International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.94.09

RNN-LSTM based Indoor Scene Classification with HoG Features

Priyanka Gupta[1], Shilpa Sharma[2], Ambica Verma[3]
[1,2,3]*Deptt. Of CSE, Lovely Professional University,Phagwara*

Abstract—The machine learning and artificial intelligence models have evolved with unenvisaged swiftness over the past decade. The machine learning and artificial intelligence models work on the basis of the mathematical models, which are capable of processing the financial, image, video, audio, audio-visual and several other forms of data. In this paper, the work has been carried upon the bulk image data representing the variety of indoor scenes. The goal is to detect the type of indoor scene, which can be utilized by variety of artificial intelligent application for various purposes. The indoor scene recognition has been performed over the computer vision & pattern recognition (CVPR), 2009 dataset, which is consisted of 15620 images. The deep learning mechanism known as recurrent neural network (RNN) has been incorporated for the classification of the indoor scene data over the histogram of oriented gradient (HoG) based features. Specifically, the forget gates based recurrent mode called Long short-term memory (LSTM) has been incorporated for the classification of the indoor scene data. The performance of the proposed model has been analyzed over the CVPR09 dataset using 50 and 100 randomly drawn test cases. The proposed model has been found 92% accurate in comparison with SVM (88%), KNN (80%) and Naïve Bayes (86%). The F1-measure based performance assessment also proves the robustness of LSTM based model with 96% accuracy over 92% of Naïve Bayes, 93% of SVM and 88% of KNN.

Keywords—Deep learning, Recurrent neural network, Long-short term memory, LSTM, indoor scene recognition, CVPR'09.

I. INTRODUCTION

Scene classification is aimed at labeling an image into semantic categories (room, office, mountain etc). It is an important task to classify, organize and understand thousands of images efficiently. From application point of view, scene classification is useful in-content based image retrieval. As accurate classification of an image, as better as it helps in better organization and browsing of the image data. Scene classification is highly valuable in remote navigation also.

Indoor scenes are cluttered with many objects. So classification techniques simply based on color, texture and intensity are not very effective to classify indoor scenes. Pioneering works used SIFT, SURF etc in combination with supervised learning. But these techniques fail to distinguish many indoor scenes. One way to bridge semantic gap between image representation and image recognition is to make use of more and more sophisticated models, but good learning and inference is extremely difficult task for such models. Alternatively semantic gap between low-level features like color, intensity, texture etc. and high-level category label can be reduced by introducing object-based representation as intermediate representation. As the performance of scene recognition is heavily dependent on feature representation, this object-based intermediate representation proves to be useful in enhancing classification results. Recently objects-based techniques for indoor scene classification have proven to be showing promising performance over other state-of-art techniques. In this work, we will review the recent and significant techniques that have been used for indoor scene classification. Besides we will identify the key approaches being used in indoor scene classification. The major contributions made by each significant work and the challenges posed to efficient classification will also be discussed.

II. RELATED WORK

Espinace, Pablo et. al. [1] has worked on the scene detection in natural imagery for the robotic objects. The robots are entirely dependent upon the computer vision based models to identify and recognize the objects in the visual objects. In this paper, the focus is kept on the detection and localization of the multiple objects in the common categories such as furniture, doors, etc in order to facilitate the robots to move around the given premises. Giannoulis, Dimitrios et. al. [2] has worked towards removing the complexity for event detection and scene classification by incorporating the multi-layered object detection methods. Antanas, Laura et. al. [3] has developed the kernel program to classify the natural scene images with application of multiple object relations. This scheme works by evaluating the combinations of various objects in the target scene in order to identify the scene. Gupta, Saurabh et. al. [4] has worked upon the RGB image database, where the perception based scene recognition techniques are applied over the given image set. This technique emphasizes the use of grouping of the objects in the specific combination in the bottom-up approach coupled with semantic behavior in order to recognize the type of scene in the target image. Juneja, Mayank et. al. [5] has developed the scene recognition approach to determine the

type of scene on by analyzing the distinctive but partial regions in the given image. Monadjemi, Amir et. al. [6] experimented upon the high definition images to determine the kind of target scene. This model uses the texture based features to determine the type of the scene on a whole, which means the focus lies upon the complete image matrix, rather than individual objects. Duda, Richard O. et. al. [7] worked towards the robotic vision (eventually computer vision) to help the robots to understand the type of scene in order to execute the location specific program to achieve a task. Fitzpatrick, Paul et. al. [8] developed the category oriented scene classification model. This model uses the supervised knowledge based model to determine the various sets of objects in order to determine the scene. Quattoni et al. [9] has analyzed the causes of failed detection of the scenes. This study remarked the detection of the target objects by using the different exemplars within the set of objects in each class. Li-Jia Li et al. [10] has worked upon the low-level feature based scene type detection. This model is empowered by the object filtering method, which recognizes several kinds of objects in the certain combination in order to correctly recognize the scene.

III. METHODS AND MATERIALS

The scene recognition model is prepared to facilitate the robots to identify their surroundings. The idea behind this project is to implement the scene or surrounding specific program selection in the robots in to complete the tasks. For example, a robot may need to clean the bedroom floor and not the lobby between 12PM to 4PM. Hence, the robot must be aware of its surroundings, and should not clean the lobby, if the time is between 12PM and 4PM, but it can clean the bedroom in the meanwhile. This model utilizes the combination of histogram of oriented gradients (HoG) and Recurrent neural networks to classify the scenes in the target image. The HoG feature descriptor is a color based feature descriptor, which describes the features of target image in block-wise pattern. The image is divided into smaller blocks, and HoG of each block is computed. The HoG matrix represents a reduced size matrix than the original images in the binarized paradigm. This matrix is further converted to the vector in order to prepare the feature matrix for multiple images altogether. HoG features obtained from all of the images in the training and testing datasets are analyzed using the HoG feature descriptor, and the final feature matrices are prepared for both training and testing data. The recurrent neural network (RNN) is utilized to classify the pattern (from images). A RNN classifier can be classified as sub-class of ANN (artificial neural network), which has the ability to create the unit-oriented connections between the neurons in order to realize the directed cycle. This behavior is known as the exhibit dynamic temporal behavior. RNNs utilize the internal memory for the processing of subjective sequences from the input data. This ability empowers RNNs to handle the unsegmented and non-connected component structures with higher accuracy than any other form of ANNs. A typical LSTM (Long short-term memory) based ANN is considered as the deep learning alternative with ability to utilize the vanishing gradients to resolve the pattern recognition problem. LSTM model is complimented by the RNNs, which can be explained as "forget" gates. These forget gates are used to process the information under the backward propagation paradigm. The algorithm of the proposed model is described below:

Algorithm 1: Indoor Scene Recognition with Recurrent Neural Networks

1. Input testing image matrix
2. Divide the image matrix to smaller blocks of MxN size, for example 2x2, 3x3 or 4x4
3. Create an empty matrix of zeros equal to the size of input image → hogM
4. Run the iteration for each of the block
    a. Acquire the pixels in the current block
    b. Compute the mean of pixels in the current block
    c. Covert the pixels with value higher than or equal to the new value of 1
    d. Covert the pixels with value lower than to the new value of 0
    e. Update the pixels accordingly in hogM matrix
    f. If it's the last iteration
        i. Exit the loop
    g. Otherwise
        i. Go to 4(a)
5. Transform hogM matrix to hogV vector
6. Acquire the training image database
7. Create a empty training matrix (2-D) → trainMat
    a. Row size equals to the number of images
    b. Column size equals to the total number of pixels
8. Run the iteration of each of training image to extract the HoG features
    a. Create an empty image matrix for trM to handle the features of current image in training database
    b. Run the following on the current training image

i. Acquire the pixels in the current block
ii. Compute the mean of pixels in the current block
iii. Covert the pixels with value higher than or equal to the new value of 1
iv. Covert the pixels with value lower than to the new value of 0
v. Update the pixels accordingly in trM matrix
vi. If it's the last iteration
    1. Exit the loop
vii. Otherwise
    1. Go to 4(a)

c. Cover the trM matrix to trV vector
d. Update the corresponding feature row in trainMat for current image with trV vector
e. If it's the last image
    i. Exit the loop
f. Otherwise
    i. Goto 8(b)(i)

9. Create the RNN-LSTM object from the corresponding library
10. Set the input parameters for RNN-LSTM to perform the classification of the given data
11. Train the RNN-LSTM classifier with the training matrix trainMat
12. Test the RNN-LSTM classifier with hogV feature
13. Analyze the output matrix returned by RNN-LSTM
14. Return the classification result

IV. RESULTS AND DISCUSSION

The results of the proposed model have been analyzed using the various performance indicators as well as the statistical type 1 and type 2 errors. The statistical type 1 and 2 errors include the true positive (TP), false positive (FP), false negative (FN) and true negative (TN) cases, which indicates all four aspects of the results to represent the false or correct acceptance and reject of the test cases. Two experiments are performed over the proposed model for the purpose of cross-validation with 50 and 100 test cases (image samples), which are drawn randomly by using the random permutation series generation for the selection of the testing candidates. The RNN-LSTM model for 50 cases has been recorded with 44 TPs, 2 TNs, 3 FPs and 1 FN, whereas the SVM, KNN and Naïve Bayes are recorded with 41, 38 and 39 TPs, 3, 2 and 4 TNs, 2, 2 and 3 FPs and 4, 8 and 4 FNs respectively.

Table 1: Performance Analysis of multiple classification models over 50 test cases

PARAMETER	SVM	KNN	NAÏVE BAYES	RNN-LSTM
TP	41	38	39	44
TN	3	2	4	2
FP	2	2	3	3
FN	4	8	4	1
TOTAL	50	50	50	50
Accuracy	0.88	0.80	0.86	0.92
Precision	0.95	0.95	0.93	0.94
Recall	0.91	0.83	0.91	0.98
F1-Measure	0.93	0.88	0.92	0.96

The RNN-LSTM model is discovered with 92% accuracy, 94% precision, 98% recall and 96% of F1-measure, which is greatest among the four classification models. The second best classification model (SVM) has been recorded with 88% accuracy, 95% precision, 91% recall and 93% F1-measure, where the proposed RNN-LSTM model outperformed the SVM with nearly 1-7% on different performance indicators. In second experiment with 100 test cases, the RNN-LSTM model has been recorded with 88 TPs, 6 TNs, 2 FPs and 4 FN, whereas the SVM, KNN and Naïve Bayes are recorded with 81, 77 and 80 TPs, 5, 4 and 6 TNs, 3, 5 and 7 FPs and 11, 14 and 7 FNs respectively.

Table 2: Performance Analysis of multiple classification models over 100 test cases

PARAMETER	SVM	KNN	NAÏVE BAYES	RNN-LSTM
TP	81	77	80	88
TN	5	4	6	6
FP	3	5	7	2
FN	11	14	7	4
TOTAL	100	100	100	100
Accuracy	0.86	0.81	0.86	0.94
Precision	0.96	0.94	0.92	0.98
Recall	0.88	0.85	0.92	0.96
F1-Measure	0.92	0.89	0.92	0.97

For 100 test cases, the RNN-LSTM model is discovered with 94% accuracy, 98% precision, 96% recall and 97% of F1-measure, which is highest from all four classification models. The nearest classification model (SVM) has been recorded with 86% accuracy, 96% precision, 88% recall and 92% F1-measure, where the proposed RNN-LSTM model outperformed the SVM with nearly 2-8% on different performance indicators.

V. CONCLUSION

The proposed model has been designed around the application of RNN, specifically long-short term memory (LSTM) based RNN to determine the indoor scenes by matching the testing image with the large pre-classified database of indoor scenes. Total 67 types of indoor scenes are present in the CVPR 2009 dataset, out of which all of the indoor scenes are incorporated for the testing and training purposes in the proposed model. The proposed model has been tested with 50 and 100 test cases, which are drawn randomly from the dataset of above 15000 indoor scene images. The performance analysis includes the statistical assessment, where proposed model has been recorded with the best results with highest true positive cases (44 and 88) and true negative cases (2 and 6) in comparison to the other algorithms, which creates the leading performance engine for indoor scene classification. The proposed model has been found highly accurate with nearly 92% and 94% of accuracy in the case of 50 and 100 test cases respectively. Also, the F1-measure of 96% and 97% is found to be highest among all of the classification algorithms for 50 and 100 test cases respectively. In the case of 100 test cases, the proposed RNN-LSTM model has been recorded with 92% accuracy and 96% of F1-measure against the SVM (88% and 93%), KNN (80% and 88%) and Naïve Bayes (86% and 92%).

VI. REFERENCES

[1] [1] Espinace, Pablo, Thomas Kollar, Nicholas Roy, and Alvaro Soto. "Indoor scene recognition by a mobile robot through adaptive object detection." Robotics and Autonomous Systems 61, no. 9 (2013): 932-947.

[2] [2] Giannoulis, Dimitrios, Dan Stowell, Emmanouil Benetos, Mathias Rossignol, Mathieu Lagrange, and Mark D. Plumbley. "A database and challenge for acoustic scene classification and event detection." In Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European, pp. 1-5. IEEE, 2013.

[3] [3] Antanas, Laura, M. Hoffmann, Paolo Frasconi, Tinne Tuytelaars, and Luc De Raedt. "A relational kernel-based approach to scene classification." InApplications of Computer Vision (WACV), 2013 IEEE Workshop on, pp. 133-139. IEEE, 2013.

[4] [4] Gupta, Saurabh, Pablo Arbelaez, and Jitendra Malik. "Perceptual organization and recognition of indoor scenes from rgb-d images." In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp. 564-571. IEEE, 2013.

[5] [5] Juneja, Mayank, Andrea Vedaldi, C. V. Jawahar, and Andrew Zisserman. "Blocks that shout: Distinctive parts for scene classification." In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp. 923-930. IEEE, 2013.

[6] [6] Monadjemi, Amir, B. T. Thomas, and Majid Mirmehdi. Experiments on high resolution images towards outdoor scene classification. Technical report, University of Bristol, Department of Computer Science, 2002.

[7] [7] Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern classification. John Wiley & Sons,, 1999.

[8] [8] Fitzpatrick, Paul. "Indoor/outdoor scene classification project." Pattern Recognition and Analysis.

[9] [9] Quattoni, Ariadna, and Antonio Torralba. "Recognizing indoor scenes." InComputer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 413-420. IEEE, 2009.

[10] [10] Li, Li-Jia, Hao Su, Yongwhan Lim, and Li Fei-Fei. "Objects as attributes for scene classification." In Trends and Topics in Computer Vision, pp. 57-69. Springer Berlin Heidelberg, 2012.

[11] [11] Antanas, Laura, Marco Hoffmann, Paolo Frasconi, Tinne Tuytelaars, and Luc De Raedt. "A relational kernel-based approach to scene classification." InApplications of Computer Vision (WACV), 2013 IEEE Workshop on, pp. 133-139. IEEE, 2013.

[12] [12] Mesnil, Grégoire, Salah Rifai, Antoine Bordes, Xavier Glorot, Yoshua Bengio, and Pascal Vincent. "Unsupervised and Transfer Learning under Uncertainty-From Object Detections to Scene Categorization." In ICPRAM, pp. 345-354. 2013.

[13] [13] Zhang, Lei, Xiantong Zhen, and Ling Shao. "Learning object-to-class kernels for scene classification." Image Processing, IEEE Transactions on 23, no. 8 (2014): 3241-3253.

[14] [14] Li, Li-Jia, Hao Su, Li Fei-Fei, and Eric P. Xing. "Object bank: A high-level image representation for scene classification & semantic feature sparsification." In Advances in neural information processing systems, pp. 1378-1386. 2010.
[15] [15] Alberti, Marina, John Folkesson, and Patric Jensfelt. "Relational approaches for joint object classification and scene similarity measurement in indoor environments." In AAAI 2014 Spring Symposia: Qualitative Representations for Robots. 2014.
[16] [16] Russakovsky, Olga, Yuanqing Lin, Kai Yu, and Li Fei-Fei. "Object-centric spatial pooling for image classification." In Computer Vision–ECCV 2012, pp. 1-15. Springer Berlin Heidelberg, 2012.
[17] [17] Espinace, Pablo, Thomas Kollar, Nicholas Roy, and Alvaro Soto. "Indoor scene recognition by a mobile robot through adaptive object detection." Robotics and Autonomous Systems 61, no. 9 (2013): 932-947.
[18] [18] C. Fredembach, M. Schroder, S. Susstrunk, Eigenregions for image classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (12) (2004) 1645–1649.
[19] [19] A. Mojsilovic, J. Gomes, B. Rogowitz, Isee: perceptual features for image library navigation, in: SPIE Human Vision and Electronic Imaging, 2002.
[20] [20] T. Kollar, N. Roy, Utilizing object–object and object–scene context when planning to find things, in: International Conference on Robotics and Automation, 2009.
[21] [21] P. Dollar, Z. Tu, P. Perona, S. Belongie, Integral channel features, in: British Machine Vision Conference, 2009.
[22] [22] S. Vasudevan, R. Siegwart, Bayesian space conceptualization and place classification for semantic maps in mobile robotics, Robotics and Autonomous Systems 56 (2008) 522–537.
[23] [23] P. Espinace, T. Kollar, A. Soto, N. Roy, Indoor scene recognition through object detection, in: IEEE International Conference on Robotics and Automation, 2010.