MULTI-LAYER LINEAR MODEL FOR TOP-DOWN MODULATION OF VISUAL ATTENTION IN NATURAL EGOCENTRIC VISION

Keng-Teck Ma[1], *Liyuan Li*[1], *Peilun Dai*[1], *Joo-Hwee Lim*[1], *Chengyao Shen*[2], *Qi Zhao*[3]

[1] Institute for Infocomm Research, A*STAR, Singapore*
[2] National University of Singapore
[3] University of Minnesota, USA

ABSTRACT

Top-down attention plays an important role in guidance of human attention in real-world scenarios, but less efforts in computational modeing of visual attention has been put on it. Inspired by the mechanisms of top-down attention in human visual perception, we propose a multi-layer linear model of top-down attention to modulate bottom-up saliency maps actively. The first layer is a linear regression model which combines the bottom-up saliency maps on various visual features and objects. A contextual dependent upper layer is introduced to tune the parameters of the lower layer model adaptively. Finally, a mask of selection history is applied to the fused attention map to bias the attention selection towards the task related regions. Efficient learning algorithm with single-pass polynomial complexity is derived. We evaluate our model on a set of natural egocentric videos captured from a wearable glass in real-world environments. Our model outperforms the baseline and state-of-the-art bottom-up saliency models.

Index Terms— ego-centric, visual attention, real-world

1. INTRODUCTION

Attention selection is a remarkable capability of human vision system. It plays a crucial role in achieving highly efficient human visual perception dealing with huge amounts of natural visual information in complex real-world environments. Computational modeling of human visual attention has always attracted much interest in physiology, neuroscience and computer vision. On one side, it could facilitate better understanding of human vision system [1]. On the other hand, it can be used to develop efficient vision systems for numerous applications such as robotics [2].

Many models of visual attention have been proposed in computer vision [2]. The research efforts have mostly focused on modeling bottom-up saliency. Computing saliency maps on various low-level image features and high-level semantic objects have been investigated, including intensity, color, and orientation [3], histogram on DoG [4], motion [5], text [6], face [7], hand [8], etc. The recent advances in deep learning has spurred a wealth of deep neural networks for bottom-up saliency detection. For example, Huang et al. who had greatly improved saliency prediction on images by leveraging on rich pools of semantic features from deep convolution neural network (2D-CNN) [9]. Models of bottom-up saliency have been very successful in predicting fixations in still images on screen in free-viewing tasks, however, they perform poorly in everyday tasks [10, 11, 8, 12].

Physiological studies have revealed that, in real-world scenarios, top-down attention plays an important role in guiding visual attention towards task-related objects in the view [13, 14]. However, modeling top-down attention is a hard task due to the limited understanding of the related mechanisms in human vision system [15, 10, 7]. A few models have been proposed. Most investigated model is linear regression [16, 17], where a linear mapping from low-level image features (e.g. gist) or bottom-up saliency maps to the fixation location is learned directly on training images. This model can be considered as single layer linear model which has limited capability to adapt to variations of tasks, scenes and viewpoints. In [15], Borji et al have investigated a few top-down models, including linear regression, kNN, and SVM. It was observed that kNN approach performs best. Borji et al have further proposed a DBN (Dynamic Bayesian Network) model to predict the fixation on the next image based on previous image sequence [10].

Inspired by the progresses in understanding of the top-down attention mechanisms [18, 13, 14], we proposed an efficient multi-layer linear model to modulate various bottom-up saliency maps adaptively for current task and scene. The lower layer is a linear regression model. It integrates bottom-up saliency maps on low-level saliency, ego-motion, exo-motion, ground, text, hand, and face. An upper layer of linear model is introduced, which is trained to learn the top-down rules to tune the weights of the lower layer according to current task. Finally, a filter of selection history for the related task is applied to the fused attention map to bias the attention selection towards task-related regions in the scene. We

*This work is supported by the REVIVE programme funded by the Joint Council Office (JCO) of A*STAR.

train and evaluate our model on natural egocentric videos capturing first-person-views of the world from a wearable glass when performing different tasks in indoor and outdoor environments, which is much more realistic to study human attention in real-world tasks [11, 12]. Our model outperforms the baseline and state-of-the-art bottom-up models by a clear margin on the publicly available egocentric video dataset with fixation records [19].

2. OUR ATTENTION MODEL

To imitate human vision system, a computational model of visual attention should integrate the bottom-up saliency and top-down attentional guidance. A full model of visual attention is implemented as the intergrative framework inspired by the Attentional Engagement Theory [18], An Integrative Framework [20], and mechanisms of human visual attention [1, 13]. In Attentional Engagement Theory, it is assumed that human visual attention deployment is a two-phase procedure. In the first phase, the physical saliences are encoded and bound together without focal attention; and in the second phase, attention is modulated by top-down factors, *e.g.* the goal of current task, to pay attention to the most behaviorally relevant locations and objects. In addition, selection history is considered as a strong influencing top-down cue [20, 2]. Hence, we establish a full model of visual attention as a three-phase sequential system, as shown in Figure 1. In the first phase, namely Pre-attentive Parallel Phase, the physical saliences, or the bottom-up saliency maps on various low-level image features and semantic objects are computed in parallel. In the second phase, namely Selective Attention Phase, the top-down modulation is performed to fuse the bottom-up saliency maps adaptively according to current task and low-bandwidth signals from the physical saliences. Finally, in the third phase, namely Gaze Deployment Phase, a filter of selection history is applied to bias the attention deployment towards the task-related regions. The multi-layer linear model for top-down modulation in phase-2 is the core novel part of this model.

2.1. Pre-attentive Parallel Phase

In the Pre-attentive Parallel phase, various bottom-up physical saliences are computed in parallel. Exploiting recent progresses in computer vision, both feature-level and object-level saliences are employed in our model. They are described briefly in the following.

Low-level saliency: Graph Based Visual Saliency (GBVS) [21] is employed for low-level saliency. It is a graph-based implementation of the Itti and Koch model [3] that uses a dissimilarity metric. It is selected due to its superior performence among the models on low-level saliences [15].

Ego-motion: An algorithm similar to method [5] is implemented. To minimize the effect of local scene motion on the estimation of ego-motion, an average global motion vector is computed along the boundaries of the Large Displacement Optical Flow (LDOF) flow field [22]. This motion vector is used to build an ego-motion saliency map [23]. This probability map provides a head-movement saliency.

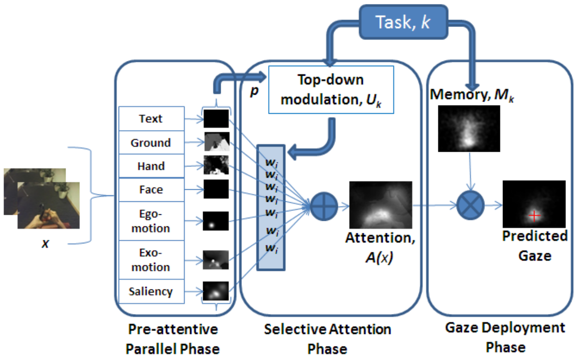

Fig. 1. The full model of visual attention inspired by human vision system. In Pre-attentive Parallel phase, the physical saliences are computed in parallel. In Selective Attention phase, top-down attention modulation is performed by a multi-layer linear model from $\mathbf{U_k}$ to $\{\mathbf{w}_i\}$ for current task k. The saliences from the first phase are linearly combined with their respectively learnt weights, w_i, Finally, in Gaze Deployment phase, a filter of selection history for current task is applied to the fused attention map $A(x)$ to deploy the gaze.

Exo-motion: First, LDOF is applied to compute the flow field of two consecutive frames. Then, by subtracting global motion vector from the flow field, the absolute values of the remaining components are normalized as the exo-motion saliency map [23]. It provides a probability map of saliency based on scene motion.

Text Detection: The method of "Class-Specific Extremal Regions for Scene Text Detection" proposed by Luks Neumann and Jiri Matas [6] is used. The implementation in OpenCV 3.0 is employed. It is trained on English alphabets only.

Face Detection: The OpenCV implementation of Haar feature-based cascade classifiers proposed by Viola and Jones [24] is employed. Faces are attracted attention during social interactions.

Hand Detection: During object manipulation, fixations are often correlated with hands. We use the hand detection algorithm for ego-centric videos [5] to generate a saliency map on hands in the view.

Ground Plane: When moving around, human pays attentions to the ground plane frequently. The Geometric Context algorithm developed by Hoiem *et al.* [25] is used to detect the ground plane and generate a saliency map.

The existing open-source implementations are used. Unless otherwise stated, the default parameters and trained models are used. The output saliency maps are normalized between 0 and 1 at each pixel, where the high value indicates the higher saliency. It is worthy involving CNN-based saliency map in future study.

2.2. Selective Attention Phase

In this phase, top-down modulation of attention is computed. A multi-layer linear model is proposed for this purpose. In the lower-layer, a linear regression model is used to integrate the bottom-up saliency maps. Instead of fixed weights used in existing models [17], an upper-layer linear model is introduced, which combines low-bandwidth signals of coarse bottom-up features and guidance of current task to tune the weights of the lower-layer online. The model and learning algorithm are described in this subsection.

For an input image $I(\mathbf{x})$ under a task C_k, Let $\{A_i(\mathbf{x})\}_{i=1}^{L}$ be the bottom-up saliency maps and L is the number of salience maps, i.e. 7., as described earlier, then let $A(\mathbf{x})$ be the fused attention map. Here and below we omit index k for task for conciseness. The fused attention map can be obtained as a weighted sum of the bottom-up saliency maps:

$$A(\mathbf{x}) = \sum\nolimits_{i=1}^{L} w_i A_i(\mathbf{x}) \tag{1}$$

However, human attends to a point according to the goal of current task and physical saliency features. That means, the weight w_i is adaptive to different scene under different task according to the mechanism of top-down attention [13]. To build a computational model of top-down modulation of attention, we propose a linear model over (1) to compute the adaptive weights. It is expressed as

$$w_i = \sum\nolimits_{j=1}^{L} u_j^i p_j + u_0^i = \mathbf{u}_i^T \mathbf{p}, \tag{2}$$

where $\mathbf{u}_i = [u_0^i, u_1^i, \cdots, u_L^i]^T$ represents the parameters of the upper-layer model, and $\mathbf{p} = [1, p_1, \cdots, p_L]^T$ represents the low-bandwidth signals of bottom-up saliences. The low-bandwidth signal provides a coarse representation of the physical saliency [13]. We use the standard deviation of a saliency map as the low-bandwidth signal, which represents how strong a visual cue against its surroundings. It is denoted as $p_j = Dev\,(A_j(\mathbf{x}))$.

Let $\mathbf{w} = [w_1, \cdots, w_L]^T$ be the weight vector of the lower-layer model. Eq. (2) can be expressed as

$$\mathbf{w} = \begin{bmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_L^T \end{bmatrix} \mathbf{p} = \begin{bmatrix} u_0^1 & u_1^1 & \cdots & u_L^1 \\ \vdots & \vdots & \ddots & \vdots \\ u_0^L & u_1^L & \cdots & u_L^L \end{bmatrix} \mathbf{p} = U\mathbf{p}. \tag{3}$$

The matrix U denotes the upper-layer model, representing cognitive knowledge of top-down attention modulation [13]. It can be learned from recordings of human fixations by eye-tracking in different scenarios under the corresponding task.

For a task C_k, we select N training images. The training set is represented as $\mathcal{T} = \{I_n(\mathbf{x}), A_n(\mathbf{x})\}_{n=1}^{N}$, where $I_n(\mathbf{x})$ is an input image and $A_n(\mathbf{x})$ is the ground truth attention map generated according to the recorded human fixation. Under the corresponding task C_k, for the n-th training image $I_n(\mathbf{x})$, we can obtain the set of bottom-up saliency maps, *i.e.*, $\{A_{ni}(\mathbf{x})\}_{i=1}^{L}$. According to (1), we have

$$A_n(\mathbf{x}) = \sum\nolimits_{i=1}^{L} w_i A_{ni}(\mathbf{x}). \tag{4}$$

Suppose the image has M pixels, we can express $\mathbf{y}_n = [A_n(\mathbf{x}_1), \cdots, A_n(\mathbf{x}_M)]^T$ as the 1-dimension vector of the ground truth map, and

$$X_n = \begin{bmatrix} A_{n1}(\mathbf{x}_1) & \cdots & A_{nL}(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ A_{n1}(\mathbf{x}_M) & \cdots & A_{nL}(\mathbf{x}_M) \end{bmatrix}, \tag{5}$$

as the data matrix of the bottom-up saliency maps. Then, Eq. (4) can be expressed as

$$\mathbf{y}_n = X_n \mathbf{w}_n. \tag{6}$$

Using Least Square Regression, we can obtain the weight vector as

$$\mathbf{w}_n = (X_n^T X_n + \alpha I)^{-1} X_n^T \mathbf{y}_n, \tag{7}$$

where I is the identity matrix and α is the parameter for regulation. We further extend to upper-level to learn the cognitive knowledge for top-down attention modulation. According to (3), on the n-th training sample, we have $\mathbf{w}_n = U\mathbf{p}_n$, where $\mathbf{p}_n$ are obtained from $\{A_{ni}(\mathbf{x})\}_{i=1}^{L}$. From all the N training samples in the training set, we can obtain

$$[\mathbf{w}_1, \cdots, \mathbf{w}_N] = U[\mathbf{p}_1, \cdots, \mathbf{p}_N], \quad \text{or} \quad W = UP. \tag{8}$$

The Least Square solution of U on (8) turns out to be

$$U = WP^T(PP^T + \alpha I). \tag{9}$$

In the application, for an input image $I(\mathbf{x})$ under task C_k, the bottom-up saliency maps are computed first. Then, through standard derivation $\mathbf{p}$ to extract low-bandwidth signals of the bottom-up saliency maps, the adaptive weights $\mathbf{w}$ for fusion are derived by matrix U. Finally, the fused map of predicted attention is obtained by weighted sum of bottom-up saliency maps.

2.3. Gaze Deployment Phase

In this phase, the top-down influence of reward or selection history [13, 20] is applied. It is implemented by applying a mask of selection history for current task to the fused attention map. Let M_k represent the selection history for task k. It is a map of prior fixation distribution generated from the training data of ground-truth fixations by convolving an isotropic Gaussian of size 4^o and $\sigma = 1^o$ [26]. The filtered attention map is the result of pixel-level product between mask M_k and

Method	BMS	GBVS	ITTI	SALICON	NS	NM	REG	kNN	CB	Ours
AAE	17.8	15.6	16.9	15.6	16.2	28.7	16.3	16.7	12.8	**12.3**
AUC	0.620	0.642	0.626	0.653	0.593	0.577	0.593	0.512	0.509	**0.677**

Table 1. Experimental Results: The average AAE and AUC scores for the different methods.

fused attention map. The image location with the maximum value of the filtered map is assigned as the deployed gaze:

$$\mathbf{x}_{gaze} = \max_{\mathbf{x}}(A(\mathbf{x})M_k(\mathbf{x})). \tag{10}$$

The constraint of selection history reduces the effects of outliers generated by the top-down linear model.

3. EXPERIMENTAL RESULTS

The bottom-up models of visual attention have achieved a great success in most of existing benchmarking datasets of free-viewing tasks on images displayed on a computer screen [7, 10]. However, they perform poorly in scenarios of everday tasks. In this study, we performed a formal evaluation of our model on a recently available dataset of natural egocentric videos [19], which capture first-person-views of the world from a wearable glass when performing daily life tasks in indoor and outdoor environments. It is much more realistic to study human attention in real-world tasks [11, 12] compared with existing datasets of viewing images on screens for free-viewing or game tasks [7].

3.1. Dataset

To our knowledge, the recent dataset of attentions in natural egocentric videos [19] contains much more daily life activities in indoor and outdoor environments compared with previous ones consisting of one or two activities (*e.g.* cooking) in a fixed place in a room [8]. The videos and eye-tracking data were recorded while six participants were engaged in daily activities, such as social interactions, object manipulations, walking in the offices, homes and public places. The 14 videos were also manually annotated with the activities (tasks) in the video segments. In our experiments, the activities are grouped into 8 tasks, of which the 7 tasks are the *combinations* of 3 simple tasks of $Social$, $Walk$ and $Object$, and another one is of $Others$.

3.2. Metrics and Protocols

Two standard and complementary metrics are used: Area Under ROC Curve (AUC) and Average Angular Error (AAE) [2]. For AUC, the saliency map is treated as a binary classifier to separate positive from negative pixels at various thresholds compared with the ground truth. It is a standard metric in the saliency prediction literature. AAE measures the angular distance between the predicted gaze point and the ground-truth. It is widely used in the gaze tracking literature [8]. The widely accepted protocol of leave-one-out cross-validation was used. The metrics for the 14 videos are then averaged and presented.

3.3. Results

We compare the results against the state-of-the-art bottom-up saliency models, *i.e.*, BMS [27], GBVS [21], Itti/Koch's [3], and a recent deep learning model SALICON [9], using the authors' own implementations. The motion cues in [21, 3] are enabled for fair comparison. In addition, we also compared our method against other well-known models, such as Normalized and Sum (NS) and Normalized and Max (NM) fusion methods described by Chevet and Meur [28]. As baseline of top-down models [15], kNN on GIST feature and linear regression (REG) model are also implemented. As suggested in their work, each top-down task (*e.g. Social*) was trained separately. We show the best results of kNN where k=1. The center bias is used as an reference [8].

The results are presented in Table 1. It is observed that the compared methods may achieve competitive performance on one metric but poor performance on the other metric. Our model with top-down modulation ranks best for both metrics. Compared to second best computational model, GBVS, it improves $3.3°$ and 5.45% for AAE and AUC respectively. For center bias baseline, it improves by $0.5°$, and 33.0% for AAE and AUC respectively.

4. CONCLUSIONS

Inspired by the mechanisms of top-down attention in human visual perception, we propose a multi-layer linear model for top-down attention modulation, where the lower-layer of linear regression integrates various bottom-up saliency maps, and the upper-layer linear model tunes the lower-layer model online according to the goal of current task and the low-bandwidth signals of bottom-up saliences. An efficient learning algorithm is derived to learn the cognitive knowledge of top-down attention modulation. The formal evaluation on a recent dataset of natural egocentric videos has shown improvements over state-of-the-art bottom-up models and baseline top-down models on FPV videos in daily life.

This work was supported by the Reverse Engineering Visual Intelligence for cognitive Enhancement (REVIVE) programme funded by the Joint Council Office of A*STAR, Singapore.

5. REFERENCES

[1] Simone Frintrop, Erich Rome, and Henrik I Christensen, “Computational visual attention systems and their cognitive foundations: A survey,” *ACM Transactions on Applied Perception (TAP)*, vol. 7, no. 1, pp. 6, 2010.

[2] Ali Borji and Laurent Itti, “State-of-the-art in visual attention modeling,” *TPAMI*, vol. 35, no. 1, pp. 185–207, 2013.

[3] Laurent Itti, Christof Koch, and Ernst Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *TPAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[4] Dashan Gao, Vijay Mahadevan, and Nuno Vasconcelos, “On the plausibility of the discriminant center-surround hypothesis for visual saliency,” *Journal of vision*, 2008.

[5] Cheng Li and Kris M Kitani, “Pixel-level hand detection in ego-centric videos,” in *CVPR*. IEEE, 2013.

[6] Lukas Neumann and Jiri Matas, “Real-time scene text localization and recognition,” in *CVPR*. IEEE, 2012.

[7] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba, “Learning to predict where humans look,” in *CVPR*. IEEE, 2009.

[8] Yin Li, Alireza Fathi, and James M Rehg, “Learning to predict gaze in egocentric video,” in *ICCV*. IEEE, 2013.

[9] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao, “Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks,” in *ICCV*, 2015.

[10] Ali Borji, Dicky N Sihite, and Laurent Itti, “What/where to look next? modeling top-down visual attention in complex interactive environments,” *Transactions on Systems, Man, and Cybernetics: Systems*, 2014.

[11] Michael F Land, “Eye movements and the control of actions in everyday life,” *Progress in retinal and eye research*, vol. 25, no. 3, pp. 296–324, 2006.

[12] Benjamin W Tatler, Mary M Hayhoe, Michael F Land, and Dana H Ballard, “Eye guidance in natural vision: Reinterpreting salience,” *Journal of vision*, vol. 11, no. 5, pp. 5, 2011.

[13] Farhan Baluch and Laurent Itti, “Mechanisms of top-down attention,” *Trends in neurosciences*, 2011.

[14] Adam Gazzaley and Anna C Nobre, “Top-down modulation: bridging selective attention and working memory,” *Trends in cognitive sciences*, vol. 16, no. 2, pp. 129–135, 2012.

[15] Ali Borji, Dicky N Sihite, and Laurent Itti, “Computational modeling of top-down visual attention in interactive environments.,” in *BMVC*, 2011, vol. 85, pp. 1–12.

[16] Robert J Peters and Laurent Itti, “Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention,” in *CVPR*. IEEE, 2007.

[17] Qi Zhao and Christof Koch, “Learning a saliency map using fixated locations in natural scenes,” *Journal of vision*, vol. 11, no. 3, pp. 9–9, 2011.

[18] John Duncan and Glyn Humphreys, “Beyond the search surface: Visual search and attentional engagement.,” 1992.

[19] Keng-Teck Ma, Rosary Lim, Peilun Dai, Liyuan Li, and Joo-Hwee Lim, “Unconstrained ego-centric videos with eye-tracking data,” in *CVPR Workshop on Scene Understanding (SUNw)*. IEEE, 2012.

[20] Edward Awh, Artem V Belopolsky, and Jan Theeuwes, “Top-down versus bottom-up attentional control: A failed theoretical dichotomy,” *Trends in cognitive sciences*, vol. 16, no. 8, pp. 437–443, 2012.

[21] Jonathan Harel, Christof Koch, and Pietro Perona, “Graph-based visual saliency,” in *Advances in neural information processing systems*, 2006, pp. 545–552.

[22] Thomas Brox, Christoph Bregler, and Jitendra Malik, “Large displacement optical flow,” *CVPR*, 2009.

[23] Kentaro Y, Yusuke S, Takahiro O, Yoichi S, Akihiro S, and Kazuo H, “Attention prediction in egocentric video using motion and visual saliency,” in *Advances in Image and Video Technology*, pp. 277–288. Springer, 2012.

[24] Paul Viola and Michael Jones, “Rapid object detection using a boosted cascade of simple features,” in *CVPR*. IEEE, 2001, vol. 1.

[25] Derek Hoiem, Alexei A Efros, and Martial Hebert, “Recovering surface layout from an image,” *IJCV*, vol. 75, no. 1, pp. 151–172, 2007.

[26] Ali Borji, Dicky N Sihite, and Laurent Itti, “Objects do not predict fixations better than early saliency: A reanalysis of einhäuser et al.’s data,” *Journal of vision*, vol. 13, no. 10, pp. 18–18, 2013.

[27] Jianming Zhang and Stan Sclaroff, “Saliency detection: A boolean map approach,” in *ICCV*. IEEE, 2013, pp. 153–160.

[28] Christel Chamaret, Jean-Claude Chevet, and Olivier Le Meur, “Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies.,” in *ICIP*, 2010, pp. 1077–1080.