Saliency Detection for Content Aware Computer Vision Applications

Manipoonchelvi Pandivalavan and Muneeswaran Karuppiah
Department of Computer Science and Engineering, Mepco Schlenk Engineering College, India

Abstract: *In recent years, there has been an increased scope for intelligent computer vision systems, which analyse the content of multimedia data. These systems are expected to process a huge quantum of image/data with high speed and without compromising on effectiveness. Such systems are benefited by reducing the amount of visual information by selectively processing only a relevant portion of the input data. The core issue in building these systems is to reduce irrelevant information and retain only a relevant subset of the input visual information. To address this issue, we propose a region-based computational visual attention model for saliency detection in images. The proposed model determines the salient object or part of the salient object without prior knowledge of its shape and color. The proposed framework has three components. First, the input image is segmented into homogeneous regions and then smaller regions are merged with neighbouring regions based on color and spatial distance between them. Second, three attributes such as spatial position, color contrast and size of each region are evaluated to distinguish salient object/parts of salient object. Finally, irrelevant background regions are suppressed and the region level saliency map is generated based on the three attributes. The generated saliency map preserves the shape and precise location of salient regions and hence it can be used to create high quality segmentation masks for high-level machine vision applications. Experimental results show that our proposed approach qualitatively better than the state-of-the-art approaches and quantitatively comparable to human perception.*

Keywords: *Content aware processing, saliency detection, computational visual attention.*

Received September 23, 2014; accepted September 15, 2014

1. Introduction

Computational visual saliency mechanism plays an important role in a variety of content aware image processing applications including salient object detection [13], image segmentation [17], image retargeting [2], multimedia content analysis and robotic control [3], compression [14] and visual search [4]. The computational visual saliency detection mechanism drives the focus of attention to appropriate regions of interest instead of processing the whole image. A number of these models implement a bottom-up mechanism which is data driven, task independent and require no a prior knowledge about the content of input image. Most of these models rely on the Feature Integration Theory (FIT) [16] which suggests that in human visual attention mechanism, features/stimulus are automatically registered in parallel and then objects are identified separately. The models based on these theory starts with extracting low-level features such as color, intensity, orientation and spatial frequency. Each of these extracted features is evaluated to compute the feature Saliency Map (SM) that is a two-dimensional grayscale image in which the brightness of a pixel is proportional to its salience. The generated feature SM s are then strategically combined into a final SM of visual attention.

Saliency values are calculated based on frequency domain analysis [1], supervised learning [13] and multi scale analysis [7, 10]. The existing saliency detection mechanisms generate a low-resolution SM, which does not highlight whole salient object regions and poorly define object boundaries. It indicates the locations of salient pixels and does not highlight a salient region. To overcome the limitation with a low-resolution SM we propose a region based model with the aim to produce a full resolution (same as that of the input image) SM which can be used by content based image processing applications. The SM generated by our approach uniformly highlights salient objects/parts of salient objects, suppresses irrelevant background regions and preserves the object's shape and size. We exploit spatial location, color contrast and size of regions to generate a high quality SM. Both subjective and objective evaluations illustrate that the proposed region based approach achieves higher quality results compared to state-of-the-art saliency detection mechanisms.

The remainder of the paper is organized as follows: section 2 explains the related research work in the field of computational visual saliency computation mechanism. Section 3 describes the proposed saliency detection mechanism. The validation methodologies and experimental results are discussed in section 4. Section 5 concludes this paper.

2. Related Work

The basis for many Computational Visual Attention (CVA) system is FIT of attention. A model of FIT is depicted in Figure 1. According to this theory, attention is driven by two mechanisms namely bottom-up and top-down mechanism. The bottom-up approach is a fast, task independent and purely data-driven mechanism. The top-down model is slower than bottom-up approach, volitionally controlled, subjective, and task dependent. Both these approaches generate SM. Since the proposed approach aims to find unknown object/region of interest, bottom-up model is used. This model does not require prior knowledge about the content in an image.

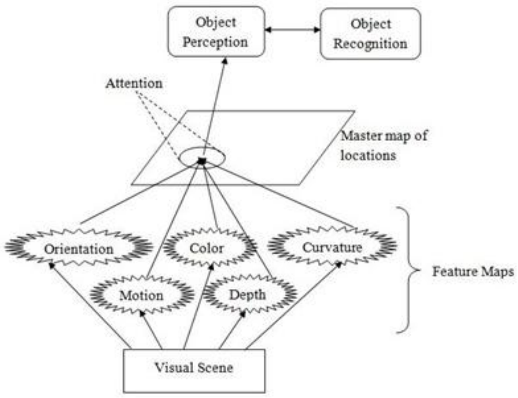


Figure 1. A model of feature integration theory.

A biologically plausible bottom-up system was first introduced by Koch and Ullman [12] based on FIT, and their theoretical model was implemented by Itti *et al.* [10]. This model hierarchically decomposes an image into multiple scales and computes visual saliency from color, intensity, and orientation. This method computes multi-scale image features using center-surround operations. The center-surround operation determines contrast by taking the differences between a fine (center) and a coarse scale (surround) for a given feature and produces feature map. The feature maps (6 for intensity, 12 for 2 chromatic channels, and 24 for orientation) are combined into three conspicuity map at scale four. These three conspicuity maps are normalized to a fixed range and then they are summed into the final SM. The SM generated by this approach is also called as spotlight SM, which can only highlight the center portion and/or the high-contrast boundaries of salient objects, but cannot suppress the high-contrast background regions. The size of the SM generated by this approach is smaller than that of the input image. Because of the hierarchical nature of the process, this model is computationally expensive.

Hou and Zhang [9] presented a model that is independent of features, categories, or other forms of prior knowledge of the objects. This model is based on the hypothesis that the spectral residual contains the novel or rare parts of an image. Their model obtains the salient locations by subtracting the log of Fourier spectrum from the general shape of log spectra. The residuals obtained by this subtraction process serve like the compressed representation of a scene. Using an inverse Fourier transform, the compressed representation is further transformed into the spatial domain resulting in the SM. The SM thus contains the nontrivial part of the scene. To improve the result, a Gaussian filter is used to smooth the SM. This method is simple to implement but produces low resolution and blurry SM.

Goferman *et al.* [7] proposed a graph-based context-aware saliency detection mechanism, which aims at detecting the image regions that represent the scene and not just the most salient object. This model identifies both fixation points and the dominant object. Their method imposes a regular grid and extracts patches at each scale. Each pixel is represented by the set of multi-scale image patches centered on it. A pixel is considered salient when its enclosing patch is highly dissimilar to all other image patches. Multiple scale processing is incorporated to further decrease the saliency of background patches, as they are more likely to repeat at multiple scales. A pixel is considered attended if its saliency value exceeds a certain threshold. Furthermore, each pixel outside the attended areas is weighted according to its Euclidean distance to the closest attended pixel. The method produces low resolution SM that highlights objects' boundaries and suppresses homogenous regions.

The low-resolution SMs are less useful for high-level vision applications as they imprecisely specify objects and their boundaries. To address the issue the region based approaches have been proposed [1, 5].

Achanta *et al.* [1] proposed a multi-scale saliency model which is based on color and luminance contrast. The underlying hypothesis of their model is that local contrast of an image region with respect to its neighbourhood at various scales derives fixation location. They compute saliency from the distance between the average feature vectors of the pixels of an image sub-region to that of its neighbourhood. This allows obtaining a combined feature map at a given scale by using feature vectors for each pixel, instead of combining separate SMs for scalar values of each feature.

The approach proposed by Cheng *et al.* [5] first, segments the input image into regions, then, computes color contrast at the region level, and defines the saliency for each region as the weighted sum of the region's contrasts to all other regions in the image. The weights are set according to the spatial distances with farther regions being assigned smaller weights. Both these region-based methods [1, 5] produce full resolution SM but fail to uniformly highlight the entire salient region.

From the detailed study, it is concluded that there is a need for a saliency detection mechanism that provides SMs with the following attributes:

- Precisely locate salient objects or parts of salient

objects in an image
- Uniformly highlight salient objects or parts of salient objects of all sizes
- Produce well-defined boundaries of salient objects
- Efficiently suppress the irrelevant background
- Produce full resolution SM.

3. Region based Saliency Detection

The major components of the proposed region based saliency detection mechanism are given in Figure 2.

We partition an image into a set of regions $R = \{R_1, R_2, R_3, \cdots R_N\}$ where N is the total number of regions. For each region R_i, a saliency value is determined from its spatial location, color contrast and size. We use spatial location, size, and region connectivity to effectively suppress the irrelevant regions. Once regions saliency is determined, the saliency value of each region is uniformly assigned to all pixels belongs to the respective region.

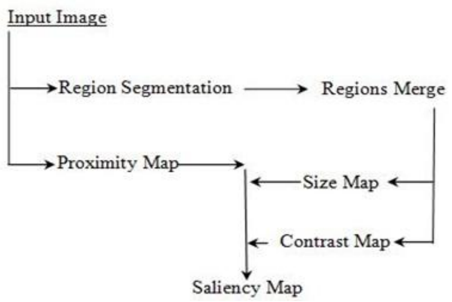


Figure 2. Components of the proposed system.

Section 3.1. details how to partition the input image into regions. Section 3.2. describes each visual attribute of a region and the procedure for computing these attributes. Section 3.3. elaborates how to effectively suppress irrelevant background region and construct the final SM.

3.1. From Pixels to Regions

The input image in the RGB color space is first, transformed into the La*b* color space. Each channel, the luminance and the two chrominance channels, is uniformly quantized into b bins and then, the three dimensional histogram H_g with b x b x b is calculated using all pixels in the image. The parameter b controls the number of quantized colors in the histogram. The optimal value of b is computed by minimizing the cost function $C(b) = \frac{2 * I_\mu - I_\sigma}{b^2}$. Here I_μ and I_σ are mean and standard deviation of the intensity of the input image. The peaks or local maxima of the histogram are fed as seeds to k-means clustering procedure to segment similar pixels. The identified regions have homogeneous color distribution in the image space. Some regions may be too small to constitute regions of interest. Hence, we merge smaller regions, whose size is less than a predefined threshold value, with bigger regions whose area is above the threshold value. A region R can be merged to its nearest neighbor if and only if:

- Area(R) < AREAThreshold
- Let ρ_i and ρ_j be the pixels on the outer boundary of a region R_i and R_j respectively. R_i can be merged with R_j iff $\rho_i \subseteq \rho_j$
- If a smaller region (R_i) has more than one neighbouring regions, it will be merged with a neighbouring region which is close in terms of color and distance. The closet neighbour is $argmin_{1 \le j \le n}$ (*Dist* (R_i, R_j)). Here n is the number of neighbours and *Dist* is the average of Euclidean spatial and color distance between the two regions.

The region segmentation and region merge results are shown in Figure 3. For illustration, each segmented region is represented using the regions' mean color in RGB space. As shown in the figure, the original image is partitioned into 586 regions and after the merging process, the number of regions is reduced to 13.

a) Input image. b) Region segmentation. c) Regions merge.

Figure 3. Segmentation result.

3.2. Saliency Computation

It is observed from a variety of images that a salient object is perceptually distinguished from any other regions in the image and has distinctive attributes, which makes the object pop out from its surroundings. The discriminating attributes could be the object's color, intensity, spatial location, texture, curvature and so on. Based on these aspects we exploit color, spatial location, and size to measure the saliency of regions.

Human observers tend to focus on known objects or center of an image or both at the same time [11, 15]. To emphasize this location prior concept in our model, we utilized Euclidean distance to measure the location-based saliency of each pixel and represented it as a proximity map. Every pixel in the proximity map indicates its physical distance to the center of the image and its value fall into the range of [0, 1]. The location-based saliency of a region is evaluated from the proximity value of each pixel p in the region.

$$S_l(R_i) = \frac{\sum_{p \in R_i} PM(x_p)}{Area(R_i)} \quad (1)$$

Where *PM* denotes the value of the proximity map at a point p and x_p denotes the coordinates of p. $S_l(R_i)$ achieves a higher value when the region R_i is near to the center of an image.

The color-based saliency is evaluated from the color contrast between a region and its surroundings.

$$S_C(R_i) = \frac{\sum_{j=1,2,\ldots,nAdjoin} D_C(R_i, R_j)}{nAdjoin} \tag{2}$$

$$D_C(R_i, R_j) = \left\| R_i(L,a,b) - R_j(L,a,b) \right\| \tag{3}$$

Where region R_i is surrounded by *nAjoin* number of regions and $R(L, a, b)$ is the mean color of the region in La*b* color space.

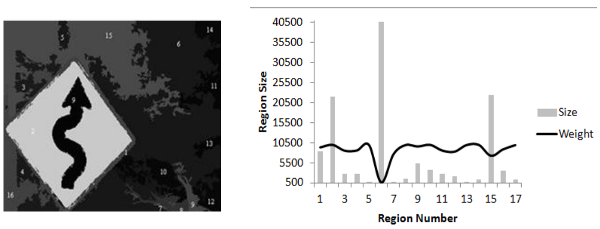
The salience of a region depends on its size relative to that of the whole image [8]. According to this theory, we consider size as one of the factors to decide the significance of a region. The size-based saliency is defined as follows.

$$A_C(R_i) = \frac{\sum_{j=1,2,\ldots,nAdjoin} D_S(R_i, R_j)}{nAdjoin} \tag{4}$$

$$D_S(R_i, R_j) = \left\| Area(R_i) - Area(R_j) \right\| \tag{5}$$

$$S_s(R_i) = e^{-\frac{(A_c(R_i) - \mu_{A_c})^2}{2\sigma^2_{A_c}}} \tag{6}$$

Where A_c is the area dissimilarity measure, μ_{A_c} is the mean of area dissimilarity measure and σ_{A_c} is the standard variance of area dissimilarity measure. The Gaussian model of normalization controls assigning higher salience value for the largest regions. The influence of the size normalization is demonstrated in Figure 4.



a) Illustration of regions. b) Weight age of regions based on area.

Figure 4. The influence of the size normalization.

Most of the bottom-up visual saliency computation model process low-level features to generate feature maps, then that are combined into a final SM. Several strategies have been proposed to combine multiple features maps into a SM. The most common strategies are normalized summation, winner takes all (i.e., maximum value among the feature maps), pixel by pixel multiplication and linear/nonlinear combination with learnt weights. In the proposed model, the region wise proximity map, color map and size map are combined into a unique saliency map S as follows.

$$S(R_i) = \frac{S_l(R_i) * S_C(R_i) * S_S(R_i)}{\max\left(S_l(R_i) * S_C(R_i) * S_S(R_i)\right)} \tag{7}$$

The denominator is for the purpose of normalization.

3.3. Background Suppression

In general, scenes are organized into perceptual groups and a set of regions are bound together to form an object. We exploit this connectedness principle to suppress irrelevant background regions. The scattered regions that are not connected to the most salient regions are assumed to be the part of the background and hence those regions are removed from the final SM. In summary, we employ both selection and elimination process to generate a SM, we retain the salient regions and remove irrelevant scattered regions to generate the SM.

The proximity map, the color map, and the size map for an example image is shown in Figure 5. From the figures, we can observe that all pixels that constitute the salient regions are uniformly highlighted and the background regions are efficiently suppressed in the SM. The highlighted salient regions are accordant with human visual attention system.

a) Input image.

b) Proximity map.

c) Size map.

d) Contrast map.

f) Integrated saliency map.

Figure 5. The size map for an example image.

4. Experiments and Results

We performed extensive experiments on the publicly available image data set [13] with the manually segmented ground truths for salient objects [1]. We compared our region based saliency detection model with four state-of-the-art saliency models including visual attention based model (Itti) [10], context aware model (CASD) [7], Frequency tuned model (FSRD) [1] and global contrast model (GCSR) [5]. We used executables, with default parameters or SMs provided by the authors for the state-of-the-art saliency models. For comparison purpose, we up sampled all SMs to the full resolution of input images.

The evaluation criterion for comparing different saliency models depends on the application. In this paper, we subjectively and objectively evaluated the quality of the generated SM for a salient object segmentation application. The experimental results on some sample images are shown in Figure 6.

The subjective evaluation indicates that the proposed model can extract salient objects more precisely than other models. Further, the proposed model uniformly highlights the parts of the salient objects and more effectively suppress the irrelevant background regions.

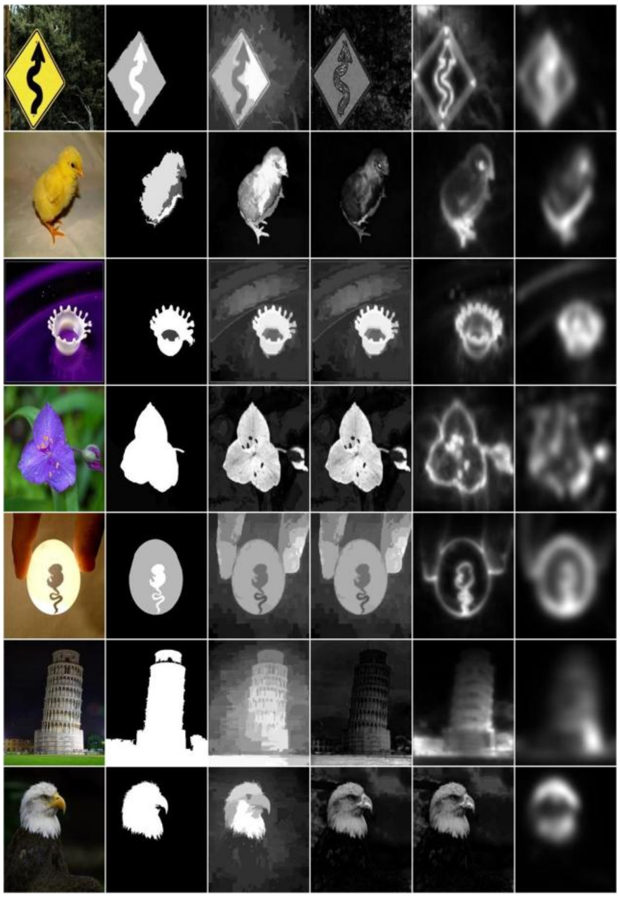


Figure 6. Saliency map comparison: From left to right: input image, proposed model, GCSR [5], FSRD [1], CASD [7], Itti [10].

In order to objectively quantify the performance of various saliency detection models we adopted SM accuracy measure (Q) proposed in [6]. The goal of the objective evaluation was to check whether the generated SM contained enough information for salient object extraction.

$$Q(GT;SM) = \frac{A(GT \cap SM)}{A(GT) + A(SM) - A(GT \cap SM)}, \tag{8}$$

Here GT is the ground truth, SM is the SM, and $A(x)$ is the area of x.

The Figure 7 illustrates the cumulative-performance curve $p(x) = [0,100] \to [0,1]$, which describes the performance distribution of all images in the database. The horizontal axis represents the percentage of total number of images. The vertical axis represents the cumulative performance of SM accuracy. A specific point (x, p(x)) on the curve indicates that in x percent of the images the common areas between the ground-truth and the identified significant regions are lower than p(x). Equivalently, this also means that in (1-x) percent of the images the common areas between the ground-truth and the identified significant regions are greater than p(x). Figure 6 illustrates that the proposed model outperforms the other saliency detection models on effectively sketching the salient object in a given image.

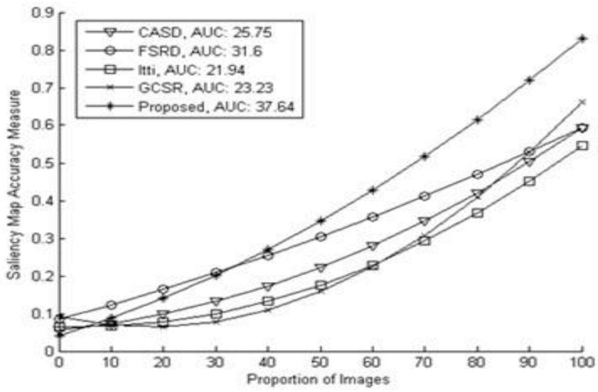


Figure 7. Object comparison for salient object detection.

5. Conclusions

In conclusion, we have presented a saliency detection model based on the region-based approach. In our model, spatial position, color contrast, and size contrast are evaluated for every region. These measures are combined to generate a SM with full resolution. The proposed model generates a SM with full resolution, which uniformly highlights the parts/whole of salient object in the input image. The subjective and objective evaluations demonstrate that the proposed model can be exploited by salient object segmentation and other content based high-level vision applications such as content-based image resizing and image coding.

References

[1] Achanta R., Hemami S., Estrada F., and Susstrunk S., "Frequency-Tuned Salient Region Detection," *in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597-1604, 2009.

[2] Avidan S. and Shamir A., "Seam Carving for Content-Aware Image Resizing," *ACM Transactions on Graphics*, vol. 26, no. 3, 2007.

[3] Begum M. and Karray F., "Visual Attention for Robotic Cognition: A Survey," *IEEE Transactions on Autonomous Mental Development*, vol. 3, no. 1, pp. 92-105, 2011.

[4] Bruce N. and Tsotsos J., "Saliency, Attention, and Visual Search: An Information Theoretic Approach," *Journal of Vision*, vol. 9, no. 3, pp. 1-24, 2009.

[5] Cheng M., Zhang G., Mitra N., Huang X., and Hu S., "Global Contrast Based Salient Region Detection," *in Proceeding of IEEE Computer Conference Computer Vision and Pattern Recognition*, Colorado Springs, pp. 409-416, 2011.

[6] Ge F., Wang S., and Liu T., "Image-Segmentation Evaluation From the Perspective of Salient Object Extraction," *in Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, pp. 1146-1153, 2006.

[7] Goferman S., Zelnik-Manor L., and Tal A., "Context-Aware Saliency Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915-1926, 2012.

[8] Hoffman D. and Singh M., "Salience of Visual Parts," *Cognition*, vol. 63, no. 1, pp. 29-78, 1997.

[9] Hou X. and Zhang L., "Saliency Detection: A Spectral Residual Approach," *in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, pp. 1-8, 2007.

[10] Itti L., Koch C., and Neibur E., "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998.

[11] Judd T., Ehinger K., Durand F., and Torralba A., "Learning to Predict Where Humans Look," *in Proceeding of IEEE 12th International Conference on Computer Vision*, Kyoto, pp. 2106-2113, 2009.

[12] Koch C. and Ullman S., "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219-227, 1985.

[13] Liu T., Yuan Z., Sun J., Wang J., Zheng N., Tang X., and Shum H., "Learning to Detect a Salient Object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353-367, 2011.

[14] Rehman A. and Saba T., "An Intelligent Model for Visual Scene Analysis and Compression," *The International Arab Journal of Information Technology*, vol. 10, no. 2, pp. 126-136, 2013.

[15] Scholl B., "Objects and Attention: The State of the Art," *Cognition*, vol. 80, no. 1-2, pp. 1-46, 2001.

[16] Treisman A. and Gelade G., "A Feature-Integration Theory of Attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97-136, 1980.

[17] Zhang Q., Gu G., and Xiao H., "Image Segmentation Based on Visual Attention Mechanism," *Journal of Multimedia*, vol. 4, no. 6, pp. 363-370, 2009.

Manipoonchelvi Pandivalavan received the bachelor of engineering degree in computer science and engineering from Bharathidasan University, Tamilnadu, India, in 1997, and the master of engineering degree in computer science and engineering from the Regional Engineering College, Tamilnadu, India, in 2001. She joined HCL Technologies, India, as Member Technical Staff in 2001 and left the company as Associate Project Manager in 2010. During her tenure at HCL Technologies she worked on image registration, rear view aid system for automobiles and traffic signal detection system. She is currently, pursuing Ph.D. degree at Department of Computer Science and Engineering in Mepco Schlenk Engineering College, India, affiliated to Anna University. Her current research interests include semantics based image segmentation, object tracking and, computer vision applications.

Muneeswaran Karuppiah received the bachelor of engineering degree in Electronics and Communication engineering from Madurai Kamarajar University, Tamilnadu, India in 1984 and the master of engineering in computer science and engineering from Bharathiyar University, Tamilnadu, India, in 1990. In 2006, he received the Ph.D. degree in computer science engineering from M.S. University, Tamilnadu, India. He is in teaching and research for the past 28 years and 12 years respectively and currently, he is working as professor in Computer Science and Engineering Department at Mepco Schlenk Engineering College, Tamilnadu. His research interests are image processing, neural networks, and semantics analysis. He has authored or co-authored about 75 publications in journal/conference level and one book on compiler design.