Relationship Aware Context Adaptive Feature Selection Framework for Image Parsing

Basim Azam
Centre for Intelligent Systems, School of Engineering and Technology,
Central Queensland University,
Brisbane, Australia
b.azam@cqu.edu.au

Ranju Mandal
Centre for Intelligent Systems, School of Engineering and Technology,
Central Queensland University,
Brisbane, Australia
r.mandal@cqu.edu.au

Brijesh Verma
Centre for Intelligent Systems, School of Engineering and Technology,
Central Queensland University,
Brisbane, Australia
b.verma@cqu.edu.au

***Abstract*— Feature selection for deep learning architectures is one of the important and challenging steps in developing an efficient image parsing application. In this paper, a novel image parsing architecture which makes use of unique feature selection is proposed. It introduces the idea of weighted relationship awareness to reduce the redundancy of features and optimally select an efficient subset of feature representations. The proposed architecture is evaluated on CamVid benchmark dataset. A comparison with state-of-the-art methods was conducted which showed significant improvements in terms of segmentation and classification accuracy.**

Keywords— Image parsing, feature selection, semantic segmentation, deep learning.

I. INTRODUCTION

Computer aided visual attention systems model the low-level impulsive human vision system. Image parsing is a vital and elementary step towards the automation of driving mechanisms and robot navigations. The computer vision techniques with aid of machine learning algorithms have enormously helped achieving sate of the art performances in object detection [1] and semantic labelling [2-4]. The evolution of algorithms has assisted analyzing the content of images. Also, the modern frameworks have shown capability for producing accurate pixel-wise labelling, however the features computed from the images exhibit variation. The state-of-the-art approaches are undermined by generalized feature selection.

Recent segmentation approaches based on Convolutional Neural Networks (CNN) are formulated for pixel-wise labelling tasks. These algorithms obtain a coarse label map by applying multiple complex pixel-wise convolutional operations on the input images. Although powerful, these techniques can be complex, combining both the deeper and expensive computations and irregular feature representation. In order to integrate the additional contextual information, some approaches make use of Conditional Random Fields (CRF) [2, 5]. The addition of CRF to architectures leads to time-consuming training inference, and addition to the increased feature representation [3]. High-dimensional features have been a reason of attraction for learning tasks, which include classification, segmentation and decision making. It is highly desirable in such cases to employ critical features for learning tasks, as features with less importance result in performance degradation and prove computationally expensive. It is necessary to eradicate inappropriate features, which may broadly be grouped as noisy features (feature with less or no contribution to performance) and redundant features (features that are similar to other features), by this means the computation complexity can be reduced. Feature selection methods have been used in the literature for applications including image based empirical medical studies [6], face recognition [7], hyperspectral image analysis [8] and computer vision-based navigation systems.

Towards this end, we present an image parsing framework, that considers additional contextual information and incorporates generalized feature selection, with an entirely novel strategy to perform segmentation. In this work, several experiments are carried out to compare the performance of the proposed architecture. The original contributions in this paper are as follows.

1) A novel feature selection technique, which makes use of mutual information between two features and the fisher criterion for optimal subset.
2) A unique image parsing framework, which considers the local and global contextual information and utilizes the proposed feature selection technique to refine the segmentation accuracy.
3) A detailed comparison and analysis with previous approaches, presenting the enhanced performance of the proposed architecture on the CAMVID benchmark dataset.

The rest of the paper is organized as follows. The background of image parsing and features selection is presented in Section II, while the proposed approach is explained in detail in Section III. Section IV presents the experiments and evaluations. Section V concludes the study.

II. BACKGROUND

This section presents the relevant work in the literature on visual feature extraction, consideration of the contextual information, the feature selection and the motivation for proposed architecture.

Image parsing deals with multiple challenges at once, the choice of patches, the attributes to learn, and optimal selection of prediction model. The initial methods to parse an image into regions computed information from each pixel [9], while the

advancements [10, 11] make the use of superpixel in such approaches. Superpixel techniques merge multiple pixels with similar information into one pixel, and thus allowing reliable feature extraction from a group of pixels. Mainly used superpixel level features include color [12], texture [13], shape and location of superpixels. Yet, the visual features are incapable to present semantic information about the images. Besides, integration of CRF to layered models not only considers the local contextual information but also increases the computational forehead.

Fully Convolutional Networks (FCNs) [14] have been deployed for semantic segmentation, making use of fully connected layers with greater receptive fields, the output network achieves impressive performance on the PASCAL VOC benchmark. Genetic programming has also been adapted for texture classification application [15] . Chen et. al [5] added fully connected CRF to the existing FCN architecture to obtain denser maps and improved the pixel-wise accuracy. The CNNs incorporate the contextual information in a hierarchical manner by down sampling the images, thus losing significant information [16]. The proposed approach derives the concept of several layers from deep neural architectures and uses the context adaptive features along with visual attributes to parse an image.

Feature extraction is an important task for many classification problems [17-21] as it can improve the accuracy. Automatic feature extraction using CNNs has been the focus of many recent studies [22]. Researchers have proven that some features might be redundant and may lead to misclassifications, so feature selection is essential part of many applications. Feature selection deals with removal of irrelevant and redundant features, ultimately proving of great importance in building models. The selection of features has been discussed enormously over the years in literature [6-8, 17-18]. Feature selection techniques reduce the effect of redundant variables by selecting a subset of existing features.

III. Proposed Approach

In this section image parsing framework is explained in detail, along with the feature selection process.

A. *Image Parsing Framework*

The proposed image parsing framework can be broadly divided out into three major layers, the Visual Prediction (VP) layer, Contextual Adaptive (CA) layer, and the integration layer. The initial visual prediction layer deals with extraction of superpixel level visual features, moreover it makes use of a novel feature selection algorithm to optimally select a subset and finally based on those features a class-specific classifier is trained to get class labels for superpixels. The second layer i.e. context adaptive CA layer anticipates the class probabilities based on short- and long-range contextual dependencies. CA considers the adjacent superpixel occurrences and the occurrences of other class labels within a given block of training set, ultimately producing probability values for the superpixel label. The final Integration layer uses the probability values computed in VP and CA layers to yield final superpixel class labels.

The proposed network architecture initially computes the visual features from the superpixels, uses a novel attribute assortment algorithm to contribute towards selection of a subset of features, and trains class-specific classifier to obtain the class probability, in parallel to this the network extracts contextual features from the images and finally these probabilities are integrated to compute the class label. An overview of the proposed image parsing architecture with feature selection is presented in Fig. 1.

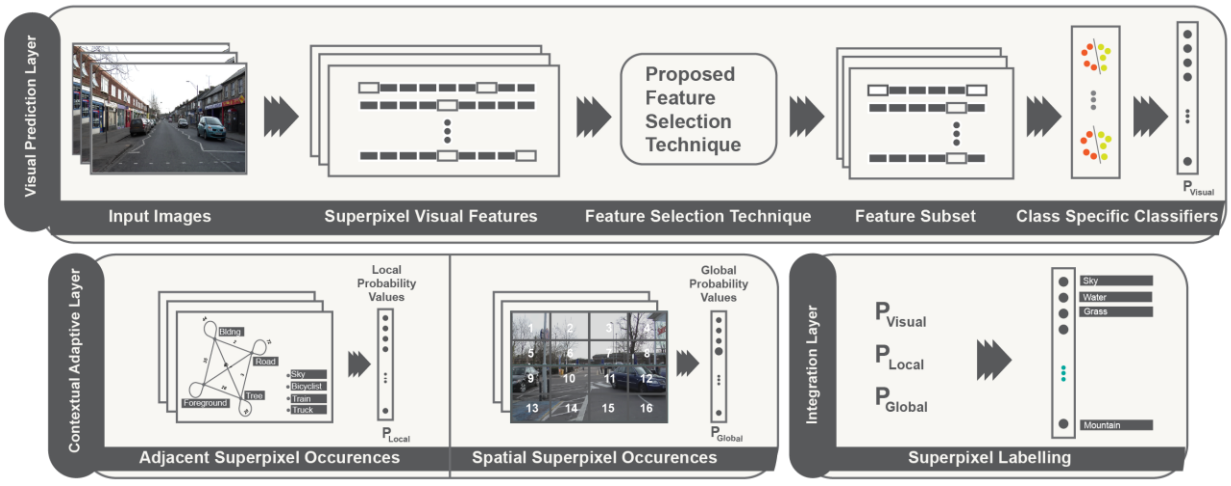

Fig. 1. Image Parsing Framework. **Visual Prediction Layer**: this layer takes a subset of optimum (chosen by novel feature selection) super pixel visual features extracted from the image and trains class specific classifier to obtain probabilities for pixels. **Contextual Adaptive Layer:** CA layer considers the probability-based votes from adjacent superpixels and the spatial block object oc-currences from the training data. **Integration Layer:** Integration layer renders final class labels for superpixels using the probabilities from VP layer and CA layer.

Algorithm 1:		**Feature Selection based Image Parsing Framework**
Input	:	$I: m \times k$ Input image,
Initialize	:	S: Number of superpixels T: Number of features to select
Output	:	Labelled image

// VP Layer

for each image in training set
do

1 : Calculate **S** superpixels
2 : Extract visual features from **S** superpixels
3 : Select **T** subset of optimum features using *proposed feature selection algorithm*
4 : Train class specific classifier using **T** features
5 : Compute probabilities P_{vis} for superpixel labels using the trained model
end for

// CA Layer

for each computed superpixel in **S**
: **do**
6 : Compute short range dependencies using adjacent superpixel information (local context)
7 : Calculate long rang dependencies using object co-occurrence in a spatial block (global context)
8 : Normalize the local and global contextual information to compute probabilities P_L and P_G
end for

// Integration Layer

for each computed superpixel in **S**
do
9 : Integrate visual, local, and global probabilities P_{vis}, P_L, P_G
10 : Assign each superpixel the class label with maximum probability
: **end for**

The image parsing problem can be formulated as, let $I(v) \in R^3$ is an image defined on a set of pixels v. The architecture can broadly be divided into three main parts as

a) The visual prediction layer computes the visual features $feat_{vis}$ from the superpixels S and uses the feature selection algorithm to compute a subset $feát_{vıs}$ from the given feature set. The subset calculated is used to train class-specific classifiers ultimately producing the class probabilities for superpixels.
b) The context adaptive layer contemplates the contextual properties of the object category and obtains contextual features based on the superpixel-to-image and overall occurrences of the object.
c) The integration layer utilizes the probabilities computed from visual prediction and context adaptive layers to produce final probability of class-labels for superpixels.

The proposed framework considers the superpixel features extracted and use the feature selection algorithm to select top ranked features, which are fed to the classifier to obtain subsequent superpixel level class labels.

1) Visual Prediction Layer

The VP layer considers the raw image sample from the training dataset and computes the superpixels at first, which is followed by the extraction of visual features from these superpixels. These visual features are fed to the proposed feature selection algorithm which optimally selects a subset of the features and feeds it to the classifier for training. The image parsing framework aims to assign every pixel v to one of the class labels $C = \{c_i | i = 1,2, \dots M\}$ while M refers to the total number of classes.

The VP layer computes superpixels instead of pixels, so the problem can be reformulated as $S(v) = \{s_i | i = 1,2, \dots N\}$ where S denotes set of N superpixels segmented from I. These superpixels are used to extract visual features, a subset of features is then optimally chosen using novel feature assortment algorithm. Class-specific classifiers are trained based on the optimal feature selected, to produce class labels for superpixels.

2) Context Adaptive Layer

The succeeding CA layer assimilates the most likely class label for the superpixels using the learned prior information of occurrences of objects. The term contextual adaptation represents the information of the objects occurring within an image in varying ranges. CA computes the local occurrences by considering the adjacent superpixels and the global occurrence by computing the occurrence of object in a block B_k of superpixels.

In a similar way we can calculate global context, as every superpixel within the block B_k votes to compute the class probability of superpixels in other blocks. Assuming a superpixel s_j in block B_{k1}, and let $S = \{s_q | q = 1,2, \dots T\}$ superpixels in all the rest blocks so the superpixel s_j receives T votes. The class label of each superpixel is produced by majority votes of labels.

Once the local and global votes are calculated they are normalized to probability values as to match the probability in the VP layer as:

$$P_L(c_i|s_j) = \frac{Votes\ for\ superpixel\ j\ belonging\ to\ class\ c\ using\ adjacent\ superpixels}{Sum\ of\ all\ votes} \quad (1)$$

$$P_G(c_i|s_j) = \frac{Votes\ for\ superpixel\ j\ belonging\ to\ class\ c\ using\ spatial\ blocks}{Sum\ of\ all\ votes} \quad (2)$$

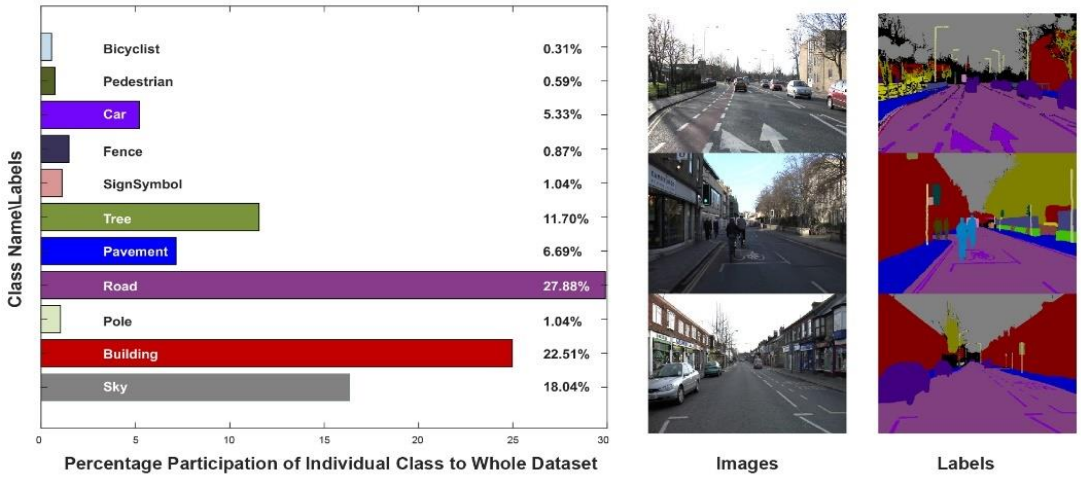

Fig. 2. The overview of the CamVid Database used in the study. Left: Percentage breakdown by category, Right: Visual sample images with labels

The probability values for CA layer can be given as:

$$P_{CA}(c_i|s_j) = w_L * P_L(c_i|s_j) + w_G * P_G(c_i|s_j) \quad (3)$$

Where w_L and w_G are weights for local and global probability values. These weights account for contextual features and are optimized using numerous linear regression models. As the CA layer considers the long- and short-range contextual properties using the spatial block and superpixels making the spatial block an important factor for the superpixel votes to be casted. The variability of number of voting superpixels and the size of spatial block presents the flexibility and importance of CA layer to achieve good results.

3) Integration Layer

The normalized probabilities estimated in integration layer can be expressed as:

$$P(c_i|s_j) = w_{CA} * P_{CA}(c_i|s_j) + +w_{VP} * P_{VP}(c_i|s_j) \quad (4)$$

Where $w_{CA}$and w_{VP} are weights for $P_{CA}(c_i|s_j)$ and $P_{VP}(c_i|s_j)$ respectively. The weights are fine-tuned by reducing the sum of squared variation of probability scores over the labelled training data and predicted labels. Finally, the class with maximum likelihood scores is assigned to the superpixel:

$$s_j \in \acute{c}\ if\ P(c_1|s_j) = \max_{1<i<M} P(c_i|s_j) \quad (5)$$

B. Feature Selection Method

The classification pipeline considers extracted attributes as input to the algorithm to classify. The number of features vary and sometimes are huge in total, ultimately affecting the computational complexity and time. The attributes with higher ability to distinguish between classes are more important in terms of performance and instead of using every attribute only high-level attributes are chosen.

The feature selection function $\varsigma(\overrightarrow{feat_x}, \overrightarrow{feat_y})$ considers mutual information between the features and the fisher criterion to rank the features in descending order. A subset of which is further chosen to feed to the classifier.

The normalized mutual information [17] for $feat_x$ can be represented as

$$mutInf_x = \sum_{c \in C} \sum_{x \in feat_x} p(c,x) \log \left(\frac{(p(c,x))}{p(c)(x)} \right) \quad (6)$$

Where C is the set of class labels and the joint probability distribution is presented with $p(.,.)$. Generally, $mutInf_x$ determines the amount by which the information provided by the feature vector decreases.

The multi-class generalized fisher criterion [18] is given as

$$Fisher_x = \frac{\sum_{k=1}^{M} (\mu_{x,k} - \acute{\mu_x})^2}{\sigma_x^2} \quad (7)$$

where μ_x and E_x denote the mean and standard deviation of the whole dataset corresponding to the $feat_x$. Fisher score helps determine the separability of the feature from the rest of the features.

Both $mutInf$ and $Fisher$ scores are weighted linearly

$$Score_x = \lambda_1 * mutInf_x + \lambda_2 * Fisher_x \quad (8)$$

where $\lambda_1$and λ_2 are learnable parameters adjusted during the experimentation by cross validation on the training set also $\sum_k \lambda_k = 1$. Finally, the weights are computed as

$$\varsigma(\overrightarrow{feat_x}, \overrightarrow{feat_y}) = Score_{feat_x} * Score_{feat_y} \quad (9)$$

The top ranked subset of features based on the defined feature selection algorithm are further fed to the classifier. As, the feature selection method optimizes and ranks the features based on the criteria defined above, the highly important features are chosen that help in improving the performance of the proposed architecture.

IV. Experiments and Analysis

The proposed architecture is experimentally validated on the Cambridge-driving Labeled Video Database (CamVid) [23]. CamVid dataset comprises of the road driving scene videos, including almost 10 minutes of best quality (970 x 720) footage which is further split into four sequences. These videos are

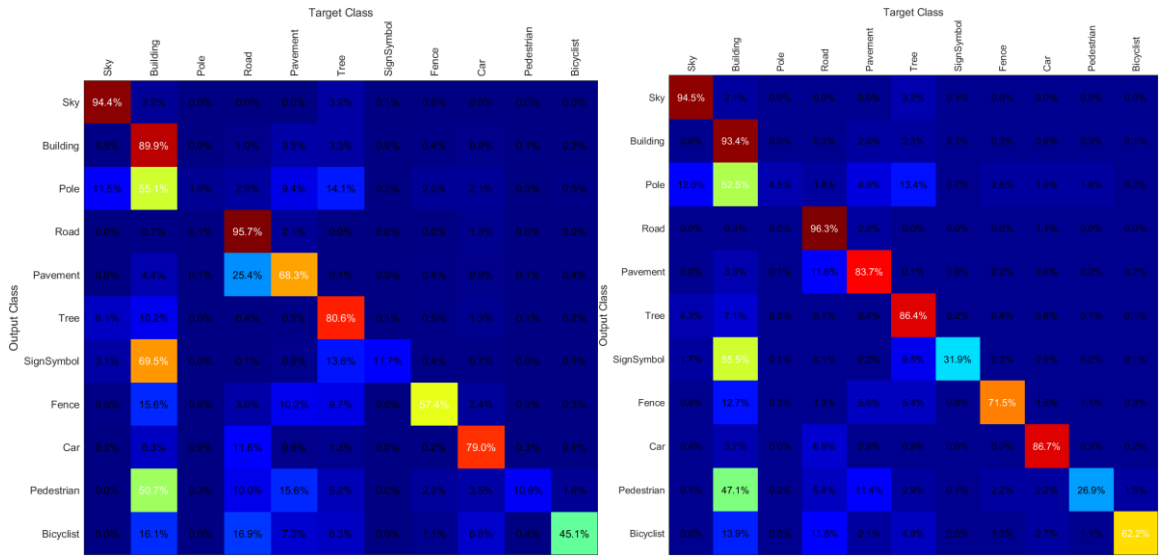

Fig. 3. Confusion matrix over the CamVid test set using Adaboost as class specific classifier and MLP in Integration Layer. *Left:* 256 Superpixels, *Right:512 Superpixels*

obtained using the camera installed on the windshield of vehicle. The scenarios consist of crowded scenes, varying illumination conditions and multiple types of roads. Among these four sequences, one is taken at dusk while the rest three are captured during sunlight. 701 images are annotated into 32 categories, however studies [24, 25] have used 11 object categories, considering most of the labelled pixels i.e. 89%. The overview of CamVid database is given in Figure 2. The experimental evaluation is setup in a way to present fair comparison with other approaches, splitting the images into 368 training and 233 test images, making use of 11 object categories [24].

Superpixels are obtained using SLIC [26]. Increase in the number of superpixels surges the overall feature extraction. For the sake of this study we generate superpixel with values 256 and 512. The algorithm generates highly uniform superpixel that can be observed in Figure 4.

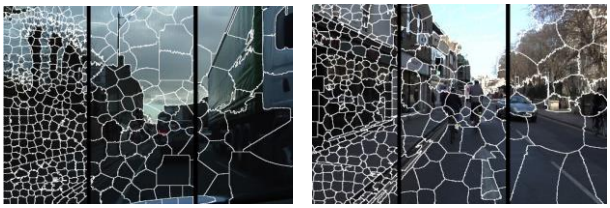

Fig. 4. Images segmented using SLIC Superpixel algorithm of (approximate) size 64, 512 and 1024 pixels.

The visual attributes computed from the superpixels include computation of color dissimilarities, change in geometry and texture. The superpixels are used to extract mean, standard deviation, mask and histograms of textons. Finally, the computation of SIFT descriptors make the feature vector of length 517. However, to overcome overfitting problem, the proposed feature selection technique is deployed to choose top ranked 50 features.

The evaluation metrics used for the comparison and result interpretation include global accuracy, average accuracy, and class accuracy. The global accuracy is biased towards the classes occurring often and thus less favorable towards classes with low presence.

TABLE I. LAYER-WISE PERFROMANCE EVALUATION IN TERMS OF ACCURACY (%) FOR 512 SUPERPIXELS

	VP Layer	CA Layer	Proposed Approach
512 Superpixels	68.25	87.61	**89.79**
Sky	95.79	92.66	**94.52**
Building	67.16	95.15	**93.44**
Pole	0.08	0	**0**
Road	95.52	98.22	**96.29**
Pavement	22.75	70.59	**83.69**
Tree	39.43	83.59	**86.42**
SignSymbol	0.06	0	**31.85**
Fence	26.76	61.43	**71.47**
Car	30.72	77.07	**86.73**
Pedestrian	11.83	0.02	**26.88**
Bicyclist	33.21	28.1	**62.16**

TABLE II. LAYER-WISE PERFROMANCE EVALUATION IN TERMS OF ACCURACY (%) FOR 256 SUPERPIXELS

	VP Layer	CA Layer	Proposed Approach
256 Superpixels	66.57	81	85.63
Sky	96.6	88.62	94.45
Building	62.15	92.35	89.93
Pole	0.05	0	0
Road	97.05	98.22	95.69
Pavement	0	38.03	68.3
Tree	46.56	76.13	80.64
SignSymbol	0.04	0	11.73
Fence	25.69	32.92	57.43
Car	27.26	61.14	79.04
Pedestrian	27.24	0	10.9
Bicyclist	0.09	0.02	45.06

TABLE III. QUANTITATIVE PERFORMANCE COMPARISONS WITH PREVIOUS APPROACHES ON THE CAMVID BG DATASET IN TERMS OF ACCURACY (%), PROPOSED FEATURE SELECTION BASE ARCHITECTURE ACHIEVES COMPETIVIE AVERAGE AND GLOBAL ACCURACY, IT ALSO PERFROMS EXTREMELY WELL FOR VARIOUS CHALLENGING CLASSES (BUILDING, CAR, TREE, FENCE, BICYCLIST)

	Building	Tree	Sky	Car	Sign Symbol	Road	Pedestrian	Fence	Pole	Pavement	Bicyclist	Global	Average
Proposed Approach (ADB-MLP)	93.4	86.42	94.5	86.7	31.9	96.3	26.8	71.47	4.5	68.3	62.16	**89.8**	**67.1**
CNN-CRF [25]	84.3	65.3	95.6	74.6	0.4	93.5	25.6	32.3	13.8	85	54.3	72.9	56.8
SfM + Appearance [24]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	69.1	53.0
Super Parsing [27]	87.0	67.1	96.9	62.7	30.1	95.9	14.7	17.9	1.7	70	19.4	83.3	51.2
Local Label Descriptors [28]	80.7	61.5	88.8	16.4	n/a	98.0	1.09	0.05	4.13	12.4	0.07	73.6	36.3
Boosting + Detector + CRF [29]	81.5	76.6	96.2	78.7	40.2	93.9	43.0	47.6	14.3	81.5	33.9	83.8	59.2
FCN+Comb [30]	79.7	77.2	85.7	86.1	45.3	94.9	45.9	69.0	25.2	86.2	57.9	88.7	63.8
ReSeg [31]	86.8	84.7	93.0	87.3	48.6	98.0	63.3	20.9	35.6	87.3	43.5	88.7	68.1

The proposed image parsing framework is evaluated on CamVid [32] dataset, the computations are compared with existing approaches in literature. Confusion matrix for 11 classes can be observed in Figure 3. The matrix elaborates the proposed technique achieves good results for classes building, tree, sky , car, road, fence and bicyclist, however the accuracy scores for sign symbol, pedestrian and pole are relatively low. The misclassification can also be detected as the superpixels of building are confused with pole, sign symbol and pedestrian. [25]

Table I and Table II present a summary of layer-wise accuracy scores for classes in dataset by varying the number of superpixels. It can be observed that overall proposed architecture achieved better and higher scores by integrating both the VP layer and CA layer. The accuracy score achieved for VP layer is by using the Adaboost classifier, while for contextual adaptive layer spatial block of size 6 is considered. Higher number of superpixels increase the overall accuracy score.

Table III presents the class-wise global and average accuracies of the proposed architecture compared with reported accuracies by state-of-the-art methods. The proposed architecture achieved global accuracy 89.8% which is highest accuracy when compared to some existing approaches. Lower class accuracies of 4.5%, 26.8% and 31.9% are observed for the classes pole, pedestrian, and sign-symbol respectively. These are relatively low scores as compared to the other approaches. The main reason for that these classes have small number of training instances, indicating that the proposed architecture focuses on frequently occurring classes.

The experimental evaluation leads us to various results and observations as:

- The proposed context adaptive layer produces noticeable improvements in terms of performance as compared to the visual feature layer individually. This helps to acknowledge that the CA layer computes the short- and long-range dependencies between the superpixels i.e. adaptable to the real context of image. So, it can be concluded that CA embeds adaptable contextual semantics to the architecture and is beneficial for producing class labels in complex scenes.
- The proposed architecture achieves higher accuracy than many existing image parsing approaches including some convolutional neural network-based algorithms on benchmark dataset. Additionally, the architecture is wide-ranging for example the classifier in VP layer can be chosen as suitable from MLP, SVM, Adaboost and Random Forest, in current study we have used Adaboost.
- The proposed feature selection algorithm helps identify the noteworthy features taking into consideration two factors, the mutual information, and the fisher score. The regularization parameters are adaptable to the real-world dataset, it is also worth mentioning here that this feature selection technique not only helps to choose an efficient subset of features but also lessens the computational complexity of the models.
- The number of superpixels calculated, to extract features from, play an important role and a smaller number of superpixels produce low accuracies. It is highly desirable to maintain a balanced number of superpixels as higher number of superpixels tend to yield complexity in the architectures.

V. CONCLUSION

We have presented a novel feature selection based deep learning architecture for image parsing. The major novelties included the introduction of short and long range contextual adaptive dependencies and the feature selection technique with adaptable parameters to the dataset. It is demonstrated that the proposed architecture brings significant improvements to the global accuracy of 89.8% on CamVid dataset. In our future research, we will evaluate our proposed approach on more benchmark datasets. The investigations will also be done to improve the proposed architecture by optimizing the learning parameters and fusion of layers.

VI. Acknowledgment

This research was supported under Australian Research Council's Discovery Projects funding scheme (project number DP200102252).

VII. References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, pp. 1-14, 2015.

[2] S. Zheng et al., "Conditional random fields as recurrent neural networks," Proceedings of the IEEE International Conference on Computer Vision, pp. 1529-1537, 2015.

[3] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," in 4th International Conference on Learning Representations, pp. 1-11, 2016.

[4] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," Proceedings of the IEEE International Conference on Computer Vision, pp. 1520-1528, 2015.

[5] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834-848, 2018.

[6] G. Li et al., "Effective breast cancer recognition based on fine-grained feature selection," IEEE Access, vol. 8, pp. 227538-227555, 2020.

[7] X. Fan and B. Verma, "Selection and fusion of facial features for face recognition," Expert Systems with Applications, vol. 36, no. 3, pp. 7157-7169, 2009.

[8] G. Taskin, H. Kaya, and L. Bruzzone, "Feature selection based on high dimensional model representation for hyperspectral images," IEEE Transactions on Image Processing, vol. 26, no. 6, pp. 2918-2928, 2017.

[9] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," International Journal of Computer Vision, vol. 81, no. 1, pp. 2-23, 2009.

[10] R. Zhang, L. Lin, G. Wang, M. Wang, and W. Zuo, "Hierarchical scene parsing by weakly supervised learning with image descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 3, pp. 596-610, 2019.

[11] A. Sharma, O. Tuzel, and D. W. Jacobs, "Deep hierarchical parsing for semantic segmentation," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 530-538, 2015.

[12] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, "Multi-class segmentation with relative location prior," International Journal of Computer Vision, vol. 80, no. 3, pp. 300-316, 2008.

[13] B. Mičušík and J. Košecká, "Semantic segmentation of street scenes by superpixel co-occurrence and 3D geometry," in IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pp. 625-632, 2010.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640-651, 2014.

[15] H. Al-Sahaf, M. Zhang, M. Johnston, and B. Verma, "Image descriptor: A genetic programming approach to multiclass texture classification," in IEEE Congress on Evolutionary Computation, pp. 2460-2467, 2015.

[16] B. Azam, R. Mandal, L. Zhang, and B. Verma, "Class probability-based visual and contextual feature integration for image parsing," in 35th International Conference on Image and Vision Computing New Zealand, pp. 1-6, 2020.

[17] N. Carrara and J. Ernst, "On the estimation of mutual information," in The 39th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, vol. 33, no. 1, p. 31, 2020.

[18] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, pp. 266-273, 2011.

[19] S. Chowdhury, B. Verma, and D. Stockwell, "A novel texture feature based multiple classifier technique for roadside vegetation classification," Expert Systems with Applications, vol. 42, no. 12, pp. 5047-5055, 2015.

[20] B. Verma, M. Blumenstein, and M. Ghosh, "A novel approach for structural feature extraction: Contour vs. direction," Pattern Recognition Letters, vol. 25, no. 9, pp. 975-988, 2004.

[21] B. Verma, J. Lu, M. Ghosh, and R. Ghosh, "A feature extraction technique for online handwriting recognition," in IEEE International Joint Conference on Neural Networks vol. 2, pp. 1-8, 2004.

[22] F. Shaheen, B. Verma, and M. Asafuddoula, "Impact of automatic feature extraction in deep learning architecture," in International Conference on Digital Image Computing: Techniques and Applications pp. 1-8, 2016.

[23] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," Pattern Recognition Letters, vol. 30, no. 2, pp. 88-97, 2009.

[24] P. Sturgess, K. Alahari, L. u. Ladický, and P. H. S. Torr, "Combining appearance and structure from motion features for road scene understanding," British Machine Vision Conference, pp. 1-11, 2009.

[25] J. M. Alvarez, Y. LeCun, T. Gevers, and A. M. Lopez, "Semantic road segmentation via multi-scale ensembles of learned features," in European Conference on Computer Vision, pp. 586-595, 2012.

[26] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, pp. 2274-2281, 2012.

[27] X. Ren, L. Bo, and D. Fox, "RGB-(D) scene labeling: Features and algorithms," in IEEE Conference on Computer Vision and Pattern Recognition pp. 2759-2766, 2012.

[28] Y. Yang, Z. Li, L. Zhang, C. Murphy, J. Ver Hoeve, and H. Jiang, "Local label descriptor for example based semantic image labeling," in European Conference on Computer Vision pp. 361-375, 2012.

[29] L. U. Ladický, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr, "What, where and how many? Combining object detectors and CRFs," in European Conference on Computer Vision vol. 6314, pp. 424-437, 2010.

[30] Y. Wu, T. Yang, J. Zhao, L. Guan, and J. Li, "Fully combined convolutional network with soft cost function for traffic scene parsing," in International Conference on Intelligent Computing, vol. 10361, pp. 725-731, 2017.

[31] F. Visin et al., "ReSeg: A recurrent neural network-based model for semantic segmentation," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 426-433, 2015.

[32] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in European Conference on Computer Vision, pp. 44-57, 2008.