Segmentation of natural images based on super pixel and graph merging

Aritra Mukherjee\textsuperscript{1} | Soumik Sarkar\textsuperscript{2} | Sanjoy K. Saha\textsuperscript{1}

\textsuperscript{1}Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
\textsuperscript{2}Department of Information Technology, Jadavpur University, Kolkata, India

Abstract
The task of natural image segmentation is one of the most researched topics of computer vision. There are mainly two principal approaches for the task, the statistical approach and the supervised approach. The proposed methodology segments natural images combining a set of statistical algorithms. First, the image is preprocessed to enhance the edges. Weighted average of the denoised image and its derivatives is the preprocessed output. Thereafter, an energy based super pixelation is applied to over segment the image. Finally, a connectivity graph is built where nodes correspond to super pixels and edges connect the adjacent super pixels. The adjacent super pixels are merged based on the confidence value defined in terms of their textural and colour similarity. Proposed methodology has been applied on the images of BSDS500 dataset. Performance of the proposed work has been compared with that of other works based on detected edge maps. Few works generate ultrametric contour maps (UCM). To compare the performance with those works, UCM is also generated by the proposed methodology. To do so images at multiple scales are considered. It is observed that the output of segmentation is better in case of the proposed methodology. Proposed methodology is much faster than others. Thus, makes it suitable for real time application in robot vision.

1 | INTRODUCTION

Segmentation of natural images is one of the most challenging problems of computer vision and still an active area of research. Semantic segmentation is quite subjective in nature. Furthermore, factors like occlusion, illumination, texture and variation in orientation, context and so on make the task further critical. Devising a solution for a variety of images and one that can handle all these issues is still in demand. Though deep networks provide good results for a set of images but it suffers from huge training data requirement and susceptibility to pixel level noise [1] and uncertainty of operations. However, Bayesian deep learning is directed towards such issues [2,3]. The statistical algorithms on the other hand are free from training and physical significance of the parameters are easily interpreted. Thus the limitations of supervised approach and flexibility of statistical approach inspired us to choose the second one to work with.

The present work is motivated from the need to generate a lightweight system to segment natural scenes for possible usage in robotic applications like semantic simultaneous localization and mapping (SLAM). In most of the cases semantically similar regions in a natural image are coherent in terms of statistical features. Our objective is to exploit this aspect in developing a set of algorithms to segment the scene. Thereafter, individual segments can be used for semantic labelling. Though deep networks can perform the task, it is quite power consuming and thus not suitable for robots with limited resources. The proposed methodology first uses a fast super pixel segmentation for over segmenting the image and then groups the statistically similar super pixels to form segments. For such grouping graph merging based methodology has been adopted.

The rest of the paper is organized as follows: Section 2 provides a survey of past efforts. Proposed methodology is described in Section 3. Experimental result and concluding remarks are put in Sections 4 and 5, respectively.
2 | PAST WORK

Past efforts can be broadly categorized as supervised learning based approach and unsupervised image statistics based approach. A brief survey on the both the approaches are outlined as follows.

2.1 | Supervised learning based approaches

An early work by Martin et al. [4] is based on supervised approach. It detects contextual boundaries using local patches and certain global information. Finally builds the segmented map. This is achieved by training a probabilistic classifier that puts fuzzy marking on boundary pixels by evaluating its context defined by surrounding patch. A two stage approach was proposed by Dong et al. [5]. At the first stage colour quantization is achieved through clustering on LUV space. Subsequently, classification on reduced colour space is done using neural network [5]. In recent times, the use of deep convolutional neural network (DCNN) has become the leading trend. To generate the labelled map of pixels, deep network either works on the whole image or it works with patches and classifies pixels according to its neighbourhood patch. Bell et al. [6] segmented the image by labelling each pixel according to its physical material predicted from its surrounding patch using DCNN. Instead of using pixel level annotation Papanastasiou et al. [7] considered bounding box of the objects as the input for training the weakly supervised CNN with expectation minimization.

Deep learning is also utilized for semantic contour detection [8]. Chen et al. [9,10] employed DCNN, atrous spatial pyramid pooling and conditional random fields in respective order to segment natural images at pixel level. Deconvolution network or deconvenet is becoming popular that takes the whole image as input rather than pixel neighbourhood patches and generates the complete segmented map in one shot. It uses an encoder-decoder type architecture for segmentation. Works by Noh et al. [11] and Badrinarayanan et al. [12] are among the pioneers in this approach. Chen et al. [13] proposed a semantic image segmentation by using multiscaled input to an attention network for refining the pixel label. Lin et al. [14] proposed the idea of using multiscaled input to parallel networks that are further tuned with multipath refinement. Loss of information due to pooling is a major concern in deep networks, which is handled by fusion of multiple parallel networks trained with different variants of loss function. Though pixel level segmentation worked well with deep networks, most of the errors were due to incorrect segment boundaries. Atrous spatial pyramid pooling coupled with deconvolution network was proposed by Chen et al. [15] to handle the fuzzyness of categorization at object boundaries. In recent times, Xian et al. [16] proposed pixel level classification by a novel deep net architecture known as semantic projection network. It can be trained with very few instances using transfer learning. Wu et al. [17] developed the idea of unseen category classification by generating annotated synthetic datasets based on trained data and further adapting the network to it.

Supervised approaches works on the idea of training a classifier that labels pixels or patches based on the features learnt during training. With considerable amount of training and complex hierarchy of the classifier, good results are achievable but designing or training a universal model is very difficult. Though recent works are trying to train models with few training instances, their performance deteriorates with increasing intraclass feature diversity.

2.2 | Unsupervised image statistics based approaches

The unsupervised statistics based approach mainly focuses on dividing the image into a collection of pixels based on the continuity in terms of properties like colour, texture and so on. It is, generally, followed by merging or splitting of those collections. Different works have followed different criteria for split/merge process. Variants of graph cut has also been used. This approach intends to segment semantically without labelling the segments or pixels. However, another stage of recognition has to be considered for categorizing or labelling the segments.

The concept of normalized cut that uses intragroup similarity and intergroup dissimilarity metric, was first proposed by Shi and Malik [18]. It formed the basis of many graph based segmentation methods. Felzenszwalb et al. [19] used a pairwise region element comparison for gradually segmenting the fully connected spatial graph and merging them iteratively. Normalized cut coupled with mean shift segmentation was proposed by Tao et al. [20]. It was to reduce the complexity of graph cut. Zhang et al. [21] proposed a hybrid scheme combining the aspects of both statistical and supervised approaches. A weak supervised learning was considered that utilized Gaussian mixture model (GMM) on structural features of super pixels. Graphlet-cut (a variant of graph cut) that segments natural images [21] was proposed.

Careful design of feature extractor is the main aspect of statistical approach. A variety of features have been proposed. Spectral signature based approach for patch categorisation and subsequent fuzzy segmentation [22], interdistance discrimination between Laplacian of marginal histogram of linear filter output [23], integrating level set with GMM [24] are some of those. Rao et al. [25] used GMM along with adaptive chain coding for detecting edges of homogeneous regions.

Apart from using the graphs, tree-based approaches are also quite popular. An approach for detecting contours by combining spectral clustering with hierarchical region tree was proposed by Arbelaez et al. [26]. Tree partitioning over initial region descriptor features was suggested by Panagiotakis et al. [27]. It is further supplemented with Bayesian flooding and region merging. A simpler version using bottom up hierarchical segmentation was proposed by Alpert et al. [28]. The work is based on a probabilistic classifier that takes geometric, spectral
and textural cues on the neighbourhood patch of a pixel to label it. It is observed that graph cut approaches mainly use correlation clustering. In this case the variation in local boundary ambiguity prevents merging or splitting of sections. This problem can be handled by higher order correlation clustering as suggested by Kim et al. [29].

K-means is used widely for initial stages in the segmentation process. Zheng et al. [30] developed hierarchical generalized K-means to improve on the results. Recently, region merging by multiscale combinatorial grouping augmented with normalized cut was proposed by Pont-Tuset et al. [31] that produces considerably good results on natural images. Considering multiple hypothesis based on different distance metrics is a recent trend. By simple minimization of cost functions over the outcome of different merging schemes on an oversegmented image achieved new benchmarks [32].

It must be noted that statistical approaches do not label the pixels or regions semantically. Supervised approaches does labelling but that is mostly restricted within the pixels whose classes are in its training set. Otherwise, the pixels are ignored as background. As a result, the segmentation also suffers.

The past study indicates that natural image segmentation is still an open area and significant works are still being reported. Most of the current works are following the image feature based approaches supplemented by merge/split operation. It acts as motivation for us to explore similar approach. A previous work of ours [33] considered K-means-based segmentation of the image into spatially connected super pixels and further used a graph-cut based merging process to refine the results. With further experiments and analysis it was understood that a faster oversegmentation method to obtain the super pixels and a robust merging technique can be explored to make the process faster and to improve the accuracy. The methodology was developed mostly to cater segmentation of scenes used for mapping in SLAM. A complete map is needed for robot vision rather than partial map of labelled segments (which is restricted to trained categories). In general, deep learning and supervised approaches tends to segment only parts of the scene that can be labelled whereas a statistical approach segments the whole scene on visual cues. Thus our motivation is fuelled by the previous experience, past studies and specific requirements. In this context, present work focuses on devising a fast methodology to segment natural scenes.

3 | PROPOSED METHODOLOGY

The proposed methodology consists of multiple stages. The broad steps are preprocessing, super pixel segmentation, neighbourhood graph-cut-based merging. The methodology can also be extended to work on multiscaled versions of an image and thereby can generate edges of varying confidence. Preprocessing step intends to meet the contradictory criteria of noise removal and strengthening the edges. Super pixel segmentation over segments the image. The extracted segments are considered as the node in the graph cut based merging step that generates the final output. The overall block diagram of the process is shown in Figure 1. Individual steps are detailed in following sections.

3.1 | Preprocessing

Super pixel segmentation forms the core part of the proposed methodology. Its performance gets affected by the noise that corrupts the edges. Hence, it is important to get rid of such noises without significant compromise with edges. To achieve this requirement an edge preserving smoothing is adopted. A fast nonlocal means denoising technique [34] is applied. It removes the grain noise and simultaneously preserves bold texture features. As Super pixel-based segmentation relies on edges, we apply measures to enhance and sharpen the edges further. First- and second-order derivatives of the denoised image are computed. An weighted average of the denoised image and two derivatives is the final preprocessed version used in super pixel based segmentation. The derivatives help to sharpen the edges that may have been weakened by denoising. On the other hand, to reduce the impact of residual noise (if any) weights of the derivatives are kept low in the process of averaging. In our work, the weights of denoised image, first and second derivatives are of the ratio 4:2:1. Figure 2 shows the preprocessing steps.
### 3.2 | Super pixel segmentation

The Super pixel segmentation stage over segments an image and extracted segments are taken as the nodes for the graph based merging. We applied the SEED Super pixel method proposed by Bergh et al. [35]. It is faster than other state-of-the-art Super pixel algorithms [35]. That has become the motivation for considering SEED in our methodology as we intend to apply segmentation in the context of real time robotics application like SLAM. For understanding we provide outline the principle of the technique.

SEED super pixel method tries to segment the image into number of super pixels by maximizing an energy function. The energy function consists of two terms. One term focuses on the colour distribution in the super pixel. Thus, maximising the energy function tends to achieve colour homogeneity. The other term is called boundary term that takes care of the shapes of the super pixel boundaries. To compute the boundary term, a patch around the boundary pixel is considered. Around the edges, the patch covers multiple super pixels. In the process of maximization, essentially it tries to redistribute the edge pixels into neighbouring super pixels. Initially, image is divided into nonoverlapped blocks of small size. Such blocks are taken as the super pixels. The process goes on iteratively following hill climbing optimization algorithm updating the super pixel boundary and maintaining the colour consistency. In comparison to traditional region growing approaches or graph cut based approach, this method utilizes a local optimizer that works in a very efficient fashion by simple memory look-up. The process is extremely fast. The parameters like initial number of super pixels and the number of iterations are to be specified. Higher the number of super pixels or iteration, more is the time requirement. The process, in general, over segments an image. In the proposed methodology, the final step is graph based merging. Use of super pixels speeds up this process. Hence, for our purpose fine tuning of parameters to achieve optimal output is not essential. Moreover, such tuning will vary for different images. To make the process simple we have considered a fixed set of parameters that are 200 and 10 for initial number of super pixels and number of iterations, respectively. SEED algorithm returns a labelled image. The label of the pixel identifies the corresponding super pixel. The value ranges from 0 to \(N_{sp} - 1\) where \(N_{sp}\) is the number of super pixels detected by the algorithm.

Figure 3 shows the overall block diagram of super pixel segmentation. In general, SEED algorithm works with HSV model. But it is observed that images with substantial grayness where R, G, B are almost equal, the outcome suffers. In our work, either RGB or HSV model is considered depending on the colour variance. RGB image is divided into number of blocks. For each block average R, G, B values are computed. Variance of block level averages of the components \((\sigma_r, \sigma_g, \sigma_b)\) are then obtained. Average of \(\sigma_r\sigma_g\) and \(\sigma_b\) denotes the overall variance of the image in RGB model. For HSV model also the variance is computed in the similar way. The model that reflects more variance is taken as the input. Computation of variance at the block level makes the process faster. In our work, an image is divided into 25 blocks in \(5 \times 5\) fashion irrespective of aspect ratio.

### 3.3 | Graph-based merging

Super pixel segmentation over segments an image. As edges have a dominating role in the process, colour variation and textures present in an object result into multiple super pixels. At this stage, we take up a merging process to reduce such effects. In the proposed methodology, graph-cut based merging is a conglomeration of different distance functions that works on different features of the image. Primarily a neighbourhood graph is created where each node are the super pixels. There is an edge between two nodes if corresponding super pixels are neighbours.
in spatial domain. The graph is represented by an adjacency matrix $G_N$ of size $N_{sp} \times N_{sp}$, $G_{N_{ij}} = 1$ if super pixels $i$ and $j$ are spatial neighbours and $G_{N_{ij}} = 0$ otherwise. Similarity between each connected pair of nodes (super pixels) is studied for possible merging. We have focused on textural and colour similarity. Figure 4 shows the block diagram for the merging process.

### 3.3.1 Distance in terms of texture

The local binary pattern (LBP) [36] of an image captures the textural features around each pixel in a small eight-neighbourhood region. The histogram of LBP values in a super pixel captures the general texture of the same with the assumption that it does not vary drastically. Suppose, $i$ and $j$ are two connected super pixels and corresponding normalized LBP histograms are $H_i$ and $H_j$, respectively. Their similarity is measured using $\chi^2$ distance [37,38] as it tests consistency between the shapes of two histograms. Moreover, the measure can be easily interpreted and is robust towards vertical scaling of histograms. Similarity between the super pixels $i$ and $j$ is computed as:

$$CS_{ij} = \sum_{k=1}^{b} \frac{(H(k)_i - H(k)_j)^2}{(H(k)_i + H(k)_j)}$$  \hspace{1cm} (1)

The LBP histograms are quantized into $b$ bins to ignore minor variation in texture. In our experiment, $b$ is taken as 16. A low value for $CS_{ij}$ corresponds to high similarity in terms of texture.

### 3.3.2 Distance in terms of colour

The LBP captures the grayscale texture patterns. Two super pixels similar in terms of LBP may vary in colour. On the other hand, super pixels with perceptually similar colour may vary in terms of LBP. Hence, we consider colour based similarity also.

---

**FIGURE 4** Graph-based merging
Global measure and detailed measures are formulated to judge the
colour similarity of neighbouring super pixels.

The simplest measure for global colour similarity between
two neighbouring super pixels \( i \) and \( j \) is computed as
\[
AC_{ij} = \frac{|\mu_i - \mu_j|}{\mu_j + \mu_i}
\]
where \( \mu_i \) and \( \mu_j \) represent average R, G, B value of the super pixel \( i \) and \( j \), respectively.

Detailed measure is based on the Bhattacharya distance
[39] between the RGB histograms of two super pixels. It is
computed as follows:
\[
BD_{ij} = \left( 1 - \frac{\sum_{c=R, G, B} h^c_i H^c_j}{\sum_{c=R, G, B} h^c_i + \sum_{c=R, G, B} h^c_j} \right)
\]

where \( h^c_i \) and \( h^c_j \) represent the normalised histograms of colour channel \( c \) (i.e., R, G or B) for super pixels \( i \) and \( j \), respectively.

3.3.3 Merging

As described, similarity between each neighbouring node (super pixel) in the graph is measured as follows:

- Combined confidence
\[
V_{ij} = (CCS_{ij} - CCS_{thres}) + (CBD_{ij} - CBD_{thres}) + (CAC_{ij} - CAC_{thres}) + (CAH_{ij} - CAH_{thres})
\]

\( CCS_{thres}, CBD_{thres}, CAC_{thres} \) and \( CAH_{thres} \) stand for the thresholds corresponding to the individual confidence measures based on texture (LBP histogram), colour (RGB) histogram, average RGB and average hue, respectively. Selection of these thresholds is an issue. Tuning of those to obtain optimal result for every individual image is nontrivial. However, to get rid of complex selection process and to make the system robust we have preset the values by experimenting with a set of images of different types. It should be noted that thresholds are not tuned for maximising accuracy for just the test images. Rather, 20 images are chosen randomly from all 500 images of BSDS500 dataset, and an exhaustive search is performed on all four thresholds, to optimize the F-measure with respect to the groundtruth. The thresholds are varied from 0 to 1 with step size of 0.1. Finally, we have considered \( CBD_{thres} \) as 0.4 and for other thresholds, it is 0.9. Actually, global colour measure based on Bhattacharya distance is more sensitive to variation rather than others. \( V_{ij} \) can have both positive and negative value. A positive value signifies overall confidence towards similarity. If \( V_{ij} > 0 \) then super pixels \( i \) and \( j \) are marked for merging. After looking into all the neighbouring super pixels, a depth first search is carried out to merge the marked neighbours.

3.4 Generation of ultrametric contour map

Segmentation approaches are broadly of two categories like region extracting or edge generating. Proposed super pixel segmentation followed by graph cut based merging extracts the regions. To compare the performance with the edge based approaches, we have also dealt with multiscale version of the input image and generated ultrametric contour map (UCM) [26]. UCM is an edge map where edge pixel value denotes the confidence that it belongs to a semantic edge.

The process described so far corresponds to scale 1. The preprocessed image is furthered scaled with the factors 0.8, 0.6, 0.4 and 0.2. For each such version, SEED super pixel and graph merging steps are applied to generate the output. Maximum number of super pixels (input to SEED) is also reduced proportionately according to the scale. Though SEEDs determines the number of super pixels for a given image dimension and type (colour/greyscale), the reduced upper bound of the number of super pixels in each scale produces larger super pixels. Edges are extracted from each version. With the increase in size of super pixels, only the pronounced edges with sharp contrast are preserved. In order to generate the UCM, all edge maps are resized to scale 1, skeletonised and averaged. In the reduced version more and
more intraregion edges or weak edges are eliminated whereas dominating edges are retained across various scales.

Thus in the UCM more contextually relevant edges are emphasized but minor edges forming due to local interference are damped. This form of output does not report individual segment rather it reflects segment boundaries in a fuzzy manner. It is useful for applications where confidence of contextual edge is important.

Figure 5 shows the stepwise output for a sample image taken from BSDS500 dataset [26]. In the preprocessed image, it can be observed that edges are quite emphasised as it was intended. On the output of super pixel segmentation, the graph has been superimposed. The red links connect the neighbouring super pixels marked for merging. The final merged regions, groundtruth and UCM have also been shown.

4 | EXPERIMENTS AND RESULTS

The proposed methodology has been implemented in C++ language and opencv libraries [40] have been used. Experiments are performed on a Linux machine using single core of Intel i5 processor clocked at 3.0 GHz, 8 GB of DDR3 RAM and without any GPU support. BSDS500 dataset [26] has been used for experiment. The dataset consists of 500 images in a mix of portrait and landscape mode. The resolution of the images are 481×321 for landscape images and 321×481 for portraits. For evaluating the proposed methodology the resolution is increased by 1.5 times. Thus the size becomes to 721×481 (481×721) for landscape (portrait) images. The reason for doing so is to make the full scale image size similar to a standard webcam resolution. As commonly this resolution is used, it enables to evaluate efficiency for real scenario. The ground truth is also scaled up accordingly. The dataset provides variety of images of different categories like natural scenes, scenes with wildlife or animals, scenes focused on human subjects and scenes with man-made objects or structures. Both indoor and outdoor images are present. Images are in either portrait or landscape modes. The groundtruth of semantic segmentation has also been provided. Each image has been manually segmented by multiple human subjects. Finally, the dataset offers single human annotation (by randomly picking one out of multiple groundtruth) and multi-human annotation (multiple groundtruth by the individuals) for each image. An edge contour map is formed by averaging the groundtruths in multi-human annotation. It is treated as the groundtruth UCM where the edge strength is real valued and lie within [0,1].

We have applied proposed methodology on each image. The performance of the proposed methodology have been compared with number of other systems in terms recall, precision, F-measure and time. These are elaborated in the following section.
4.1 Comparison of performance

We have compared the performance of proposed methodology with the works of Felzenszwalb et al. [19], Arbelaez et al. [26], Pont et al. [31] and Bosch et al. [32]. Two of these works [19,32] have generated segmented image as output and other two [26,31] provide the output in UCM form. We have also provided the output in region form (i.e., scale is 1) and also in UCM form (i.e., averaging the edge maps at various scale). For comparison relevant one is considered accordingly.

Figure 6 shows a visual comparison of the output generated by the methodologies of Felzenszwalb et al. [19] and Bosch et al. [32]. For the groundtruth, the segments are shown in average of the colours of corresponding pixels. The
output segments of Felzenszwalb et al. [19] are randomly colour assigned by their source code. Hence, it should not be compared with the groundtruth based on the colour. A colour corresponds to a region and comparison with the groundtruth has to be made based on the regions. The output of the proposed methodology and that of Bosch et al. [32] are rendered in the same manner as the groundtruth. It can be observed that the images are quite oversegmented by the method of Felzenszwalb et al. [19]. The methodology does not perform any global optimization. Moreover, the local optimization is also affected by the size of components. Whereas, the methodology of Bosch et al. [32] optimizes the segmentation by comparing various different techniques and balances each other's local vulnerabilities. Finally, it results into a globally optimized output. Proposed methodology uses various measures for comparing only neighbouring super pixels for graph merging. So, in a way, it is the best of both. It excels in accuracy without compromising the speed. Visually, it is observed that output of the proposed methodology is better. It should be noted that comparison is performed on the basis of binary edge map of the groundtruth images and region output of Felzenszwalb et al. [19], Bosch et al. [32] and proposed methodology. As the segments are not labelled semantically so a region overlapping metric is not applicable in this case and thus pixel to pixel comparison between binary edge maps are performed instead. Figure 7 shows the visual comparison with the methodologies of Arbelaez et al. [26] and Pont et al. [31]. In this case the outputs are considered in UCM form. Visually, it is clear that the proposed methodology outperforms others.

For quantitative comparison of performance recall, precision and $F$-measure are considered as metrics. Once we have considered single human annotation as the groundtruth and again the UCM groundtruth obtained from multi-human annotation as described earlier. Table 1 shows the comparison with the works of Felzenszwalb et al. [19] and Bosch et al. [32]. Considering all the images in the dataset the average values are shown in the table. In terms of $F$-measure and recall, proposed methodology is better than others. However, precision is less. It is so because of under-segmentation, that is, missing the regions and hence the edges are considered costlier in our work. Proposed methodology oversegments the image. It results in a fall in precision. However, thresholds may be tuned to achieve higher precision but that will compromise with the recall. The $F$-measure of Felzenszwalb et al. is significantly lower due to its over simplification of local features and thus it generates too many segments. Bosch et al. performs close to our approach in terms of $F$-measure and its precision and recall values are also balanced. It uses a combination of multiple segmentation approaches with subsequent voting of the best result. Thus, its results are quite good in terms of accuracy. But, it is computationally heavy.

**TABLE 1** Comparison of performance based on edge map

| Methodology               | Average $F$-measure | Average precision | Average recall |
|---------------------------|---------------------|-------------------|----------------|
| Felzenszwalb et al. [19]   | 0.38                | 0.49              | 0.34           |
| Bosch et al. [32]         | 0.53                | 0.56              | 0.52           |
| Proposed methodology     | 0.56                | 0.48              | 0.73           |

**FIGURE 8** Comparison of $F$-measure based on UCM. UCM, ultrametric contour map
Although the time required for segmenting a full scale image does not qualify for real-time robotic applications, at slightly reduced scale the latency decreases in proportion to the area of the image. Hence, proposed methodology is much faster and can be utilized in robot vision application.

Table 2 shows the time required by the proposed methodology for images at different scales. With decrease in scale, the latency decreases in proportion to the area of the image. Although the time required for segmenting a full scale image does not qualify for real time robotic applications, at slightly reduced scale it is good enough. However, with higher computational setup it can achieve speed suitable for real-time application. Table 3 shows the comparison between the methodologies in terms of the time requirement. Felzenszwalb et al. [19] have dealt with large graph that makes the process slow. The methodology of Arbelaez et al. [26] first prepares contour image by using multiple local cues with nonlinear complexity. Pont et al. [31] have applied a fast normalized cut approach on in a hierarchical graph of super pixels to generate UCM at a scale. For both the works, process is repeated at multiple scales. Relatively, complex processing at each scale has been adopted in those works. Hence, proposed methodology is faster than those. Bosch et al. [32] applied multiple popular image segmentation schemes to keep the accuracy high and that makes the process slow. Thus, it is clear that proposed methodology is much faster than the others in both single scale and multiscale mode. At a reduced scale and/or with multicore implementation the methodology can handle six or more frames per second and can be utilized in robot vision application.

Table 3

| Methodology                  | Minimum time (ms) | Maximum time (ms) | Average time (ms) |
|------------------------------|-------------------|-------------------|------------------|
| Proposed work (full scale)   | 1525              | 2145              | 1685             |
| Proposed work (multi scale)  | 100              | 14,102            | 12,689           |
| Bosch et al. [32]            | 11,003            | 14,102            | 12,689           |
| Pont et al. [31]             | 8896              | 9986              | 9702             |
| Arbelaez et al. [26]         | 8579              | 12,054            | 9342             |
| Felzenszwalb et al. [19]     | 1525              | 2145              | 1685             |

Figure 8 shows the variation of $F$-measure with threshold for the methods of Arbelaez et al. [26], Pont et al. [31] and the proposed one. In this case, comparison is based on UCMs generated by the methods and groundtruth UCM. Threshold (i.e., confidence) on the UCMs including the groundtruth are applied. For each UCM, the confidence values are normalized and varied from 0.5 to 0.95 to generate the binary edge images for comparison. Considering all the images in the dataset, average $F$-measure has been shown. The plot shows that $F$-measure for the proposed methodology is much higher than the others over a major range. It indicates that proposed methodology detects the true edges with considerably high confidence and hence survives even at higher threshold.

Table 2

| Scale of image | Minimum time (ms) | Maximum time (ms) | Average time (ms) |
|----------------|-------------------|-------------------|------------------|
| 1.0 (full scale) | 287               | 359               | 317              |
| 0.8            | 182               | 234               | 204              |
| 0.6            | 103               | 131               | 112              |
| 0.4            | 43                | 57                | 48               |
| 0.2            | 12                | 20                | 17               |
| All scales     | 633               | 788               | 691              |

5 | CONCLUSION

In this a work, a novel methodology has been presented for natural image segmentation that elegantly combines different techniques. In the preprocessing stage weighted average of denoised image and its derivatives helps to enhance the edges along with noise suppression. SEED super pixel method is used to generate the super pixels. On the basis of their adjacency, the super pixels are represented in a graph and a merging scheme is applied based on the similarity confidence of the connected super pixels. The confidence has been defined based on the textural, global and detailed colour similarity between the adjacent super pixels. The performance of the proposed methodology on BSDS500 dataset has been compared with number of other works and found superior. Proposed methodology is found to be much faster and can be utilized in real-time application.

REFERENCES

1. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 427–436 (2015)

2. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision?. Adv. Neural Inf. Process. Syst. 5580–5590 (2017)

3. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. arXiv preprint arXiv:170507115, (2017)

4. Martin, D.R., Fowlkes, C.C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Trans. Pattern Anal. Mach. Intell. 26(5), 530–549 (2004)

5. Dong, G., Xie, M.: Color clustering and learning for image segmentation based on neural networks. IEEE Trans. Neural Network. 16(4), 925–936 (2005)

6. Bell, S., Upchurch, P., Snavely, N., Bala, K.: Material recognition in the wild with the materials in context database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3479–3487 (2015)

7. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. Proceedings of the IEEE international Conference on Computer Vision, pp. 1742–1750 (2015)

8. Shen, W., Wang, X., Wang, Y., Bai, X., Zhang, Z.: DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3092–3091 (2015)

9. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv preprint arXiv:160600915, (2016)
10. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40(4), 834–848 (2018)
11. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. Proceedings of the IEEE International Conference on Computer Vision, pp. 1520–1528 (2015)
12. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:151100561, (2015)
13. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attenion to scale: Scale-aware semantic image segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3640–3649 (2016)
14. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1925–1934 (2017)
15. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
16. Xian, Y., Choudhury, S., He, Y., Schiele, B., Akata, Z.: Attention to scale: Scale-aware semantic image segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3640–3649 (2016)
17. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1925–1934 (2017)
18. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 888–905 (2000)
19. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vis. 59(2), 167–181 (2004)
20. Tao, W., Jin, H., Zhang, Y.: Color image segmentation based on mean shift and normalized cuts. IEEE Trans. Syst. Man Cybernetics, Part B (Cybernetics). 37(5), 1382–1389 (2007)
21. Zhang, L., Song, M., Liu, Z., Liu, X., Bu, J., Chen, C.: Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1908–1915 (2013)
22. Torralba, A., Oliva, A.: Statistics of natural image categories. Network Comput. Neural Syst. 14(5), 391–412 (2003)
23. Heller, M., Schnörr, C.: Natural image statistics for natural image segmentation. Int. J. Comput. Vis. 63(1), 5–19 (2005)
24. Xie, Z., Wang, S., Hu, D.: New insight at level set & Gaussian mixture model for natural image segmentation. Signal Image Video Process. 7, 521–536 (2013)
25. Rao, S.R., Mohabi, H., Yang, A.Y., Sastry, S.S., Ma, Y.: Natural image segmentation with adaptive texture and boundary encoding. Asian Conference on Computer Vision, pp. 135–146 (2009)
26. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 33(5), 898–916 (2011)
27. Panagiotakis, C., Grinias, I., Tziritis, G.: Natural image segmentation based on tree equipartition, Bayesian flooding and region merging. IEEE Trans. Image Process. 20(8), 2276–2287 (2011)
28. Alpert, S., Galun, M., Brandt, A., Basri, R.: Image segmentation by probabilistic bottom-up aggregation and cue integration. IEEE Trans. Pattern Anal. Mach. Intell. 34(2), 515–527 (2012)
29. Kim, S., Yoo, C.D., Noowzin, S., Kohli, P.: Image segmentation using higher-order correlation clustering. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1761–1774 (2014)
30. Zhang, Y., Leon, B., Xu, D., Wu, Q., Zhang, H.: Image segmentation by generalized hierarchical fuzzy c-means algorithm. J. Intell. Fuzzy Syst. 28(2), 961–973 (2015)
31. Pont, J., Arbelaez, P., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE Trans. Pattern Anal. Mach. Intell. 39(1), 128–140 (2017)
32. Bosch, M., Gifford, C.M., Dress, A.G., Lau, C.W., Skibo, J.G., Christie, G.A.: Improved image segmentation via cost minimization of multiple hypotheses. arXiv preprint arXiv:180200088. (2018)
33. Mukherjee, A., Sarkar, S., Saha, S.K.: Segmentation of natural image based on colour cohesion and spatial criteria. Ninth International Conference on Advances in Pattern Recognition (ICAPR), pp. 1–7 (2017)
34. Buades, A., Coll, B., Morel, J.M.: Non-local means denoising. Image Processing On Line. 1, 208–212 (2011)
35. den Bergh, M.V., Boix, X., Roig, G., Gool, L.J.V.: Seeds: Super pixels extracted via energy-driven sampling. CoRR, 2013:3848. abs/1309.
36. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recogn. 29(1), 51–59 (1996)
37. Rubner, Y., Tomasi, C.: Perceptual metrics for image database navigation. Springer Science and Business Media; 2001:594.
38. Zhang, D., Lu, G.: Proceedings of the 2003 International Conference on Neural Networks and Signal Processing. vol. 2. IEEE; 2003:928–931. Evaluation of similarity measurement for image retrieval
39. Aherne, F.J., Thacker, N.A., Rockett, P.I.: The Bhattacharyya metric as an absolute similarity measure for frequency coded data. Kybernetika. 34(4), 363–368 (1998)
40. Bradski, G.: Opencv library. Dr Dobb’s Journal of Software Tools. (2000)

How to cite this article: Mukherjee A, Sarkar S, Saha SK. Segmentation of natural images based on super pixel and graph merging. IET Comput. Vis. 2021;15:1–11. https://doi.org/10.1049/cvi2.12008