Learning Visual Saliency

Qi Zhao and Christof Koch
Computation and Neural Systems
California Institute of Technology, Pasadena, CA 91125
Email: {qzhao,koch}@klab.caltech.edu

***Abstract*—Inspired by the primate visual system, computational saliency models decompose the visual input into a set of feature maps across spatial scales. In the standard approach, the feature maps of the pre-specified channels are summed to yield the final saliency map. We study the feature integration problem and propose two improved strategies: first, we learn a weighted linear combination of features using the constraint linear regression algorithm. We further propose an AdaBoost based algorithm to approach the feature selection, thresholding, weight assignment, and nonlinear integration in a single principled framework. Extensive quantitative evaluations of the new models are conducted using four public datasets, and improvements on model predictability power are shown.**

I. Introduction

Humans and other primates move their eyes to select visual information from any one visual scene. This allows them to bring the high-resolution part of their retina, the fovea, onto relevant parts of the image, thereby focusing processing resources on the most relevant visual information and interpreting complex scenes in real time. Besides being able to understand the mechanism that drives this selection of interesting parts in the image, predicting interesting locations as well as locations where people are likely to look has tremendous real-world applications.

Commonalities between fixation patterns in synthetic or natural scenes from different individuals allow computational models to predict where people look. Starting from the *Feature Integration Theory* of Treisman and Gelade [1] and the proposal by Koch and Ullman [2] for a map in the primate visual system that encodes the extent to which any location in the field of view is conspicuous or salient, based on bottom-up, task-independent factors, a series of ever refined algorithms have been designed to predict where subjects will fixate in synthetic or natural scenes [3], [4], [5], [6], [7], [8], [9].

On top of all the recent advances in image features and normalization methods for computational saliency models, the *linear summation* of different feature channels into the final saliency map remains the norm [3], [10], [11], [12]. Linear summation has some psychophysical support [13] and is simple to apply. However, some psychophysical arguments [14], [15] have been raised against linear summation strategies. In addition, prior work [16], [17] has been aware of the different strengths contributed by different features to perceptual saliency.

We study feature integration strategies of computational saliency models using eye movement data from four recent datasets [12], [18], [19], [20]. Based on the conventional linear summation approach, we first investigate the importance of different bottom-up features in driving gaze allocation and improve the linear method by learning a set of optimal feature weights using the constraint linear least square algorithm. In this method, we retain the basic structure of the standard saliency model [3], [12] by using a linear integration scheme and considering a small number of bottom-up feature channels - color, intensity, orientation, and face. A considerable advantage of learning weights for a basic set of feature channels is its compatibility with a large psychophysical and physiological literature.

We further propose an AdaBoost [21], [22], [23], [24] based algorithm that provides a principled computational framework for feature selection and integration. The AdaBoost based model has the following key advantages. First, the new model selects from a feature pool the most informative features that nevertheless have significant variety. The framework could easily incorporate any interesting features proposed by researchers in the community and selects the best ones in a greedy manner. Second, it finds the optimal threshold for each feature, which is consistent with the neuron firing properties [14]. Third, it makes no assumption of linear superposition or equal weights of features. Indeed, we explicitly demonstrate that certain types of nonlinear combination consistently outperform linear combination. This raises the question of the extent to which the primate brain takes advantages of such nonlinear integration strategies. Biological neurons are highly nonlinear devices [25]. Thus implementing the type of nonlinearities inherent in AdaBoost is not particular problematic for the brain. Future psychophysical and neurophysiological research will be needed to untangle this question.

The rest of the paper is organized as follows: Section II presents a linear integration algorithm where optimal weights of feature channels are obtained using linear regression with constraints. Section III discusses a nonlinear feature integration method based on AdaBoost learning. Section IV demonstrates promising comparative and quantitative results and Section V concludes the paper.

II. Learning Optimal Weights

To quantify the relevance of different features in deciding where to look, we use linear, least square regression with constraints to learn the weights from eye movement data.

Formally, let $\mathbf{C}$, $\mathbf{I}$, $\mathbf{O}$, $\mathbf{F}$ be the stacked vectors of the color, intensity, orientation, and face values at all image locations (see Section IV-A1 for details of features), denote $\mathbf{V} =$

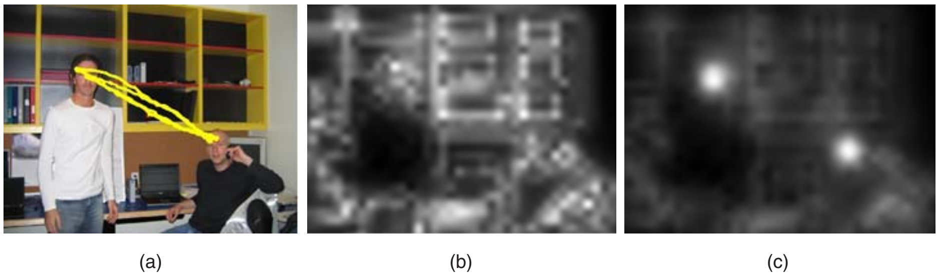

Fig. 1. Illustration of how feature weights affect the final saliency maps. (a) Original image with eye movements of one subject (fixations denoted as red circles). Image and eye movement data are from the FIFA dataset [12]. (b) Saliency map from linear combination with equal weights. (c) Saliency map from linear combination with optimal weights (Table I), where the face channel is emphasized.

$[\mathbf{C}\ \mathbf{I}\ \mathbf{O}\ \mathbf{F}]$, $\mathbf{M}_{fix}$ vectorized fixation map that is represented as the recorded fixations convolved with an isotropic Gaussian kernel (see Figure 3(b) for an example of a fixation map), and $\mathbf{w} = [w_C\ w_I\ w_O\ w_F]^T$ the weights of the feature channels. The objective function is

$$\arg\min_{\mathbf{w}} \|\mathbf{V} \times \mathbf{w} - \mathbf{M}_{fix}\|^2, \tag{1}$$

subject to

$$\mathbf{w} \geq \mathbf{0}.$$

The problem is solved using an active set method similar to that described in [26].

Figure 1 provides an illustration of how feature weights affect the final saliency maps. The weight of a feature indicates the importance of that particular feature in deciding where to look at. We see that the map with optimal weights (Figure 1(c)) is more consistent with the eye movement data (Figure 1(a)).

To investigate the level of inter-subject variability, we learn optimal weights for each individual as well as for the entire population of subjects. The only difference are the fixation maps we feed the algorithm (eq. 1).

III. Learning Nonlinear Feature Integration Using AdaBoost

To quantify the relevance of different features across multiple scales in deciding where to look, we learn using AdaBoost a nonlinear integration of features $G(\mathbf{f}) : R^d \to R$, where d is the dimension of features. The AdaBoost algorithm [21], [22], [23], [24] has been shown to be one of the most effective method for object detection [27], [28]. As a special case of boosting, the strong classifier is a weighted combination of weak classifiers that are iteratively built, and subsequent weak classifiers are tweaked in favor of the misclassified instances.

Formally,

$$G(\mathbf{f}) = \sum_{t=1}^{T} \alpha_t g_t(\mathbf{f}), \tag{2}$$

where $g_t(\cdot)$ denotes the weak learner ($g_t(\mathbf{f}) \in \{0,1\}$) and $G(\cdot)$ the final saliency model ($G(\mathbf{f}) \in [0,+\infty)$). α_t is the weight of $g_t(\cdot)$, as would be described in Algorithm 1 below. T is the number of weak learners. Instead of taking the sign of the AdaBoost output, as conventionally used to solve classification problems, we use the real values of this $G(\mathbf{f})$ to construct a saliency map. Algorithm details are given in Algorithm 1, and Figure 2 illustrates the training and testing stages of the AdaBoost based saliency model.

Algorithm 1 Learning A Nonlinear Feature Integration Using AdaBoost

Input: Training dataset with N images and eye movement data from M subjects. A testing image Im.

Output: Saliency map of Im.

Training stage (from eye movement data of M subjects viewing N images):

1. For all locations in N images, sample $\{\mathbf{x}_s\}_{s=1}^S$ with labels $\{y_s\}_{s=1}^S$. See Section IV-A1 for details of sampling. Compute $\{\mathbf{f}_s\}_{s=1}^S = \{\mathbf{f}(\mathbf{x}_s)\}_{s=1}^S$ as a stack of features for the sample at location $\mathbf{x}_s$.

2. Initialize weights to be $\{w_s = \frac{1}{S}\}_{s=1}^S$.

3. For $t = 1, .., T$

a. Train a weak classifier g_t: $R^d \to \{0,1\}$, which minimizes the weighted error function $g_t = \arg\min_{g_u \in \mathcal{G}} \epsilon_u$, where $\epsilon_u = \sum_{s=1}^S w_t(s)[y_s \neq g_u(\mathbf{f}_s)]$.

b. Set the weight of g_t as $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$.

c. Update sample weights $w_{t+1}(s) = \frac{w_t(s)\exp(-\alpha_t \cdot y_s \cdot g_t(\mathbf{f}_s))}{Z_t}$, where Z is a normalization factor.

4. The final model is defined as $G(\mathbf{f}) = \sum_{t=1}^T \alpha_t g_t(\mathbf{f})$.

Testing stage (for a new image Im): for each location $\mathbf{x}$ in Im, compute the feature vector $\mathbf{f}(\mathbf{x})$, then apply the strong classifier $G(\mathbf{f}) : R^d \to R$ (eq. 2) to obtain the saliency value of $\mathbf{f}(\mathbf{x})$.

IV. Experimentals

A. Experiment Paradigm

1) Features and Fixation Map: To focus on the comparisons of different combination algorithms, we use a simple set of biologically plausible features [3], [12]. Specifically we consider two color channels (blue/yellow and red/green), one intensity channel, four orientation channels $(0°, 45°, 90°, 135°)$ at (1) six levels of pyramid $l = \{2,3,4,5,6,7\}$, (2) six center-surround (c-s) levels (center level $c = \{2,3,4\}$, surround level $s = c+\delta$, where $\delta = \{2,3\}$),

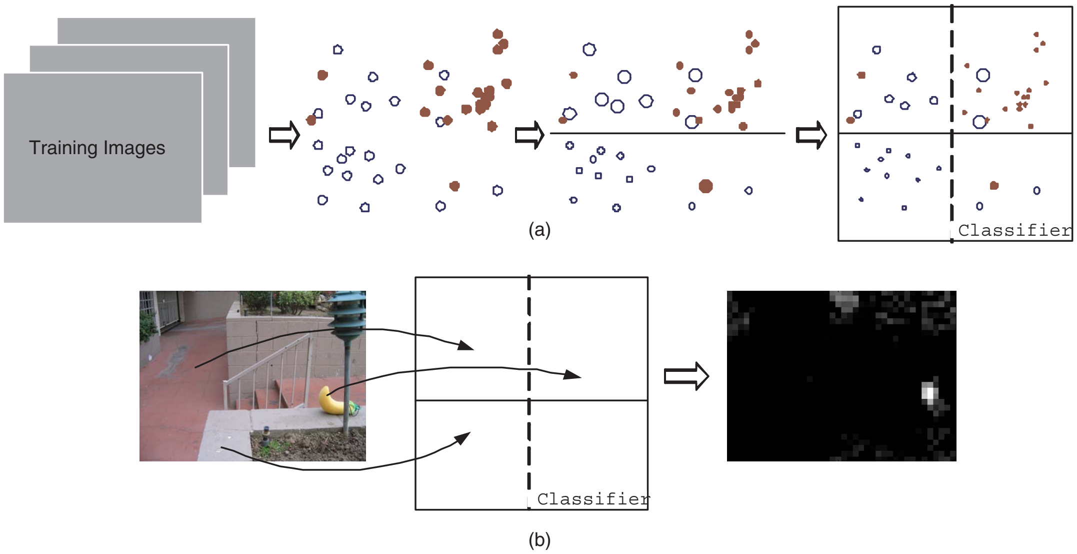

Fig. 2. Illustration of the AdaBoost based saliency model. (a) Training stage: using samples from training images, weak classifiers are trained through iterations and combined to form a strong classifier (we use a n-dimensional ($n \geq 4$) space but plot a 2-dimensional one here for illustration). (b) Testing stage: for a new image, the feature vectors of image locations are calculated and input to the strong classifier to obtain saliency scores. The rightmost map is the resulting saliency map, where brighter regions denote more salient areas (output from the final model (eq. 2) is fed into a sigmoid function for visualization purposes).

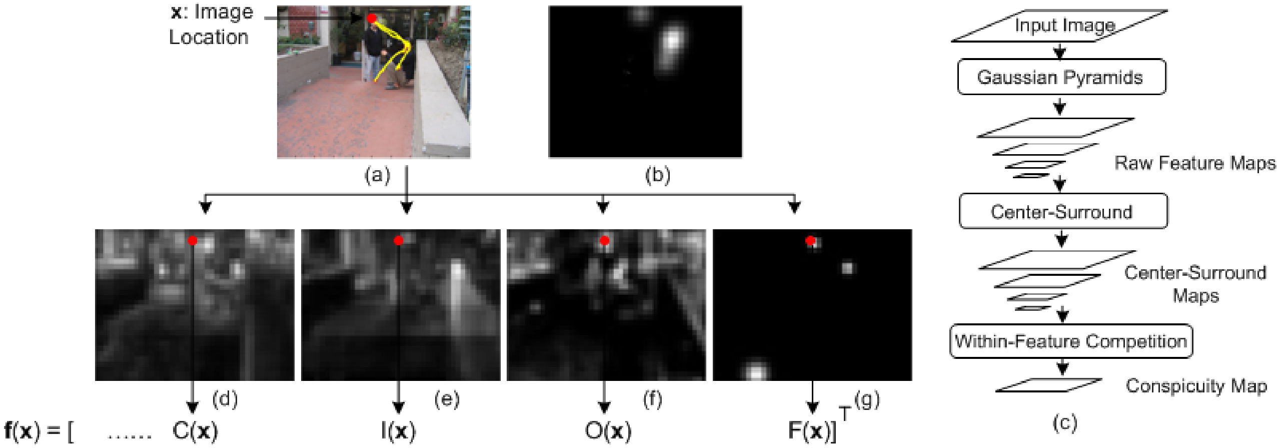

Fig. 3. Sample Illustration. (a) Original image with eye movements of one subject. (b) Fixation map of the same subject. (c) An illustration of generating a conspicuity map from the input image. For the color and orientation features, each has multiple channels, and the conspicuity map in this case is the summation of all the channels of that particular feature. (d-g) Color, intensity, orientation conspicuity maps and face map. A sample data is a feature vector comprising the feature values of all relevant maps at a particular location.

and (3) three conspicuity maps capturing the “conspicuity” of color, intensity, and orientation channels, which are built through across-scale addition of the center-surround difference maps (represented at scale 4). Specifically as shown in Figure 3(c), for an input image, raw feature maps at different spatial scales are created using dyadic Gaussian pyramids [29], then center-surround is implemented as the point-wise difference between fine and coarse scales to construct center-surround maps, finally for each low-level feature (i.e., color, intensity, and orientation), a conspicuity map is generated after within-feature normalization and summation. For the face channel, the conspicuity map is generated by running the Viola & Jones face detector [27]. Although different from early visual features such as color, intensity, and orientation, face attracts attention strongly and rapidly, independent of task, therefore is also considered part of the bottom-up saliency pathway [12]. An illustration of the conspicuity maps and the face map is given in Figure 3(d)-(g).

In learning optimal weights, a sample vector at an image location $\mathbf{x}$ is a feature vector that comprises the values of the color, intensity, orientation, and face conspicuity maps at this particular location. This way using values directly from the conspicuity maps, the most compact information is used for regression. Differently for nonlinear integration, since the AdaBoost framework automatically selects good features, we consider all the intermediate maps (i.e., raw feature maps and center-surround maps) as well as the conspiciuty maps for learning, where the feature vector for a particular image location **x** comprises values from all these maps at **x**.

Fixation maps are constructed by convolving recorded fixations with an isotropic Gaussian kernel (an example shown in Figure 3(b)). Note that the first fixation of each image is not used as it is always the center of the image. In learning optimal weights for linear combination, such fixation maps are directly fed into the constraint linear least square algorithm. For learning a nonlinear combination, we use a sampling technique to convert the continuous map into binary labels: locations of positive examples are sampled from the fixation maps (i.e., an image location with a larger value in fixation maps has a higher probability of being sampled as a positive sample), and locations of negative examples are sampled uniformly from non-activated areas (i.e., with zero values) of the fixation maps.

2) Datasets: This study analyzes eye movements from four recent datasets [12], [18], [19], [20]. Briefly In the FIFA dataset [12], fixation data were collected from 8 subjects performing a 2 second long "free-viewing" task on 180 color natural images (28° × 21°). Images included faces in various skin colors, age groups, gender, positions, and sizes. The second dataset from Bruce & Tsotsos [18] (referred to as Toronto dataset) contains data from 11 subjects viewing 120 color images of outdoor and indoor scenes, where participants were given no particular instructions except to observe the images (32°×24°), 4 second each. The third dataset published by Judd *et al.* [19] (referred to as MIT database) includes 1003 images collected from *Flickr creative commons* and *LabelMe*. Eye movement data were recorded from 15 users who free viewed these images (36° × 27°) for 3 second. The fourth dataset we use is published by Subramanian *et al.*, which includes 758 images containing semantically affective objects/scenes such as expressive faces, nudes, unpleasant concepts, and interactive actions. Images are from *Flickr*, *Photo.net*, *Google*, and *emotion-evoking IAPS* [30]. In total 75 subjects free-viewed (26° × 19°) part of the image set for 5 second each (each image was viewed by an average of 25 subjects).

3) Similarity Measure: Unlike most saliency papers that use solely the Area Under the ROC Curve (AUC) to quantify model performance, we found in our previous work [31] that in practice, as long as the hit rates are high, the AUC is always high regardless of the false alarm rate. Therefore, an ROC analysis, while very useful, is by itself insufficient to describe the deviation of predicted fixation patterns from the actual fixation map. In this work we continue to use three complementary similarity measures to provide a comprehensive assessment metric - AUC is about order while the Normalized Scanpath Saliency (NSS) [4], [17] and the Earth Mover's Distance (EMD) [32] measure differences in value. In addition, both AUC and NSS compare maps primarily at the exact locations of fixation while EMD accommodates shifts in location and reflects the overall discrepancy between two maps on a more global scale.

Given the extant variability among different subjects looking at the same image, no saliency algorithm can perform better (on average) than the AUC dictated by inter-subject variability. We compute an ideal AUC by measuring how well the fixations of one subject can be predicted by those of the other $n-1$ subjects, iterating over all n subjects and averaging the result. These AUC values are 78.6% for the FIFA dataset, 87.8% for the Toronto dataset, 90.8% for the MIT dataset, and 85.7% for the NUS dataset. In general, we express the performance of saliency algorithms in terms of normalized AUC (nAUC) values, which is the AUC using the saliency algorithm normalized by the ideal AUC.

A strong saliency model should have an nAUC value close to 1, a large NSS and a small EMD value.

B. Performance

Table I shows, for each of the above four datasets (Section IV-A2), optimal weights obtained for each of color, intensity, orientation, and face feature channels, using linear regression with constraints. Experiments show that the weights learned from all four datasets are consistent - except the Toronto dataset which includes few frontal faces, we see that face is the most important, followed by orientation, color, and intensity.

	Color	Intensity	Orientation	Face
FIFA [12]	0.027	0.024	0.222	0.727
Toronto [18]	0.403	0.067	0.530	0
MIT [19]	0.123	0.071	0.276	0.530
NUS [20]	0.054	0.049	0.256	0.641

TABLE I
Optimal weights learned from four datasets [12], [18], [19], [20]). Face is the most important (except the Toronto dataset [18] which include few frontal faces), followed by orientation, color, and intensity.

We then perform quantitative evaluations of different saliency models on the datasets. Due to space limitation we in this paper present experimental results and discussions on the FIFA and the Toronto datasets.

In the first experiment we compare linear and nonlinear integration on the FIFA dataset [12]. The dataset of 180 images is divided into 130 training images and 50 testing ones. The subject-specific models are learned using eye movement data from one specific observer, while the general model is trained using data from all 8 subjects. In this experiment we limit the nonlinear integration to the level of conspicuity maps and include three conspicuity maps and a face map in the feature pool. The measurements are averaged over the 8 subjects.

From Table II, we make the following observations: (1) by setting proper weights to different feature channels, the linear model improves significantly. Predictions are more consistent with eye movement data using the optimal weights obtained from constrained linear least squares (Section II). This suggests that we do rely on certain features more than others in deciding where to look at and such features

	Linear Integration			Nonlinear Integration	
	Equal Weights	Optimal Weights			
		Subject-Specific	General	Subject-Specific	General
nAUC	0.924	0.945	0.944	0.959	0.953
NSS	0.845	1.35	1.32	1.47	1.42
EMD	5.26	4.33	4.41	2.68	2.87

TABLE II
Quantitative comparisons of linear and nonlinear feature integrations on the FIFA dataset [12]. The NSS of the linear models with optimal weights are noticeably larger than those with equal weights, suggesting a greater correspondence between fixation locations and the salient points predicted by the model. The EMD is considerably smaller using optimal weights, reflecting a better global consistency between the saliency maps and the fixation maps. The nonlinear integration further improves predictability.

	Linear Integration		Nonlinear Integration		
	Equal Weights	Optimal Weights	Conspicuity Level	Raw/C-S Feature Level	
			4 channels	88 Channels	Top 10 / 88 Channels
nAUC	0.828	0.834	0.836	0.916	0.912
NSS	0.872	0.920	0.913	1.37	1.35
EMD	4.85	4.50	3.66	3.20	3.28

TABLE III
Quantitative comparisons of linear integrations with equal and optimal weights, and nonlinear integration at different levels on the Toronto dataset [18]. Including more feature channels in the feature pool improves model performance. However after the AdaBoost algorithm selects a number of “best” features, model performance does not improve much by selecting more features.

Models	Itti *et al.* [3], [12]	Gao *et al.* [33]	Bruce & Tsotsos [18]	Hou & Zhang [34]	Our Model
nAUC	0.828	0.880	0.890	0.903	0.916

TABLE IV
Normalized AUC for different saliency models on the Toronto dataset [18]. Our model is based on nonlinear integration.

should be emphasized in the final saliency map. (2) The nonlinear integration learned using AdaBoost further improves predictability. Essentially, the AdaBoost algorithm automatically finds the optimal weights for each feature through iterations. This implies that the way humans combine features is nonlinear and complex. Though a set of trained optimal weights improve the conventional linear summation strategy, the linear approximation is still insufficient. (3) Compared with AUC, the performance difference is better reflected by NSS and EMD, as discussed in Section IV-A3. (4) There is no significant improvement using subject-specific models over the general models, suggesting unremarkable inter-subject variability in terms of feature preference.

Compared with the FIFA dataset [12], the Toronto dataset [18] contains less frontal faces or other distinct large objects in an image therefore are considered a more difficult dataset. We divide the image set of 120 images into 80 training images and 40 testing ones. As there are less fixation data than the FIFA dataset we build only general models in this experiment and focus on the comparisons on linear integrations, and nonlinear integration at different levels. When the conspicuity maps and the face map are used as candidate features as in the previous experiments on the FIFA dataset, nonlinearity is performed at a coarser level as conspicuity maps are constructed by linear combinations of the raw feature maps and center-surround maps. In contrast, when all feature channels described in Section IV-A1 are included we are exploiting nonlinearity at a deeper level. Comparing the 4^{th} column of Table III to the last column of Table II where both use 4 candidate features, the results on the FIFA dataset is better, consistent with the aforementioned fact that the FIFA dataset is relatively easier due to the presence of large frontal faces and objects in most of the images.

The AdaBoost algorithm can be easily interpreted as a greedy feature selection process [27], which selects from a feature set good ones that nevertheless have significant variety. We run AdaBoost and select the top 10 features from the feature pool. The last two columns in Table III indicate that after selecting the most discriminative features, the rest do not improve performance much. In fact this demonstrates the advantage of feature selection capability of AdaBoost: it opens a framework where different features could be included, no matter independent or correlated, and the algorithm selects the best ones without efforts from the domain experts.

The Toronto dataset has been used as a benchmark in several recent works. Table IV summarizes a performance comparison of different saliency models.

V. Conclusion

This paper studies feature integration strategies of computational saliency models using eye movement data. Based on the conventional linear summation approach, we first improve the linear method by learning a set of optimal feature weights using linear regression with constraints. We further propose an AdaBoost based algorithm to learn a general nonlinear feature combination. It naturally approaches the feature selection, thresholding, weight assignment, and integration problems in a single computational framework.

Experiments demonstrate that the people depend more on certain features than others in deciding where to look at, and the predictivity of the saliency model improves significantly by addressing such differences by learning optimal weights. The advantage of the nonlinear integration over the linear ones are also shown, indicating that linear approximation is still insufficient to model the complex way that humans use to integrate different features. Furthermore, through AdaBoost feature selection, different features on various scales, with all sorts of operations (e.g., Difference of Gaussian, max-ave normalization [3]) could all be assessed in a principled manner. Insights could be derived from the outputs of the computational models, and hopefully interact with psychophysical and neurophysiological research to understand the visual system.

Acknowledgment

The authors would like to thank Jonathan Harel for helpful discussions on weighting different features. This research was supported by the NeoVision program at DARPA, by the ONR, the Mathers foundation, and the WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-10008).

References

[1] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[2] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

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

[4] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[5] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson, "Top-down control of visual attention in object detection," in *International Conference on Image Processing*, 2003, pp. I:253–256.

[6] D. Walther, T. Serre, T. Poggio, and C. Koch, "Modeling feature sharing between object detection and top-down attention," *Journal of Vision*, vol. 5, no. 8, pp. 1041–1041, 2005.

[7] T. Foulsham and G. Underwood, "What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition," *Journal of Vision*, vol. 8, no. 2, pp. 601–617, 2008.

[8] W. Einhauser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, no. 14, pp. 1–26, 2008.

[9] C. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur, "Everyone knows what is interesting: Salient locations which should be fixated," *Journal of Vision*, vol. 9, no. 11, pp. 25:1–22, 2009.

[10] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems*, 2006, pp. 547–554.

[11] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2007, pp. 545–552.

[12] M. Cerf, E. 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, pp. 10:1–15, 2009.

[13] H. Nothdurft, "Salience from feature contrast: Additivity across dimensions," *Vision Research*, vol. 40, no. 10-12, pp. 1183–1201, 2000.

[14] Z. Li, "A saliency map in primary visual cortex," *Trends in Cognitive Sciences*, vol. 6, no. 1, pp. 9–16, 2002.

[15] A. Koene and L. Zhaoping, "Feature-specific interactions in salience from combined feature contrasts: Evidence for a bottomcup saliency map in v1," *Journal of Vision*, vol. 7, no. 7, pp. 6:1–14, 2007.

[16] L. Itti, "Models of bottom-up attention and saliency," in *Neurobiology of Attention*, L. Itti, G. Rees, and J. Tsotsos, Eds. Elsevier, 2005, pp. 576–582.

[17] R. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[18] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, pp. 1–24, 2009.

[19] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision*, 2009.

[20] R. Subramanian, H. Katti, N. Sebe, M. Kankanhalli, and T. Chua, "An eye fixation database for saliency detection in images," in *European Conference on Computer Vision*, 2010, pp. 6314:30–43.

[21] Y. Freund and R. Schapire, "Game theory, on-line prediction and boosting," in *Conference on Computational Learning Theory*, 1996, pp. 325–332.

[22] J. Friedman, T. Hastle, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *The Annals of Statistics*, vol. 38, no. 2, pp. 337–374, 2000.

[23] R. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999.

[24] A. Vezhnevets and V. Vezhnevets, "Modest adaboost - teaching adaboost to generalize better," in *Graphicon*, 2005.

[25] C. Koch, *Biophysics of Computation: Information Processing in Single Neurons*. Oxford University Press, New York, New York, 1999.

[26] P. Gill, W. Murray, and M. Wright, *Practical Optimization*. Academic Press, London, UK, 1981.

[27] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. I:511–518.

[28] X. Chen and A. Yuille, "Detecting and reading text in natural scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2004, pp. 366–373.

[29] H. Greenspan, S. Belongie, R. Goodman, R. Perona, S. Rakshit, and C. Anderson, "Overcomplete steerable pyramid filters and rotation invariance," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 222–228.

[30] P. Lang, M. Bradley, and B. Cuthbert, "(iaps): Affective ratings of pictures and instruction manual," in *Technical Report, University of Florida*, 2008.

[31] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of Vision*, 2011.

[32] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[33] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," in *Advances in Neural Information Processing Systems*, 2007, pp. 497–504.

[34] X. Hou and L. Zhang, "Dynamic visual attention: searching for coding length increments," in *Advances in Neural Information Processing Systems*, 2008.