Classroom Scene Recognition for Monitoring

Prajakta Survase
Electronics and Telecommunication Department
Vishwakarma Institute of Technology
Pune, India
prajakta.survase17@vit.edu

Sakshi Rathi
Electronics and Telecommunication Department)
Vishwakarma Institute of Technology)
Pune, India
sakshi.rathi17@vit.edu

Almas Sangle
Electronics and Telecommunication Department
Vishwakarma Institute of Technology
Pune, India
almas.sangle17@vit.edu

Priyanka Sargam
Electronics and Telecommunication Department
Vishwakarma Institute of Technology
Pune, India
priyanka.sargam17@vit.edu

Jyoti Madake
Electronics and Telecommunication Department
Vishwakarma Institute of Technology
Pune, India
jyoti.madake@vit.edu

***Abstract*—Indoor scene recognition is a challenging problem in computer vision. Although, many approaches have been proposed for indoor scene recognition, there is not any significant progress when it comes to Indoor scene recognition. Unlike outdoor scenes, indoor scenes lack distinctive local or global visual substance patterns. In this paper, first we propose different approaches based on the method of feature extraction and classification and compare them. This traditional method of scene recognition fails to give good accuracy, so CNN approach with 3 layers is considered. This yielded not more than 70% accuracy and therefore, a Deep CNN approach with transfer learning is implemented using MobileNet and Keras. All the algorithms are trained on MIT67 Indoor dataset with broad classification between 4 main classes classroom, concert hall, meeting hall and auditorium. With the proposed deep CNN accuracy of to 90% is achieved on MIT67 Indoor dataset. For the crowd analysis in a classroom, face detection along with the counting the faces and identifying the gender is proposed using Haar Classifier and CNN respectively.**

***Keywords*—Indoor Scene, SVM, HOG, Daisy, Gist, Deep CNN, Transfer Learning, MobileNet, Keras**

I. Introduction

Classroom scene recognition is based on different approaches of indoor scene recognition. Various approaches combining different feature descriptors and classifiers can be looked into for further discussions. The number of students in a classroom is detected using face detection algorithm and in addition to that gender of detected students is identified. Firstly, the GIST is computed as a global feature vector, providing a holistic representation of the image. Further, a suitable classifier like SVM can be used to classify the images based on the features obtained. The second approach that is looked into is an end-to-end pipeline for recognizing scene categories consisting of feature extraction, encoding, pooling and classification steps. HOG, a global descriptor is combined with SVM classifier for the scene recognition. For improving the classification accuracy further, this pipeline is made using a hybrid feature descriptor by combining local feature descriptor Daisy, along with the global descriptor, HOG. A 2-level pooling scheme to combine DAISY and HOG features is trained by SVM classifier for enhancing the results. A third approach of CNN is looked into, which is convolution neural network consisting of 3 layers. This approach fails due to limited size of the dataset of each class the accuracy doesn't go beyond 70% for classroom detection. Hence, a Deep learning provides an intelligent agent that analyses it's environment and acts accordingly. Deep convolutional neural networks (CNNs) achieve impressive performance in scene recognition. Face detection is done using Haar Cascade Classifiers in Python and this will give a count of total heads present in a classroom in a study place. For determining gender, a CNN approach is used and trained, for accurate results.

II. Related Work

Classroom scene recognition is based on different approaches of indoor scene recognition. Various approaches combining different feature descriptors and classifiers were looked into for analysis and accuracy purpose. Firstly, the GIST was computed as a global feature vector, providing a holistic representation of the image. A suitable classifier is

then used to classify the images based on the features obtained. . The GIST descriptor for an image can be computed by convolving it with 32 Gabor filters at 8 orientations and 4 scales that produces 32 feature maps of the same resolution as the original image. GIST vectors are very robust for natural scene categorization as compared to indoor scene classification. The second approach that is looked into is an end-to-end pipeline for recognizing scene categories consisting of feature extraction, encoding, pooling and classification steps. HOG, a global descriptor is combined with SVM classifier for the scene recognition. Local features such as DAISY features use skimage library for feature extraction. A standard HOG descriptor corresponding to the whole image effectively allows us to choose features at different scales. The DAISY descriptors form a bag-of-visual-words by suitable clustering. SVM Classifier uses encoding and pooling is followed by classification. Standard SVM classifier is used for classification. For improving the classification accuracy further, this pipeline can be made using a hybrid feature descriptor by combining the use of local and global feature descriptor A 2-level pooling scheme to combine DAISY and HOG features can be trained by SVM classifier for enhancing the results.

Feature extractor	Classifier	Accuracy(%)
GIST	SVM	52
HOG	SVM	40
HOG + Daisy	SVM	65

TABLE I: Feature Extraction and their accuracy

From the given table, the accuracy of GIST feature with SVM for indoor scene is very average and better results are yielded for natural scene recognition. Similarly, using only HOG, global descriptor does not give a better accuracy either but when combined with a local descriptor DAISY, the accuracy definitely shoots up. But the overall accuracy of the system doesn't go beyond 65%. Therefore, in order to increase accuracy a CNN model is trained according to the system's specifications. This CNN module had 3 layers and the accuracy measure doesn't go beyond 70%. In this paper we shall look into approaches based on deep learning to improve the performance of the model.

III. Experimental Setup

The technology stack involved in the implementation of this project include keras for training and validating the model along with MobileNet ssd model. MobileNet is a CNN architecture model for Image Classification and Mobile Vision. There are other models as well but what makes MobileNet special that it very less computation power to run or apply transfer learning to. The core layer of MobileNet is depthwise separable filters, named as Depthwise Separable Convolution. The network structure is another factor to boost the performance. SSD is Single Shot Multibox Detector. It is a part of the family of networks which predict the bounding boxes of objects in a given image. It is a simple, end to end single network, removing many steps involved in other networks which tries to achieve the same task, at the time of its publishing. SSD works better than the state of the art Faster-RCNN in cases of higher dimensional images. Implement the model on keras is an open-source neural-network library written in Python. MIT67 indoor dataset is used for training purpose. It contains 67 types of indoor scenes and 15,620 RGB images totally. For training purpose only 4 classes are considered.

IV. Model Implementation

1) Classroom Detection:
Various scenes are classified on the basis of four broad categories from the MIT Indoor 67 dataset. These four categories are classroom, auditorium, meeting room, concert hall, for any computer it is particularly difficult to identify scenes which are closely linked, hence, a machine is trained with adequate and appropriate data, which distinguishes the four categories.Here, a deep convolutional neural network is used since it yields a better performance as compared to the methods of feature extraction like HoG and Daisy, GIST etc, and classifiers such as SVM. Models such as MobileNet has been used here, display further improvements. In order to enhance the results and accuracy, a mechanism of transfer learning, using MobileNet and Keras, and deep CNN is implementing using a fully-trained module for a set of classes like classroom, auditorium, meeting hall and concert hall and retrained from the existing weights for the required classes. The final layer is retrained from scratch, while keeping the remaining layers intact. The Neural Network applies a series of filters to the dataset in order to extract them and learn more features.

   a) MobileNet: It is a small, low-latency, low-power model. It has a lightweight architecture, which uses depth wise separable convolutions. It has the effect of filtering the input channels. The pointwise convolution applies a 1×1 convolution to combine the outputs the depth wise convolution. MobileNet are Deepwise Separable Convolution Neural Networks, counting depthwise and pointwise convolutions, having several layers. After the output layer is generated, any image can be tested on the trained model by using the classification file. This file inputs the test image and outputs the classes it classifies into along with the percentage of accuracy against each class.
   b) Transfer Learning: It is an advanced technique of Deep Learning where a model developed for a task is used as a starting point for a model and training a Convolutional Neural Network (CNN).

In Transfer Learning technique, the last layers or some of the last layers of the pre-trained models (i.e. MobileNet) depending on the data being used are unfrozen and retrained on the new dataset. If only the last layer in the model is unfrozen, the resulting partial model then gives the probabilities of the outputs against each other rather than giving the most probable output.

MobileNet comprises of 86 layers of convolutional layers comprised of depthwise layer, pointwise layer that doubles the number of channels, depthwise layer with stride 2, pointwise layer that doubles the number of channels. The entire system comprises of total 91 layers, consisting of MobileNet layers along with 5 dense layers.

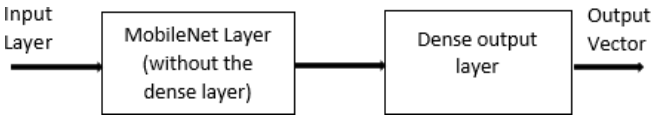

Fig. 1. Architecture of System

The algorithm is implemented using the MobileNet model which is pre-trained on the selected classes' dataset by only training the new top layer to identify the required classes. First step is to split the dataset to: training, validation and testing set. Image list is created on the basis of the three sets and the value of the layer run before the output layer is calculated. This makes use of the list of images with the labels, input tensor for the jpg data and bottleneck tensor which is the final layer. A fully connected layer is added and here the classification of the classes of the required module is done and trained with labels of the classes examples label[0] defines it as a classroom, a label[1] for auditorium and so on. The training is carried out for many steps as specified in the input. To get the accuracy of the model, evaluation is done on the set images that weren't a part of the dataset. From the observations, an accuracy value of around 90% to 95% is obtained. This is defined by the fraction of the images in the test data that are assigned the correct label after the model is fully trained.

2) Students' Analysis: If the assigned label matches to that of a classroom, this step is carried out. It involves further two parts that is face detection and count and the second part is identifying the gender of all the faces detected.
   a) Face detection: OpenCV has pre-trained classifiers for detection of face, in the form of XML files. By loading the appropriate XML file and the input image in grayscale model. Theses parameters image, scaleFactor, minNeighbours, flags, minSize and maxSize are set appropriately. If faces are found, they are bounded by drawing a rectangle Rect(x,y,w,h). Also, the number of faces are counted based on the number of rectangles detected.

   b) Gender :
   For determining gender of the detected faces, two binary classes are required, male and female. A 3 layered CNN model is trained from scratch and gender is identified for every face detected.
   In the figure below, the entire working of the system is shown pictorially. This sums up the entire flow of the system wherein when an input image is given, is classified into either of the four classes. For classroom, the presence of students is identified by face detction and their number is analysed along with the gender.

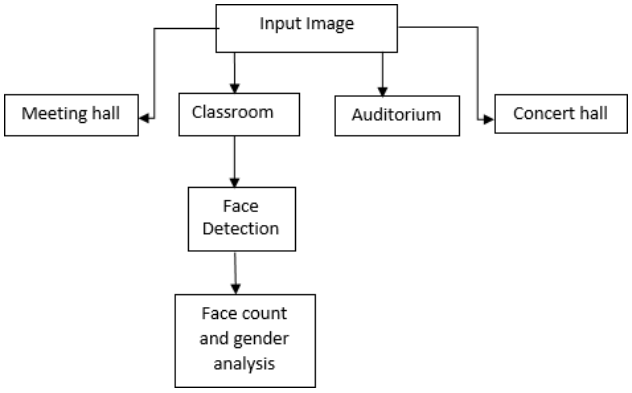

Fig. 2. Proposed System

V. RESULTS

For training 4 classes are considered with 652 images in total. After every training level, a part validation is done by the model and in the end a large test set is validated which is considered as the accuracy score of the model, which is 90.3% for classroom recognition. For gender classification, accuracy is 94.44%.

1) Classroom is detected.

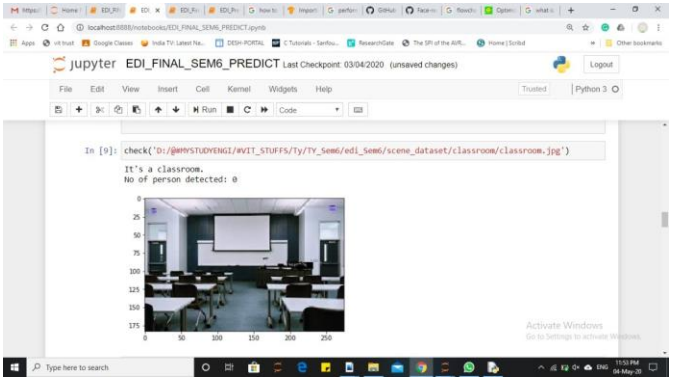

Fig. 3. Empty classroom

2) Auditorium is detected.

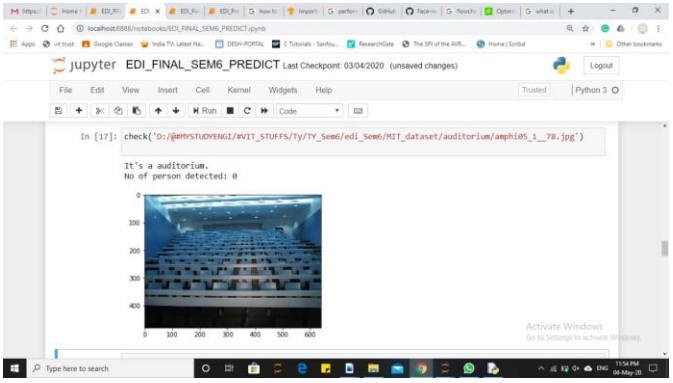

Fig. 4. Empty auditorium

3) Meeting room is detected.

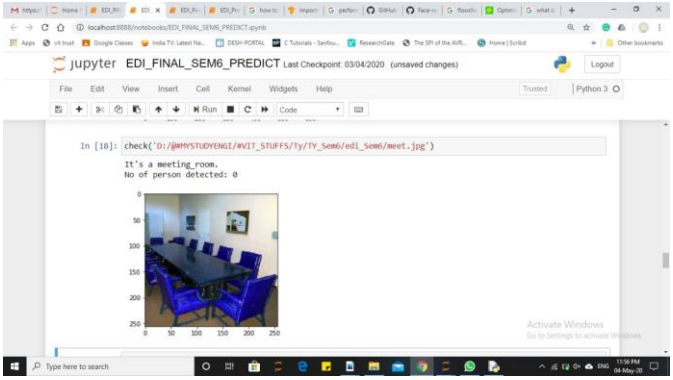

Fig. 5. Empty meeting room

4) Concert Hall is detected.

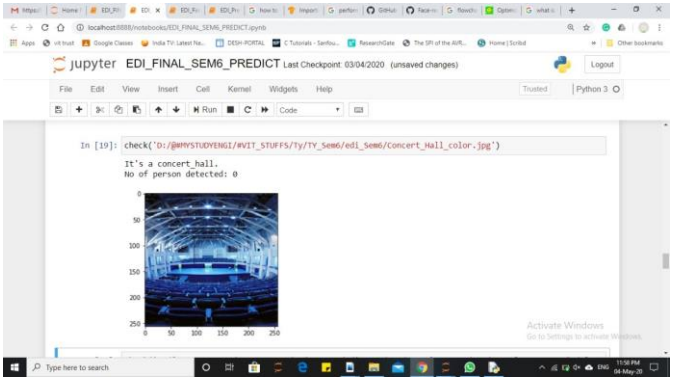

Fig. 6. Empty concert hal

5) Classroom is detected.

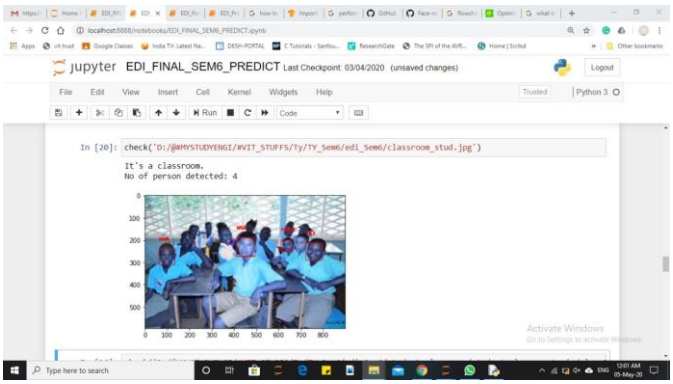

Fig. 7. Classroom with students

VI. Future Work

The future work would include usage of indoor scenes with 3D datasets for classification. Also, more number of classifiers can be used over a wide range of data for various applications. The entire class monitoring system can be enhanced by recognising the students present or absent instead of giving a number of present students, by feeding the dataset of every student.

VII. Conclusion

This paper proposes a classification for indoor scenes, based on transfer learning approach implemented using Keras. Face detection for classroom particularly is also implemented and it is limited by the fact that, a few overlapping faces may not be detected and this limits our scope of project. This approach has been taken into consideration after comparison with traditional methods of feature extraction like GIST, HOG and DAISY with classification by SVM, and also with convolution neural network. Since, there is a limited number of dataset images for each class, the accuracy is not promising and therefore, a deep CNN approach has been used as it has inbuilt feature extraction and has contributed to a large decrease in error rate. MobileNet with Transfer Learning is used since it is one of the lightest method and quickest approach to implement transfer learning for CNN's.

REFERENCES

[1] N. Sun, X. Zhu, J. Liu and G. Han, ”Indoor scene recognition based on deep learning and sparse representation”, 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2017. Available: 10.1109/fskd.2017.8393385

[2] A. Hanni, S. Chickerur and I. Bidari, ”Deep learning framework for scene based indoor location recognition”, 2017 International Conference on Technological Advancements in Power and Energy (TAP Energy), 2017. Available: 10.1109/tapenergy.2017.8397254

[3] A. Ramakrishnan, E. Ottmar, J. LoCasale-Crouch and J. Whitehill, ”Toward Automated Classroom Observation: Predicting Positive and Negative Climate”, 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), 2019. Available: 10.1109/fg.2019.8756529.

[4] Lup.lub.lu.se,2020.[Online].Available:http://lup.lub.lu.se/student-papers/record/8971320/file/8971323.pdf.

[5] Arxiv.org,2020.[Online].Available:https://arxiv.org/ftp/arxiv/papers/1805/1805.06618.pdf.

[6] W. Tahir, A. Majeed and T. Rehman, ”Indoor/outdoor image classification using GIST image features and neural network classifiers”, 2015 12th International Conference on High-capacity Optical Networks and Enabling/Emerging Technologies (HONET), 2015. Available: 10.1109/honet.2015.7395428

[7] Changting He, Ya Wang and Ming Zhu, ”A class participation enrollment system based on face recognition”, 2017 2nd International Conference on Image, Vision and Computing (ICIVC), 2017. Available: 10.1109/icivc.2017.7984556

[8] L. Herranz, S. Jiang and X. Li, ”Scene Recognition with CNNs: Objects, Scales and Dataset Bias”, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. Available: 10.1109/cvpr.2016.68