESULTS

The Dataset was divided into 7 categories of 200 images each. Out of these 200 images, 140 were used for training the model and 60 were used for testing the model. Accuracy for the code was found out to be 85% for training set and 75% for the test set.

7×2 table

Label	Count
bedroom	200
coast	200
forest	200
industrial	200
livingroom	200
mountain	200
store	200

Fig.3 Dataset count as seen in the program

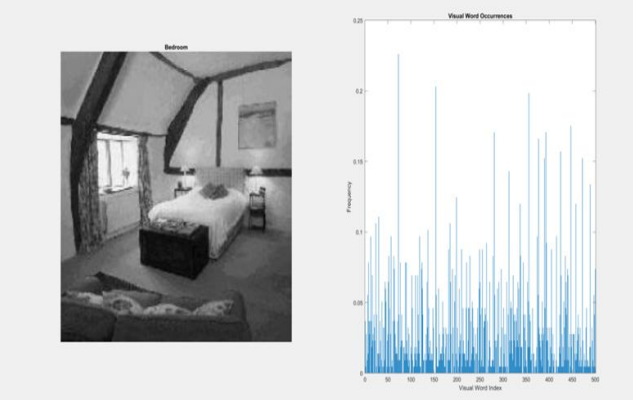


Fig.4 Example of an Feature Vector of an image

Fig.5 Correct Recognition by the code

A. Confusion Matrix

Average accuracy after evaluation of confusion matrix was found out to be 75% for test set.

	PREDICTED						
KNOWN	Bedroom	Coast	Forest	Industrial	Living room	Mountain	Store
Bedroom	**0.55**	0.00	0.00	0.13	0.23	0.05	0.03
Coast	0.00	**0.88**	0.00	0.00	0.00	0.12	0.00
Forest	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00
Industrial	0.05	0.08	0.02	**0.58**	0.07	0.07	0.13
Living room	0.20	0.00	0.00	0.03	**0.63**	0.00	0.13
Mountain	0.02	0.05	0.02	0.00	0.00	**0.90**	0.02
Store	0.02	0.02	0.17	0.05	0.07	0.02	**0.65**

Fig.6 Confusion matrix after evaluation

V. CONCLUSION AND FUTURE WORK

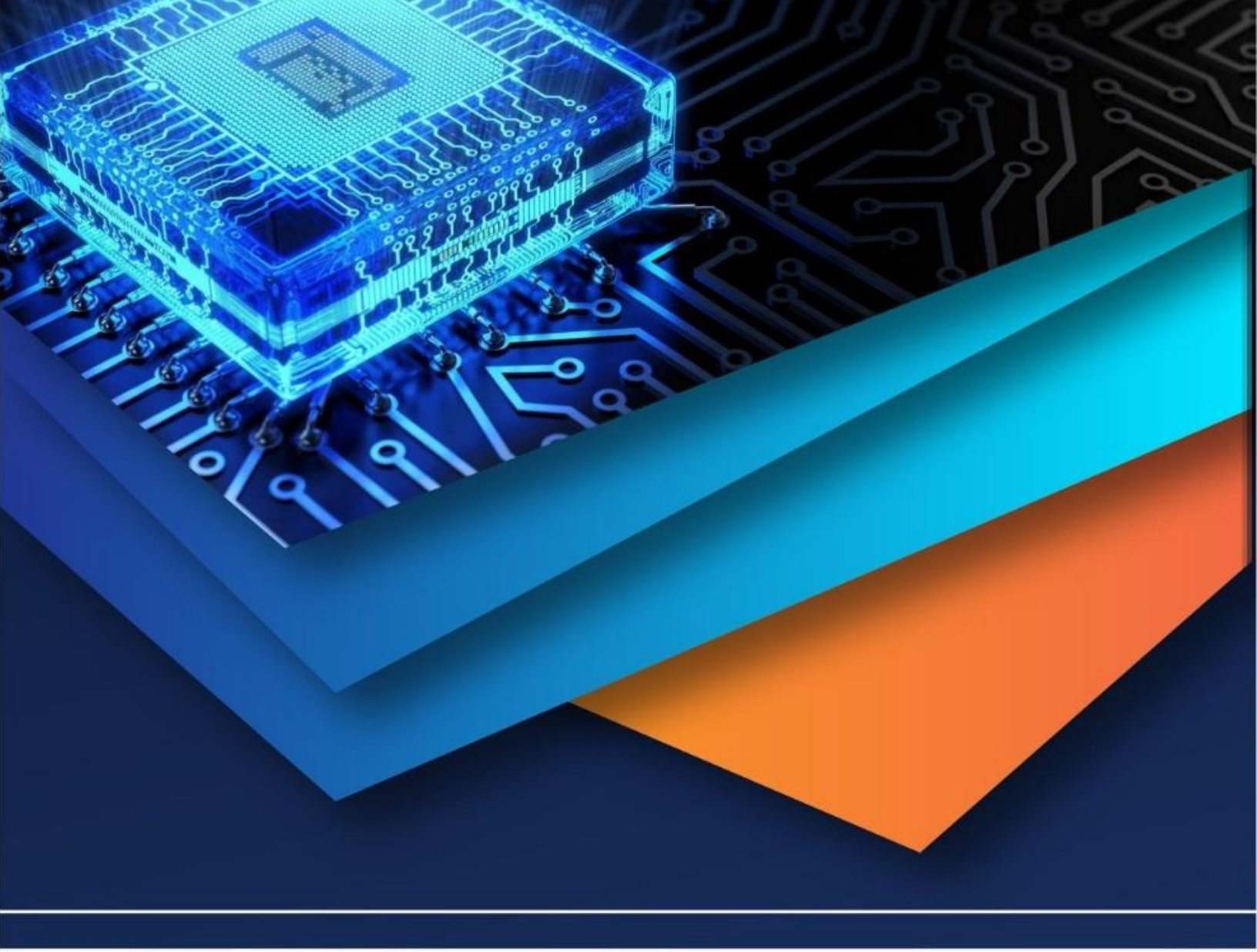
The accuracy of computer vision and machine learning algorithms depend hugely on the feature extractor used and dataset provided, and also it's correlation with the actual object we are trying to recognize. If the data has points which correlate to other scenes both scenes can easily be confused. For example, there is not a lot of difference between Living room and bedroom in general sense. Some of the various conclusions are as follows:

A. The computer can easily get confused between the two as there is a high correlation between Living room and bedroom.
B. Then there are indoor and outdoor scenarios and lighting conditions vary between day and night.
C. A lot of factors are involved in training, testing and predicting the data. But the main point to take away her is that better is the quality of data better is the predication. More examples means more is the chance for computer to create better understanding of the scenes.

Research and Development in computer vision and machine learning is ongoing. As we get more advanced in this field more applications are going to prop up. This program in general could be further improved by using and comparing different extractors (e.x FAST,HOG,SIFT) and neural nets. More and diverse data could also be trained with so the program could get better at predicting the scene. Various ways to improve the program

REFERENCES

[1] Torralba and Freeman.
[2] R. F. L. Fei-Fei and P. Perona., "Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories."
[3] 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.
[4] F.F. Li, P. Perona. "A bayesian hierarchical model for learning natural scene categories," Proc. IEEE Conf. Computer Vision and Patter Recognition, vol. 2, pp. 524–531, 2005
[5] F.F. Li,R.Fergus,and A.Torralba.Recognizing and learning object categories. Proc. IEEE Conf. Computer Vision, short course, 200
[6] 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.
[7] Vidal-Naquet, Michel, and Shimon Ullman. "Object Recognition with Informative Features and Linear Classification," ICCV. Vol. 3. 2003.
[8] Leung, Thomas, and Jitendra Malik. "Representing and recognizing the visual appearance of materials using three-dimensional textons," International journal of computer vision 43.1 (2001). 29-44.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY