Class Probabilities Based Visual Features with Contextual Features for Image Parsing

Basim Azam
Centre for Intelligent Systems, School of Engineering and Technology
Central Queensland University, Brisbane,
Queensland 4000, Australia

Ranju Mandal
Centre for Intelligent Systems, School of Engineering and Technology
Central Queensland University, Brisbane,
Queensland 4000, Australia

Ligang Zhang
Centre for Intelligent Systems, School of Engineering and Technology
Central Queensland University, Brisbane,
Queensland 4000, Australia

Brijesh Verma
Centre for Intelligent Systems, School of Engineering and Technology
Central Queensland University, Brisbane,
Queensland 4000, Australia

***Abstract*— Deep networks have become one of the most promising architectures for im-age parsing tasks. Although existing deep networks consider global and local contextual information of the images to learn coarse features individually, they lack automatic adaptation to the contextual properties of scenes. In this work we present a visual and contextual feature based deep network for scene parsing. The network exploits the contextual features along with the visual features for class label prediction with class specific classifiers. The contextual features consider the prior information learnt by calculating co-occurrence of object labels both within a whole scene and between neighboring superpixels. The class-specific classifier deals with imbalance of data for various object categories and learns the coarse features for every category individually. A series of weak classifiers in combination with boosting algo-rithms are used as classifiers along with the aggregated contextual features. The experiments were conducted on the Stanford background benchmark dataset which showed that the proposed network can achieve better acc4uracy than many existing approaches.**

***Keywords*— Image parsing, scene understanding, semantic segmentation, object detection, deep learning.**

I. Introduction

Image parsing is a fundamental and important step for autonomous driving, robot navigation and traffic scene understanding. It is also known as scene parsing. Image parsing refers to recognizing and segmenting objects and stuff in an image. Robust and precise image parsing is a challenging problem. In pursuit of understanding an image, computer vision algorithms have evolved over the last decades. A heap of machine learning and deep learning algorithms have been proposed to analyze scene content in images. These approaches have helped immensely in robot vision, security, autonomous driving, human computer interaction and augmented reality. The modern deep learning-based frameworks have great potential for detailed semantic segmentation along with pixel-wise labelling, however the context representations are not explicit and yet scene parsing remains a challenging task.

State of the art image parsing architectures are mostly based on fully convolutional networks (FCNs), and they have boosted the performance for scene parsing. The state-of-the-art architectures comprise of two major components: multi-scale context module and neural network design. The context information is crucial for pixel-wise labelling tasks.

The paper presents a novel approach which learns global and local contextual features for scene parsing. Existing approaches lack the ability to capture global contextual information from the images. Also, the existing approaches do not focus on the labels in surrounding of object category in focus. The contextual features help classifying the object categories achieving high accuracies by taking into consideration the neighborhood of the superpixel. The original contributions of this paper are as follows:

1. An architecture is proposed that exploits the extraction and integration of the visual features and the contextual features to generate accurate class labels for every image pixel in realistic scene data. The contextual features are obtained by counting object occurrence priors of each superpixel to represent both local and global contextual information.
2. A class-specific classifier is designed that takes into account both multiple weak classifiers and boosting techniques to attain high accuracy of class labelling.
3. A detailed comparison to previous approaches is presented, confirming a superior performance of the proposed approach on the Stanford background benchmark dataset.

The rest of the paper is organized as follows. Section 2 briefly presents relevant work in the literature. Section 3 describes the architecture of the proposed network. The experiments are discussed in section 4 while section 5 concludes the paper.

II. Related Work

This section reviews recent related work for scene parsing. Convolutional Neural Networks (CNNs) have achieved remarkable performance on several semantic scene parsing benchmarks. The state of the art comprises of two major components: multi-scale context module and neural network design. The context information is crucial for pixel-wise labelling tasks. The early methods for scene parsing tasks include region proposal based methods [1], which take region proposals into account to generate class label results. Pyramid Scene Parsing Network (PSPNet) [2] exploits global contextual information with the help of pyramid pooling module and effectively produces quality results for scene parsing tasks.

Nguyen et al. [3] proposed a hybrid Deep Learning-Gaussian Process (DL-GP) network to segment a scene image into lane and background regions. In contrast to the existing deep learning approaches the proposed architecture combines a compact network having small number of parameters with a powerful nonparametric GP classifier. In both quantitative and visual comparisons, the proposed technique outperforms SegNet [4] and Bayesian SegNet [5]. However, the paper considers only single class (pedestrian lane) for evaluation and comparison with other techniques.

A novel neural network approach introduced by Zhang et al. [6] addresses scene understanding, parsing an input image into a structured configuration. The proposed technique consists of two networks initial being a convolutional neural network which extracts the image representation for pixel-wise object labelling and the second network is Recursive Neural Network (RNN) that discovers the hierarchical object structures and the inter-object relations. The experiments show the model performs well in producing meaningful scene configurations.

Residual Atrous Pyramid Network (RAPNet), a novel deep learning framework proposed by Zhang et al. [7], deals with importance-aware street scene parsing. The method incorporates Importance-Aware Feature Selection (IAFS) mechanism which selects important features for label predictions, the labelling is further enhanced by Residual Atrous Spatial Pyramid (RASP) module to sequentially aggregate global-to-local context information in a residual refinement manner. A Unified Panoptic Segmentation Network (UPSNet) was proposed by Xiong et. al. [8]. The network tackles the panoptic segmentation task. The architecture consists of a single backbone residual network followed by deformable convolution based semantic segmentation head and Mask R-CNN style segmentation head, resolving subtask simultaneously. Finally, a panoptic head solves the panoptic segmentation via pixel wise classification.

Zhang et al. [9] proposed deep gated attention networks for street scene parsing. The proposed spatial gated attention (SGA) module automatically highlights the attentive regions for pixel-wise labelling. The proposed module takes as input the multi-scale feature map based on a fully convolution network (FCN) backbone and produces the attention mask for each feature map. The multiscale features combine with Attentive Feature Interaction (AFI) module are used to generate final predication maps.

A single-shot, bottom-up approach for whole image parsing proposed by [10] Yang et al. is referred as DeeperLab. The proposed DeeperLab image parser performs whole image parsing with a significantly simpler, fully convolutional approach that jointly addresses the semantic and instance segmentation tasks in a single-shot manner. The encoder of the network is built on efficient depth wise separable convolution for faster inference, the backbone is additionally augmented with ASPP module for Atrous convolutions. The decoder recovers detailed object boundaries, using two large kernel depth wise convolutions to further increase the receptive field. Finally, the network contains five prediction heads, one specific for semantic segmentation while other four for class-agnostic instance segmentation. The technique achieves nearly real-time speed and comparably better results, evaluated on Mapillary Vistas dataset [11].

RefineNet [12] exploits features at various levels of abstractions for high resolution pixel-wise predictions The pixelwise labelling with FCN frameworks have improved by employing dilated convolutions, multi scale features and refining boundaries. Zhang et al. [13] explored the global contextual information by introducing the Context Encoding Module which captures class dependent feature map, and it improves semantic segmentation results with marginal extra computation cost over FCN.

Yuhui et al. from The Microsoft Research team [14] proposed Object Context Network (OCNet) for scene parsing. The OCNet architecture considers an object context pooling layer. Its implementation is inspired self-attention and considers i) similarities between each pixel and all other pixels, and ii) the pixel representation based on aggregation features from all other pixels. The technique achieves robust results compared to other existing scene parsing algorithms on the Cityscapes and AED20K datasets.

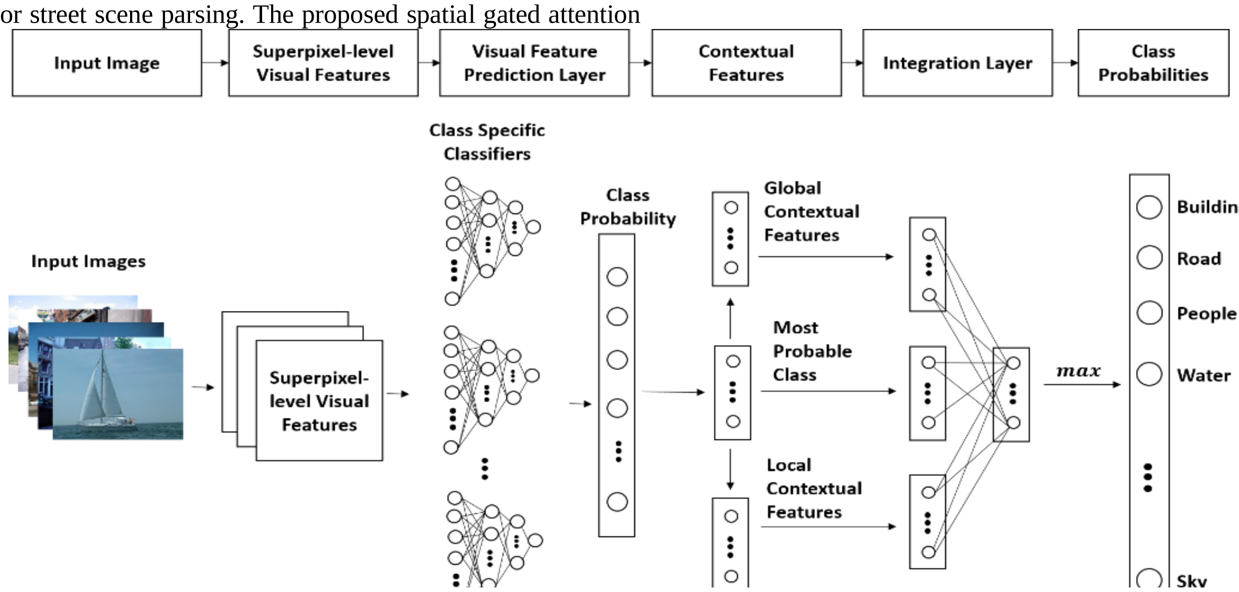
Input Image
Superpixel-level Visual Features
Visual Feature Prediction Layer
Contextual Features
Integration Layer
Class Probabilities
Class Specific Classifiers
Input Images
Superpixel-level Visual Features
Class Probability
Global Contextual Features
Most Probable Class
Local Contextual Features
max
Building
Road
People
Water
Sky

Fig. 1. Contextual and Visual Feature based Network Architecture for Scene Parsing

III. Proposed Network Architecture

In this section, we present the proposed network framework for effective scene parsing using image data. Traditional deep learning architectures are difficult to train and hard to optimize. The proposed architecture consists of multiple pipelined layers to deal with the scene parsing problem. The procedure for training the proposed network is summarized in Algorithm 1.

Algorithm 1 Training of the network

Input: Image, Super Pixels

Output: Class Labels

```
// Initially learn visual feature based class probability

  Foreach Super Pixel $s_j \in S$
      Foreach Class $c_i \in C$
              Extract visual features $f_{i,j}{}^{vis}$
              Train object category specific classifier $P^{vis}(c_j|s_j) = \phi_i(f_{i,j}^{vis})$
      End
              Get most probable class. $\hat{c} = \max_{1<i<M} P(c_i|s_j)$
  End

// Calculate contextual features of super pixels
Foreach Super Pixel $s_j \in S$
      Foreach Class $c_i \in C$
              Extract Contextual Features $V^{con}(C|s_j)$
      End
End

// Unify the contextual features with visual features and obtain probability
Foreach Super Pixel $s_j \in S$
      Foreach Class $c_i \in C$
              Combine Visual and Contextual Features $P(c_i|s_j)$
      End
              Assign superpixel class with maximum probability.
              $s_j \in \hat{c}\ if\ P(\hat{c}|s_j) = \max_{1<i<M} P(c_i|s_j)$
End
```

The architecture initially extracts visual features from superpixels and trains class specific classifiers to classify every superpixel into a correct class category. An overview of the proposed architecture is shown in Fig. 1.

The architecture is divided into three main parts as follows:

a) The visual feature extraction layer.

This layer extracts the visual features from the super pixels and builds class semantic supervised classifiers to predict class probabilities of superpixels.

b) The contextual exploitation layer.

This layer considers the contextual properties of the object and obtains context features for each superpixel, calculating the co-occurrence of object categories in different ranges of images.

c) The unification layer.

The unification layer jointly models the correlation of visual and contextual features to derive a final class label for each superpixel.

The proposed network takes superpixel-level visual features as input and outputs a class label for each superpixel. The next layer receives class probabilities predicted in the first layer and integrates the most probable class with object co-occurrence priors to learn image dependent contextual features. The contextual features reflect the prior local and global context information learnt for each superpixel and they are obtained by casting votes to every superpixel in a test image as shown in Equations (1-3).

$$V^l(C|s_j) = \psi^l(P^v(C|s_j), OCP) \quad (1)$$

$$V^g(C|s_j) = \psi^g(P^v(C|s_j), OCP) \quad (1)$$

$$V^{con}(C|s_j) = \zeta\left(\overbrace{(V^l(C|s_j),}^{local}\overbrace{(V^g(C|s_j)}^{global}\right) \quad (2)$$

Where, $\psi^l and\, \psi^g$ represent the voting functions for global and local context, respectively, and ζ represents the function to join the contextual features. The contextual features adaptively attain the dependencies of superpixels within an image.

It has been proved that both visual features and contextual features convey critical information for scene parsing. To combine the visual features and the contextual features, contextual features are normalized to class probabilities. The correlation of class probabilities of visual and contextual features is modeled in the unification layer which already has a set of neurons.

$$P(C|s_j) = \mathcal{H}\left(\overbrace{(P^{vis}(C|s_j),}^{visual\ features} \quad \overbrace{(P^{con}(C|s_j)}^{contextual\ features}\right) \quad (3)$$

$$P^T(C|s_j) = U_{1<i<M} P(c_i|f_i^T; W^T) \quad (4)$$

where $\mathcal{H}$ represents the function for jointly modeling both of class probabilities $P^T(C|s_j)$ of s_j based on features $f_i^T$and the corresponding weights $W^T, T \epsilon \{vis, con\}$.

1) The visual feature extraction layer extracts the visual superpixel features and trains multiple class specific classifiers to obtain an approximate prediction of the probabilities of all superpixels belonging to each class based on the set of visual features. For the j^{th} superpixel s_j, we will obtain its class probability for i^{th} class s_j:

$$P^{vis}(c_j|s_j) = \phi_i(f_{i,j}^{vis}) = fn(w_{1,i} f_{i,j}^{vis} + b_{1,i}) \quad (5)$$

where $f_{i,j}^{vis}$is the visual features of s_j extracted for the ith class c_i , ϕ_i is initial trained binary classifier for c_i , fn indicates that the prediction function of ϕ_i , and $w_{1,i}$ and $b_{1,i}$ are trainable weights and constants parameters respectively. For all M classes we can get a probability vector for s_j:

$$P^{vis}(c_j|s_j) = [P^{vis}(c_1|s_j), \dots, P^{vis}(c_i|s_j), \dots, P^{vis}(c_M|s_j)] \quad (6)$$

The above vector contains the likelihood of each superpixel s_j belonging to all classes C, then we assign s_j to the class which has maximum probability:

$$s_j \epsilon\ \hat{c} \quad if\ P^{vis}(c_1|s_j) = \max_{1<i<M} (P^v is(c_i|s_j) \quad (7)$$

Class-specific classifiers are advantageous to exploit the discriminative features for each class, and in addition it is effective and desirable for datasets with a large number of classes. The class-specific classifiers handle the problem of unbalanced training data for natural data where various rarely occurring, but vital classes may have less data. In such cases, training a multi-class classifier has the risk of completely ignoring rare classes and being favorably biased towards common classes.

2) The contextual features indicating contextual label votes of all classes for s_j can be obtained based on the contextual features:

$$V^{con}(C|s_j) = [V^{con}(c_1|s_j), \dots, V^{con}(c_i|s_j), \dots, V^{con}(c_M|s_j)] \quad (8)$$

3) The unification layer integrates both the contextual features and the visual features to produce context-sensitive classification of each superpixel. The contextual features are normalized to obtain normalized probabilities.

$$P^{con}(c_i|s_j) = \frac{V^{con}(c_1|s_j)}{\sum_{1<i<M} V^{con}(c_M|s_j)} \quad (9)$$

For each of all classes, a connection of its corresponding neurons from P^{vis} and P^{con} will be established to integrate both types of probabilities for the prediction of the probability of the i^{th} class c_i for superpixel s_j:

$$P(c_i|s_j) = b_{2,i}^{c} + w_{2,i}^{vis} * P^{vis}(c_i|s_j) + w_{2,i}^{con} * P^{con}(c_i|s_j) \quad (10)$$

For all M classes, a series of connected neuron models with different weights will be learnt:

$$P(C|s_j) = [P(c_1|s_j), \dots, P(c_i|s_j), \dots, P(c_M|s_j)] \quad (11)$$

Finally, the superpixel s_j will be assigned to the class $\hat{c}$ which has the maximum probability across all classes using a majority voting strategy:

$$s_j \in \hat{c}\ if\ P(\hat{c}|s_j) = \max_{1<i<M} P(c_i|s_j) \quad (12)$$

The visual feature-based probability is obtained using SVR classifier and Adaboost. The SVR takes help from RBF kernel while Adaboost considers four weak classifiers including generative discriminate analysis, k-nearest neighbors, naïve bayes classifier and the logistic regression in combination to the adaptive boosting algorithm (or Adaboost), an iterative boosting process that primarily focuses on the objects that are not classified easily. Initially, every object is assigned the same weight and then iteratively the weights for both the incorrect and correct classifications are adjusted.

$$H(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \quad (13)$$

Where h_m is the mth weak classifier and α_m are the corresponding weights for the classifier.

Experiments and Analysis

In this section, we present the evaluation results of the proposed technique on the widely used Stanford background benchmark dataset.

A. The Stanford Background Dataset

The Stanford background dataset [15] consists of 715 images, which includes varying outdoor public scenes extracted from already existing databases that such as MSRC, PASCAL and LabelMe. This dataset comprises of 8 object classes that include sky, tree, road, grass, water, building, mountain, and foreground objects. The annotation of the image pixels is done manually using Amazon Mechanical Turk, the size of the images is 320*240 pixels on average. Five-fold cross validation is used to obtain classification accuracy.

TABLE I. CONFUSION MATRIX FOR EIGHT OBJECTS ON THE STANFORD BACKGROUND DATASET (ADB, GLOBAL ACCURACY = 80.57%)

	Sky	Tree	Road	Grass	Water	Bldng.	Mtn.	Fgnd
Sky	**94.2**	2.2	0.16	0.18	0.03	1.74	0.1	1.35
Tree	6.67	**64.7**	3.1	0.95	0.16	17.84	0.22	6.26
Road	0.16	0.54	**91.19**	0.33	0.06	2.42	0.02	5.28
Grass	0.37	9.54	18.17	**53.33**	0.13	7.12	0.49	10.8
Water	0.57	0.53	36.06	1.85	**28.67**	0.24	0.29	31.7
Bldng.	2.09	2.67	2.52	1.6	0.02	**87.43**	0.13	3.55
Mtn.	11.5	3.46	2.24	3.07	0.25	6.41	**64.1**	8.88
Fgnd.	3.37	2.87	8	1.64	0.68	10.34	0.26	**72.8**

TABLE II. CONFUSION MATRIX FOR EIGHT OBJECTS ON THE STANFORD BACKGROUND DATASET (SVR, GLOBAL AC-CURACY = 82.02%)

	Sky	Tree	Road	Grass	Water	Bldng.	Mtn.	Fgnd
Sky	**91**	3.27	0.11	0.14	0.14	3.41	0.08	2.12
Tree	2.5	**70.5**	2.06	1.36	0.26	18.4	0.13	4.74
Road	0.0	0.46	**87.8**	2.45	0.07	2.81	0.01	6.33
Grass	0.1	6.03	18.6	**59.62**	0.26	4.81	0.05	10.4
Water	0.5	0.41	16.53	6.32	**56.54**	0.24	0.06	19.3
Bldng.	1.05	2.89	1.49	0.32	0.4	**91.16**	0.01	2.68
Mtn.	4.98	21.09	3.02	3.41	0.32	8.16	**47.24**	11.78
Fgnd.	2.6	3.44	5.31	1.43	1.06	13.23	0.11	**72.83**

B. Evaluation Metrics

The evaluation metrics used for the experiments include pixel accuracy, class accuracy and mean accuracy.

- **Global Pixel Accuracy:** Ratio of correct pixels and total pixels. The percentage of pixels classified correctly among all the pixels of the dataset. It considers the less occurring classes as well.
- **Mean Accuracy:** Mean of category wise pixel accuracy, it is the correctly classified pixels percentage of every class irrespective of the number of times the class has occurred.
- **Class Accuracy:** Pixel accuracy for each individual class. Class accuracy helps in identifying the performance for each object category.

C. Experimental Setup

Initially, the super pixels are obtained using the fast graph-based algorithm [16] from the input images, followed by the extraction of contextual features comprising of:

a. Geometric, color variations and texture features.

b. The mean and standard deviation of RGB colors within a superpixel are calculated.
c. The top height of superpixel bounding box to the image height.
d. The mask of the super pixel shape over the image
e. Histograms of RGB colors.
f. Histogram of textons over the superpixel region.
g. Dilated RGB, SIFT and texton histograms over the superpixel regions.

The feature vector calculated consists of 537 elements. The SIFT descriptors are computed using 8 orientation and 4 scale filters while the textons are the clustered 8-dimensional response filter bank Maximum Response 8 (MR8) (rotationally invariant) using K-means algorithm. To avoid overfitting, reduce redundancy and for robust representation of the images, class-specific subsets of top 50 features are obtained using the minimum redundancy and maximum relevance (mrmr) algorithm [17].

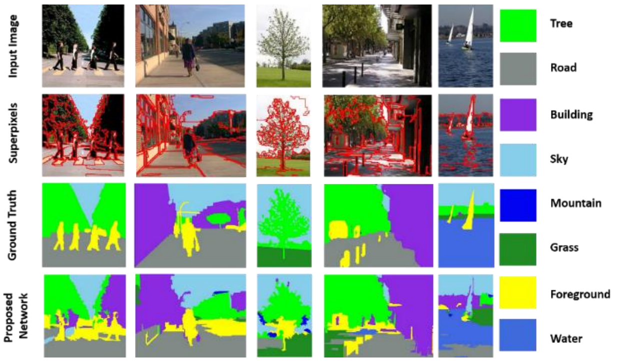

Fig. 2. Quantitative Results of Proposed Approach on Stanford Background Dataset.

D. *Experimental Results on Stanford Background Dataset*

The proposed architecture is evaluated on Stanford background dataset [18] and compared to the previous approaches. The performance comparison is based on the global, mean and class accuracies. The confusion matrix for eight object classes by using boosting classifier on the stated network can be observed in Table 1. It is evident from the table that sky, road and building are being classified easily, while the object category water and mountains are least classified. A minor confusion between grass and tree gives birth to misclassification in these two classes as well.

TABLE III. . PERFORMANCE (%) COMPARISON WITH PREVIOUS APPROACHES ON THE STANFORD BACKGROUND DATASET.

Ref	Global Accuracy	Average Accuracy
Gould et al. [18]	76.4	NA
Kumar et al. [19]	79.4	NA
Lemptisky et al. [20]	81.9	72.4
Farabet et al. [21]	78.5	50.8
Munoz et al. [22]	76.9	66.2
Sharma et al. [23]	**82.3**	79.1
Proposed Contextual Features (Adaboost)	**83.1**	**69.9**
Proposed Contextual Features (SVR)	**80.3**	**72.2**

Table 2 displays the confusion matrix using the SVR classification, it follows a similar trend to the boosting algorithm as well. The proposed network achieves very high accuracies for classes occurring in one scene including sky, road, and buildings. The architecture significantly outperforms the individual classifiers Adaboost and SVR and achieves 17-28% better accuracy. Table 3 presents a comparison with previous approaches and elaborates evidently the better performance of proposed network. It can also be noted in the confusion matrix that the approach performs best for mountain, one of the difficult classes to train in the dataset. As the approach considers global and local contextual features, counting the prior occurrence of objects, helps achieve better results for least occurring classes as well.

A portion of the tree object category pixels are misclassified as grass possibly due to the similar color feature maps. The water pixels are misclassified most using the contextual features with Adaboost classifier, also the combination of SVR and contextual features helped classifying the mountain pixels least.

TABLE IV. PERFORMANCE COMAPRISN IN TERMS OF GLOBAL ACCURACY USING VARIOUS CLASSICATION TECHNIQUES FOR TRAINING CLASS SPECIFIC CLASSIFIER AND REGRESSION TECHNIQUES IN INTEGRATION LAYER

	MLP	**SVR**	**Non-Linear**	**R-Ensemble**
Adaboost	83.1	80.3	80.2	79.81

The integration layer takes into account the probability map produced by global contextual features, local features and the class-specific classifiers, and produce the final probability of each super-pixel. We have incorporated different variations in the integration layer to compute the probability using MLP, SVR, Regression ensemble and Non-linear regression algorithm. Table 3 presents a brief summary of the combinations of classifiers used in visual feature prediction layer and regression technique used in the integration layer. Adaboost in combination with MLP achieved best accuracy on the test set of Stanford background dataset.

CONCLUSION

We presented a new deep network architecture-based technique for scene parsing. The main novelty of the architecture is that it combines visual and contextual features using an effective ingeneration mechanism to improve the performance. The class-specific classifier such as Adaboost and SVR have been considered and investigated. The proposed technique performed well on the Stanford background dataset. It achieved 83% accuracy which is the highest among published results on Stanford background dataset.

In our future work, we will extend the evaluation of the proposed technique on more benchmark datasets. We will also extend the technique by optimizing the learning parameters for the network training.

REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Sep. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proceedings - 30th IEEE Conference on Computer Vision

and Pattern Recognition, CVPR 2017, Nov. 2017, vol. 2017-January, pp. 6230–6239, doi: 10.1109/CVPR.2017.660.

[3] T. N. A. Nguyen, S. L. Phung, and A. Bouzerdoum, “Hybrid Deep Learning-Gaussian Process Network for Pedestrian Lane Detection in Unstructured Scenes,” IEEE Trans. Neural Networks Learn. Syst., pp. 1–15, Feb. 2020, doi: 10.1109/tnnls.2020.2966246.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[5] A. Kendall, V. Badrinarayanan, and R. Cipolla, “Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding,” Br. Mach. Vis. Conf. 2017, BMVC 2017, Nov. 2015, Accessed: Apr. 23, 2020. [Online]. Available: http://arxiv.org/abs/1511.02680.

[6] R. Zhang, L. Lin, G. Wang, M. Wang, and W. Zuo, “Hierarchical Scene Parsing by Weakly Supervised Learning with Image Descriptions,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 3, pp. 596–610, Mar. 2019, doi: 10.1109/TPAMI.2018.2799846.

[7] P. Zhang, W. Liu, Y. Lei, H. Wang, and H. Lu, “RAPNet: Residual Atrous Pyramid Network for Importance-Aware Street Scene Parsing,” IEEE Trans. Image Process., vol. 29, pp. 5010–5021, 2020, doi: 10.1109/TIP.2020.2978339.

[8] Y. Xiong et al., “UPSNet: A Unified Panoptic Segmentation Network,” pp. 8818–8826, Jan. 2019, Accessed: Apr. 27, 2020. [Online]. Available: http://arxiv.org/abs/1901.03784.

[9] P. Zhang, W. Liu, H. Wang, Y. Lei, and H. Lu, “Deep gated attention networks for large-scale street-level scene segmentation,” Pattern Recognit., vol. 88, pp. 702–714, Apr. 2019, doi: 10.1016/j.patcog.2018.12.021.

[10] T.-J. Yang et al., “DeeperLab: Single-Shot Image Parser,” Feb. 2019, Accessed: Apr. 26, 2020. [Online]. Available: http://arxiv.org/abs/1902.05093.

[11] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, “The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes,” in Proceedings of the IEEE International Conference on Computer Vision, Dec. 2017, vol. 2017-October, pp. 5000–5009, doi: 10.1109/ICCV.2017.534.

[12] G. Lin, F. Liu, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-Path Refinement Networks for Dense Prediction,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 5, pp. 1228–1242, May 2020, doi: 10.1109/TPAMI.2019.2893630.

[13] H. Zhang et al., “Context Encoding for Semantic Segmentation,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 7151–7160, Mar. 2018, Accessed: May 11, 2020. [Online]. Available: http://arxiv.org/abs/1803.08904.

[14] Y. Yuan and J. Wang, “OCNet: Object Context Network for Scene Parsing,” Sep. 2018, Accessed: May 26, 2020. [Online]. Available: http://arxiv.org/abs/1809.00916.

[15] G. Heitz, S. Gould, A. Saxena, and D. Koller, “Cascaded Classification Models: Combining Models for Holistic Scene Understanding,” 2009.

[16] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient Graph-Based Image Segmentation.”

[17] H. Peng, F. Long, and C. Ding, “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy.”

[18] S. Gould, R. Fulton, and D. Koller, “Decomposing a scene into geometric and semantically consistent regions,” in Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 1–8, doi: 10.1109/ICCV.2009.5459211.

[19] M. P. Kumar and D. Koller, “Efficiently selecting regions for scene understanding,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3217–3224, doi: 10.1109/CVPR.2010.5540072.

[20] V. Lempitsky, A. Vedaldi, and A. Zisserman, “Pylon Model for Semantic Segmentation.” pp. 1485–1493, 2011.

[21] C. Farabet, C. Couprie, L. Najman, and Y. Lecun, “Learning hierarchical features for scene labeling,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1915–1929, 2013, doi: 10.1109/TPAMI.2012.231.

[22] D. Munoz, J. A. Bagnell, and M. Hebert, “Stacked hierarchical labeling,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, vol. 6316 LNCS, no. PART 6, pp. 57–70, doi: 10.1007/978-3-642-15567-3_5.

[23] A. Sharma, O. Tuzel, and D. W. Jacobs, “Deep Hierarchical Parsing for Semantic Segmentation,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June-2015, pp. 530–538, Mar. 2015, Accessed: May 31, 2020. [Online]. Available: http://arxiv.org/abs/1503.02725.