Vol 8, Issue 3, 2025 **Impact Factor: 8.535** DOI: https://doi.org/10.31426/ijamsr.2025.8.3.8144

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

A Deep Recurrent Framework for Robust Scene Labeling and Image Segmentation

Mahesh D [1]

Research Scholar, Department of Computer Science Engineering,
Sri Satya Sai University of Technology & Medical Sciences, Sehore, M.P., India.

Dr. Harsh Lohiya [2]

Research Supervisor, Department of Computer Science Engineering,
Sri Satya Sai University of Technology & Medical Sciences, Sehore, M.P., India.

ABSTRACT

The process of scene labelling involves assigning a category to each pixel in a picture. In order to identify objects in images, this work employs the ReNet architecture. The four RNNs that make up this ReNet architecture take the role of each convolutional layer (CNN) and combine features from lower layers in various ways. The picture is first oversegmented into superpixels after feature extraction, and then it is labelled into individual superpixels. The Conditional Random Field statistical method will be used to investigate and take advantage of the dependencies to the neighbouring superpixel labels. Using this method on the SIFT Flow and Stanford Background Dataset datasets results in more precise pixels, but a little longer time to segment and label the pictures.

Keywords: *Superpixel, Full Scene labeling, Convolutional Layer, Conditional Random Field.*

I. INTRODUCTION

There are many real-world uses for image understanding, making it a crucial job. Performing full-scene labelling, sometimes called scene parsing, is a crucial step in interpreting a picture. It involves assigning a category to each pixel in the image. Delineating and labelling each item and territory follows flawless scene parsing. The fact that scene parsing incorporates multi-label identification, detection, and segmentation into a single operation is one of its challenges.

A basic subject that has been explored extensively is scene labelling, which aims to densely identify everything in a scene. Outside settings have been the primary focus of scene labelling studies. Indoor scene labelling has been mostly disregarded, maybe with the exception of the Manhattan world layout, despite the fact that most people spend their time inside. This is due in part to the fact that interior scenes provide more of a problem, what with factors like inadequate lighting, a lack of distinguishing characteristics, and a wide variety of scene kinds.

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

In order to comprehend images, accurate scene labelling is necessary. Scene labelling involves dividing a picture into its relevant areas and then associating each pixel with its respective location. Choosing pixel labels using low-level attributes, such colour or texture, taken from a tiny window around pixels is probably not the only option. Differentiating between "grass" and "tree" or "forest" would be a real pain in such environment, for example. In reality, it is the spatial interdependence between areas that allow people to perceptually differentiate them. For example, based on their location with relation to the top or bottom of a picture, visually identical sections may be projected as "sky" or "ocean" respectively.

Based on the similarities of the pixels' low-level properties and their spatial connections, a graphical model is often used to generate a higher-level representation of scenes, which is their global context. The graphical models use the similarities between nearby segments to build the global dependencies. Two of the most well-known methods based on graphs are conditional random fields (CRF) and markov random fields (MRF). Presegmentation, superpixels, or candidate regions are often necessary for most of these approaches.

II. REVIEW OF LITERATURE

Fernandes, Rodrigo et al., (2024). Accurate image and video annotation is foundational to computer vision and AI, and it is dependably used to build trustworthy machine learning models. Important to automatic annotation are techniques for object tracking and image retrieval, which greatly improve the process's efficiency and accuracy. This paper delves into the methods for picture acquisition and object tracking. It explores how these technologies may collaborate to enhance the efficacy and precision of image and video dataset annotation processes. The annotation process may be automated using object tracking, which follows moving items in a video. Image retrieval then uses this data to suggest annotations for fresh photographs. Various techniques are included in the assessment, demonstrating their effectiveness in various contexts including urban surveillance and medical investigations. It makes use of advanced machine learning and neural network methods. Despite significant advancements, challenges remain related to algorithm robustness and effective human-AI interaction. This study sheds light on the current and future state of these technologies, which aids in the enhancement of photo annotation techniques. It also demonstrates the current applications of these methods and how they might be fully used when combined.

Feng, Xuanang. (2023). Computer vision has a significant challenge with object identification due to the fact that it must locate and quantify all the interesting things in an image. With so many variables that may impact image quality (such as lighting, occlusion, and other types of interference), target recognition has always been the biggest obstacle for machine vision researchers. In light of recent developments in target detection technology, this article reviews many articles on the subject. Various target detection networks are discussed and examined. The evaluation criteria and test dataset summaries are also part of this. A summary and expectation of the whole content are included at the end of the work.

Hoeser, Thorsten & Kuenzer, Claudia. (2020). Adaptive approaches to emerging problems in Earth observation (EO) are increasingly being tackled via deep learning (DL), which is having a profound effect on several fields of study. It is challenging for new researchers to break into EO since the field

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

is both highly populated and rapidly growing, driven mostly by advances in computer vision (CV). This study summarises the progress of DL with a focus on CNN-based photo segmentation and object identification, with the intention of helping EO researchers. The survey will go on until the end of 2019, starting in 2012 when a CNN set new standards for image identification. Our goal is to facilitate the evaluation of new DL models by highlighting the connections between the most popular CNN designs and CV pillars. We also provide a short history of the most famous DL frameworks and summarise EO datasets. By demonstrating high-performing DL architectures on these datasets and commenting on breakthroughs in CV and their implications for future echolocation research, we want to close the gap between the theoretical concepts addressed in CV and their practical implementation in EO.

Aamir, Muhammad et al., (2018). Object classification using image content is one of the most challenging issues in computer vision. Data recorded in the superpixels enables object identification and location-based picture categorisation. The goal of this study was to provide a method for detecting and classifying image pixel locations using an upgraded bag of words (BOW). Once the beginning points of each image segment have been determined using superpixels, the results are sorted according to region score. In addition, both global and local features are extracted from this data using a hybrid technique that combines GIST and Scale Invariant Feature Transform (SIFT). Improved classification accuracy is achieved by the feature fusion approach's use of a weight parameter to combine local and global feature vectors. We use a supervised classification method called the support vector machine classifier to assess the proposed method. The experiment uses the VOC2007 dataset, which stands for Pascal Visual Object Classes Challenge 2007, to test the results. Because it generated high-quality classes for the positions of independent objects, the proposed technique enhanced the detection rate. Average best overlap (MABO) was 0.833 at 1,500 sites. In terms of non-rigid class categorisation accuracy, the results demonstrate that it outperformed previous techniques.

III. PROPOSED ARCHITECTURE

Dataset

Two separate fully labelled datasets, the SIFT Flow Dataset and the Stanford Background, are used to evaluate the suggested technique. Eight classes, totalling 820 photos, cover both urban and rural settings in the Stanford collection. About 320 × 240 pixels make up the sceneries. This bigger dataset, known as the SIFT Flow, has 3566 pictures, each with 256 × 256 pixels and 33 semantic labels.

Every one of these networks learned its trade secrets by randomly selecting a pixel from the training picture set to encase in a sampling patch. There are two methods used to determine the image's accuracy. I. Accuracy on a pixel level and (ii) accuracy on a class level. Pixel-wise accuracy is the proportion of accurately predicted pixels, while class-wise IoU is the average of the intersection of union of pixels over all 150 semantic categories. But in scene labelling (particularly in datasets with a high number of classes), this metric is more affected by classes that appear more often than others; for example, the class'sky' appears more frequently than'moon.

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

System Description

Together, the following preprocessing techniques—Superpixel Segmentation, Multiscale Conditional Random Field defined with reference to a set of superpixels, ReNet Architecture—capture the global contexts and exhibit its property of efficient parallelization—and allow the Scene Labelling system to extract pertinent contextual information from raw pixels. The ReNet architecture is used to extract features and segment superpixels at the same time. Then, the pixels are correctly categorised and labelled using the Conditional Random Field.

- **ReNET Architecture**

The network architectures are ideal for sequence labelling for a number of reasons, including their adaptability to different kinds of data and representations, their ability to discern sequential patterns even when sequential distortions are present, and their flexibility in using context information to decide what to store and what not to store.

The architecture of the ReNet is defined by several important parameters, such as the number of ReNet layers (NRE), the receptive field sizes (wp hp) and feature dimensionality (dRE) of each layer, the number of fully-connected layers (NFC) and the types (fFC) and numbers (dFC) of hidden units associated with each layer.

After Convolutional Neural Networks (CNNs) showed promise, particularly in computer vision, many people turned to Recurrent Neural Networks (RNNs) to represent sequential data, such audio and text. While extracting characteristics from a particular area within the whole picture, the recurrent layers comparatively take the entire image into account. Contrarily, while extracting picture features, CNN just takes the immediate context window into account.

Recurrent Neural Networks (RNNs) work by starting with an input picture and building a hierarchical representation of it using the output of each successive layer's operations on the representations derived from the layer below.

- **Superpixel Segmentation**

So that no single superpixel covers more than one item, the optimal number of superpixels to use for detection is dependent on how efficiently inference can be performed. Feature and category extraction from every segment, as well as from various combinations of nearby segments, is achieved by pre-segmentation utilising superpixels. Noisy predictions result from trying to forecast pixel attributes and categories apart from nearby segments. Therefore, it is sufficient to do a basic cleanup by making use of labelled areas of uniform colour intensity. Afterwards, we undertake location-based picture classification, superpixel-level prediction aggregation, and average superpixel-level class distribution computation.

- **Multiscale Conditional Random Field (mCRF)**

There are two types of feature functions used by standard CRFs. The first type is the state feature function, which is defined in a 2D image as f (li, X, i) and involves the label at site i and the observed image. The second type is the transition feature function, which is defined in the image and involves the labels at site i and a neighbouring site j. On the subset of label variables, label features often encode certain patterns. The label feature encodes the precise restrictions between the labels and the picture within the same area. It is a kind of potential function.

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

There is a binary hidden variable that toggles between each label feature and its corresponding value. A parametrised Conditional Probability Table (CPT) is used to encode the characteristics inside the area in order to detect the pattern of labels. Every site's label values are distributed according to a multinomial probability distribution using this CPT. Both the hidden and label variables are conditionally independent of each other.

The CRF specifies a multiplicatively combined method of combining feature predictions. To start, not every site in the same area has to have its label specified. No meaningful findings are produced by combining data that are homogeneous within the same area. Predicting characteristics of a certain location in the area is the purpose of this so-called "don't care" forecast. The second point is that the component distributions could not be as crisp as the label of any given location. Here, the product will be substantially increased depending on the value, which happens when two multinomials share a certain value. Consequently, confident labelling is achieved by the uncertain predictions.

Specifically, the mCRF framework consists of a local classifier, regional features, and global features, all of which function at various sizes.

IV. RESULT AND DISCUSSION

As shown in Figure 8, this method is applied to a selection of randomly selected photographs from the Stanford and Sift Flow datasets. A pre-segmentation step utilising Superpixels and multiscale Conditional Random Field is performed before the feature extraction of the whole picture is executed using the ReNet. Finding out whether some pixels are part of a certain group is what pre-segmentation is all about. All pixels in the same group get the same average score, which is calculated in the pre-segmentation layer. Figure 8 displays the experimental results of the suggested method for the two datasets' random pictures.

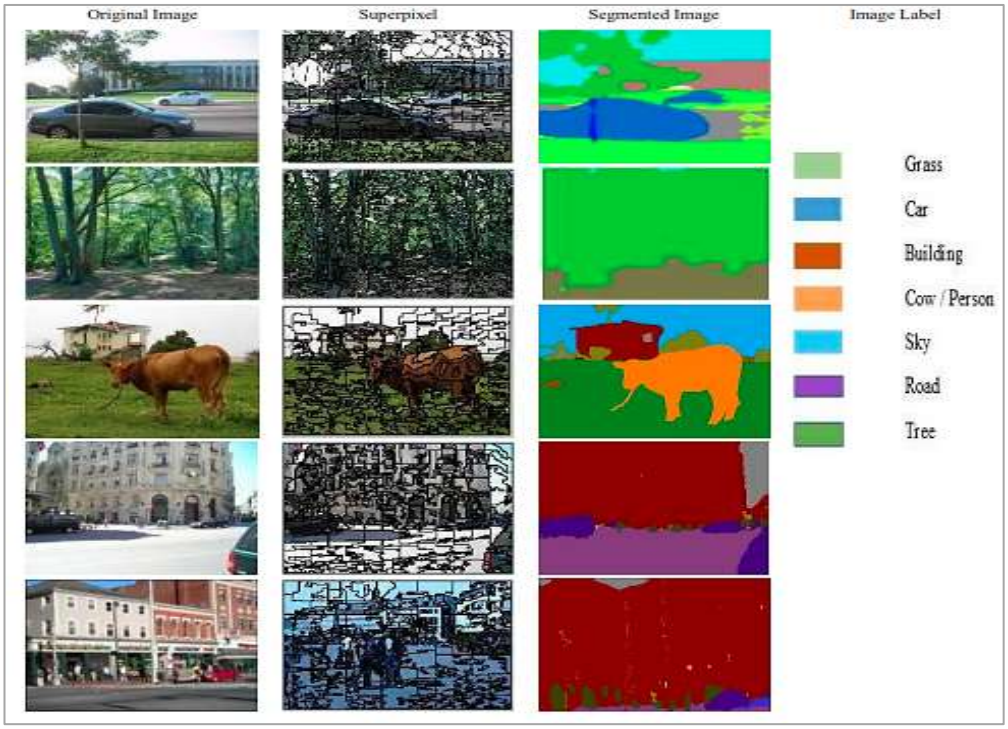

Figure 1: Experimental Output of The Proposed Approach

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

Table 1 please elaborate on the findings by comparing the suggested method with other current techniques in terms of calculation time and pixel precision. The pixel precision is great, but the calculation time is somewhat higher, when compared to all the approaches using our approach. This occurs as a result of the preprocessing procedures used prior to labelling.

Table 1: Accuracy Comparison

Method	Pixel Accuracy (%)	Computation Time (s)
Region-based Model	74.8	120–520
SuperParsing	76.9	95–260
Stacked Hierarchical Labeling	77.6	18
Relationship Prediction Model	79.1	< 480
Learning Hierarchical Features	78.4	1.2
Our Approach	82.3	9

V. CONCLUSION

By merging global contextual feature extraction with superpixel-based segmentation and multiscale Conditional Random Field (mCRF) modelling, the suggested method sought to enhance pixel-level classification accuracy. To reduce noise and preserve object boundaries, superpixel segmentation was a key component. The approach made sure that labelling judgements were more consistent and computationally meaningful by grouping pixels with comparable features. To top it all off, the multiscale CRF improved the model's capacity to leverage linkages between nearby areas and across various spatial scales, which led to more accurate predictions and better semantic coherence in the labelled output. This method shines in situations where precise pixel-level interpretation is of the utmost importance, such as autonomous navigation, intelligent surveillance, medical image analysis, and robotic vision.

REFERENCES

1. R. Fernandes, A. Pessoa, M. Salgado, A. Paiva, I. Pacal, and A. Cunha, “Enhancing image annotation with object tracking and image retrieval: A systematic review,” *IEEE Access*, vol. 6, no. 2, pp. 1–1, 2024.
2. Sharma and S. Rautaray, “Efficient object detection, segmentation, and recognition using YOLO model,” in *Lecture Notes in Computer Science*, vol. 1, no,2 pp. 145–156, 2024.
3. X. Feng, “Research of image object detection using deep learning,” *Applied and Computational Engineering*, vol. 6, no. 2, pp. 1269–1275, 2023.
4. P. Sharma, S. Gupta, S. Vyas, and M. Shabaz, “Object detection and recognition using deep learning-based techniques,” *IET Communications*, vol. 17, no.2 pp. 15–16, 2022.
5. T. Hoeser and C. Kuenzer, “Object detection and image segmentation with deep learning on Earth observation data: A review—Part I: Evolution and recent trends,” *Remote Sensing*, vol. 12, no. 2, pp. 76–94, 2020.
6. M. Aamir, Y.-F. Pu, Z. Rahman, W. Abro, H. Naeem, F. Ullah, and A. Badr, “A hybrid proposed framework for object detection and classification,” *Journal of Information Processing Systems*, vol. 14, no. 2, pp. 1176–1194, 2018.

IJAMSR

International Journal of Advanced Multidisciplinary Scientific Research (IJAMSR) ISSN:2581-4281

7. K. Sharma and N. Thakur, "A review and an approach for object detection in images," *Int. J. Comput. Vis. Robot.*, vol. 7, no. 2, pp. 196–210, 2017.
8. M. Pordel and T. Hellström, "Semi-automatic image labeling using depth information," *Computers*, vol. 4, no. 2, pp. 142–154, 2015.
9. Y. A. Yushkevich *et al*., "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 2, pp. 1116–1128, 2006.
10. P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.
11. Y. Xu, E. Saber, and A. Tekalp, "Object segmentation and labeling by learning from examples," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 627–638, 2003.
12. L. A. Vese and T. F. Chan, "A multiphase level set framework for image segmentation using the Mumford and Shah model," *Int. J. Comput. Vis.*, vol. 50 no. 2, pp. 271–293, 2002.