Signal Processing 93 (2013) 1401–1407

Contents lists available at SciVerse ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Learning saliency-based visual attention: A review

Qi Zhao [a,*], Christof Koch [b,c]

[a] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[b] Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA
[c] Allen Institute for Brain Science, Seattle, WA, USA

ARTICLE INFO

Article history:
Received 28 January 2012
Received in revised form
5 June 2012
Accepted 9 June 2012
Available online 27 June 2012

Keywords:
Visual attention
Machine learning
Feature representation
Central fixation bias
Public eye tracking datasets

ABSTRACT

Humans and other primates shift their gaze to allocate processing resources to a subset of the visual input. Understanding and emulating the way that human observers free-view a natural scene has both scientific and economic impact. It has therefore attracted the attention from researchers in a wide range of science and engineering disciplines. With the ever increasing computational power, machine learning has become a popular tool to mine human data in the exploration of how people direct their gaze when inspecting a visual scene. This paper reviews recent advances in learning saliency-based visual attention and discusses several key issues in this topic.

1. Introduction

Besides understanding 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 many real-world applications. Computational models can be applied to various computer vision tasks such as navigational assistance, robot control, surveillance systems, object detection and recognition, and scene understanding. Such predictions also find applications in other areas including advertising design, image and video compression, pictorial database querying, and gaze animation.

In the past decade, a large body of computational models [33,57,54,81,18,13,49,10,47] have been proposed to predict gaze allocation, some of which were inspired by neural mechanisms.

Broadly, a saliency detection approach includes the following components:

(1) Extract visual features: Commonly used features include contrast [62], edge content [2], intensity bispectra [40], color [35], and symmetry [60], as well as higher-level ones such as faces and text [9]. Image processing techniques (e.g., [69,68]) could be applied to enhance or transform low-level image features, while higher-level ones are generally more invariant.

(2) Compute individual feature maps to quantify saliency in that particular feature dimension: This step uses biologically plausible filters such as Gabor or Difference of Gaussian filters, or more sophisticated methods. For example, Itti and Baldi [31] hypothesize that the information-theoretical concept of spatio-temporal surprise is central to saliency, and compute saliency using Bayesian statistics. Gao et al. [23] and Mahadevan and Vasconcelos [47] quantify saliency based on a discriminant center-surround hypothesis. Raj et al. [61] derive an entropy minimization algorithm to select fixations. Seo and Milanfar [66] compute saliency using a "self-resemblance" measure, where each pixel of the saliency map indicates the statistical likelihood of saliency of a feature matrix given its surrounding feature matrices. Bruce and Tsotsos [4] present a model based on "self-information" after independent component analysis (ICA) decomposition [28] that is in line with the sparseness of the response of cortical

* Corresponding author. Tel.: +65 65166658.
E-mail addresses: eleqiz@nus.edu.sg (Q. Zhao), koch@klab.caltech.edu (C. Koch).

0165-1684/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2012.06.014

cells to visual input [17]. Wang et al. [82] calculate the site entropy rate to quantify saliency also based on ICA decomposition. Avraham and Lindenbaum [1] use a stochastic model to estimate the probability that an image part is of interest. In Harel et al. [25], an activation map within each feature channel is generated based on graph computations. In Carbone and Pirri [6], a Bernouli mixture model is proposed to capture context dependency.

(3) Integrate these maps to generate a final map of a scalar variable termed saliency: In the saliency literature, there have been preliminary physiological and psychophysical studies in support of "linear summation" (i.e., linear integration with equal weights) [75,52,55,14], "max" [44,39], or "MAP" [78] type of integration, where the former one has been commonly employed in computational modeling [33,9]. Later, under the linear assumption, Itti and Koch [32] suggest various ways to normalize the feature maps based on map distributions. Hu et al. [27] compute feature contributions to saliency using "composite saliency indicator", a measure based on spatial compactness and saliency density.

Unfortunately, this conventional structure requires many design parameters such as the number and type of features, the shape and size of the filters, the choice of feature weights and normalization schemes, and so on. Various assumptions are often included for modeling. For many years, the choices of these parameters or assumptions are either ad hoc or are chosen to mimic biological visual system. In many cases, however, the biological plausibility is ambiguous. While there is more to be explored in the understanding of biological visual system as well as effectively designing biologically plausible artificial systems, a readily useful computational solution is to mine human data and "learn" from them in deciding where people look at in a scene.

By characterizing the underlying distributions, recognizing complex patterns, and making intelligent decisions, machine learning provides one of the most powerful sources of insight into machine intelligence. The understanding of saliency and visual attention draws inspirations from learning outcomes from the biological data. Using such an approach, Zhao and Koch [88] show that observers weight different early visual features differently when deciding where to look. Further, feature integration is nonlinear [89,90]. This raises the question of the extent to which the primate brain takes advantages of such nonlinear integration strategies. Future psychophysical and neurophysiological research are needed to untangle this question. Another advantage of learning is that it provides a unified framework for analyzing data and making comparisons under different conditions (e.g., with different populations, or with different tasks). For example, in the same framework [88], data from all participating subjects are used to infer group properties, while individual data are used to derive individual characteristics.

There are several challenges particular to learning saliency-based visual attention using supervised machine learning techniques:

(a) Obtaining ground truth is labor intensive: as for many supervised learning applications, obtaining ground truth data is essential yet usually requires a very large effort. Examples of such image databases are LableMe [63], ImageNet [12], and Amazon Mechanic Turk. Learning where people look at, however, is less straightforward—eye tracking devices are required to record eye positions when subjects view the visual input, which greatly limits the data collection process. There have been recent advances in gaze and eye modeling and detection (e.g., [24,70]), yet large-scale accurate eye tracking experiments are still difficult without customized eye tracking hardwares. As will be introduced in Section 4, the sizes of the current datasets are at the order of hundreds images and tens subjects, much smaller than those for object detection, categorization, or scene understanding.

(b) Laboratory experimental setup is constrained: under standard experimental conditions, a strong central bias is seen, that is, photorgaphers and subjects tend to look at the center of the image. This is largely due to the experimental setup [72,87,86,36] and the feature distributions of the image sets [62,57,73,13,36], as will be detailed in Section 3. In order to effectively use the data collected in laboratory settings, compensations for the spatial bias need to be incorporated. An alternative is to conduct unrestrained eye tracking experiments with full-field-of-view (e.g., while subjects are walking) and collect data where limitations of laboratory settings are avoided [65,11,74].

(c) The problem is loosely defined: unlike typical computer vision tasks such as image segmentation or object detection where the goal is clearly specified, for predicting where people look at, the paradigm is more ambiguous. Some studies [33,57,59,35,25,83,84] focus on stimulus-dependent factors while others [51,36,10] argue that task and subject-dependent influences are no less important. Further, although it is widely accepted that saliency depends on context, the unit of information that is selected by attention – be regular shaped regions [33,38,88–90], or proto-objects [50], or objects [13] – is still a controversial topic in the neuroscience community. Open questions relating to this problem tend to lead to a mixture of findings in this literature [74]. Thus, it depends upon the readers to identify relevant design assumptions and paradigms. For example, if a model is to be built for predicting gazes in free-viewing, then data collected for different tasks may not be applicable.

In the following sections, we will be describing recent advances in learning saliency-based visual attention: Section 2 reviews related methods in saliency detection; Section 3 discusses methods to compensate the spatial bias induced in laboratory settings; Section 4 introduces public eye tracking datasets; and Section 5 concludes the paper.

2. Learning saliency-based visual attention

The problem of saliency learning is formulated as a classification problem [38,36,88–90]. Formally, a mapping function $G(f): R^d \rightarrow R$ (d is the dimension of the feature vector) is trained using learning algorithms to map a high dimensional feature vector to a scalar saliency value. To train the mapping functions, positive and negative samples are extracted from training images. Particularly, a positive sample comprises a feature vector at fixated locations and a label of "1", while a negative sample is a feature vector at nonfixated (or background) regions

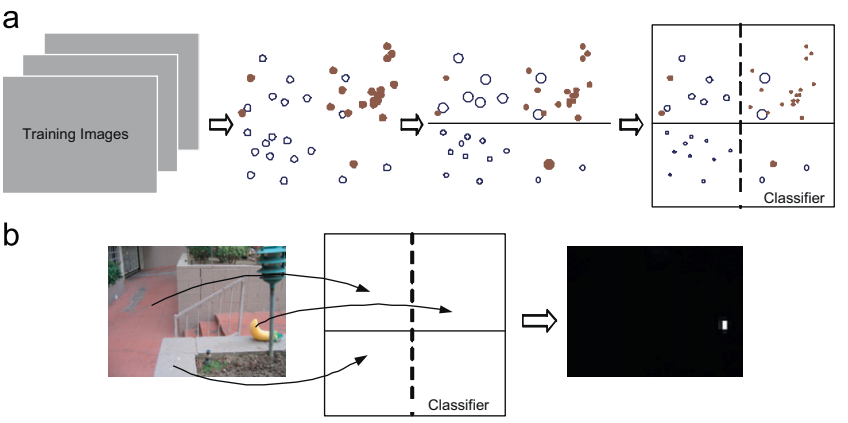

Fig. 1. Illustration of learning saliency-based attention. (a) Training stage: a saliency predictor (classifier) is trained using samples from training images in which observers fixated within the scene. The dimension of the feature vector of each sample is usually much higher than 2. We use 2 here for pedagogical purposes. (b) Testing stage: for a new image, the feature vectors of image locations are calculated and provided to the trained classifier to obtain saliency scores. The rightmost map is the output of the classifier, where brighter regions denote more salient areas.

together with a label of "-1". A typical saliency learning algorithm including a training stage and a testing stage as illustrated in Fig. 1.

2.1. Feature representation in learning saliency-based attention

To represent positive and negative samples, Kienzle et al. [38] directly cut out a square image patch at each fixated location and concatenate the raw pixel values inside the patch to form a feature vector. Determining the size and resolution of the patches is not straightforward and compromises have to be made between computational tractability and generality: in their implementation, the resolution is fixed to 13×13 pixels, leading to a 169 dimensional feature. The high dimensional feature vector of an image patch requires a large number of training samples. Experimental results [38] show a comparable performance with the conventional saliency model of Itti et al. [33], although no prior is used in Kienzle et al. [38].

Given the very large input vectors if using raw image data, an alternative is to perform a feature extraction step before learning. Now, positive samples correspond to the extracted features at a fixated location and negative samples to extracted features at nonfixated locations. For example, Judd et al. [36] use "low-, middle- and high-level" image features, where "low-level" features include a mixed of hand-crafted or biologically inspired image features such as local energy of the steerable pyramid filters [67], RGB values, and so on; "mid-level" feature aims to capture scene context, which is a horizon line detector from mid-level gist features [53]; and "high-level" features are semantic ones including a face detector [79] and a pedestrian detector [16]. In Zhao and Koch [88–90], simple and biologically plausible features [33,57,9] are extracted for learning. Although different from early visual features such as color, intensity, and orientation, objects such as faces attract gaze in an automatic and task-independent manner [9], and including them fills some of the gaps between the predictive powers of the current saliency models. The research and engineering efforts in the computer vision community concerned with successful face detection algorithms [79] have made this incorporation feasible in computational saliency models. Generally, progress in identifying and incorporating such features for saliency requires efforts from both the human vision and computer vision communities. Further neurophysiological or psychophysical experiments are needed to justify features that are in the stimulus-dependent pathway while building saliency models including an automatic detection of such features requires consistent efforts from computational experts.

With extracted features as priors, feature dimensions are substantially reduced [36,88–90] compared with training on raw image data, and better performance is achieved by learning in the lower-dimensional feature spaces. The tradeoff is design efforts from experts, and weak features tend to lead to unsuccessful saliency models. To approach this dilemma, Zhao and Koch [89,90] propose an AdaBoost based framework for saliency learning which automatically selects from a feature pool the most informative features that nevertheless have significant variety. The framework allows an easy incorporation of any candidate features and a natural selection of the best ones in a principled manner. Given the abundant studies on low-level image-based features, future explorations on good higher-level features to fill the semantic gaps of model predictions and human behaviors would be important.

2.2. Learning saliency-based attention

With training samples (i.e., feature vectors and labels), the saliency predictor $G(f)$ can be learned using machine learning techniques.

Ideally any design parameters relating to features, inferences, and integrations (as described in Section 1) can be learned from human data, yet the availability of reliable ground truth data and the computational power of existing learning algorithms impose practical limits on the learning process. Kienzle et al. [38] aim to learn a

completely parameter-free model directly from raw data using support vector machine (SVM) [5] with Gaussian radial basis functions (RBF). Unfortunately, the high dimensional vector concatenated from raw image patch raises a high demand on the sample numbers. Further, even if future efforts make the data collection procedure easier and more samples accessible, the scaling issues and computational bottlenecks may still prohibit the learning of all parameters.

Different priors are thus used to make saliency learning computationally tractable. Besides feature extraction for dimension reduction [36,88–90], Zhao and Koch [88] use the linear integration assumption as a prior and learn saliency using constrained linear regression. The simple structure makes the learned outcomes applicable to numerous studies in psychophysics and physiology and leads to an extremely easy implementation for real-world applications. Similarly, using a set of pre-defined features, Judd et al. [36] learn the saliency model with liblinear SVM [15] which is used to achieve performance no worse than models with RBF kernels as proposed by Kienzle et al. [38]. Zhao and Koch [89,90] propose an AdaBoost [19,64,20,77] based model to approach feature selection, thresholding, weight assignment, and integration in a principled, nonlinear learning framework. The AdaBoost based method combines a series of base classifiers to model the complex input data. With an ensemble of sequentially learned models, each base model covers a different aspect of the dataset [34]. In some of the methods [36,88–90], parameters of the spatial prior (as will be discussed in detail in the next section) are also directly learned from data and integrated into the models to compensate the bias shown in human data.

Alternative approaches employ learning based saliency models based on objects rather than image features [37,45]. To make the object detection step robust and consistent, pixel neighborhood information is included. Thus, Khuwuthyakorn et al. [37] extend generic image descriptors of Itti et al. [33] and Liu et al. [45] to a neighborhood-based descriptor setting by considering the interaction of image pixels with neighboring pixels. In other efforts, conditional random field (CRF) [41] that encodes interaction of neighboring pixels effectively detects salient objects in images [45] and videos [46], although CRF learning and inference are quite slow.

3. Central fixation bias

Under standard testing conditions, a strong central bias is seen, that is, subjects tend to look at the center of the image.

Several explanations for this phenomenon have been suggested. Some attribute the center bias to the drop in visual system sensitivity in the periphery [57,59] and to a motor bias in the saccadic system that favors small amplitude saccades over large amplitude ones [3,58,22]. These two factors, combined with the fact that scene viewing experiments typically start in the center, result in a central fixation bias. The experimental setup (users are placed centrally in front of the screen; [72,87,86,36]) and the bias toward centering the eyeball within its orbit reinforce the tendency to look toward the center [85,21,56,72]. However, Vitu et al. [80] demonstrate that it is the screen center rather than the straight-head position – the orbital center – that produces the central fixation tendency. Many [62,57,73,13,36] assume that the bias arises from a central bias of image feature distributions. As human photographers place objects of interest in the center, it is not surprising that subjects will look at such centrally placed objects. Lastly, Le Meur et al. [43] and Tatler [72] suggest that the center of the scene offers strategic advantages—it is an optimal location for extracting information from the scene and a convenient location for the efficient exploration of the scene.

One way to better account for the center bias is not to use nonfixated locations as negative samples but to use the location of fixated location from randomly shuffled image locations [73,7,87]. This ensures that both categories have identical spatial distributions and therefore the same spatial bias. Hence, the saliency difference cannot arise from any spatial difference (i.e., the positive samples are closer to the center [73]). Of course, as pointed out by Tatler et al. [73] and later by Carmi and Itti [7], in cases when central image feature distributions is one cause for central fixation bias, which is true for most static images, using the above method would underestimate the magnitude of saliency effects.

Another popular remedy [57,59,36] to compensate this bias is to use a single Gaussian or exponential spatial filter. The Gaussian/exponential type prior is ad hoc but effective. Recently, Zhao and Koch [88] present a computational model that takes into account different causes of center bias including both time-varying and constant factors, and for the first time models the center bias as a dynamic process—the model considers the possibility that the center bias may be stronger early on and then diminish over time (or vice versa). It has been shown that the saccade sequence follows a Gaussian process and that the distribution of fixations is a mixture of Gaussians. Furthermore, by proving the convergence of the Gaussian covariance matrix, approximating this time-varying process via a single kernel is justified. The Gaussian parameters are learned from human data.

4. Public eye tracking datasets

With the growing interests in the neuroscience as well as computer science communities to understand how humans and other animals interact with visual scenes and to build artificial visual models, in recent years, several eye tracking datasets have been constructed and made publicly available to facilitate vision research.

An eye tracking dataset includes natural images (or videos) as the visual stimuli and eye movement data recorded using eye tracking devices when human subjects view the stimuli. A typical image set is at the order of hundreds or a thousand of images. Different from the conventional laboratory psychophysics/eye tracking experiments based on synthetic stimuli, natural stimuli reflect realistic visual input and offer a better platform for the study of vision and cognition. On the other hand, the natural stimuli are less controlled thus requiring more

sophisticated computational techniques for analysis. Usually tens of subjects are asked to view the stimuli while locations of their eyes in the image coordinates are recorded. In some datasets, Matlab codes are also available for basic operations such as calculating fixations, visualizing eye traces, and so on.

In learning saliency-based attention, a dataset is divided into training sets and testing sets, where the former is used to train the classifier while the latter for performance assessment. In the following we briefly list several examples of public datasets—five sets with static scenes (images) and one with dynamic scenes (videos):

In the FIFA dataset [9], fixation data are collected from eight subjects performing a 2-s-long free-viewing task on 180 color natural images ($28^\circ \times 21^\circ$). They are asked to rate, on a scale of 1–10, how interesting each image is. Scenes are indoor and outdoor still images in color. Most of the images include faces in various skin colors, age groups, gender, positions, and sizes.

The second dataset from [4] (referred here as the Toronto database) contains data from 11 subjects viewing 120 color images of outdoor and indoor scenes. Participants are given no particular instructions except to observe the images ($32^\circ \times 24^\circ$), 4 s each. One distinction between this dataset and that of the FIFA [9] is that a large portion of images here do not contain particular regions of interest, while in the FIFA dataset most contain very salient regions (e.g., faces or salient nonface objects).

The eye tracking dataset from Judd et al. [36] (referred to as MIT database) includes 1003 images collected from Flickr and LabelMe. The image set is considered general due to its relatively large size and the generality of the image source. Eye movement data are recorded from 15 users who free-view these images ($36^\circ \times 27^\circ$) for 3 s. A memory test motivates subjects to pay attention to the images: they look at 100 images and need to indicate which ones they have seen before.

The NUS database is recently published by Subramanian et al. [71]. A big feature of this dataset compared with others is that the 758 images in the dataset contains a large number of semantically affective objects/scenes such as expressive faces, nudes, unpleasant concepts, and interactive actions, thus providing a good source to study social and emotion related topics. Images are from Flickr, Photo.net, Google, and emotion-evoking IAPS [42]. In total, 75 subjects free-view ($26^\circ \times 19^\circ$) part of the image set for 5 s each (each image is viewed by an average of 25 subjects).

The last image dataset listed here is the DOVES dataset [76]. It includes 101 natural grayscale images that are selected from the Natural Stimuli Collection created by Hans van Hateren, and cropped at 1024×768 pixels. Eye movements from 29 human observers as they free-view the images ($17^\circ \times 13^\circ$) are collected. To discourage observers from fixating at only one location, and to ensure a somewhat similar cognitive state across observers, a simple memory task is also used: following the display of each image, observers are shown a small image patch (about $1^\circ \times 1^\circ$) and asked to indicate whether the image patch is from the image they just viewed.

Compared with eye tracking datasets with static scenes, there are much fewer resources on dynamic scenes: the Itti dataset consists of a body of 520 human eye tracking data traces obtained while normal, young adult human volunteers freely watch complex video stimuli (TV programs, outdoors videos, video games). It comprises eye movement recordings from eight distinct subjects watching 50 different video clips ($\sim$25 min of total playtime; [29,30]), and from another eight subjects watching the same set of video clips after scrambling them into randomly re-ordered sets of 1–3 s clippets [7,8].

Besides being valuable recourses for saliency research, the public datasets allow a fair comparison of different computational models. For example the Toronto dataset has been used as a benchmark for several recent saliency algorithms (e.g., [23,87,26]), and we expect that more comparative works would come out using other datasets as well. With the different nature and size of the datasets, researchers can either select specific ones to study particular problems (e.g., using the FIFA dataset to study face and the NUS dataset for emotion related topics), or make a comprehensive comparisons on all the datasets for general issues that should not vary across datasets. For example, Zhao and Koch [88–90] have tested their learning algorithms on the datasets with color images—the FIFA, Toronto, MIT, and NUS datasets, and show consistent conclusions: for the four feature channels of interest, face is the most important, followed by orientation, color, and intensity.

The current public datasets all record "free-viewing" eye movements. Thus, they may not generalize to other tasks, such as search tasks. However, temporal changes in viewing strategies can be exploited using the data which typically record eye locations for several seconds. For example, Zhao and Koch [88] show that saliency decreases with time, consistent with the findings [48] that initial fixations are purely driven by stimulus-dependent saliency (such as feature contrast) compared to later ones. Further, by making comparisons over time [88], it is found that face attracts attention faster than other visual features.

5. Summary

This paper reviews several issues relating to learning saliency-based attention. Unlike the conventional structure of computational saliency modeling that relies heavily on assumptions and parameters to build the models, learning based methods apply modern machine learning techniques to analyze eye movement data and derive conclusions. Saliency predictors (classifiers) are directly trained from human data and free domain experts from efforts in designing the model structure and parameters that are often ad hoc to some extent. As an important component in these data-driven approaches, a steady progress is being made on data collection and sharing in the community. Access to large datasets and use of standard similarity measures allow an objective evaluation and comparison of saliency models. Lastly we expect a tighter coupling of machine learning and saliency detection, where domain-specific learning techniques are developed to better utilize the limited, and sometimes noisy human data to predict where people look at.

References

[1] T. Avraham, M. Lindenbaum, Esaliency (extended saliency): meaningful attention using stochastic image modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 99 (1) (2009) 693–708.

[2] R. Baddeley, B. Tatler, High frequency edges (but not contrast) predict where we fixate: a Bayesian system identification analysis, Vision Research 46 (18) (2006) 2824–2833.

[3] A. Bahill, D. Adler, L. Stark, Most naturally occurring human saccades have magnitudes of 15 degrees or less, Investigative Ophthalmology & Visual Science 14 (6) (1975) 468–469.

[4] N. Bruce, J. Tsotsos, Saliency, attention, and visual search: an information theoretic approach, Journal of Vision 9 (3) (2009) 1–24.

[5] C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (2) (1998) 121–167.

[6] A. Carbone, F. Pirri, Learning saliency. An ica based model using bernoulli mixtures, in: Proceedings of Brain Inspired Cognitive Systems, 2010.

[7] R. Carmi, L. Itti, The role of memory in guiding attention during natural vision, Journal of Vision 6 (9) (2006) 898–914.

[8] R. Carmi, L. Itti, Visual causes versus correlates of attentional selection in dynamic scenes, Vision Research 46 (26) (2006) 4333–4345.

[9] M. Cerf, E. Frady, C. Koch, Faces and text attract gaze independent of the task: experimental data and computer model, Journal of Vision 9 (12) (2009). 10:1–15.

[10] S. Chikkerur, T. Serre, C. Tan, T. Poggio, What and where: a Bayesian inference theory of attention, Vision Research 50 (22) (2010) 2233–2247.

[11] F. Cristino, R. Baddeley, The nature of the visual representations involved in eye movements when walking down the street, Visual Cognition 17 (6) (2009) 880–903.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[13] W. Einhauser, M. Spain, P. Perona, Objects predict fixations better than early saliency, Journal of Vision 8 (14) (2008). 18:1–26.

[14] S. Engmann, B. 't Hart, T. Sieren, S. Onat, P. König, W. Einhäuser, Saliency on a natural scene background: effects of color and luminance contrast add linearly, Attention, Perception, & Psychophysics 71 (6) (2009) 1337–1352.

[15] R. Fan, K. Chang, C. Hsieh, X. Wang, C. Lin, Liblinear: a library for large linear classification, Journal of Machine Learning Research 9 (2008) 1871–1874.

[16] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[17] D. Field, What is the goal of sensory coding? Neural Computation 6 (1994) 559–601.

[18] T. Foulsham, G. Underwood, What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition, Journal of Vision 8 (2) (2008) 601–617.

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

[20] J. Friedman, T. Hastle, R. Tibshirani, Additive logistic regression: a statistical view of boosting, Annals of Statistics 38 (2) (2000) 337–374.

[21] J. Fuller, Eye position and target amplitude effects on human visual saccadic latencies, Experimental Brain Research 109 (3) (1996) 457–466.

[22] D. Gajewski, A. Pearson, M. Mack, F. Bartlett, J. Henderson, Human gaze control in real world search, in: L. Paletta, J. Tsotsos, E. Rome, G. Humphreys (Eds.), Attention and Performance in Computational Vision, Springer-Verlag, New York, 2005., pp. 3368:83–99.

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

[24] D. Hansen, Q. Ji, In the eye of the beholder: a survey of models for eyes and gaze, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (3) (2010) 478–500.

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

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

[27] Y. Hu, X. Xie, W. Ma, L. Chia, D. Rajan, Salient region detection using weighted feature maps based on the human visual attention model, in: IEEE Pacific-Rim Conference on Multimedia, 2004, pp. 993–1000.

[28] A. Hyvarinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (4–5) (2000) 411–430.

[29] L. Itti, Automatic foveation for video compression using a neurobiological model of visual attention, IEEE Transactions on Image Processing 13 (10) (2004) 1304–1318.

[30] L. Itti, Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes, Visual Cognition 12 (6) (2005) 1093–1123.

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

[32] L. Itti, C. Koch, Comparison of feature combination strategies for saliency-based visual attention systems. In: Proceedings of SPIE Human Vision and Electronic Imaging, 1999, pp. 3644:473–82.

[33] L. Itti, C. Koch, E. Niebur, A model for saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1254–1259.

[34] R. Jin, Y. Liu, L. Si, J. Carbonell, H. A.G., A new boosting algorithm using input-dependent regularizer, in: International Conference on Machine Learning, 2003.

[35] T. Jost, N. Ouerhani, R. von Wartburg, R. Muri, H. Hugli, Assessing the contribution of color in visual attention, Computer Vision and Image Understanding 100 (1–2) (2005) 107–123.

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

[37] P. Khuwuthyakorn, A. Robles-Kelly, J. Zhou, Object of interest detection by saliency learning, in: European Conference on Computer Vision, 2010, pp. 6312:636–649.

[38] W. Kienzle, F. Wichmann, B. Scholkopf, M. Franz, A nonparametric approach to bottom-up visual saliency, in: Advances in Neural Information Processing Systems, 2006, pp. 689–696.

[39] A. Koene, L. Zhaoping, Feature-specific interactions in salience from combined feature contrasts: evidence for a bottom-up saliency map in v1, Journal of Vision 7 (7) (2007). 6:1–14.

[40] G. Krieger, I. Rentschler, G. Hauske, K. Schill, C. Zetzsche, Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics, Spatial Vision 13 (2–3) (2000) 201–214.

[41] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: International Conference on Machine Learning, 2001, pp. 282–289.

[42] P. Lang, M. Bradley, B. Cuthbert, (iaps): Affective Ratings of Pictures and Instruction Manual, Technical Report, University of Florida, 2008.

[43] O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, A coherent computational approach to model the bottom-up visual attention, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (5) (2006) 802–817.

[44] Z. Li, A saliency map in primary visual cortex, Trends in Cognitive Sciences 6 (1) (2002) 9–16.

[45] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H. Shum, Learning to detect a salient object, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (2) (2011) 353–367.

[46] T. Liu, N. Zheng, W. Ding, Z. Yuan, Video attention: learning to detect a salient object sequence, in: IEEE Conference on Pattern Recognition, 2008, pp. 1–4.

[47] V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in highly dynamic scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (1) (2010) 171–177.

[48] S. Mannan, C. Kennard, M. Husain, The role of visual salience in directing eye movements in visual object agnosia, Current Biology 19 (6) (2009) 247–248.

[49] C. Masciocchi, S. Mihalas, D. Parkhurst, E. Niebur, Everyone knows what is interesting: salient locations which should be fixated, Journal of Vision 9 (11) (2009). 25:1–22.

[50] S. Mihalas, Y. Dong, R. von der Heydt, E. Niebur, Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects, Proceedings of the National Academy of Sciences 108 (18) (2011) 75–83.

[51] V. Navalpakkam, L. Itti, Modeling the influence of task on attention, Vision Research 45 (2) (2005) 205–231.

[52] H. Nothdurft, Salience from feature contrast: additivity across dimensions, Vision Research 40 (10–12) (2000) 1183–1201.
[53] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, International Journal of Computer Vision 42 (2001) 145–175.
[54] A. Oliva, A. Torralba, M. Castelhano, J. Henderson, Top-down control of visual attention in object detection, in: International Conference on Image Processing, 2003, pp. I:253–256.
[55] S. Onat, K. Libertus, P. König, Integrating audiovisual information for the control of overt attention, Journal of Vision 7 (10) (2007). 11:1-6.
[56] M. Pare, D. Munoz, Expression of a recentering bias in saccade regulation by superior colliculus neurons, Experimental Brain Research 137 (3–4) (2001) 354–368.
[57] D. Parkhurst, K. Law, E. Niebur, Modeling the role of salience in the allocation of overt visual attention, Vision Research 42 (1) (2002) 107–123.
[58] J. Pelz, R. Canosa, Oculomotor behavior and perceptual strategies in complex tasks, Vision Research 41 (25–26) (2001) 3587–3596.
[59] R. Peters, A. Iyer, L. Itti, C. Koch, Components of bottom-up gaze allocation in natural images, Vision Research 45 (18) (2005) 2397–2416.
[60] C. Privitera, L. Stark, Algorithms for defining visual regions-of-interest: comparison with eye fixations, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (9) (2000) 970–982.
[61] R. Raj, W. Geisler, R. Frazor, A. Bovik, Contrast statistics for foveated visual systems: fixation selection by minimizing contrast entropy, Journal of the Optical Society of America A 22 (10) (2005) 2039–2049.
[62] P. Reinagel, A. Zador, Natural scene statistics at the center of gaze, Network: Computation in Neural Systems 10 (4) (1999) 341–350.
[63] B. Russell, A. Torralba, K. Murphy, W. Freeman, LabelMe: a database and web-based tool for image annotation, International Journal of Computer Vision 77 (1–3) (2008) 157–173.
[64] R. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, Machine Learning 37 (3) (1999) 297–336.
[65] F. Schumann, W. Einhauser, J. Vockeroth, K. Bartl, E. Schneider, P. Konig, Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments, Journal of Vision 8 (14) (2008). 12:1–17.
[66] H. Seo, P. Milanfar, Static and space–time visual saliency detection by self-resemblance, Journal of Vision 9 (12) (2009). 15:1–27.
[67] E. Simoncelli, W. Freeman, The steerable pyramid: a flexible architecture for multi-scale derivative computation. in: International Conference on Image Processing, 1995, pp. III:444–447.
[68] M. Song, D. Tao, C. Chen, J. Bu, J. Luo, C. Zhang, Probabilistic exposure fusion, IEEE Transactions on Image Processing 21 (1) (2012) 341–357.
[69] M. Song, D. Tao, C. Chen, X. Li, C. Chen, Colour to grey: visual cue preservation, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1537–1552.
[70] M. Song, D. Tao, Z. Sun, X. Li, Visual context boosting for eye detection, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 40 (6) (2010) 1460–1467.
[71] R. Subramanian, H. Katti, N. Sebe, M. Kankanhalli, T. Chua, An eye fixation database for saliency detection in images, in: European Conference on Computer Vision, 2010, pp. 6314:30–43.
[72] B. Tatler, The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions, Journal of Vision 7 (14) (2007) 1–17.
[73] B. Tatler, R. Baddeley, I. Gilchrist, Visual correlates of fixation selection: effects of scale and time, Vision Research 45 (5) (2005) 643–659.
[74] B. Tatler, M. Hayhoe, M. Land, D. Ballard, Eye guidance in natural vision: reinterpreting salience, Journal of Vision 11 (5) (2011). 5:1–23.
[75] A. Treisman, G. Gelade, A feature-integration theory of attention, Cognitive Psychology 12 (1) (1980) 97–136.
[76] I. van der Linde, U. Rajashekar, A. Bovik, L. Cormack, Doves: a database of visual eye movements, Spatial Vision 22 (2) (2009) 161–177.
[77] A. Vezhnevets, V. Vezhnevets, Modest adaboost—teaching adaboost to generalize better, in: Graphicon, 2005.
[78] B. Vincent, R. Baddeley, T. Troscianko, I. Gilchrist, Optimal feature integration in visual search, Journal of Vision 9 (5) (2009). 15:1–11.
[79] P. Viola, 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.
[80] F. Vitu, Z. Kapoula, D. Lancelin, F. Lavigne, Eye movements in reading isolated words: evidence for strong biases towards the center of the screen, Vision Research 44 (3) (2004) 321–338.
[81] D. Walther, T. Serre, T. Poggio, C. Koch, Modeling feature sharing between object detection and top-down attention, Journal of Vision 5 (8) (2005) 1041.
[82] W. Wang, Y. Wang, Q. Huang, W. Gao, Measuring visual saliency by site entropy rate, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2368–2375.
[83] J. Xu, Z. Yang, J. Tsien, Emergence of visual saliency from natural scenes via context-mediated probability distributions coding, PLoS ONE 5 (12) (2010) e15796.
[84] V. Yanulevskaya, J. Marsman, F. Cornelissen, J. Geusebroek, An image statistics-based model for fixation prediction, Cognitive Computation 3 (1) (2010) 94–104.
[85] D. Zambarbieri, G. Beltrami, M. Versino, Saccade latency toward auditory targets depends on the relative position of the sound source with respect to the eyes, Vision Research 35 (23–24) (1995) 3305–3312.
[86] L. Zhang, M. Tong, G. Cottrell, Sunday: saliency using natural statistics for dynamic analysis of scenes, in: Proceedings of the 31st Annual Cognitive Science Conference, 2009, pp. 2944–2949.
[87] L. Zhang, M. Tong, T. Marks, H. Shan, G. Cottrell, Sun: a Bayesian framework for saliency using natural statistics, Journal of Vision 8 (7) (2008) 1–20.
[88] Q. Zhao, C. Koch, Learning a saliency map using fixated locations in natural scenes, Journal of Vision 11 (3) (2011). 9:1–15.
[89] Q. Zhao, C. Koch, Learning visual saliency, in: Conference on Information Sciences and Systems, 2011, pp. 1–6.
[90] Q. Zhao, C. Koch, Learning visual saliency by combining feature maps in a nonlinear manner using Adaboost, Journal of Vision 12 (6) (2012). 22:1–15.