A Visual Attention Based Object Detection Model beyond Top-Down and Bottom-up Mechanism

Du-Zhen Zhang[1], Chuan-Cai Liu [2]

[1]*School of Computer Science and Technology, Jiangsu Normal University, Xuzhou 221116, Jiangsu, P.R. China*
[2]*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu, P.R. China*
[1]*zhduzhen@aliyun.com*

Abstract: Traditional saliency-based attention theory supposed that bottom-up and top-down factors combine to direct attentional behavior. This dichotomy fails to explain a growing number of cases in which neither bottom-up nor top-down can account for strong selection biases. Thus, the top-down versus bottom-up dichotomy is an inadequate taxonomy of attentional control. In our previous study, we presented a general computational framework for detecting task-oriented salient objects in images beyond top-down and bottom-up mechanism. It possesses three parts: selection history, current goal and physical salience. Selection history is integrated with current goal and physical salience to compose an integrative framework. In this extended version, our ameliorated model is applied to face and car detection and simulates task-dependent reasonable eye trajectories (visual scan paths). Experimental results demonstrate that selection history (reward) is shown to influence saccade trajectories. Our findings support the idea that attention and gaze can be directed voluntarily to regions of interest (by selection history and current goal) and can be captured by local features of an object that stand out from the background (by physical salience).

1. Introduction

Visual search plays a key role in our everyday activities; the visual system pays attention to the salient objects for efficient search. Visual saliency plays important roles in natural vision in that saliency can direct eye movements, deploy attention, and facilitate tasks like object detection and scene understanding. Attention determines which among multiple competing stimuli are represented in the brain. There are two conventional categories of factors that drive attention: bottom-up and top-down factors [11]. Bottom-up factors are derived solely from the visual scene. Regions of interest that attract our attention are in a bottom-up way and the responsible feature for this reaction must be sufficiently discriminative with respect to surrounding features. Top-down methods [14] [23] are task-driven or goal-driven. Top-down and bottom-up factors should be combined to direct attentional behavior. Many models have been built to compute saliency maps. A recent review of attention models from a computational perspective can be found in [6] [7]. Saliency models have been developed for eye fixation prediction and salient object detection. The former focuses on identifying a few fixation locations on natural images, which is important for understanding human attention. The latter, also called salient object segmentation, is used to accurately detect where the salient object should be, which is useful for many high-level vision tasks [24] [6].

Recently, the theoretical dichotomy of attentional control between top-down and bottom-up is challenged. The dichotomy fails to explain a growing number of cases in which neither bottom-up nor top-down can account for strong selection biases [4]. Thus, the top-down versus bottom-up dichotomy is an inadequate taxonomy of attentional control. A wealth of research has demonstrated that reward information provides a third source of input to the attention system. When a large reward is received, attention is strongly primed to select the rewarded target [12] [10].

Awh et al. [4] proposed selection history (including two classes of 'history' effects, i.e., selection and reward history) as a third category of control by explicitly distinguishing current goals from selection history effects. A 'priority map' which they still highlighted integrates three distinct categories of selection bias: the observer's current selection goals, selection history, and physical

salience of the items competing for attention. Acknowledging selection history as a third category of control can clarify many ongoing debates and can make clear large swaths of selection phenomena that are unrelated to current selection goals and physical salience. This concept model is a breakthrough to traditional prominent models of attentional control.

In the paper [25], we proposed a general computational framework for detecting specific salient objects in images beyond top-down and bottom-up mechanisms and verified qualitative and quantitative effects of current selection goals and selection history in our experiments. Salient objects are detected by directly measuring the saliency of an image window in the original image and the well established sliding window based object detection paradigm is adopted. Our experimental results on challenging object detection datasets demonstrate that physical salience generates a bottom-up saliency map for highlighting the salient regions of an image. The main effect of the selection history is to concentrate on salient objects, the current goal has a strong effect in detecting correct salient objects. Experiments also indicate that there is competition among selection history, current goal and physical salience to detect correct salient objects.

In this paper, we will extend our previous work. Our model is applied to salient object detection and generates task-dependent eye trajectories.

The rest of this paper is organized as follows. Section II introduces related works. Our extended computational framework is described in Section III. Experimental results and comparisons are presented in Section IV, and conclusions are given in Section V.

2. Related Works

The objectness measure [2] quantifies how likely an image window contains an object of any class. Each outputting image window is endowed with an objectness score to measure how likely this window contains a salient object. It uses several existing image saliency cues (including a novel 'superpixels straddling' cue to capture the closed boundary characteristic of objects), and greatly reduces the number of windows from an image according to their objectness distribution. We use it to generate physical salience and quantify how likely an image window contains a salient object [25].

LabelMe [16] is a web-based image annotation tool that is used to label the identity of objects and where they occur in images. We use the HOG (Histograms of Oriented Gradient) descriptor from the LabelMe toolbox and extend it to extract image features of current goal, selection history and sampled image windows.

Fig. 1 is a schematic diagram of our framework [25]. Selection history, physical salience, and current goal are three distinct sources of selection biases to accomplish salient object detection. Our primary goal is to present a general computational framework for detecting salient objects in images. Selection history is integrated with current goals and physical salience to compose an integrative framework.

1) *Selection history*. This category of control is intended to represent 'history' effects which shape the overall landscape of the observer's selection biases. Umemoto et al. [21] demonstrated that likely targets are more likely to be encoded into working memory, even when observers lack explicit knowledge of this contingency. We extract the mean HOG features of a specific image class as the selection history used to indicate reward.

2) *Physical salience*. This category represents the fact that selection is sometimes biased in a manner that depends only on the properties of the stimulus display itself, it has been described as a 'bottom-up' (stimulus-driven) process. In our study, we use the objectness measure [2] to generate physical salience and quantify how likely a sampled image window contains a salient object.

3) *Current goal*. This category of attentional control acknowledges the critical role of goal-driven selection, it has been described as a 'top-down' (goal-driven) process. Whenever there is a search task, goal-driven processes tend to dominate guidance, as indicated by attention being systematically biased toward image features that resemble those of the search target.

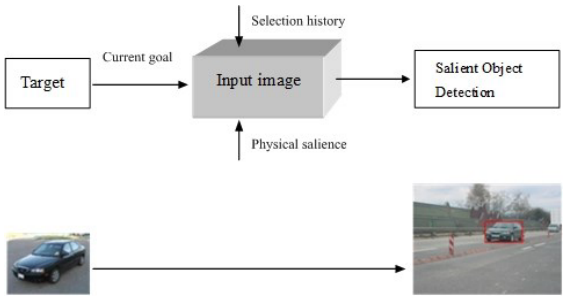

Figure 1. A schematic diagram of our framework [25].

In the present study, we devise a computational model of visual attention during search in complex scenes based on similarity between the target and regions of the search scene. Similarity is defined using a histogram-matching technique.

An image window saliency is defined as the objectness score of the window [2]. The objectness measure [2] is executed and outputs a number of windows.

The most promising candidate window is defined as:

$$\max_{score} \min_{w} (\| f_{in} - f_w \|)^{\|f_w - f_h\|} \quad (1)$$

Where f_{in} , f_w and f_h denote the features of target (current goal), sampled window and selection history respectively, $\|\bullet\|$ is *L-1* norm of features.

If there is no target, the most promising candidate window is defined as:

$$\max_{score} \min_{w}(\| f_w - f_h \|) \quad (2)$$

We use (2) to verify the effect of selection history.

Each image may have several objects of the same class, we output the best 4 detected windows in our experiments.

Fig. 2 is an example for image detection results using our detector. On each image the best 4 detected windows are superimposed, brighter windows have higher objectness and saliency.

Fig. 2(b) shows that the history has a strong influence on deployment of attention. The selection history biases the characteristic of visual attention during search toward scene regions that resemble target features. This is in line with the research by Umemoto et al. [21]. Fig. 2(c) demonstrates that the task has a strong influence on deployment of attention too. The salient object within the image is correctly detected.

Figure 2. (best viewed in color) Example detection results using objectness [2] and our detector. On each image the best 4 detected windows are superimposed, brighter windows have higher objectness. From left to right: (a) objectness [2] based detetctor, (b) history effect, (c) our integrative detector [25].

3. An Extended Algorithm Implementation for Salient Object Detection

The limitations of the model are unsteady and time-consuming due to objectness [2]. In this extended version, objectness [2] is superseded by the algorithm LARK (Locally Adaptive Regression Kernels) proposed by Seo and Milanfar [19].

LARK is a generic detection/localization algorithm capable of searching for a visual object of interest without training. The proposed method operates using a single example of an object of interest to find similar matches, does not require prior knowledge (learning) about objects being sought, and does not require any preprocessing step or segmentation of a target image either. LARK is based on the computation of local regression kernels as descriptors from a query (target), which measure the likeness of a pixel to its surroundings. Salient features are extracted from said descriptors and compared against analogous features from the target image.

The detection framework of LARK can also be useful for solving the bottom-up saliency detection [18]. Therefore, the schematic diagram of our framework modified as shown in Fig. 3.

The most promising candidate window is defined as:

$$\min_{w}(\| f_{in} - f_w \|)^{\|f_w - f_h\|} \quad (3)$$

Where f_{in} , f_w and f_h denote the features of target (current goal), sampled window and selection history respectively, $\|\bullet\|$ is *L-1* norm of features.

If there is no target, the most promising candidate window is defined as:

$$\min_{w}(\| f_w - f_h \|) \quad (4)$$

We use (4) to verify the effect of selection history.

Each image may have several objects of the same class. The number of the detected objects m satisfies:

$$D_m < 1.5\bar{D}_n = \frac{1.5}{n}\sum_{i=1}^{n} D_i \quad (5)$$

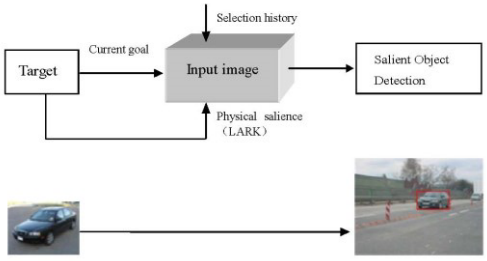

Figure 3. A modified schematic diagram of our framework.

where $D = (\| f_{in} - f_w \|)^{\|f_w - f_h\|}$, n is the number of output windows running LARK.

4. Experimental Results

Attention and saliency play important roles in visual perception. Fixations were found around faces, cars, and text [8] [13]. In our experiments, the model is applied to face and car detection and generate task-dependent eye trajectories.

Dalal and Triggs [9] showed that HOG descriptors significantly outperformed existing feature sets for human detection. In our experiments, HOG descriptors [15] are extracted from current goal, selection history and sampled image windows.

4.1 Face Detection and Task-Dependent Eye Trajectories

We extract mean HOG descriptors as the selection history from the ORL face database provided by Samaria and Harter [17]. There are 10 different images of 40 distinct

subjects. For some of the subjects, the images were taken at different times, varying lighting slightly, facial expressions (open/closed eyes, smiling/non-smiling) and facial details (glasses/no-glasses). All the images are taken against a dark homogeneous background and the subjects are in up-right, frontal position (with tolerance for some side movement). The size of each image is 92x112, 8-bit grey levels. Fig. 4 shows images of faces randomly selected from the ORL face database.

Figure 4. Example images randomly selected from the ORL face database.

Fig. 5 shows our result on the image from Seo and Milanfar [19]. Eye trajectory (visual scan path) is superimposed on the image. Red represents higher resemblance.

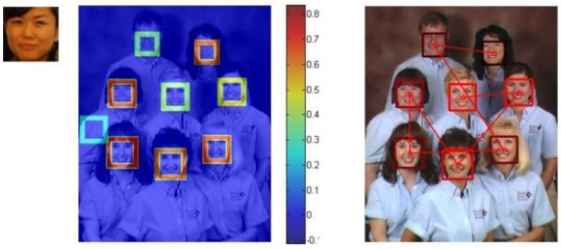

Figure 5. (Left) The original query. (Middle) The LARK detection result by Seo and Milanfar [19]. (Right) Our result. Eye trajectory superimposed on the image (solid line).

From Fig. 6, it can be seen that selection history (reward) guides search automatically and interferes with the performance of a specific visual search task. It has a direct, non-volitional impact on human perception and attention that is independent of its impact on endogenous attentional control. The anterior cingulate cortex (a cortical expression of the mesolimbic dopamine system) plays a crucial role in this source of attentional control [12]. The results demonstrate that selection history acts to change visual salience and thus plays an important and undervalued role in attentional control. A recent review by Anderson [3] also demonstrated that learned value plays a distinct role in the guidance of attention (referred to as value-driven attention).

The more similar matches have higher priority to attend to and have higher reward. Eye trajectories generated by our model are consistent with the study assuming that eye movements are selected to maximize reward by reducing uncertainty that could result in suboptimal actions and that framing the decision about where to look in terms of uncertainty reduction has been effective in explaining aspects of static scene viewing as well as dynamic scene viewing [20].

Fig. 7 illustrates the comparison between our results and [22]. It is obviously that our model have better effect on face detection and eye trajectories.

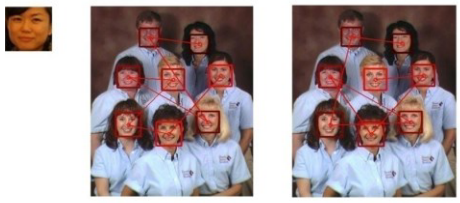

Figure 6. (Left) Query. (Middle) Only query effect. (Right) Only history effect. Eye trajectories superimposed on the images (solid line).

Figure 7. (Left) Query. (Middle column) Salient objects segmentation results and scan paths are generated using the latest version of the saliency tool box downloaded from http://www.saliencytoolbox.net using the default parameters. Full details of the model can be found in [22]. (Right) Our result eye trajectories superimposed on the images (solid line).

4.2 Car Detection on the UIUC Car Dataset

The UIUC car dataset [1] consists of learning and test sets. The learning set contains 550 positive (car) images and 500 negative (non car) images. The test set is divided into two parts: 170 gray-scale images containing 200 side views of cars of size 100*40, and 108 gray-scale images containing 139 cars at various sizes with a ratio between the largest and smallest cars of about 2.5. We use only one query image at a time from the 550 positive examples and extract mean HOG descriptors as the selection history. Fig. 8 shows images of cars randomly selected from the UIUC car dataset.

Fig. 9 shows our results of correct detections on the UIUC car test set. Eye trajectory is superimposed on the

image (solid line) if the image has more than one car. We compute detection precision as:

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

where TP is the number of true positives, FP is the number of false positives.

The overall detection precision improvement of our proposed method over LARK [19] is from 71.7% to 79.5% on the UIUC car test set. Our method increases 8% by using (5) to filter out false positives.

Figure 8. Example images randomly selected from the UIUC car dataset.

Figure 9. Examples of correct detections on the UIUC car test set. Eye trajectories superimposed on the images (solid line).

4.3 Discussions

LARK is a powerful training-free nonparametric object detection framework by employing local steering kernel (LSK), which well captures image underlying structure, and by using the Matrix Cosine Similarity (MCS) measure [19]. The method can automatically detect in the target image the presence, the number, as well as location of similar objects to the given query image. To deal with more general scenarios, accounting for large variations in scale and rotation, multiscale and multirotation approach were also proposed. It is faster (3.655 second per image of our integrative detector MATLAB implementation on the UIUC car dataset with Intel(R) Core(TM) i3-2330 2.2 GHz CPU and 4GB RAM), while objectness [2] is time-consuming. In this study, objectness [2] is superseded by the algorithm LARK and (5) is used to filter out false outputs.

We apply our model to face and car detection and generate task-dependent eye trajectory. Experimental results demonstrate that selection history (reward) is shown to influence saccade trajectories. This is consistent with the study by Belopolsky [5].

As in [19], the performance of our method is little affected by the choice of similar query images. Semantic feature extraction and semantic similarity measure will be employed to further depress the effect in our future study.

5. Conclusions

Attention is a central question for vision science. The traditional theory of top-down versus bottom-up is an inadequate taxonomy of attentional control. In our previous study [25], we presented a general computational framework for detecting salient objects in images beyond top-down and bottom-up mechanism. Its three parts (selection history, current goal and physical salience) are integrated to compose an integrative framework. We verified qualitative and quantitative effects of current selection goals and selection history in experiments. Experimental results suggest that physical salience generates bottom-up saliency maps for highlighting the salient regions of images. The main effect of the selection history is to concentrate on salient objects although this might cause errors. The current goal has a strong effect on detecting correct salient objects. Our experiments indicated that there is competition among selection history, current goal and physical salience to detect correct salient objects. The limitations of the model are unsteady and time-consuming due to objectness [2]. In this extended version, objectness [2] is superseded by the algorithm LARK. Our ameliorated model is applied to face and car detection and simulates task-dependent reasonable eye trajectories. Experimental results demonstrate that selection history (reward) is shown to influence saccade trajectories. Our findings support the idea that attention and gaze can be directed voluntarily to regions of interest (by selection history and current goal) and can be captured by local features of an object that stand out from the background (by physical salience).

The main contributions of this study are our integrative computational framework and experimental conclusions. Our combined computational model currently remains limited to processing of low-level features, and as such it is unable to reflect eye movement influences that depend on higher-level visual features (such as objects). Our current detector implementation is simple and there is much room for improvement, e.g., using more sophisticated visual features.

Acknowledgment

The authors would like to thank the reviewers for their constructive comments and suggestions. This work is supported by the National Natural Science Fund of China (Grant Nos. 61373063, 61373062), the project of Ministry of Industry and Information Technology of China (Grant No. E0310/1112/02-1), and the Natural Science Fund of JSNU (Grant No. 15XLA09).

References

[1] Agarwal, S., Awan, A. and Roth, D., "Learning to Detect Objects in Images via a Sparse Part-Based Representation," IEEE Trans. Pattern Analysis and Machine Intelligence, 26(11), pp.1475-1490, 2004.

[2] Alexe, B., Deselaers, T., and Ferrari, V., "What is an object?" In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 73–80, 2010.

[3] Anderson, B. A., "A value-driven mechanism of attentional selection," Journal of Vision, vol. 13, no. 3, pp. 1–16, 2010.

[4] Awh, E., Belopolsky, A. V., and Theeuwes, J., "Top-down versus bottom-up attentional control: A failed theoretical dichotomy," Trends in cognitive sciences, 16(8), pp.437–443, 2012.

[5] Belopolsky, A. V., "Common priority map for selection history, reward and emotionin the oculomotor system," Perception, vol. 0, no. 0, pp.1–14, 2015.

[6] Borji, A., Sihite, D. N., and Itti. L., "Salient object detection: a benchmark," In Proc. European Conf. on Computer Vision (ECCV), pp.414–429, 2012.

[7] Borji, A., and Itti, L., "State-of-the-art in Visual Attention Modeling," IEEE Trans. Pattern Anal. Mach. Intell., PAMI 35(1), pp.185–207, 2013.

[8] Cerf, M., Harel, J., Einhauser, W., and Koch, C., "Predicting human gaze using low-level saliency combined with face detection," Neural Information Processing Systems, vol. 20, pp. 241–248, 2007.

[9] Dalal, N., and Triggs, B., "Histograms of oriented gradients for human detection," In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Vol. 1, pp. 886-893, 2005.

[10] Della Libera, C., and Chelazzi, L., "Visual selective attention and the effects of monetary reward," Psychological Science, vol. 17, pp. 222–227, 2006.

[11] Desimone, R., and Duncan, J., "Neural mechanisms of selective visual attention," Annual Reviews of Neuroscience. 18, 193–222, 1995.

[12] Hickey, C., Chelazzi, L., Theeuwes, J., "Reward changes salience in human vision via the anterior cingulated," The Journal of Neuroscience, 30(33), pp.11096-11103, 2010.

[13] Judd, T., Ehinger, K., Durand, F., and Torralba, A., "Learning to predict where humans look," in Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV '09), pp.2106–2113, Kyoto, Japan, September 2009.

[14] Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., and Shum, H., "Learning to detect a salient object," Pattern Anal. Mach. Intell. PAMI 33(2), pp.353–367, 2011.

[15] Ludwig, O., Delgado, D., Goncalves,V., Nunes, U.,"Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection," in: 12th Internat. Conf. on Intelligent Transportation Systems, pp.432-437, 2009.

[16] Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T., "LabelMe: a database and web-based tool for image annotation," International Journal of Computer Vision, 77(1-3), pp.157-173, 2008.

[17] Samaria, F., Harter, A., "Parameterisation of a stochastic model for human face identification," in 2nd IEEE Workshop on Applications of Computer Vision December, Sarasota (Florida), 1994.

[18] Seo, H. J., Milanfar, P., "Static and Space-Time Visual Saliency Detection by Self-Resemblance," J. Vision, vol. 9, no. 12, pp. 1-27, 2009.

[19] Seo, H. J., Milanfar, P., "Training-free, generic object detection using locally adaptive regression kernels," Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(9), pp. 1688-1704, 2010.

[20] Tatler, B. W., Hayhoe, M. M., Land, M. F., et al.. "Eye guidance in natural vision: Reinterpreting salience," Journal of Vision, 11(5), pp. 1–23, 2011.

[21] Umemoto, A., Scolari, M., Vogel, E. K., and Awh, E., "Statistical learning induces discrete shifts in the allocation of working memory resources," Journal of Experimental Psychology: Human Perception and Performance, 36(6), pp.1419-1429, 2010.

[22] Walther, D., Koch, C., "Modeling attention to salient proto-objects," Neural Networks, vol. 19, pp.1395–1407, 2006.

[23] Yang, J., and Yang, M., "Top-down visual saliency via joint crf and dictionary learning," In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2296–2303, 2012.

[24] Yang, C., Zhang, L.H, Lu, H.C, & Ruan, X., "Saliency detection via graph-based manifold ranking," In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.3166–3173, 2013.

[25] Zhang, D. Z., Liu, C. C., "A salient object detection framework beyond top-down and bottom-up mechanism," Biologically Inspired Cognitive Architectures, vol.9, pp. 1-8. 2014.