International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 12 (2015) pp. 31585-31594

http://www.ripublication.com

Top-Down Approach for Modelling Visual Attention using Scene Context Features in Machine Vision

Manjunath R Kounte and Dr. B.K. Sujatha

Research Scholar, JAIN University
School of Electronics and Communication Engineering
Bengaluru, India
manjunath.kounte@gmail.com
Professor, Department of Telecommunication Engineering
M.S. Ramaiah Institute of Technology and Management
Bengaluru, India
bksujatha@msrit.edu

Abstract

Present state-of-the-art approaches to model human visual attention include high level object detections demonstrating top down image semantics in a separate channel along with other bottom up saliency channels. Top down image approach play a major role in predicting and analyzing where people look in images. However, multiple objects in a scene are competing to attract our attention and the literature survey shows that this interaction is snubbed in most of the current models. To overcome this limitation, we review the object context based visual attention model which incorporates the co-occurrence of multiple objects in a scene for visual attention modeling. The proposed review is based on regression based algorithm and uses several high level object detectors for faces, people, cars, text and understands how their joint presence affects visual attention.

Index Terms— Computational Cognitive Neuroscience, Saliency Map, Visual Attention, Visual Attention Region, top-down approach.

I. INTRODUCTION

Humans are able to swiftly process a rich stream of visual data and extract informative regions suitable for high level cognitive tasks. Therefore, understanding the manner in which human's process visual stimuli in a free viewing scenario has been an interesting problem in the scientific and engineering community. Several

applications in computer vision (object recognition [1], visual tracking [2], text detection [3]), graphics (non-photo realistic rendering [4]

), multimedia (video summarization [5], video compression [6]) and robotics (robot localization [7]) can benefit from better under- standing of human visual attention. A detailed overview of various saliency algorithms and its applications are presented in [8].

Early visual attention models [9, 10] are pure bottom up approaches and use multiple low level image features such as intensity, color, orientation, texture and motion to determine regions of interest in natural images. In these approaches feature specific saliency maps are computed for every low level feature and the final master map is a linear or nonlinear combination of individual feature specific saliency maps. However, in meaningful scenes, top down factors such as task at hand and image semantics play a major role in capturing attention.

Recent research [11] suggests that when subjects view natural scenes, faces and text primarily attract attention. A mathematical model using this information to improve human attention prediction was proposed in [4] which utilizes multiple object detectors (car, person and face) and low level saliency maps. A linear SVM is trained on these features to predict human attention regions in an image. This approach essentially learns a single weight for each feature vector. In practice, a single weight for each object irrespective of the scene content can be a severely limiting assumption. It is known that human visual attention, irrespective of top-down task is biased towards faces and text. The first step towards obtaining scene semantic prior from eye tracking information alone is to build models that predict face and text regions in images, which is the primary focus of the paper. This information is useful to improve the speed and precision of state-of-the-art detectors for challenging categories such as text, cats and dogs. We note that the performance of state-of-the-art cat and dog detectors in turn depends on head (face) detection algorithm which can be enhanced using eye movement information. However, presence of interesting semantic objects initiates change in visual attention from low-level to high level context and current visual attention models do not explicitly model this transition.

Also, recent research using controlled experiments [12, 13] highlight the importance of object co-occurrence and context for visual search tasks. These works indicate that other objects in a scene can provide a distracting (sometimes positive) effect for visual search of a specific object using reaction time studies. In a similar perspective, our effort aims to model the effect of object co-occurrence for a free viewing task and helps in creating a better organization of interesting regions in a scene.

In addition, previous learning based approach [4] artificially generates a classification problem by thresholding the attention map to estimate visual saliency. However, as the attention maps are continuous, it naturally presents itself as a regression problem. To summarize, the primary contribution of our work is to create a novel framework which predicts visual attention by modeling object co- occurrence in a scene using a regression approach. A comparison of our algorithm with existing state of the art visual attention algorithms in the MIT eye tracking dataset yields encouraging results.

Section II has a detailed assessment of Visual attention modelling using various conceivable methods which include Psychophysics, Computational Methods and Neurophysiology. We will emphasize on computational modelling which can be further classified into filter models, neural models and computational cognitive neuroscience models.

Section III gives information on extracting early visual features and visual processing from the retina.

Section IV describes Scene Context Feature hypothesis for extraction of saliency map and identifying the visual attention region followed by Conclusion and future work in last section

.

II. MODELLING VISUAL ATTENTION

A. Taxonomy

The advantages of Modelling Visual Attention includes in the field of Human-robot interaction, Robotics like Active vision, Robot Navigation, Robot Localization, Synthetic vision for simulated actors. In various other fields like Advertising, Finding tumor's in mammograms, Retinal prostheses etc.

In the field of Computer Vision and Graphics like Image segmentation, Image re-targeting, Image matching, Image rendering, Image and video compression, Image thumb nailing, Image quality assessment, Image super-resolution, Image super resolution, Video summarization, Scene classification, Object detection, Salient object detection, Object recognition, Visual tracking, Dynamic lighting, Video shot detection, Interest point detection, Automatic collage creation Face segmentation and tracking.

In human beings, the attention is facilitated by a retina that has evolved a high-resolution central fovea and a low resolution periphery. The important parts of scene gathering and collection of important information is guided by the anatomical structure of the retina. We focus on the Computational Modelling of this interesting field.

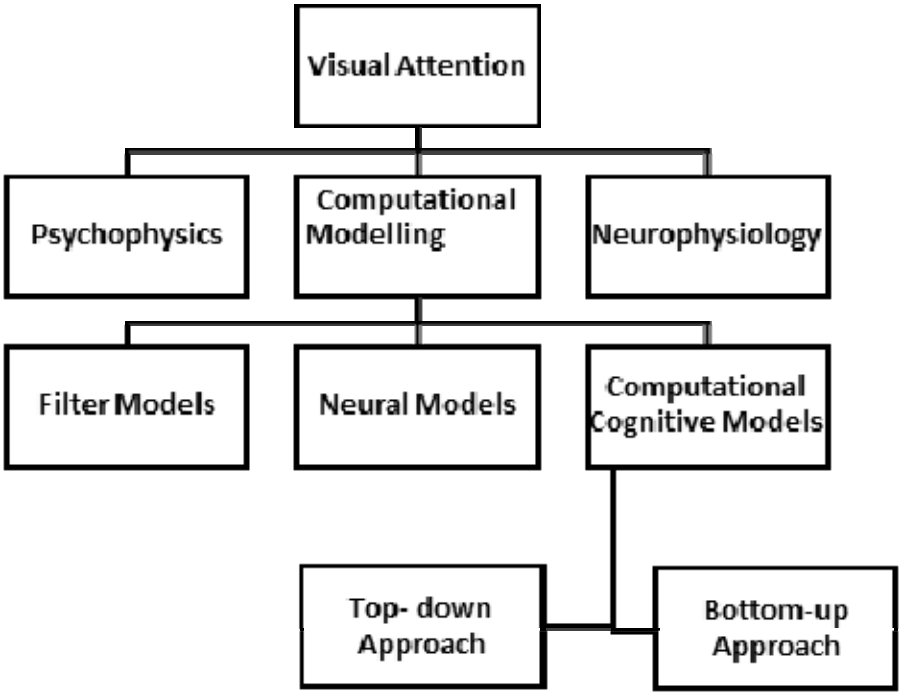

Fig. 1. Taxonomy of Visual Attention Models.

As Shown in the figure 1, there are various approaches to the modelling of the visual attention. Research is being carried out by scientists from numerous domain of science. Foremost among them are the Psychologists who have studied behavioral correlates of visual attention such as change blindness, inattentional blindness, and attentional blink. Whereas Neurophysiologists have shown that how the neurons accommodate themselves to better represent objects of interest.

In Computational Modelling approach, the filter models have built models that can compute saliency maps and realize the Visual attention based on Top-Down Attentional Models and Bottom-Up Models. In neural network models, the approach is to simulate and explain attentional behaviors. [27]

Although there are many models available now in the research areas mentioned above, here we propose a new model based on computational Cognitive Neuroscience to build systems capable of working in real-time.

.

III. EXTRACTING EARLY VISUAL FEATURES

A. Visual Processing

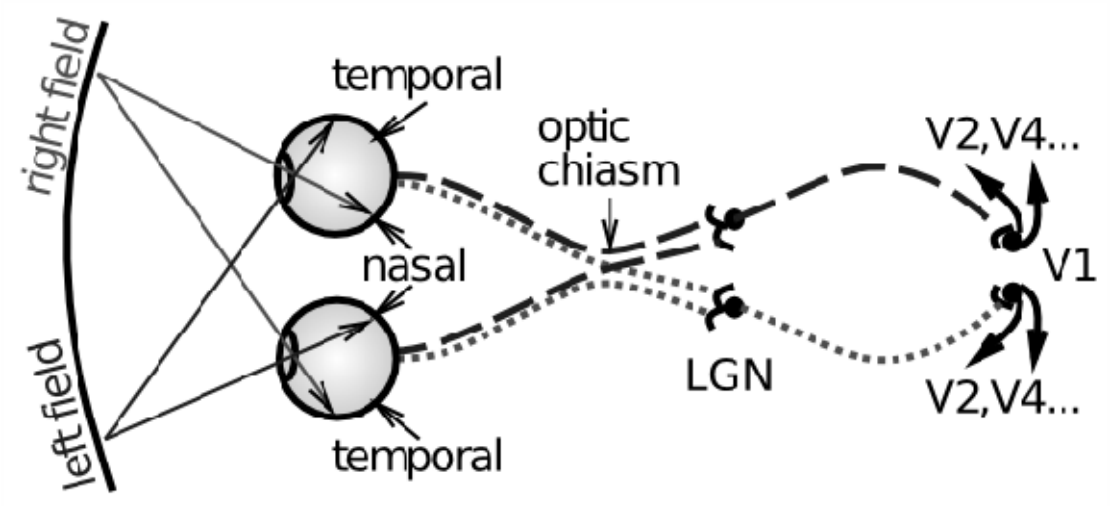

Fig. 2. Visual Processing from the retina through lateral geniculate nucleus of the thalamus to primary visual cortex.

Figure 2, shows the graphical representation of basic optics and transmission corridors of input visual signals, which entre through the retina, and headway to the lateral geniculate nucleus of the thalamus (LGN), and further to primary visual cortex (V1). The primary organizing principles at work here, and in other perceptual modalities and perceptual areas more generally, are: i) Transduction of different information -- in the retina, photoreceptors are sensitive to different wavelengths of light (red = long wavelengths, green = medium wavelengths, and blue = short wavelengths), giving us color vision, but the retinal signals also differ in their spatial frequency (how coarse or fine of a feature they detect -- photoreceptors in the central fovea region can have high spatial frequency = fine resolution, while those in the periphery are lower resolution), and in their temporal response (fast vs. slow responding, including differential sensitivity to motion).

The brain regions that participate in the deployment of visual attention include most of the early visual processing area. Visual information enters the primary visual cortex via the lateral geniculate nucleus, although smaller pathways, for example, to the superior colliculus (SC), also exist. From there, visual information progresses along two parallel hierarchical streams. Cortical areas along the 'dorsal stream' (including the posterior parietal cortex; PPC) are primarily concerned with spatial localization and directing attention and gaze towards objects of interest in the scene. The control of attentional deployment is consequently believed to mostly take place in the dorsal stream. Cortical areas along the 'ventral stream' (including the inferotemporal cortex; IT) are mainly concerned with the recognition and identification of visual stimuli. Although probably not directly concerned with the

control of attention, these ventral stream areas have indeed been shown to receive attentional feedback modulation, and are involved in the representation of attended locations and objects (that is, in what passes through the attentional bottleneck).In addition, several higher-function areas are thought to contribute to attentional guidance, in that lesions in those areas can cause a condition of 'neglect' in which patients seem unaware of parts of their visual environment. From a computational view point, the dorsal and ventral streams must interact, as scene understanding involves both recognition and spatial deployment of attention. One region where such interaction has been extensively studied is the prefrontal cortex (PFC).Areas within the PFC are bidirectionally connected to both the PPC and the IT. So, in addition to being responsible for planning action (such as the execution of eye movements through the SC), the PFC also has an important role in modulating, via feedback, the dorsal and ventral processing streams.

IV. HYPOTHESIS FOR TOP-DOWN APPROACH

A. Low level features

The model utilizes the following low level features due to their importance in bottom up saliency.

Itti and Koch saliency: Early saliency model [9] motivated by linear filtering and center surround operation provides intensity, color and orientation channels which are suitable for bottom up visual attention modeling.

Steerable pyramid filters: Provides filter responses which correlate well with visual attention and therefore local energy of steerable pyramid filters [14] in four orientations and three scales are used. Torralba Saliency: Provides a holistic representation of a scene [15] using spectral and coarsely localized information.

Color histogram features: The values of the red, green and blue channels and the probabilities of each of these channels are used according to [4].

Signature Saliency: Provides a saliency map [16] using the theoretical framework of sparse signal mixing which spatially approximates image foreground.

Graph Based Visual Saliency: Jointly models feature extraction and activation map creation in a unified manner by defining edge weights using saliency and dissimilarity [17].

B. Mid level features

Horizon detection is performed using using gist descriptor [15]. It is especially important in outdoor scenes where salient objects are present near the ground plane .

C. High level features

High level objects such as faces and text have high visually saliency. We utilize automatic object detectors for face [18], person [19], car [19] and text [20] in our model. The source code for [20, 28] is not publicly available and we used our implementation in this paper.

D. Scene Context features

In addition to high level features, we propose a novel set of features which model the pairwise interaction between multiple high level features. High level features typically model attention gain in the locality of semantic objects. However, presence of interesting objects in a scene also incurs attention loss in other objects (in general other high, mid and low level features) in a scene. This attention loss scene context features can be described using a cause-effect mechanism. Let there be N possible objects in a scene and $\mathbf{f}^{SC}$ (x, y) denote the scene context feature between object **i** and **j** at position(x, y). The scene context vector models the attention loss in the scene incurred on the pixels of object **i** (effect) due to presence of object **j** (cause). Now let the total number of objects corresponding to label **i** in the scene be denoted by $\mathbf{N_i}$ and the number of object **i**'s in position (x, y) be $\mathbf{n_i}$ (x, y).

Cause effect clustering: The modeling the factors affecting cause and effect of per pixel attention loss (attention density loss) which will be predicted by the learning algorithm is proposed

E. Learning

The features are pre-computed in all the images and a regression model is used to predict where subjects look in new images. For this purpose, the dataset is divided into training and test sets in a 10-fold cross validation setting. From each image in the training dataset, we randomly pick equal number of pixels from the top 20% and bottom 80% attention regions (to have adequate representation for high attention regions) to create a pixel level training subset. A regression model is learnt from this subset and pixel wise attention density in new images are predicted using this regression model.

V. CONCLUSION AND FUTURE WORK

In this paper, we have given a hypothesis for Computational Cognitive Neuroscience based Saliency Map identification and highlighting of Visual Attention Region's (VAR's) in Machine vision using scene context based top-down approach.

Scene context plays a crucial role in determining where people look in images. This paper is a pioneering effort to understand the role of scene context in task free viewing scenario. Apart from low, mid and high level features we propose scene context features which communicates to each object (and few other features) the presence of other objects in a scene. This results in loss of attention in the object of interest and is automatically learnt as negative weights in a linear regression setting. The Model also compares classification of linear and non-linear regression techniques for learning attention maps.

Acknowledgment

We would like to thank JAIN University, REVA University and M S Ramaiah Institute of Technology for providing the necessary support and infrastructure to carry our research work.

References

[1] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," Advances in neural infor- mation processing systems, vol. 17, no. 481-488, pp. 1, 2004.

[2] V. Mahadevan and N. Vasconcelos, "Saliency-based discrim- inant tracking," in Computer Vision and Pattern Recogni- tion, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp.1007–1013.

[3] S. Karthikeyan, V. Jagadeesh, and BS. Manjunath, "Learning bottom-up text attention maps for text detection using stroke width transform," IEEE International Conference on Image Processing, 2013.

[4] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in Computer Vision, 2009 IEEE 12th international conference on. IEEE, 2009, pp. 2106–2113.

[5] Y.F. Ma, L. Lu, H.J. Zhang, and M. Li, "A user attention model for video summarization," in Proceedings of the tenth ACM international conference on Multimedia. ACM, 2002, pp. 533–542.

[6] C. Guo and L. Zhang, "A novel multiresolution spatiotempo- ral saliency detection model and its applications in image and video compression," Image Processing, IEEE Transactions on, vol. 19, no. 1, pp. 185–198, 2010.

[7] C. Siagian and L. Itti, "Biologically inspired mobile robot vi- sion localization," Robotics, IEEE Transactions on, vol. 25, no. 4, pp. 861–873, 2009.

[8] A. Borji, L. Itti, J. Liu, P. Musialski, P. Wonka, J. Ye, S. Ji, W. Xu, M. Yang, K. Yu, et al., "State-of-the-art in visual at- tention modeling.," Transactions on Pattern Analysis and Ma- chine Intelligence.

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, no. 11, pp. 1254–1259, 1998.

[10] R. Rosenholtz, "A simple saliency model predicts a number of motion popout phenomena," Vision research, vol. 39, no. 19, pp. 3157–3163, 1999.

[11] M. Cerf, E.P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," Journal of vision, vol. 9, no. 12, 2009.

[12] C. Hickey and J. Theeuwes, "Context and competition in the capture of visual attention," Attention, Perception, & Psy- chophysics, vol. 73, no. 7, pp. 2053–2064, 2011.

[13] S.C. Mack and M.P. Eckstein, "Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment," Journal of Vision, vol. 11, no.9, 2011.

[14] E.P. Simoncelli and W.T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in Image Processing, 1995. Proceedings., International Con- ference on. IEEE, 1995, vol. 3, pp. 444–447.

[15] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International Journal of Computer Vision, vol. 42, no. 3, pp. 145–175, 2001.

[16] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," Pattern Analysis and Machine Intel- ligence, IEEE Transactions on, vol. 34, no. 1, pp. 194–201,2012.

[17] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," Advances in Neural Information Processing Sys- tems, 2007.

[18] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. IEEE, 2001, vol. 1, pp. I–511.

[19] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discrim- inatively trained, multiscale, deformable part model," in Com- puter Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008, pp. 1–8.

[20] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE,2010, pp. 2963–2970.

[21] J. Cohen and P. Cohen, Applied multiple regression/correlation analysis for the behavioral sciences., Lawrence Erlbaum,1975.

[22] L. Breiman, "Random forests," Machine learning, vol. 45, no.1, pp. 5–32, 2001.

[23] S. Wold, M. Sjö strö m, and L. Eriksson, "Pls-regression: a ba- sic tool of chemometrics," Chemometrics and intelligent labo- ratory systems, vol. 58, no. 2, pp. 109–130, 2001.

[24] Manjunath R Kounte, Dr. B K Sujatha, "A Review of Modelling Visual Attention using Computational Cognitive Neuroscience for Machine Vision " , *International Journal of Advanced Research in Computer and Communication Engineering*,Vol. 2, Issue 9, pp 3558-3563,September 2013.

[25] Dirk Walther and Christof Koch, "Modeling attention to salient protoobjects", *Neural Networks 19*, 1395-1407,2006

[26] S. Karthikeyan and Vignesh Jagadeesh and B. S. Manjunath, "Learning Top Down Scene Context For Visual Attention Modelling In Natural Images", IEEE International Conference on Image Processing, Sep 2013.

[27] Manjunath R Kounte, Dr. B K Sujatha, "Bottom up Approach for Modelling Visual Attention using Saliency Map in Machine Vision", *International Journal of Applied Engineering Research*,Vol. 10, No.10, 2015.

[28] Mohammed Riyaz Ahmed, Dr. B.K.Sujatha, "Memory Modelling Schemes In Neuromorphic VLSI Chips Using Reinforcement Learning Based on Cognition", International Journal of Applied Engineering Research,ISSN 0973-4562 Volume 10, Number 9 (2015) pp. 23769-23778.