Current and Future Trends of Deep Learning based Visual Attention

Mostafa E. A. Ibrahim[1], Qaisar Abbas[†]

[†]College of Computer and Information Sciences, Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

[1]Electrical Engineering Dept. - Benha Faculty of Engineering - Benha University Benha – Egypt

Summary

Human handles the huge amount of information in a complex scene by focusing on a certain portion of the information which is known as visual attention. Human visual system (HVS) is having very strong reasoning compare to current development of automatic computerize systems. Nowadays, computer vision systems stimulate the behavior of human visual attention that has many applications in practice such as objects recognition, tracking, and image cropping. Those systems were designed and implemented to accelerate an automatic processing. Among all those fields, the visual attention domain is a very complicated task to process the objects in real-time. To perform real-time processing, many deep learning techniques have been developed in the past try to simulate the visual attention. In this review article, first we introduce the top-most variants of deep learning algorithms such as convolutional and recurrent neural network models. A state-of-the-art survey is also presented about the advances in the field of employing deep learning models in the field of visual attention. Especially, a comparison is also presented in terms in this paper to show the importance of this field in terms of visual attention. The current and future trends are also described to attract the researchers in this field.

Key words:

Visual Attention, Machine Learning, Deep Learning, Convolutional Neural Networks, Recurrent Neural Networks.

1. Introduction

Each day a person comes across numerous visual scenes and gets overflowed with enormous information that is automatically realized by human brain. Computer vision systems stimulate the way a human brain analyzes a visual scene. With the huge amount of input data offered by visual scenes, visual attention is utilized to limit the search for subset of the input information to process and hence accelerates the data processing.

Visual attention as defined by Evans et al. [1] is the process that "describes a set of mechanisms that limit some processing to a subset of incoming stimuli". It involves the assortment of spatial and temporal region of interest, restriction of features and features dimensions, prevailing the stream of data, and moving between different selected regions of interest. Visual attention is involved in a variety of computer vision tasks such as objects detection and recognition, image/video captioning and description.

Deep learning (DL) is a fast-growing area which models high-level patterns in data as complex multilayered networks. Recently, DL architectures are utilized in low-level image and video processing tasks such as objects detection, recognition and tracking [2]. DL techniques are also employed in high-level video processing tasks such as 1 Human actions recognition, gesture and emotion recognition, pedestrian detection, scene interpretation, and video captioning [3]. DL architectures outperform other image processing techniques when the learning dataset is large enough.

This review article reveals the latest accomplishments in visual attention by means of deep learning techniques. These deep learning architectures are detailed described in the sub-sequent sections.

2. Deep Learning Architectures

This section explores two important DL architectures that are frequently employed for visual attention task.

2.1 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is an advanced version of feed forward neural network architecture in traditional machine learning [4]. CNNs consist of numerous small processing units called neurons structured in several cascaded layers. A CNN at least has an input layer that accept input streams as raw images. It also has at least single convolutional layer that divide the input image in windows of certain number of pixels such as 3 × 3 or 5 × 5 which are convolved with filters of the same size. The outputs of these convolutions are lined in such a way that there is an overlapping of the input regions that gives a clear representation of the original input image. That process is repeated for all the convolution layers with different window sizes. CNN models are recently used in varsity of video and image applications such as objects

Manuscript received January 5, 2019
Manuscript revised January 20, 2019

recognition, natural language processing and scene analysis.

A convolution layer correlates a bank of *K* filters with *C* channels and size *R* × *S* against a small set of N images with *C* channels and size *H* × *W*. We represent filter elements as $G_{k,c,u,v}$ and image elements as $D_{i,c,x+u,y+v}$. The computation of a single convolution layer output $Y_{i,k,x,y}$ is given by Eq. 1:

$$Y_{i,k,x,y} = \sum_{c=1}^{C} \sum_{v=1}^{R} \sum_{u=1}^{S} \dots D_{i,c,x+u,y+v}\, G_{k,c,u,v} \quad (1)$$

and the output of an entire image/filter pair is represented as follows in Eq. 2:

$$Y_{i,k} = \sum_{c=1}^{C} D_{i,c} * G_{k,c} \quad (2)$$

Where $*$ represents 2D correlation. Figure 1 illustrates an example of a CNN architecture with a single convolution layer.

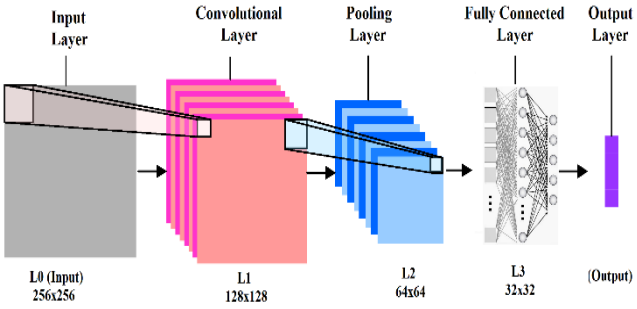

Fig. 1 Typical CNN model architecture

Among variants of deep learning algorithms, the CNN is a well-known, frequently used, and an entirely supervised learning model. A huge training set, millions of semantically labeled imageries, is necessary for accurate learning and thus accurate results. The training set is prepared by categorizing millions of instances generating tens of millions of training samples. Next, the CNN architecture is trained using gradient descent and back-propagation methods for faster convergence [5].

The main merits of the CNN models are as follows. First, the use of common weights in convolution layers facilitates the usage of the same filter for each pixel in the layer. Second, CNNs directly learn from the raw input which is the image itself without any requirement for extracting features as pre-request for the learning process. While traditional algorithms require preprocessing and features extraction stages. Third, CNNs are easy to train and require less human understanding and effort. CNNs easily get 2D structure of the input image by local connections and weights followed by a pooling technique which results in translation invariant features. Fourth, the most distinguishing feature of CNN model is that it has the 3D volume of neurons in which the neurons are arranged in three dimensions namely weight, height, and depth.

The main cons of CNNs are 1) CNNs require huge amount, may reach millions, of semantically labeled imageries for training the network in order to get highly accurate results. 2) CNNs involve large memory requirements to hold the in-between results of the convolution layers to the back-propagation layer.

2.2 Recurrent Neural Networks

The Recurrent Neural Network (RNN) is a stochastic multiple layer model which is used to identify objects in video scene, music, text and motion capture [6]. RNN handles the inconstant length sequence using a recurrent hidden state whose activation at each time depends on that of the previous time. Hereafter, Eq. 3 presents how a RNN updates its recurrent state ht; where S is a sequence S = {s1, s2 …, sT}, ϕ is a nonlinear function:

$$h_t = \begin{cases} 0, & t = 0 \\ \emptyset(h_{t-1}, s_t & otherwise \end{cases} \quad (3)$$

RNNs are capable of processing real data sequences step at a time and then based on the training dataset it predicts a new sequence. Practically, a RNN produce the new sequence by using fuzzy rules relying on hidden layers to avoid lengthy sequences. Consequently, RNN models are usually used in optimization algorithms.

The training of a RNN model is based on few inputs, and these inputs were themselves predicted by the network then it has little opportunity to recover from past mistakes. If we increase the network size or sample size, then the network cannot look further back in the past to formulate its predictions. However, if we add the noise then the RNN model can go and forth to learn more effective solution during training process. Moreover, RNN is hard to identify long term dependencies since the gradients are mostly vanish or occasionally explode but with strong impacts. To avoid this issue some researchers, provide a little bit complex learning algorithm than the stochastic gradient one. While others, designed alternative activation functions that use gating units [7].

Two main recurrent units have been developed namely; Long Short-Term Memory (LSTM) [8] and Gated Recurrent Unit (GRU) [9]. LSTM is usually used for better storing and representing information [6]. While in GRU

"each recurrent unit to adaptively capture dependencies of different time scales" [7]. Figure 2 shows the internal structure of LSTM and GRU unit of an RNN architecture.

3. Deep learning Visual Attention models

This section presents the latest research efforts for employing different DL techniques for visual attention.

Liu and Milanova [10] categorized the DL based visual attention models as bottom-up and top-down models. In bottom-up models, fixation points, where it stands out from its surrounding and grab our attention at first glance, are located. While top-down models require a pre-knowledge of the visual scene.

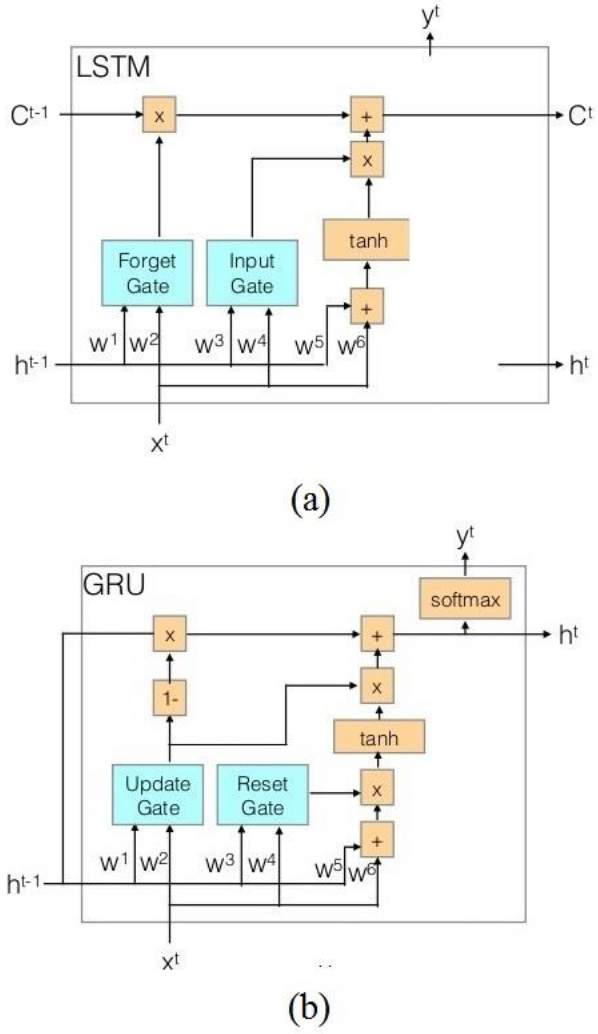

Fig. 2 RNN Models: (a) LSTM-RNN architecture (b) GRU-RNN architecture

Fig. 3 An early attempt for using DL for visual attention was proposed by Bazzani et al. [11]. They proposed a visual attentional model for multiple objects identification and tracking that is obsessed by theories of human vision system. They used Restricted Boltzmann Machine (RBM) as a DL algorithm to model the appearance of objects and accomplish object classification. They evaluated their approach using the tracking accuracy that is measured in pixels and expressed in two metrices, the mean and standard deviation, over time of the distance between the targeted ground truth and the calculated.

Singh et al. [12] suggested a deep NN based technique to search landmarks in images of objects by means of both appearance and spatial context. Their technique is employed without any modification to two issues: parsing human body layouts, and finding landmarks in images of birds. Their proposed technique understand a sequential search for localizing landmarks, repeatedly finding novel landmarks provided the appearance and contextual information from the previously found ones. Proposed technique shows a new spatial model for the kinematics of sets of landmarks, and shows durable performance on two diverse model difficulties.

Mnih et al. [13] proposed a new visual attention framework that is expressed as a single RNN which is known as Recurrent Attention Model (RAM). It employed prevue window as its input and the internal state of the network chose the next position to attention in addition to produce control signals in a dynamic environment. Their suggested integrated architecture is trained end-to-end from pixel inputs to activities using a model of gradient technique. The framework has many attractive characteristics such as it reduces the computational cost compared to methods that use convolutional filtering.

Ba et al. [14] extended the research work in [13] to recognize multiple objects using visual attention and RNN. They proposed a deep recurrent-based attention framework to select where to emphasize its computation and displayed how it can be trained end-to-end to sequentially categories numerous objects in an image. Their recurrent attention architecture consisted of multiple NNs. The framework overtook the up-to-date ConvNets on a multi-digit house number classification job although spending mutually fewer parameters and fewer computation than the best ConvNets, thus concluding that attention schemes can improve both the accuracy and efficiency of ConvNets on a real-world task. Thus, they recommended the Deep Recurrent Attention Model (DRAM) as a flexible and powerful method.

Whereas, Liu et al. [15] have presented a new CNN framework that consists of eye fixation prediction framework. Proposed framework has attained the superlative performance with important enhancement to state-of-the-art saliency models. The better performance of proposed technique shows that the human visual system (HVS) is more expected to execute low-level contrast and high-level semantics mutually rather than individually.

Wang and Jianbing [16] proposed a CNN based bottom-up visual attention framework to forecast human-eye focuses in scenes. Their approach integrates multi-level attention forecasts within one CNN. Hence, it reduces the repeated employment of learning several network streams with different input scales. They assessed their approach using five datasets and their results showed that their approach outperforms other visual attention techniques.

Kümmerer et al. [17] addressed the issue of "to what extent human fixations can be predicted by low-level (contrast) compared to high-level (presence of objects) image features". They built a CNN model based on the VGG19 to predict human fixations. Their results showed that their proposed model outperforms other methods and achieved a Normalized Scanpath Saliency (NSS) of 2.34 and Area under the Curve (AUC) of 88% on the MIT300 dataset. S. Jia [18] presented a DL based saliency prediction framework. Their framework is based on CNN and is considered a modular system. The encoder and decoder have the option to be disjointedly trained to allow for further scalability. Besides, the encoder is capable of employing multiple CNN models with different architectures that can be pre-trained on different datasets. Consequently, it allows for better feature extraction.

Vision, Data Mining, Biomedical Image and Signal processing, Bioinformatics and video tracking. To segment regions from retinal images, competition-based region growing algorithm [22] (CRGA) modified to segment only red lesions instead of blood vessels lines. To detect deep candidate regions (DCRs), the unsupervised region-based convolutional neural networks (R-CNN) model [23] is used and afterward, the stack-based autoencoders (SAEs) [24] is utilized to finally select MAs from DCRs candidate regions. These steps are explained in the following sub-sections.

Kümmerer et al. [17] addressed the issue of "to what extent human fixations can be predicted by low-level (contrast) compared to high-level (presence of objects) image features". They built a CNN model based on the VGG19 to predict human fixations. Their results showed that their proposed model outperforms other methods and achieved a Normalized Scanpath Saliency (NSS) of 2.34 and Area under the Curve (AUC) of 88% on the MIT300 dataset. S. Jia [18] presented a DL based saliency prediction framework. Their framework is based on CNN and is considered a modular system. The encoder and decoder have the option to be disjointedly trained to allow for further scalability. Besides, the encoder is capable of employing multiple CNN models with different architectures that can be pre-trained on different datasets. Consequently, it allows for better feature extraction.

Huang et al. [19] proposed a CNN based visual attention system for saliency prediction. Their DL model is based on three famous CNN models namely; GoogleNet, AlexNet and VGG-16. They evaluated their system using six datasets including the MIT1003, PASCAL-S, and NUSEF. They achieved NSS of 2.12 and AUC of 87%.

Kruthiventi et al. [20] developed a DeepFix, a completely CNN-based framework to precisely predict saliency. Dissimilar from traditional saliency prediction systems that use hand-crafted features for the saliency map, their model learns features directly in a hierarchical style. DeepFix can "capture semantics at multiple scales while taking global context into

4. Discussions

This section recapitulates the cutting-edge research in the DL based visual attention field, which signalizes the restrictions and sets apart auspicious remarks to address these restrictions.

In this review article, we have briefly described the authors reveal the latest accomplishments in visual attention by means of deep learning techniques. This information is described in Section III. Visual attention methods endeavor to bind processing to vital information which is presently relevant to behaviors or computer vision tasks. An early attempt of using RBM as a DL technique in visual attention was proposed in [11]. Later, a deep NN technique was introduced to seek landmarks in the image sequence of objects using both appearance and spatial contexts [12].

Recurrent models have been used for visual attention in [13, 14]. The model used in [13] composed of a single recurrent network to select where to emphasize the computation. While in [14] the model employed multiple networks. Convolutional neural networks have been also used for visual attention as in [15-20].

Table I exemplifies the appointed deep learning algorithms, the application, data sources, results and designates whether the results were judged against other state-of-the-art algorithms. While Table II explores the main features of the mostly utilized datasets in visual attention and saliency prediction fields.

Generally, DL algorithms achieved better results in almost all computer vision tasks, but they required the huge training data sources. Besides, they are also required highly efficient processing processors along with the aids of GPUs to handle such massive input data. However, this is increased the cost of development application in domains of visual attention. Despite this problem, the researchers are still trying to increase the efficiency and accuracy in terms of recognizing the visual objects.

Table 1: State-of-the-art deep-learning methods for visual attention

Cited	Year	Application	Method	Data	Results	Cmp.
[11]	2011	multiple objects identification and tracking	RBM	MNIST, Youtube videos	Tracking Acc. Mean: 2.5 SD:1.6	Yes
[12]	2015	Learning a Sequential Search	ANN	LSP, Fashion Pose dataset, Caltech-UCSD Birds	Rec. Acc.: 65.2%	Yess
[13]	2014	Visual Attention	Recurrent Attention Model (RAM)	MNIST	Error 1.07%	Yes
[14]	2015	Object Recognition	DRAM	SVHN	Error 4.6%	Yes
[15]	2015	Predicting Eye Fixations	MR-CNN	MIT, Toronto, Cerf, NUSEF	Score 0.7227	Yes
[16]	2017	Human eye fixation prediction with view-free scenes	CNN	MIT300, MIT1003, DUT-OMRON, PASCAL, Toronto	Normalized Scanpath Saliency (NSS): 2.38	Yes
[17]	2017	Fixation prediction	CNN (VGG19)	MIT300	NSS: 1.29	Yes
[18]	2018	Saliency Prediction	EML-NET (Multiple CNN)	MIT300	NSS: 2.47	Yes
[19]	2015	Saliency Prediction	CNN	MIT1003 NUSEF, PASCAL-S, FIFA	NSS: 2.12 (MIT1003)	Yes
[20]	2015	Human eye fixation prediction	CNN	MIT300, CAT2000	NSS: 2.26 (MIT300)	Yes

Table 2: State-of-the-art Datasets used for visual attention

Cited	Dataset	# images	# eye-observers	View-time	Resolution
[21]	MIT300	300	39	3 sec.	1024x1024
[22]	MIT1003	1003	15	3 sec.	1024x1024
[23]	NUSEF	714	25	5 sec.	1024x728
[24]	PASCAL-S	850	8	-	500x500
[25]	FIFA	200	8	2 sec.	1024x768
[26]	TORONTO	120	20	4 sec.	681x511
[27]	OSIE	700	15	3 sec.	800x600
[28]	CAT2000	4000	24	5 sec.	1920x1080

5. Conclusions

This paper explores the utilization of deep learning architectures in the hot topic of visual attention or saliency prediction. Two main variants of deep learning architectures are mostly employed in the state-of-the-art visual attention field such as convolutional neural network (CNN) and recurrent neural network (RNN) models. This paper is firstly explained the theory behind these two deep learning models. Then, it reviews the most recent advances in research work related to the deep learning based visual attention. At the end, the authors described in the discussion section on the reviewed research articles indicating the datasets utilized and briefly summarizing the main features of each dataset. From those studies, it should be noticed that the researchers did not obtain up-to-the-mark results in terms of efficiency and effectiveness of the results. Therefore, it is still a huge research gap is presented in this field.

Acknowledgment

The authors would like to express their cordial thanks to the department of Research and Development (R&D) of IMAM, university for research grant no: 360915.

References

[1] K. K. Evans, T. S. Horowitz, P. Howe, R. Pedersini, E. Reijnen, Y. Pinto, Y. Kuzmova, J. M. Wolfe. Visual Attention. In: Wiley Interdisciplinary Reviews: Cognitive Science, 2 (2011): 503-514.

[2] Q. Abbas, M. E. A. Ibrahim and M. Jaffar. A comprehensive review of recent advances on deep vision systems. In: Artificial Intelligence Review Journal, (2018). Doi: doi.org/10.1007/s10462-018-9633-3

[3] Q. Abbas, M. E. A. Ibrahim and M. Jaffar. Video scene analysis: an overview and challenges on deep learning algorithms. In: Multimedia Tools and Applications, 77(16): 20415-20453, 2018. Doi: doi.org/10.1007/s11042-017-5438-7

[4] [90] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks, in: Proceedings Part I of the 13th European Conference Computer Vision (ECCV'14), Zurich, Switzerland, 2014, pp. 818-833. doi:10.1007/978-3-319-10590-1_53.

[5] [91] R. Poppe, A Survey on Vision-based Human Action Recognition, Journal of Image Vision Computing, 28 (2010): 976-990.

[6] [92] A. Graves, A. r. Mohamed, G. Hinton, Speech Recognition with Deep Recurrent Neural Networks, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645-6649. doi:10.1109/ICASSP.2013.6638947.

[7] [89] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: Neural Information Processing Systems (NIPS'14) Workshop on Deep Learning, December 2014.

[8] [93] S. Hochreiter and J. Schmidhuber. Long short-term memory. In: Journal on Neural Computation, 9(8):1735–1780, 1997.

[9] [94] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In: arXiv preprint arXiv:1409.1259, 2014.

[10] [87] X. Liu, M. Milanova. Visual attention in deep learning: a review. In: International Robotics & Automation Journal. 2018;4(3):154–155. DOI: 10.15406/iratj.2018.04.00113.
[11] [95] L. Bazzani, N. de Freitas, H. Larochelle, V. Murino, and J. Ting. Learning attentional policies for tracking and recognition in video with deep networks. In: proceedings of the 28th International Conference on Machine Learning (ICML'11), 2011.
[12] [85] S. Singh, D. Hoiem, D. Forsyth, Learning a sequential search for landmarks. In: proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15), 2015, pp. 3422-3430. doi:10.1109/CVPR.2015.7298964.
[13] [83] V. Mnih, N. Heess, A. Graves, k. kavukcuoglu, Recurrent models of visual attention. In: Collections of Advances in Neural Information Processing Systems, No. 27, Curran Associates, Inc., 2014, pp. 2204-2212.
[14] [84] J. Ba, V. Mnih, K. Kavukcuoglu. Multiple Object Recognition with Visual Attention. In: Proceedings of International Conference on Learning Representations (ICLR'15), San Diego, California, USA, 2015.
[15] [86] N. Liu, J. Han, D. Zhang, S. Wen, T. Liu, Predicting Eye Fixations Using Convolutional Neural Networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15), 2015, pp. 362-370. doi:10.1109/CVPR.2015.7298633.
[16] [88] W. Wang and J. Shen. Deep Visual Attention Prediction. In: IEEE Transactions on Image Processing, vol. 27, no. 5, pp. 2368-2378, May 2018. doi: 10.1109/TIP.2017.2787612
[17] M. Kümmerer, T. S. A. Wallis, L. A. Gatys and M. Bethge, Understanding Low- and High-Level Contributions to Fixation Prediction. In: Proceedings of IEEE International Conference on Computer Vision (ICCV'17), Venice, 2017, pp. 4799-4808. doi: 10.1109/ICCV.2017.513
[18] S. Jia. EML-NET:An Expandable Multi-Layer Network for Saliency Prediction. eprint arXiv:1805.01047. 2018.
[19] X. Huang, C. Shen, X. Boix and Q. Zhao. SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. In: proceedings of IEEE International Conference on Computer Vision (ICCV'15), Santiago, 2015, pp. 262-270. doi: 10.1109/ICCV.2015.38.
[20] S. Kruthiventi, K. Ayush, B. Venkatesh. DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations. eprint arXiv:1510.02927. 2015.
[21] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. In: MIT Technical Report, 2012.
[22] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV'09), 2009, pp. 2106–2113.
[23] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.S. Chua. An eye fixation database for saliency detection in images. In: proceedings of 11th European Conference on Computer Vision (ECCV'10), 2010, pp 30-43.
[24] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICCVPR'14), 2014, pp. 280–287.
[25] M. Cerf, J. Harel, W. Einhauser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In: Neural Information Processing Systems (NIPS'08), 2008, pp. 241–248.
[26] N. Bruce and J. Tsotsos. Saliency based on information maximization. In: Advances in Neural Information Processing Systems, 8:155, 2006.
[27] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. In: Journal of Vision January (JoV), vol 14, 2014. doi: 10.1167/14.1.28.
[28] A. Borji, L. Itti. CAT2000: A Large-Scale Fixation Dataset for Boosting Saliency Research. eprint arXiv:1505.03581. 2015.

Qaisar Abbas received his BSc and MSc degrees in Computer Science from Bahauddin Zakariya University(BZU), Pakistan in 2001, 2005; respectively. He then became the Lecturer and Software developer in the same department. He has completed PhD in the school of Computer Science and Technology at the Huazhong University of Science and Technology (HUST), Wuhan China. He is now working as an assistant professor in College of Computer and Information Sciences, Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia. His research interests include: image processing, medical image analysis, Genetic programming and pattern classification.

Mostafa Ibrahim received his BS degree in 1997, in computer engineering from High Institute of Technology - Benha University, Egypt. He received his MSc degree in 2004 from the same university. In November 2009, he received his PhD in electronics and comm. Eng. from Cairo University, Egypt in conjunction with Vienna University of Technology, Vienna, Austria under a channel supervision grant. He is now an assistant professor at Faculty of Engineering - Benha University, Egypt. His research interests include embedded code optimization from energy consumption perspective, Software Defined Radio, image and video processing, and Computer Architecture.