A Knowledge Driven Computational Visual Attention Model

Amudha J [1], Soman. K.P [2] and Padmakar Reddy. S [3]

**[1] Department of computer science, Amrita School of Engineering
Bangalore, Karnataka, India**

**[2] Department of Computational Engineering and Networking, Amrita School of Engineering
Coimbatore, Tamilnadu, India**

**[3] Department of computer science, Amrita School of Engineering
Bangalore, Karnataka, India**

Abstract

Computational Visual System face complex processing problems as there is a large amount of information to be processed and it is difficult to achieve higher efficiency in par with human system. In order to reduce the complexity involved in determining the saliency region, decomposition of image into several parts based on specific location is done and decomposed part is passed for higher level computations in determining the saliency region with assigning priority to the specific color in RGB model depending on application. These properties are interpreted from the user using the Natural Language Processing and then interfaced with vision using Language Perceptional Translator (LPT). The model is designed for a robot to search a specific object in a real time environment without compromising the computational speed in determining the Most Salient Region.

Keywords: *Visual Attention, Saliency, Language Perceptional Translator, Vision.*

1. Introduction

Visual attention is a mechanism in human perception which selects relevant regions from a scene and provides these regions for higher-level processing as object recognition. This enables humans to act effectively in their environment despite the complexity of perceivable sensor data. Computational vision systems face the same problem as humans as there is a large amount of information to be processed. To achieve computational efficiency, may be even in real-time Robotic applications, the order in which a scene is investigated must be determined in an intelligent way. The term attention is common in everyday language and familiar to everyone. Visual attention is an important biological mechanism which can rapidly help human to capture the interested region within eye view and filter out the minor part of image. By means of visual attention, checking for every detail in image is unnecessary due to the property of selective processing. Computational Visual Attention (CVA) is an artificial intelligence for simulating this biometric mechanism. With this mechanism, the difference feature between region centre and surround would be emphasized and integrated in a conspicuity map. Given the complexity of natural language processing and computer vision, few researchers have attempted to integrate them under one approach. Natural language can be used as a source of disambiguation in images since natural language concepts guide the interpretation of what humans can see. Interface between natural language and vision is through a noun phrase recognition systems. A noun phrase recognition system is a system that given a noun phrase and an image is able to find an area in an image where what the noun phrase refers to is located. One of the main challenges in developing a noun phrase recognition system is to transform noun phrases (low level of natural language description) in to conceptual units of a higher level of abstraction that are suitable for image search. The goal is to understand how linguistic information can be used to reduce the complexity of the task of object recognition. However, integrating natural language processing and vision might be useful for solving individual tasks like resolving ambiguous sentences through the use of visual information.

The various related works in the field of computational visual attention model are discussed in Section 2. Section 3 explains the system architecture and Language Processing model. The Section 4 gives the implementation details with analysis of the model followed by conclusion in section 5.

2. Related Work

The various models which identify the salient region are analyzed in this section. Frintrop proposed a Visual Attention System for Object Detection and Goal directed search (VOCUS) [1]. Laurent Itti, Christof Koch and Ernst Niebur [5] proposed an algorithm to identify the saliency region in an image using linear filtering. The authors describe in detail how the feature maps for intensity, orientation, and colour are computed. All computations are performed on image pyramids that enable the detection of features on different scales. Additionally, they propose a weighting function for the weighted combination of the different feature maps by promoting feature maps with few peaks and suppressing those with many ones. Simone Frintrop, Maria Klodt and Erich Rome [6] proposed a bottom-up approach algorithm for detection of region of interest (ROI) in a hierarchical way. The method involves smart feature computation techniques based on integral images without compromise on computational speed. Simone Frintrop, Gerriet Bracker and Erich Rome [2] proposed an algorithm where both top-down and bottom-up approaches are combined in detection of ROI by enabling the weighting of features. The weight's are derived from both target and back ground properties**.** The task is to build a map of the environment and to simultaneously stay localized within the map which serves as visual landmarks for the Robot. Simone Frintrop and Markus Kessel proposed a model for Most Salient Region tracking [10] and Ariadna Quattoni [3] has proposed a model for detection of object using natural language processing, which is used in system discussed here.

In psychophysics, top-down influences are often investigated by so called cuing experiments. In these experiments, a "cue" directs the attention to the target. Cues may have different characteristics: they may indicate *where* the target will be, or *what* the target will be. A cue speeds up the search if it matches the target exactly and slows down the search if it is invalid. Deviations from the exact match slow down search speed, although they lead to faster speed compared with a neutral cue or a semantic cue. This is the main motivation behind integrating the verbal cues to the attention model to enhance the search speed which is experimentally verified.

3. System Architecture

The block diagram in Fig.1 describes the flow of the system. The system architecture describes two major modules. 1) Language Perceptional Translator (LPT) [3] 2) Visual Attention Model (VAM) [1, 4, 7, 8, 9].

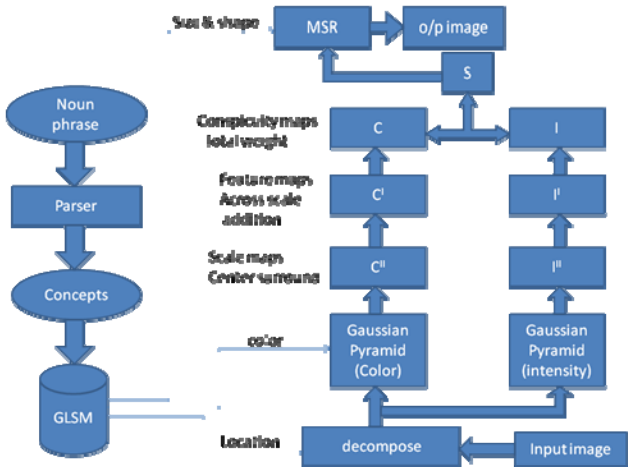


Fig. 1 Visual Attention Model with NLP.

1) LPT: One of the main challenges in developing a noun phrase recognition system is to transform noun phrases (low level of natural language description) into conceptual units of a higher level of abstraction that are suitable for image search. That is, the challenge is to come up with a representation that mediates between noun phrases and low-level image input. The Parser processes the sentence and it outputs the corresponding properties like location, Color, Size, Shape and for the Thing (object). We must construct a "grounded" lexicon semantic memory that includes perceptual knowledge about how to recognize the things that words refer to in the environment. A "grounded" lexical semantic memory would therefore connect concepts to the physical world enabling machines to use that knowledge for object recognition. A GLSM (Grounded Lexical Semantic Memory) is a data-structure that stores know- ledge about words and their relationships. Since the goal of LPT is to transform a noun-phrase into perceptual constraints that can be applied to visual stimuli to locate objects in an image. The outputs of GLSM is given to the VAM at different processing levels like location property at decomposition level, Color property at Gaussian pyramid construction and Size and Shape property after detecting of salient region to identify the required object in an image.

2) The Visual Attention model (VAM) identifies the most attended region in the image. The following sections present the algorithm in detail.

3.1 Visual Attention Model

The 1st level of bottom-up visual attention shown in fig.1 is decomposition of an image based on location property. We divided the image based on index method as shown in Fig .2 as Top, Left, Right, etc.

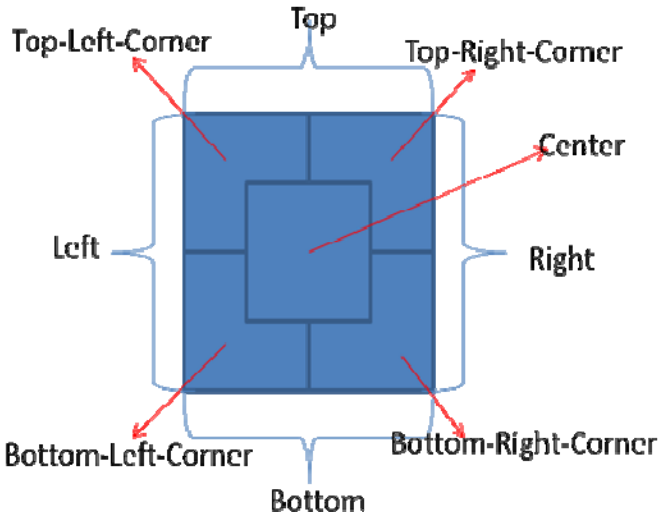


Fig. 2 Dividing Image by Index Method.

The image I is divided into 9 different parts and the default option is an entire image. Here the location cue property determines the search region to detect an object which reduces the possibility to shift the focus of attention to other objects in an entire image due to intensity or color is reduced when we crop the image based on location property.

In our approach we used to detect the sign boards which uses the prior knowledge of location has Top-Left-Corner or Top-Right-Corner. Before decomposing the image based on location cue matrix is converted in to NxN square Matrix by resizing the image I. I is divided in to 9 parts with different Location Cues are shown in Fig.3.

$$I = \begin{pmatrix} I^{(1,1)} \ldots\ldots\ldots\ldots I^{(1,N)} \\ \ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots \\ I^{(N,1)} \ldots\ldots\ldots\ldots I^{(N,N)} \end{pmatrix}_{N \times N}$$

Original Image

$$I_1 = \begin{pmatrix} I^{(1,1)} \ldots\ldots\ldots\ldots I^{(1,N/2)} \\ \ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots \\ I^{(N/2,1)} \ldots\ldots\ldots\ldots I^{(N/2,N/2)} \end{pmatrix}_{N/2 \times N/2}$$

Left Top Corner

$$I_2 = \begin{pmatrix} I^{(1,N/2+1)} \ldots\ldots\ldots\ldots I^{(1,N)} \\ \ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots \\ I^{(N/2+1,N/2)} \ldots\ldots I^{(N/2,N)} \end{pmatrix}_{N/2 \times N/2}$$

Right Top Corner

$$I_3 = \begin{pmatrix} I^{(N/2+1,1)} \ldots\ldots\ldots I^{(N/2+1,N/2)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ I^{(N,1)} \ldots\ldots\ldots\ldots\ldots I^{(N,N/2)} \end{pmatrix}_{N/2 \times N/2}$$

Left Bottom Corner

$$I_4 = \begin{pmatrix} I^{(N/2+1,N/2+1)} \ldots\ldots\ldots I^{(N/2+1,N)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ I^{(N,N/2+1)} \ldots\ldots\ldots\ldots I^{(N,N)} \end{pmatrix}_{N/2 \times N/2}$$

Right Bottom Corner

$$I_5 = \begin{pmatrix} I^{(N/4+1,N/4+1)} \ldots\ldots\ldots I^{(N/4+1,3N/4)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ I^{(3N/4,N/4+1)} \ldots\ldots\ldots I^{(3N/4,3N/4)} \end{pmatrix}_{N/2 \times N/2}$$

Center

$$I_6 = \begin{pmatrix} I^{(1,1)} \ldots\ldots\ldots\ldots\ldots I^{(1,N)} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ I^{(N/2,1)} \ldots\ldots\ldots\ldots I^{(N/2,N)} \end{pmatrix}_{N/2 \times N}$$

Top

$$I_7 = \begin{pmatrix} I^{(N/2+1,1)} \ldots\ldots\ldots\ldots I^{(N/2+1,N)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ I^{(N,1)} \ldots\ldots\ldots\ldots\ldots I^{(N,N)} \end{pmatrix}_{N/2 \times N}$$

Bottom

$$I_8 = \begin{pmatrix} I^{(1,1)} \ldots\ldots\ldots\ldots I^{(1,N/2)} \\ \ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots \\ I^{(N,1)} \ldots\ldots\ldots\ldots I^{(N,N/2)} \end{pmatrix}_{N \times N/2}$$

Left

$$I_9 = \begin{pmatrix} I^{(1,N/2+1)} \ldots\ldots\ldots\ldots I^{(1,N)} \\ \ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots \\ I^{(N,N/2+1)} \ldots\ldots\ldots\ldots I^{(N,N)} \end{pmatrix}_{N \times N/2}$$

Right

Fig. 3 Different Locations of Image (I) with respective Matrices.

The input image I is sub-sampled into a Gaussian pyramid on 4 different scales, and each pyramid level is decomposed into channels for red(R), green (G), blue (B), yellow (Y), intensity (I) using (1), (2), (3), (4) and (5) .

Fig. 4 Opponent colors.

$$I = (r + g + b)/3 \quad (1)$$

$$R = r - (g + b)/2 \quad (2)$$

$$G = g - (r + b)/2 \quad (3)$$

$$B = b - (r + g)/2 \quad (4)$$

$$Y = r + g - 2(|r - g| + b) \quad (5)$$

Depending on color property cue from the GLSM the priority of which color is high and which is low is set on different color channels of red(R), green (G), Blue (B), and Yellow(Y). The Color opponent process is a color theory that states that the human visual system interprets information about color by processing signals from cones and rods [in an antagonistic manner]. Opponency is thought to reduce redundant information by de-correlating the photoreceptor signals. It suggests that there are three opponent channels Red Vs Green, Blue Vs Yellow, Dark Vs White. Response to one color of an opponent channel are antagonistic to those to the other color, i.e. one color produces an excitatory effect and the other produces an inhibitory effect, the opponent colors are never perceived at the same time (the visual system can't be simultaneously excited and inhibited).The decision on which color channel to be used is based on the color cue. The output of the feature maps are then fed to the center-surround. These 5 channels are fed to the center surround differences after resizing all the surround images to the center image. Center-Surround operations are implemented in the model as difference between a fine and a coarse scale for a given feature. The center of the receptive feature corresponds to the pixel at the level c ∈ {2, 3} in the pyramid and the surround corresponds to the pixel at the level s = c+1. Hence we compute three feature maps in general case. One feature type encodes for on/off image intensity contrast, two encodes for red/green and blue/yellow double component channels. The intensity feature type encodes for the modulus of image luminance contrast. That is the absolute value of the difference between the intensity at the center and the intensity in the surround as given in (6).

$$I^{II}_{(I,C,S)} = N(|I(c)\Theta I(s)|) \quad (6)$$

The quantity corresponding to the double opponency cells in primary visual context are then computed by center surround differences across the normalized color channels. Each of the three-red /green Feature map is created by first computing (red-green) at the center, then subtracting (green-red) from the surround and finally outputting the absolute value. Accordingly maps RG(c,s) are created in the model to simultaneously account for red/green and green/red double opponency and BY(c,s) for blue/yellow and yellow/blue double opponency using (7) and(8).

$$C^{I}_{RG,C,S} = N(|R(c) - G(c))\Theta(R(s) - G(s))|) \quad (7)$$

$$C^{I}_{BG,C,S} = N(|B(c) - Y(c))\Theta(B(s) - Y(s))|) \quad (8)$$

The feature maps are then combined into two conspicuity maps, intensity $\bar{I}$ (9), color $\bar{C}$ (10), at the saliency map's scale (σ= 4). These maps are computed through across-scale addition (⊕), where each map is reduced to scale four and added point-by-point:

$$I = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(I(c,s)) \quad (9)$$

$$C = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [N(RG(c,s) + N(BY(c,s))] \quad (10)$$

The two conspicuity maps are then normalized and summed into the input S to the saliency map (11).

$$S = (N(I) + N(C)) \quad (11)$$

The N(.) represents the non-linear Normalization operator. From the saliency map the most attention region is identified by finding the maximum pixel value in the salient region. The identification of the segmented region can be made based on size and shape property.

4. Results and Analysis

The system developed is tested on a dataset where the attention object is a signboard. The various signs in the dataset are bike, crossing and pedestrian symbols. The number of testing samples used for analysis is as shown in Table 1. The cues that are used in the dataset are the location cues, the color cue, the size and shape cue pertaining to the object signboard. In table 2 the verbal cues that mostly suit for the chosen dataset is shown.

Table 1: Testing samples for signboard detection

Type of Image	*Total No. of Images*
Bike	*16*
Crossing	*16*
Pedestrian	*16*

Table 2: Cues for data set

Location	*Color*	*Size*	*Shape*	*Thing*
Right top Corner	Red	Large/Small	Triangle	Sign board
Right top Corner	Blue	Large/Small	Rectangle	Sign board
Right top Corner	Blue	Large/Small	Circle	Sign board

The analysis is done with and without cues. Visual attention model without cues has NxN i.e. N^2 computations at each level, where as with cues depending on Location Property the number of computations is reduced to $N^2/4$ or $N^2/2$ at each level to get Region of Interest. Priority for color is chosen by trial and error method with different combinations of inhibiting and exhibiting channels. The system developed is tested under various cases scenarios like

a) No verbal cues are given to the system.
b) Only the color property is obtained.
c) Only the location (region information available).
d) Both color and location information.

VAM is tested and compared with the different combination of cues like only color, only location, both color and location and without cues as shown in Table 3.

Table 3: VAM with different combinations of Cues.

Images	*Total No. of Images*	*No. of images Detected with different combinations*			
		No Cues	*Only Color*	*Only Location*	*Both Color and Location*
Bike	16	3	10	4	15
Crossing	16	8	7	9	12
pedestrian	16	3	15	5	15

In Table 4 VAM is tested with both location and color cues for the same data set with varying the color priority.
The VAM decides excite the weights to frame channels to enhance the color information in the image in the following ways. For identifying the Red color Signboards.

1. Increment Red and decrement Green component by a factor of 0.5.
2. Increment Red and Green component by a factor of 0.7.
3. Increment Red component by 0.7 and decrement Green, Blue, and Yellow components by a factor of 0.3.
4. Double the Red component and decrement Green, Blue and Yellow by a factor of 0.3.

For identifying the Blue color Signboards replace the Red color with Blue and Blue color with the Red and repeat the above 4 steps and the same as shown in Table 4.

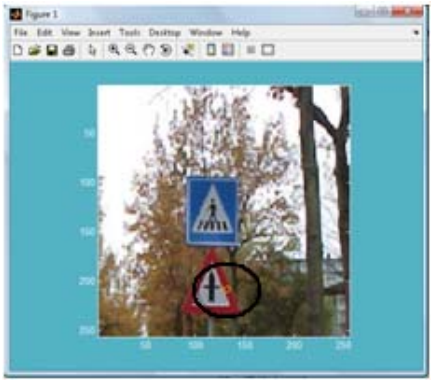
Table 4: Testing sign board data set with different Priority levels.

Images	*No. of Correctly detected images with color priority.*			
	R_i& G_d by 50%	R_i by 70% & G_d by 30%	R_i by 70% &(G,B,Y)_d by30%	Double R & (G, B,Y)_d by 30%
Crossing Priority RED color	6	4	10	12
	B_i&Y_d by 50%	B_i by 70% & Y_d by 30%	B_i by 70% & (R,G,Y)_d by 30%	Double B & (R,G,Y)_d by 30%
Bike Priority BLUE color	10	13	12	15
Pedestrian Priority BLUE color	10	12	13	14

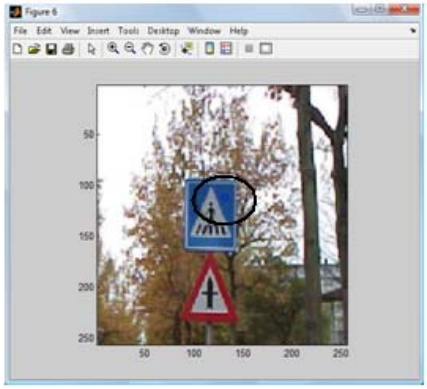
The Symbol's R/B/G/Y_i indicates Red/Blue/Green/ Yellow color priority increased and R/G/B/Y_d indicates Red/Green/Blue/Yellow color priority decreased.

To the VAM system the input Sign board image shown in Fig.4 (a) is given as input to the VAM and input to the LPT is noun phrase which is "Find the Red color Sign board on "Right_top_corner". So, here the desired color cue is Red, location cue is Right_top_corner and the object is Sign board. The result of VAM is shown in Fig.4 (b) and when the color cue is Blue is shown in Fig.4(c). The performance with different priority levels shown in Table 4 and for the same color cue is shown in Fig (5).

(a)



(b)



(c)

Fig.5 Image with both Crossing and pedestrian sign boards (a) Input Image to the system (b) Output of VAM with Color and Location cues. (c) Result of VAM with Color and Location cues.

In Fig. 6 **Type 1** indicates, increment R, B and decrement G, Y by a factor of 0.5. **Type 2** indicates, increment R, B by a factor of 0.7 and decrement G, Y by a factor of 0.3. **Type 3** indicates, increment R/B by a factor of 0.7 decrement G, B/R, Y by a factor of 0.3. **Type 4** indicates double R/B component and decrement G, B/R, Y by a factor of 0.3.

The **Type 4** system performance is much better than other systems, hence the system assigns color cue weightage based on Type 4. Comparison between the various visual attention models on computation of the number of maps computed for identifying the salient region is shown in Table 5. The statistics clearly depict the computation of the map which is less in case of VAM in comparison with VOCUS and Itti's Model. In case of VAM with verbal cue color it is only 52 maps. In case of VAM with location cue the computation of the number of maps remains the same but the image size is reduced to half or quarter of the original size which reduces the time taken for computation.

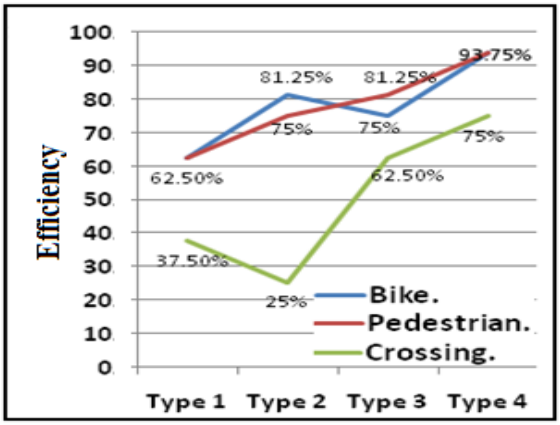


Fig.6 Performance with different sign board images and with different types of priority color Cues.

Table.5. Comparison of Maps in various models

Various Maps in the Architecture	***Various Visual Attention Architecture Maps at different levels.***				
	Itti's	VOCUS	VAM	VAM with verbal cue color	VAM with verbal cue location
Pyramid Maps	24	28	45	30	45
Scale Maps	42	48	14	12	14
Feature Maps	7	10	7	6	7
Conspicuity Maps	3	3	3	3	3
Saliency Map	1	1	1	1	1
Total Maps	77	100	65	52	65

Comparison between the various visual attention models on computation of the number of maps computed for identifying the salient region is shown in table5. The statistics clearly depict the computation of the map which is less in case of VAM in comparison with VOCUS and Itti's Model. In case of VAM with verbal cue color it is only 52 maps. In case of VAM with location cue the computation of the number of maps remains the same but the image size is reduced to half or quarter of the original size which reduces the time taken for computation

5. Conclusion

The computation of saliency region is determined with and without decomposing the image and the time taken to compute the most salient region with decomposition takes less time in comparison without decomposition. The verbal cue also reduces the number of maps computed for determining the saliency. The other cues for size and shape which reduces the time taken to identity the object hasn't been implemented and is left for future scope of the system. The various other issues like

combination of the verbal cues which will result in a flexible architecture for visual attention has to be studied extensively with a language interface.

6. References

[1] Frintrop, S. VOCUS: A Visual Attention System for Object Detection and Goal directed Search. PhD thesis Rheinische Friedrich-Wilhelms-University at Bonn Germany (2005). Published 2006 in Lecture Notes in Artificial Intelligence (LNAI), Vol. 3899, Springer Verlag Berlin/ Heidelberg.

[2] Frintrop, S., Backer, G. and Rome, E. Goal-directed Search with a Top-down Modulated Computational Attention System. In: Proc. of the Annual meeting of the German Association for Pattern Recognition DAGM 2005 Lecture Notes in Computer Science (LNCS) Springer (2005) 117–124.

[3] Ariadna Quattoni, Using Natural Language Descriptions to aid object Recognition. PhD thesis University of Massachusetts, Amherst Massachusetts, 2003.

[4] Frintrop, S., Jensfelt, P. and Christensen, H. Attentional Landmark selection for Visual SLAM. In: Proc. of the International Conference on Intelligent Robots and Systems (IROS '06) (2006).

[5] Itti, L., Koch, C. and Niebur, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE* Transactions on Pattern Analysis and Machine Intelligence 20 (11, 1998) 1254–1259.

[6] Simon Frintrop, Maria Klodt, and Erich Rome. A Real time Visual Attention System Using Integral Images, in proc of the 5th international conference on ICVS 2007, Bielefeld, Germany,March 2007.

[7] Wei-song Lin and Yu-Wei Huang. Intention-oriented Computational Visual Attention Model for learning and seeking image Content. Department of Electrical engineering National Tiwan University . 2009 IEEE Tranaction.

[8] Simone Frintrop, Patric jensfelt and Henrik Christensen. Attentional Robot Localization and Mapping at the ICVS Workshop on Computational Attention and Applications, (WCAA), Bielefeld, Germany, March 2007.

[9] Cairong Zhao,ChuanCai Liu, Zhihui Lai, Yue Sui, and Zuoyong Li**.** Sparse Embedding Visual Attention Model IEEE Transaction 2009.

[10] Simone Frintrop and Markus Kessel," Most Salient Region Tracking", IEEE 2009 international conference on Robotics and Automation (ICRA'09), Kobe, Japan, May 2009.

Amudha Joseph is an assistant Professor in Amrita School of Engineering, Bangalore. Her qualifications include B.E., Electrical and Electronics Engineering from Bharathiyar University, Coimbatore, India in 1998 and M.E. Computer Sceince and Engineering , Anna University, Chennai, India in 2003. Her research interests include image processing, computer vision and soft computing.

Dr. K.P Soman is the head, CEN, Amrita Vishwa Vidyapeetham Amrita Vishwa Vidyapeetham, Ettimadai, Coimbatore-641105. His qualifications include B.Sc. Engg. in Electrical engineering from REC, Calicut.P.M. Diploma in SQC and OR from ISI, Calcutta.M.Tech (Reliability engineering) from IIT, KharagpurPhD (Reliability engineering) from IIT, Kharagpur.Dr. Soman held the first rank and institute silver medal for M.Tech at IIT Kharagpur. His areas of research include optimization, data mining, signal and image processing, neural networks, support vector machines, cryptography and bio-informatics. He has over 55 papers in national and international journals and proceedings. He has conducted various technical workshops in India and abroad.

Padmakar Reddy.S is a Post graduate student in Amrita School of Engineering, Bangalore, Karnataka. His qualifications include B.Tech. in Electronics and Communication Engineering in Madanapalli Institute of Technology & Sciences, Madanapalli, Andhra Pradesh, India. His research interests include image processing and Embedded Systems.