Scene Labeling using H-LSTM by Predicting the Pixels using Various Functions

N. Shanmugapriya, D. Chitra

Abstract: ***Scene Labeling plays an important role in Scene understanding in which the pixels are classified and grouped together to form a label of an image. For this concept, so many neural networks are applied and they produce fine results. Without any preprocessing methods, the system works very well compared to methods which are using preprocessing and some graphical models. Here the neural network used to extract the features is Hierarchical LSTM method, which already gives greater result in Scene parsing in the existing method. In order to reduce the computation time and increase the Pixel accuracy H-LSTM is used with Makecform and Softmax functions were applied. The color transformation is applied using the Makecform function. The color enhancement of images has given object as input to H-LSTM function to identify the objects based on the referential shape and color. H-LSTM constructs the neural network by taking the reference pattern and the corresponding label as input. The pixels present in the neighbourhood identified with the help of neural network. In this method, the color image is converted into greyscale and then the Hierarchical LSTM method is applied. Therefore, this method gives greater results when it is implemented in Matlab tool, based on pixel accuracy and computation time when compared to other methods.***

Keywords: Scene Labeling, RNN, CNN, LSTM, P-LSTM and MS-LSTM

I. INTRODUCTION

Nowadays Internet grows rapidly and the amount of image data also increases and this makes the Image analysis as an important task. The Image analysis task includes various tasks such as image classification and Image segmentation to extract the desired visual information from various size, qualities, and semantics images. In this Image analysis study, scene understanding can be performed using various steps like Scene parsing, Scene Labeling and 3D construction of images. A Scene will be defined in various methods and it has details like objects, regions, location and non-visual attributes also. On the whole the entire image which consists of all these attributes comprises a scene. For the purpose of identifying the objects present in the scene understanding is used which leads to Scene Labeling.

Manuscript published on 30 September 2019
* Correspondence Author

N. Shanmugapriya*, Assistant Professor, Oxford Engineering College, Trichy, Tmailnadu, India. E-mail: shanmugapriyavinod@gmail.com
Dr. D. Chitra, Professor and Head, Department of Computer Science and Engineering, P.A. College of Engineering and Technology, Pollachi, Tmailnadu, India. E-mail: chitrapacet@gmail.com

Scene Labeling can be performed by dividing the regions of the image by their meaningful information and labeling pixels based on their regions. Humans can understand the regions by distinguishing them visually by the spatial dependencies between them. For example, visually identical regions like “sky” and “Ocean” can be predicted by assuming their place present on the scene.

This can be useful only for the image which is considered in global context. So the low level features can also be considered. The pixel labels cannot be identified by low level features alone; further they are mainly used for the construction of similarity of pixels and on their spatial dependencies using some graphical model like Markov Random Field (MRF) [1] and Conditional Random Field (CRF)[2]with the help of various neural networks.

In Recurrent Neural Networks [3], the color and texture features from over segmented regions are merged. Of all Deep learning methods, Convolutional Neural Networks (CNNs) [4] using the concept of end-to-end learning is the successful one. The main applications of this method are image segmentation, recognition of the object, and face and scene labeling of natural images.

Classification CNNs are smoothly transformed into Fully Convolutional Networks (FCNs) [5] by replacing fully-connected layers with 1x1 Convolutional layers and this process involves taking an image of arbitrary size and predicting a semantic label map.

Though FCNs have provided a near-perfect result in semantic segmentation, it has its own limitations including issues in modeling distant contextual regions.

In the method of Feedforward networks, inputs are applied to the network and using some functions it is then converted into an output with the help of supervised learning. Then this the output becomes a label and this name will be applied to the input. This method maps the raw data with each category to recognize the patterns.

After Feedforward Network, the Recurrent networks(RNN) are introduced. The key difference between Feedforward and RNN is that feedback loops are connected ingesting their own outputs moment after moment as input. Other than this difference the Recurrent Networks have memory. The sequence of information is stored into the memory for further use is the purpose of adding this memory to neural but it is not needed by Feedforward networks.

Long Short-Term Memory units, or LSTMs [6], the technique which was developed by the German researchers for the purpose of solving the vanishing gradient problem. The main advantage of this network is that it is used to preserve the error, which can be backpropagated through time and layers.

The preserved error makes the recurrent network work step by step and it can be linked to causes and effects. This will be the more serious problem of machine learning.

LSTMs[6]stores the information in the gated cell in the form of normal flow. This information may be in the form of readable or writable form present in the cell like memory.

Long Short Term Memory (LSTM)[6] recurrent neural network architecture considers the local dependency and global dependency that is pixel-by-pixel and label-by-label making into a single process for the purpose of scene labeling. It does not include other methods of multiscale or different patch sizes. It solves the scene labeling problem without the help of human or machine.

For the purpose of sequence learning only the network LSTM[6] is proposed. It consists of cells that are recurrently connected for the purpose of learning the dependencies. This can happen between two time frames. It is then transferred to the next frame. This information is stored and retrieved with the help of memory for short or longer time. The main applications of LSTM networks are handwriting and speech recognition.

In order to make the LSTM for Labeling of images present in the Scenes, a novel Hierarchical LSTM (H-LSTM)[7] recurrent network is used. This network simultaneously parses a still image, makes that into a series of geometric regions, and predicts the interaction relations among these regions.

II. RELATED WORK

A new method for Full Scene Labeling or Scene Parsing [4] used a Multiscale Convolutional Network to extract dense feature vectors and a tree of segments is computed from a graph of pixel dissimilarities. The segment is encoded by feature vectors and the classifier is then applied to all the feature vectors. It produces the categories of segments and its impurity by measuring the entropy. Using this each and every pixel is labeled and segmented to identify the class of the pixel.

Later, Kekec¸ et al. [8] make changes to CNNswith the help of combining two CNN models. This model learns context information and visual features by using separate networks. This approach improves the accuracy. It involves pre-processing steps for the help of learning. This method selects the random patches for every class and a specified color for each input data to do the learning.

Recurrent convolutional neural network [9] introduced the concept of considering a large input context but it limits the capacity of the model. It combines the Recurrent architecture with that of convolutional neural networks by sharing the same parameters. This method does not require any engineered features. Because, this type of architectures trains filters in complete manner. Here this type of network does not depend on the prediction phase but it needs only the forward evaluation. The above two are the advantages of the network.

The use of deep learning techniques [10] was considered to deal with scene labeling. In this technique segments are recursively merged for the purpose of allocating a label category wise. Here the architecture is different, that is ReNet architecture, which is used to parse the scene.

A deep learning strategy [11] was proposed to do the scene parsing. Scene parsing is the technique of allocating a label for each and every pixel based on their class. The network used for this purpose is a deep convolutional network. This network labels the complex scenes with the help of supervised learning technique. This method involves engineered features not the hand crafted features. When comparing the CRF method this is the main advantage. Apart from this advantage two more advantages are given below. (i) it requires the forward evaluation not the label space (ii) it does not involve the calculation of normalization factor because of the use of Stochastic Gradient Descent (SGD).

In recent years Byeon et al., 2015 [12] developed the multi-dimensional LSTM for the purpose of capturing the dependencies locally. It considers the natural scene images and labels them by the use of spatial dependencies. This network can be applied for the concept of classification and segmentation by learning the spatial model parameters. Over the raw RGB values the network captures the local contextual information as well as the global and it can work well for complex images. It gives less computational complexity when compared to previous methods in the Stanford Background and the SIFT Flow datasets.

A hierarchical approach [13] was proposed for labeling semantic objects and regions in scenes. This approach used a decomposition of the image in order to encode relational and spatial information. It bypassed a global probabilistic model and instead directly trained a hierarchical inference procedure inspired by the message passing mechanics of some approximate inference procedures in graphical models.

Dense RNNs [17]identifies contextual dependencies between the images using dense connections that is used to increase the power. For this purpose, this method propose Dense RNN, the model which gives more weightage to contextual dependencies and gives less importance to unconcerned dependencies.

III. SYSTEM DESCRIPTION

The H-LSTM[7]differs from the above papers in the method of training an integrated network to solve the problem of geometric region labeling and relation prediction. The two phases of H-LSTM is P-LSTM and MS-LSTM. With these two phases the long-range spatial dependencies can be captured for the hierarchical feature representation which can be applied on the pixels and multi-scale super-pixels. The proposed system contains the following sections:3.1 Introduction about H-LSTM,3.2 P-LSTM Pixel LSTM for Labeling of Image, 3.3 MS-LSTM Multiscale Superpixel LSTM for Interaction Relation Prediction, and 3.4 Proposed System for Scene Labeling.

The P-LSTM gets the picture information from neighbouring pixels and these contextual information are stored and are further used for feature extraction in the later layer. The MS-LSTM using this information hierarchically reduces the redundancy and extracts the relations in different layers. The proposed system uses the H-LSTM concept of layers to produce the labeling of the images.

3.1 Introduction to H-LSTM

The Hierarchical LSTM contains two phases: P-LSTM and MS-LSTM. This system consists of a stack of convolutional and pooling layers. Through this the input image is passed. The output gives the feature maps. These feature maps are then given as inputs in P-LSTM [7] and MS-LSTM [7] in a shared mode, and the final result is provided in the form of the pixel-wise labeling of images

In this P-LSTM technique the last two layers are fully connected layers and they it produce the feature maps for the input image. The feature maps are then placed into another layer called transition layer. For each and every position hidden and memory cells were produced by this layer. By producing this it ensures that that the first P-LSTM layer input states is same as that of next P-LSTM layer.

The output will be inserted into the five stacked P-LSTM layers which increases the receptive field of each position and sense the large contextual region.

3.3 Multiscale Superpixel LSTM for Interaction Relation Prediction

The MS-LSTM [7]used to find out the interaction relation between the identified pair-wise super-pixels and after that finds out the functional boundaries of the pixels which are present in the same group.

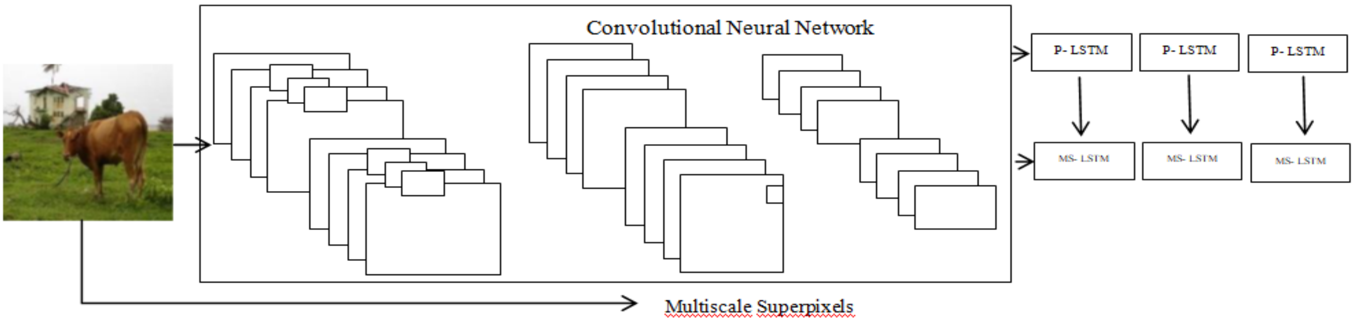


Figure 1: Hierarchical LSTM architecture diagram. In this first input image is fed into a set of convolutional layers. It produces feature maps. This output is passed into the both layers namely Pixel LSTM layers and Multi-scale Super-pixel LSTM. This method generates the pixel labeling and dependence between the regions for further evaluation.

and its interaction relation between adjacent regions, respectively. scale. After that these features are passed onto the LSTM units. This method exploits the spatial dependencies. Other than this the mapping of superpixel can be used to extract the higher-level contextual dependencies.

3.2 Pixel LSTM for Labeling of Image

In Hierarchical LSTM the first part is finding out the local information about each and every individual pixels and further identify the interactions between the pixel based on the context. Here the term j represents the features of each position. There are 8 spatial and one depth hidden cells from local and previous layer also involved. When considering the previous layer the above hidden cells produce the features which are identified by the term "depth". The layers which are present in MS-LSTM [6]are mapped with the feature maps of convolutional layer with the layer of P-LSTM. There are five stacked layers of MS-LSTM that can be applied for feature extraction. It is possible with the help of contextual dependencies of different scales. The different scales (i.e. 16,32, 48, 64 and 128) can be extracted by the over-segmentation algorithm.

Here in each superpixel scale represents the average count of pixels present. This can be done by different types of MS-LSTM layers. Apart from this the hidden cells present in each layer will be enhanced by the previous output taken from P-LSTM layer. This can be inserted into MS-LSTM layers and it is affected by various degrees of context. This makes the model consider the semantic information locally. Atlast, based on the prediction of relation the optimization of adjacent superpixels was done.

3.4 PROPOSED SYSTEM

Given a set of inputs, pixel value of the image of each and every column is stored in reading pixel variables. The stored variables are converted into grayscale to get the linear color differentiation. The grey scale color enhancement is applied to remove the noise and distortion from the image. And in order to differentiate the object in the image, color transformation is applied using the Makecform function.

The Makecform function supports conversions between members of the family of device-independent color spaces. Makecform also supports conversions to and from the sRGB and CMYK color spaces. To perform a color space transformation, pass the color transformation structure created by Makecform asan argument to the Applycform function.

The color enhancement of images has given object as input to H-LSTM [7] function to identify the objects based on the referential shape and color. H-LSTM constructs the neural network by taking the reference pattern and the corresponding label as input.

The initial neurons are created with the reference to random sample patterns with the labels in the input layer.

In the Hidden layer the activation structure is defined, through the learning progress which relatively memorizes the context information. In the hierarchical structure processing, the input image is handled level by level in each iteration.

In the first iteration the pixel representations are parsed with minimum correlation to the nearest correlation pixel points. The correlation is activated using the Softmax[15]function which increases the learning rate of the hidden layer. Once the relativity is identified, the interaction relation is invoked to determine the spatial and depth dimension in each layer.

$$y_j = \mathrm{softmax}(F(h_j ;W_{label})) \qquad \ldots (3)$$

where y_j is the term used for probability of geometric surface which is predicted by the j-th pixel, and W_{label} is the network parameter. The function for transformation is denoted by F().

Then the propagated function generates the inputs in Hidden cells in each neighbor positions. Then by mapping the initial connection in geometric representations, the geometric surface is predicted.

During the prediction process, the functional boundaries are identified to connect with the similar hidden cells. In each iteration, the hidden layer is validated to check the probability of changes in the hidden cell input.

If there is a change in the current layer then the output of the hidden layer is normalized using Softmax in terms of generalized of logistic function with the exponential function.

If the iterative process of the hidden layer reached the maximum, the termination of the iteration is then invoked to identify the output current context.

The activation function is iteratively used in each iteration, which determines the local inference of each pixel point with neighbor pixel point with the same or similar category. The interaction relation of the pixel point is validated using the centroid by checking the replicated pixels.

If the pixel point possesses the relativity, the current output is activated with the linear mapping of the corresponding labels.

In the activation process, the spatial structure of the object is determined and projecting geometric with texture mapping is combined together to get the output. After mapping the projecting geometric, the next level mapping is applied by considering the pixel value of the current level with the best correlation. Then the same operations are repeated in each iterative process of the hidden layer to obtain accurate output and it is presented in the output.

Using this method the Mean Square Error(MSE), Peak Signal to Noise Ratio, Specificity, Classifier Accuracy, and Pixel Accuracy can be calculated. Here the PSNR and MSEare used to compare the squared error between the original image and the reconstructed image. There is an inverse relationship between PSNR and MSE. The formula for calculating the PSNR and the MSE are given below[18].

$$\mathrm{MSE} = \sum_{y=1}^{M}\sum_{x=1}^{N}[I(x,y) - I'(x,y)]^2$$

where, I(x,y) is the original image, I'(x,y) is its noisy approximated version and M,N are the dimensions of the images value for MSE implies lesser error.

$$\mathrm{PSNR} = 10\log_{10}(\mathrm{MAXi}^2/\mathrm{MSE})$$

where, MAXi is the maximum possible pixel value of the image. A higher value of PSNR is always preferred as it implies the ratio of Signal to Noise will be higher. 'signal' here is the original image, and the 'noise' is the error in reconstruction.

IV. EXPERIMENT AND RESULT

The proposed method has taken into account the two datasets for the purpose of scene labeling, namely the SIFT Flow Dataset [16] and Stanford Background Dataset[19]. The SIFT Flow is a larger dataset which consists of 2688 images with 256 X 256 pixels and 33 semantic labels. The Stanford dataset contains classification of 8 different classes which contains totally 715 images. The scenes have approximately 320 X 240 pixels. All these networks are trained by sampling patches which are surrounded by a pixel which is chosen randomly from a randomly chosen image from the training image set. The following is the table of values for the MSE, PSNR, Specificity, and Classifier accuracy for the random images chosen.

Table 4.1 comparison values of images chosen from the dataset

IMAGE	MSE	PSNR	SPECIFICITY	CLASSIFIER ACCURACY	PIXEL ACCURACY
Image 1	1.22	47.31	99.54	98.48	99.77
Image 2	1.27	47.11	99.49	50.00	99.74
Image 3	1.31	47.00	99.93	96.42	99.96
Image 4	1.19	47.41	98.64	98.00	99.31
Image 5	1.10	47.75	99.94	99.00	99.97

The above Table 4.1, explicates the results based on the comparison between the proposed and other existing approaches with reference to MSE, PSNR, Specificity, Classifier Accuracy, and pixel accuracy. When compared to existing approach H-LSTM techniques for Scene Parsing the proposed method gives greater accuracy.

From the above values it is clearly visible that the proposed approach gives low MSE values and high PSNR values. And it is understood that a low MSE value and a high PSNR value produce object detection in accurate manner. So, the proposed approach has succeeded to produce a better result.

The experimental result of the proposed approach is given below by considering the basic objects such as Sky, Building, Road, and Tree. Before labeling, those objects of the input images are first converted into Grayscale and then further enhanced for labeling purpose.

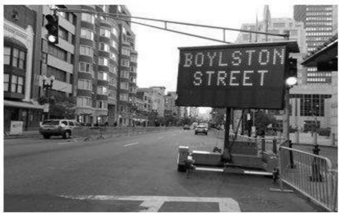
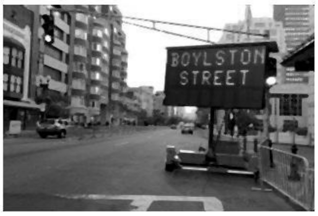
Original Image	Greyscale Image	Contrast Enhanced Image
1-original image 	1-Grey Scale 	1-Contrast Enhanced Image
2-original image 	2-Grey Scale 	2-Contrast Enhanced Image
3-original image 	3-Grey Scale 	3-Contrast Enhanced Image
4-original image 	4-Grey Scale 	4-Contrast Enhanced Image
10-original image 	10-Grey Scale 	10-Contrast Enhanced Image 

Fig. 4.1 Original image, Grayscale and Contrast Enhanced Image.

Here the Original image is first converted into Grayscale to get the linear color differentiation and then to contrast enhanced image to remove the noise and distortion from the image.

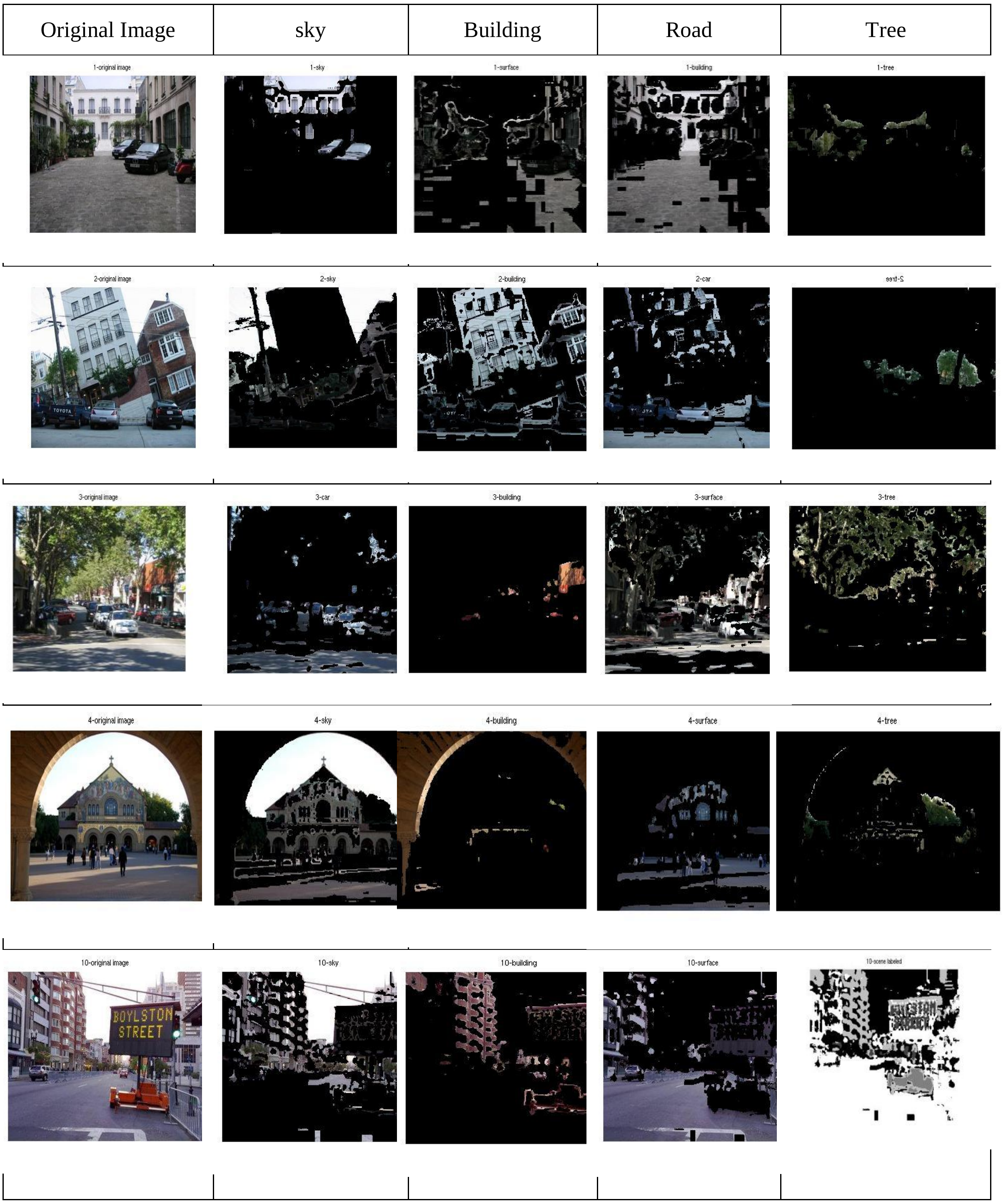


Fig. 4.2 Labeling of images with its object identification

Color transformation is applied using the Makecform function in order to differentiate the objects present in the image and then H-LSTM constructs the neural network by taking the reference pattern and the corresponding label as input.

V. CONCLUSION

This paper uses the existing approach H-LSTM with Makecform and Applycform with the help of Softmax function to get the object identification and labeled. The results of the experiment prove that the proposed method provides higher pixel-accuracy when compared with other methods. The proposed method convert the color image into Grayscale image and H-LSTM method is applied with the functions Makecform, Applycform, and softmaxfunctions. When comparing the MSE, PSNR and Pixel Accuracy with the existing approaches, it gives the optimal result. This method has proved to be a successful method to detect the objects in natural scenes, more effectively in analyzing the images and comparing their presence. As a future scope, the method can be modified without converting the image into Graysale, applying the H-LSTM method directly, which reduces the running time. Moreover this paper deals with some of the random images of the Sift Flow and Stanford dataset only. In future, it can be extended to full dataset and for the datasets which contain more images like Barcelona Dataset.

Instead of applying the above concept on the images which are converted into Grayscale, this method can be directly applied to color images.

REFERENCES

1. R. C. Dubes, A. K. Jain, S. G. Nadabar, C. C. Chen, "MRF model based algorithms for image segmentation", *Proc. Pattern Recognition*, vol. 1, pp. 808-814, 1990.
2. Xuming He, Richard S. Zemel, and Miguel A. Carreira-Perpin˜a´n, “Multiscale Conditional Random Fields for Image Labeling,” in CVPR'04 Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition, Pages 695-703.
3. S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, “Conditional random fields as recurrent neural networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1529–1537.
4. C. Farabett, C. Couprie, L. Najman, “Learnings hierarchical features for scene labeling,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1915– 1929, 2013.
5. Jonathan Long, Evan Shelhamer and Trevor Darrell, “Fully convolutional networks for semantic segmentation”, In CVPR, 2015.
6. Sepp Hochreiter and J¨urgen Schmidhuber, “Long short-term memory. Neural computation”, 9(8):1735–1780, 1997.
7. Zhanglin Peng, Ruimao Zhang, Xiaodan Liang, Xiaobai Liu and Liang Lin, “Geometric Scene Parsing with Hierarchical LSTM”, in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), 2016.
8. T. Kekec¸, R. Emonet, E. Fromont, A. Tr´emeau, C. Wolf, and F. Saint-Etienne, “Contextually constrained deep networks for scene labeling”, In Proceedings of the British Machine Vision Conference, 2014, 2014.
9. Pedro O. Pinheiro and Ronan Collobert, “Recurrent Convolutional Neural Networks for Scene Labeling,” International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.
10. Francesco Visin, Kyle Kastner, Kyunghyun Cho, Matteo Matteucci, Aaron Courville, Yoshua Bengio, “ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015.
11. D. Grangier, L. Bottou, and R. Collobert, “Deep Convolutional Networks for Scene Parsing,” In ICML 2009 Deep Learning Workshop, 2009.
12. Wonmin Byeon, Thomas M. Breuel,Federico Raue, and Marcus Liwicki, “Scene labeling with LSTM recurrent neural networks”, In CVPR, 2015.
13. D. Munoz, J. Bagnell, and M. Hebert, “Stacked hierarchical labeling,” ECCV 2010, Jan 2010.
14. Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji and Yichen Wei, “Fully Convolutional Instance-aware Semantic Segmentation”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
15. Junjie Yan, Yinan Yu, Xiangyu Zhu, Zhen Lei and Stan Z. Li, “Object detection by Labeling Superpixels,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Oct 2015.
16. Ce Liu, Jenny Yuen and Antonio Torralba, “SIFT Flow: Dense Correspondence across Scenes and its Applications,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011.
17. Heng Fan, Peng Chu,Longin Jan Latecki, Haibin Ling, “Scene Parsing Via Dense Recurrent Neural Networks With Attentional Selection,” in IEEE Winter conference on Computer Vision, Jan 2019.
18. Dibya Jyoti Bora, Anil Kumar Gupta, Fayaz Ahmad Khan, “An Efficient Approach toward Color Image Segmentation with Combined Effort of Soft Clustering and Region Based Techniques using L Channel of LAB Color Space,” International Journal of Computer Applications, Volume 128 – No.12, October 2015.
19. S. Gould, R. Fulton, D. Koller, “ Decomposing a Scene into Geometric and Semantically Consistent Regions,” Proceedings of International Conference on Computer Vision (ICCV), 2009

AUTHOR PROFILE

Mrs. N. SHANMUGAPRIYA is working as an Associate Professor in department of IT, Oxford Engineering college, Trichy. She is pursuing her doctorate in Anna University, Chennai and obtained her M.E in Computer Science and Engineering. She has published 6 papers so far in International and national Conferences and Journals. She has Guided 15 projects in both under graduate and postgraduate students towards their project work. She has 12 years of teaching experience. Her area of interest is Image Processing and Pattern Recognition.

Dr. D. CHITRA is working as a Professor and Head in the department of CSE, P. A. College of Engineering and Technology. She received her Doctor of Philosophy from Anna University, Chennai and Master’s Degree in Computer Science and Engineering. Her areas of interest include Digital Image Processing, Pattern Recognition, Computer Vision, Data Mining and Grid & Cloud Computing. She has 17 years of experience in teaching and published 70 papers in National and International Conferences and Journals. She is a member of IEEE, ISTE, CSI, IAENG and IRED. She has guided 67 projects in both UG and PG, and currently 9 research scholars pursuing Ph.D. She is a reviewer for many Journals and Conferences. She attended 25 national and International seminars/conferences/workshops. She has received awards such as Best Circuit Faculty Award SIAA (ASDF), Shri. P. K. Das Best Faculty Award, Best Faculty Award in Kongu Engineering College and Best Faculty Award in P. A. College of Engineering and Technology. She also organized 21 programmes sponsored by AICTE, Anna University, CSIR, DRDO, ICMR, and INSA.

Retrieval Number C4285098319/19©BEIESP
DOI: 10.35940/ijrte.C4285.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication