Computational Visual Attention Models

Milind S. Gide
Arizona State University
mgide@asu.edu

Lina J. Karam
Arizona State University
karam@asu.edu

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Signal Processing

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

M. S. Gide and L. J. Karam. *Computational Visual Attention Models*. Foundations and Trends® in Signal Processing, vol. 10, no. 4, pp. 347–427, 2016.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-280-8

Foundations and Trends® in Signal Processing
Volume 10, Issue 4, 2016

Editorial Scope

Topics

Foundations and Trends® in Signal Processing publishes survey and tutorial articles in the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multirate signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing

Information for Librarians

Foundations and Trends® in Signal Processing, 2016, Volume 10, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.

Foundations and Trends® in Signal Processing
Vol. 10, No. 4 (2016) 347–427

DOI: 10.1561/2000000055

Computational Visual Attention Models

Milind S. Gide
Arizona State University
mgide@asu.edu

Lina J. Karam
Arizona State University
karam@asu.edu

Contents

Abstract

The human visual system (HVS) has evolved to have the ability to selectively focus on the most relevant parts of a visual scene. This mechanism, referred to as visual attention (VA), has been the focus of several neurological and psychological studies in the past few decades. These studies have inspired several computational VA models which have been successfully applied to problems in computer vision and robotics. In this paper we provide a comprehensive survey of the state-of-the-art in computational VA modeling with a special focus on the latest trends. We review several models published since 2012. We also discuss theoretical advantages and disadvantages of each approach. In addition, we describe existing methodologies to evaluate computational models through the use of eye-tracking data along with the VA performance metrics used. We also discuss shortcomings in existing approaches and describe approaches to overcome these shortcomings. A recent subjective evaluation for benchmarking existing VA metrics is also presented and open problems in VA are discussed.

M. S. Gide and L. J. Karam *Computational Visual Attention Models*. Foundations and Trends® in Signal Processing, vol. 10, no. 4, pp. 347–427, 2016.
DOI: 10.1561/2000000055.

1

Introduction

1.1 What is visual attention?

> Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought.
>
> William James, 1890.

Every time we open our eyes, the human visual system (HVS) is bombarded with vast amounts of visual information. It is estimated that this information is in the order of 10^9 bits per second [35, 48]. This information is so vast that the neuronal "hardware" in our brain (specifically the visual cortex) is not capable of processing it all at once. As a result, our brains have evolved certain mechanisms that allow us to selectively process relevant portions of the information by using the available limited resources. The broad area of research that involves the study of the neuro-physiological underpinnings and computational modeling of these mechanisms is known as visual attention (VA).

Figure 1.1 illustrates how humans employ the VA mechanism in real life. Eye-tracking maps obtained over 15 human observers while

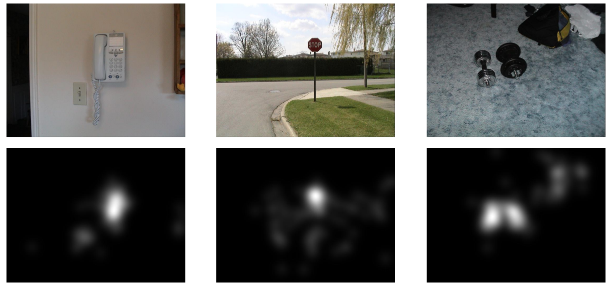

Figure 1.1: Demonstration of VA while viewing stationary scenes. On casual viewing of these images most humans focus on selective regions in them like the telephone, the stop sign and the weights. Images and eye-tracking data have been taken from the Toronto dataset [5].

viewing the stimulus images show that humans tend to notice only the main objects of interest, i.e., the phone, the stop sign and the weights in these images, and tend to ignore the background when the images are viewed casually as we would a photo-album. In the eye-tracking maps a brighter pixel value denotes a higher probability of the corresponding pixel being fixated by humans.

1.2 Different aspects of VA

VA has been an active topic of research over the past few decades and researchers from diverse scientific backgrounds such as psychology, physiology and neuroscience have expounded different theories to explain different aspects of VA such as pre-attentive and attentive VA, bottom-up and top-down components of VA, serial and parallel search in VA, overt and covert VA and so on. Although these different aspects of VA have some overlap, it is useful to delve into them to understand VA in more details.

1.2.1 Preattentive and attentive stages of visual attention

Neisser and Hoffman [53, 24] proposed a theory in which the VA process is looked at from a signal processing view point, and is divided into two stages. The first stage is a pre-attentive stage in which basic features like color, orientation, edge-information, or motion are extracted from the scene. This extraction of features occurs over the entire scene independent of attention. This theory is based on the fact that, in the primary visual cortex, there are several simple cells that extract these features based on their receptive fields by applying different filters on the input stimulus. The pre-attentive stage consists of high speed and parallelized operations that are involuntary in nature. Once these pre-attentive features are extracted, they are integrated by an attention stage that identifies the regions with the most "relevant" information and fixates on these regions to observe them in greater detail. Due to the integration, different features in the pre-attentive stage may be bound together or the dominant features may be selected. The pre-attentive stage and the attention stage are also referred to as "vision before attention" and "vision with attention" attention stage called vision with attention in some works such as [81], respectively.

1.2.2 Bottom-up and top-down mechanisms of visual attention

A number of experiments [81, 10] conducted in the past few decades point to a two-component framework for explaining how attention is deployed. According to this framework, VA mechanisms can be separated into bottom-up and top-down components that are inter-related but conceptually complementary to each other. Bottom-up attention usually occurs in the pre-attention stage and is a result of simple center-surround operations on basic features extracted in the pre-attentive stage like color, orientation, motion, etc. The bottom-up component of VA is attracted to visually conspicuous areas in the scene automatically irrespective of task and hence is also known as the stimulus-driven attention component. The bottom-up component is a very fast, almost instantaneous component, as it is handled by early vision regions in the primary visual cortex that operate in parallel. On the other hand, the top-down component is highly dependent on the task at hand as well as the mental state and prior experiences of the observer. In a famous experiment conducted by Yarbus [83], a complex scene of people in a family room was shown to several human observers and they were either asked no questions or were asked questions of a varying nature like estimating the age of the people in the scene, or remembering the position of certain objects in the scene. The results showed that the eye tracking data of the observers varied significantly depending on whether a question was asked and if a question was asked, also on the type of question. For example, when the observers were asked to estimate the age of the people in the scene, most eye-movements were located on the faces. When they were asked to remember the position of an object, the fixations were located on or near the objects. As the top-down component depends on the task in question, it is also called the task-driven component. The top-down component is believed to be processed in the higher visual cortex and is a much slower component than the bottom-up component. In general, the top-down component is not totally independent of the bottom-up component, and the VA mechanism is considered to be the result of an interplay of both these components.

1.2.3 Parallel vs serial processing in VA

The vast network of interconnected neurons in the human brain allows visual information incident on the retina to be processed in parallel. This is true specially in certain areas of the primary visual cortex that are part of the pre-attentive processing described earlier. However, the shifts in gaze that are guided by attention which helps humans focus on different objects in a complex scene, take place serially. Triesman and Gelade [74] constructed certain psychovisual examples of serial and parallel processing similar to those seen in Figure 1.2. The results showed that when the target differs from the distractors in a single feature, it is identified instantaneously through a parallel search mechanism as seen in Figure 1.2(a) where the target differs from the distractors in only the color dimension. Also, in this case, the speed with which the target is identified does not change with an increasing number of distractors. On the other hand, when the target differs from the distractors in more than one feature, the search is serial and the time taken to find the target is much more and increases with an increase in the number of distractors. This is seen in Figure 1.2(b) where the

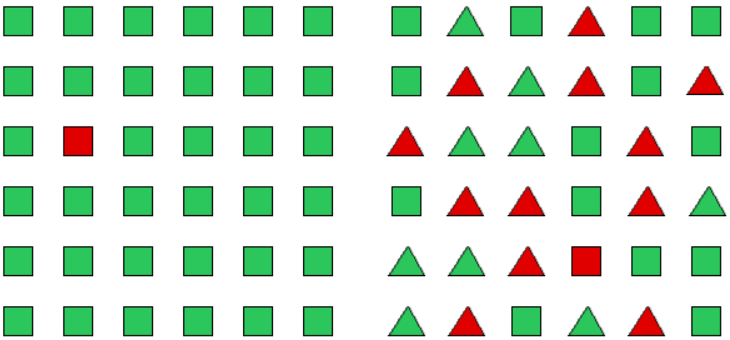
(a) Parallel Search. (b) Serial Search.

Figure 1.2: Illustration of parallel vs serial search.

target differs from the distractors in both color and shape. As a result, for real-world complex scenes, the search for the target is mostly serial in nature.

1.2.4 Overt and covert VA

The human visual system (HVS) is constantly seeking relevant information from a visual scene by shifting the gaze from one interesting region to another through a process known as attention shift [37]. As part of this process, the uniqueness of an already fixated upon region weakens and the next interesting region is fixated upon. This shift in gaze involves eye-movements to the next interesting location and is known as overt attention. Most studies in VA use eye-tracking devices to track the eye-movements of humans while viewing stimuli images. As a result, most of the computational models are geared towards overt attention.

The HVS also has an ability to attend to regions in a scene without explicit eye-movements. This type of attention is known as covert attention. An example of covert attention is when a driver notices and understands traffic signs without explicitly moving his eyes towards them. Covert attention is an important evolutionary trait that helps humans attend to important changes in the visual environment in the periphery without loosing focus of the current attended object.

1.3 Psychological and physiological theories explaining visual attention

1.3.1 Gestalt principles

Gestalt principles are rules of perceptual organization formulated by a group of researchers in the early 20th century to explain how humans group multiple elements in a complex visual scene. Gestalt is the German word for "shape" or "form". These rules dictate how humans perceive certain objects as individual items whereas in other cases a group of objects with common features are thought of as a single entity. Some of the basic principles that are exploited in computational VA models follow:

- Figure-ground articulation: In the case of a uniform image with no variation, according to the Gestalt principles there is no internal organization. However, in the case of an inhomogenous field with a patch of color surrounded by a different background color as shown in Figure 1.3, the field is considered to be composed of two distinct components, the figure (colored patch) on ground (surrounding background). The difference in figure-ground could be in any other dimension apart from color. The figure is assigned object-like properties and receives more attention, whereas the ground is treated as background and is not considered salient. This leads to the important property of surroundedness of salient objects that is used in VA models like BMS ([84]) as discussed later in Section 2.1.2.

- Proximity: In a scene, objects close to each other are usually grouped together as one single entity. For example in Figure 1.4, in the image to the left, the group of circles is taken to be a single object (a square), whereas in the image to the left, three different "columns" of circles are perceived.

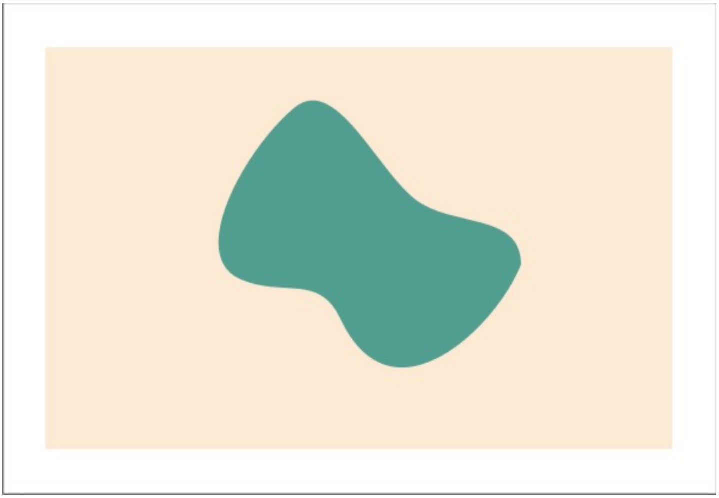

Figure 1.3: Figure-ground articulation.

- Similarity: In a scene, objects similar to each other in some respect are also grouped as one single entity. For example in Figure 1.5, the rows of dark and light circles are considered as different entities even though according to the proximity principle they could be considered as a single square entity.

- Symmetry: According to this principle the HVS has a tendency to be sensitive towards objects that possess symmetry. As a result, two unconnected elements which are symmetric about a certain axis will be perceived as a single object. This is illustrated in Figure 1.6. The image shown is interpreted as three sets of parentheses instead of six different ones. The symmetry principle is applied in the VA model developed by Kootstra [40] that equates saliency of a region to how symmetric it is.

1.3.2 Feature integration theory

The feature integration theory introduced by Triesman and Gelade [74] is based on the notion of pre-attentive vision (Section 1.2.1) in which features are extracted early, involuntarily, and in parallel over the entire

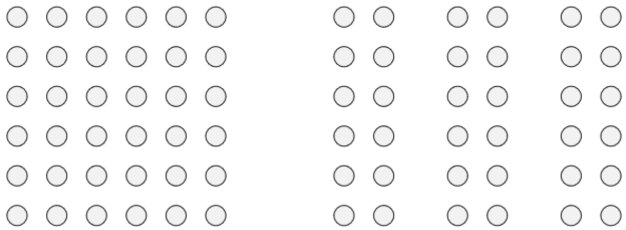

Figure 1.4: Proximity.

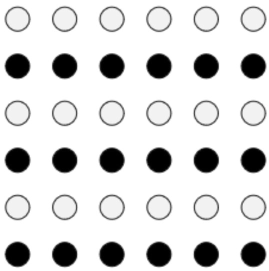

Figure 1.5: Similarity.

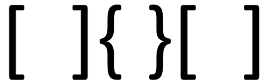

Figure 1.6: Symmetry.

scene, before objects are recognized. The recognition of objects happens at a much later stage and in a separate process that requires focused attention. Basic separable features like color, orientation, spatial frequency, and motion are extracted at the early stage to give feature maps. According to this hypothesis, at this stage, the feature maps float free, in that though they are perceived, they do not contribute to knowledge about location of objects as such. In the attentive stage, these features are combined by stimulus location and features that are present for a specific attentive fixation are combined to form an object, the focal attention providing a glue that binds together the initially independent features. Once the objects have been recognized, they are stored and remembered for some time before memory decay or interference may cause the features to go into a free-floating state again. According to this theory, without focused attention, the features cannot be related to each other and stay independent and separable. The feature maps can be treated as binary maps, which signal the presence or absence of a certain feature. If the presence of a single feature is enough to complete the task (i.e., identify the target from the distractors in the experiments conducted in [74]), the attention stage is not required and the task is completed in parallel and in a rapid manner. However, if the task requires conjunction and relies on more than one feature, the attention stage is called upon and fixated regions are scanned serially to complete the task.

1.3.3 Boolean map theory

A competitive theory to the feature integration theory is the Boolean Map Theory proposed by Huang and Pashler [28]. This theory deals with the aspects of "access" and "selection" in VA. "Access" defines what an observer can visually apprehend in the scene at any given moment whereas "selection" represents the mechanism of VA that control what regions are accessed. A boolean map is considered to be a spatial representation that partitions the visual scene into two distinct regions, a selected region and a non-selected region, based on a single featural label per dimension. A featural label provides an overall featural description of the entire map. For example, in Figure 1.7, for the Boolean map, there could be a label that covers the two shapes but this label does not define the greenness or redness of the objects as that would not cover both the objects. There can be independent featural labels that can comprise a Boolean map that belong to different dimensions; for example, a Boolean map can have redness as a color label and verticalness as an orientation label. A single Boolean map

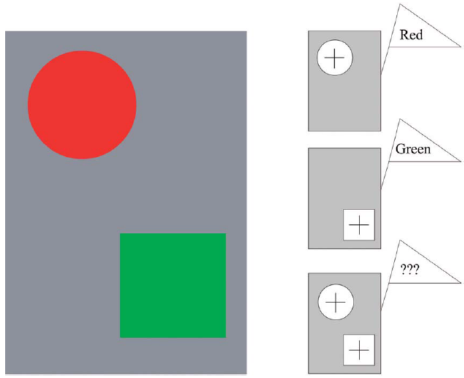

Figure 1.7: Figure illustrating the concept of a boolean map. The three possible boolean maps are (top) map describing the shape and color of the red disk, (middle) map the shape and color of the green-square, (bottom) map describing only the shapes of the two circle and square objects but not their color. Image reproduced from Huang and Pashler [28].

describes the visual awareness of an observer of a scene at any given time instant. For complex scenes, different Boolean maps are combined through operations of intersection and union to direct attention. The boolean map theory is used by the Boolean Map Saliency (BMS) [84] algorithm described in Section 2.1.2.

1.3.4 Computational modeling of VA

Concept of saliency map

Koch and Ulman [37] developed the first biologically inspired VA model that was based on the Feature Integration Theory. In this work, the concept of a saliency map was introduced for the first time. The saliency map is a two dimensional topographic map that denotes the visual conspicuousness of a pixel. The higher the value, the more conspicuous or salient a pixel will be. In Koch and Ulman [37], first low-level features are first extracted in parallel similar to that in the pre-attentive stage to obtain several topographic feature maps. These feature maps are then combined to give a global topographic saliency maps. All the other VA models that have been developed since then use a similar concept of a saliency map.

Main stages in computational modeling

Most computational VA models that are in some way based on the feature integration theory and the concept of a "saliency map" consist of the following stages in their processing pipeline [21]:

1. Feature Extraction

   In this stage, features based on color, orientation, depth, motion and other low-level properties of images are extracted over the entire spatial extent of the image. These feature-extraction operations mimic those performed by the simple cells in the primary visual cortex and usually include some level of multi-resolution analysis in the form of pyramidal decompositions.

2. Feature Activation

   This stage performs the center-surround difference operations that correspond to those performed by the receptive-fields of the neurons in the HVS which helps in identifying regions that "pop-out" from their surroundings.

3. Normalization/Combination

   In this stage, the different activation maps are combined after normalization to give the final saliency map which denotes how salient each pixel in the image is.

1.4 Eye-tracking data and evaluation of VA models

Ideally, the saliency map that is produced by a computational VA model should highlight the regions that are attended to by humans. Thus, to evaluate the performance of VA models, first, a set of images varying in their content are shown to a number of humans under an experimental setup and the humans' fixations are recorded by instruments known as eye-trackers. The eye-trackers work on the principle

(a) Fixation points (b)Fixation density map

Figure 1.8: Fixation points and fixation density map for an image from the Toronto database [5].

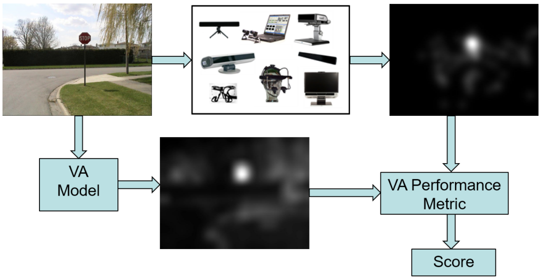

Figure 1.9: Block diagram for the typical process of VA model evaluation using eye-tracking data.

of Purkinje reflections in which infra-red light incident on the eyeballs of the subject gets reflected in three different ways. The angle of the reflected light can then be used to compute the location which was fixated upon on the screen. The data obtained is averaged over several subjects to get eye-tracking data that is made available for the research community to use along with the stimuli images in the form of a dataset. There are several such datasets that are covered in detail in Section 3.1. This data is available in two forms: (1) as fixation locations or (2) as fixation density maps which are obtained by placing 2D Gaussian kernels on the fixation locations and normalizing the resultant maps. The standard deviation of the Gaussian kernels is set such that the full width at half maximum of the Gaussian is equal to the visual angle subtended by the fovea on the screen surface. Figure 1.8 shows an image along with the fixation locations based on 15 subjects along with the corresponding fixation density map. The eye-tracking data is then compared with the predicted saliency maps output by computational VA models through a comparison measure called a performance metric or VA metric (used interchangeably here). The process can be summarized by the block diagram shown in Figure 1.9. Section 3.2 describes existing popular and newly proposed VA metrics that are used in the research community.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1597–1604, June 2009.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P.l Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[3] J. Bergstra, D. Yamins, and D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of The 30th International Conference on Machine Learning*, pages 115–123. IMLS, 2013.

[4] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, January 2013.

[5] N. D. B. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):5, 2009.

[6] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. MIT Saliency Benchmark. http://saliency.mit.edu/.

[7] M. Cerf, J. Harel, W. Einhaeuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems*, pages 241–248. Curran Associates, Inc., 2008.

[8] D. Cox and N. Pinto. Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*, pages 8–15, 2011.

[9] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[10] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222, 1995.

[11] E. Erdem and A. Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*, 13(4):11, March 2013.

[12] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[13] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[14] W. Förstner and B. Moonen. A metric for covariance matrices. In *Geodesy–The Challenge of the 3rd Millennium*, pages 299–309. Springer, 2003.

[15] A. Garcia-Diaz, V. Leborán, X. R. Fdez-Vidal, and X. M. Pardo. On the relationship between optical variability, visual saliency, and eye fixations: a computational approach. *Journal of Vision*, 12(6):17, January 2012.

[16] M. S. Gide, S. F. Dodge, and L. J. Karam. Visual attention quality database for benchmarking performance evaluation metrics. In *IEEE International Conference on Image Processing (ICIP)*, pages 2792–2796, September 2016.

[17] M. S. Gide and L. J. Karam. A locally weighted fixation density-based metric for assessing the quality of visual saliency predictions. *IEEE Transactions on Image Processing*, 25(8):3852–3861, 2016.

[18] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, 2012.

[19] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[20] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing*, 19(1):185–198, 2010.

[21] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in Neural Information Processing Systems*, 19:545–552, 2007.

[22] G. Heidemann. Focus-of-attention from local color symmetries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):817–830, 2004.

[23] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[24] J. E. Hoffman. Hierarchical stages in the processing of visual information. *Perception and Psychophysics*, 18(5):348–354, 1975.

[25] X. Hou, J. Harel, and C. Koch. Image Signature: Highlighting Sparse Salient Regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34, July 2011.

[26] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[27] X. Hou and L. Zhang. Dynamic visual attention: searching for coding length increments. In *Advances in Neural Information Processing Systems*, pages 681–688, 2008.

[28] L. Huang and H. Pashler. A boolean map theory of visual attention. *Psychological Review*, 114(3):599–631, July 2007.

[29] L. Itti, N. Dhavale, and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Optical Science and Technology, SPIE's 48th Annual Meeting*, pages 64–78. International Society for Optics and Photonics, 2004.

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

[31] ITU. Methodology for the subjective assessment of the quality of television pictures. Recommendation BT. 500-11. *International Telecommunication Union, Geneva, Switzerland*, 2002.

[32] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2083–2090. IEEE, 2013.

[33] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. *MIT Tech Report*, 2012.

[34] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *12th IEEE International Conference on Computer Vision*, pages 2106–2113, October 2009.

[35] D. Kelly. Information capacity of a single retinal channel. *IRE Transactions on Information Theory*, 8(3):221–226, 1962.

[36] J. Kim, J. Sim, and C. Kim. Multiscale saliency detection using random walk with restart. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(2):198–210, 2014.

[37] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Matters of Intelligence*, 1987.

[38] G. Kootstra, B. de Boer, and L. R. B. Schomaker. Predicting Eye Fixations on Complex Visual Stimuli Using Local Symmetry. *Cognitive Computation*, 3(1):223–240, March 2011.

[39] G. Kootstra, A. Nederveen, and B. De Boer. Paying attention to symmetry. In *Proceedings of the British Machine Vision Cconference (BMVC)*, pages 1115–1125. The British Machine Vision Association and Society for Pattern Recognition, 2008.

[40] G. Kootstra and L. R. B. Schomaker. Prediction of human eye fixations using symmetry. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society (CogSci09)*, pages 56–61. Cognitive Science Society, 2009.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[42] M. Kümmerer, L. Theis, and M. Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.

[43] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (iaps): Technical manual and affective ratings, 1999.

[44] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–66, March 2013.

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

[46] T. Lee, M. Girolami, and T. J. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, 11(2):417–441, 1999.

[47] L. Li, H. Su, F. Li, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in Neural Information Processing Systems*, pages 1378–1386, 2010.

[48] Z. Li and J. J. Atick. Toward a theory of the striate cortex. *Neural Computation*, 6(1):127–146, 1994.

[49] Z. Liu, O. Le Meur, and S. Luo. Superpixel-based saliency detection. In *14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4, July 2013.

[50] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[51] S. Lu and J. Lim. Saliency modeling from image histograms. In *European Conference on Computer Vision*, pages 321–332. Springer, 2012.

[52] M. Mancas, C. Mancas-Thillou, B. Gosselin, and B. Macq. A rarity-based visual attention map-application to texture description. In *IEEE International Conference on Image Processing*, pages 445–448, October 2006.

[53] U. Neisser. *Cognitive Psychology: Classic Edition*. Psychology Press, 2014.

[54] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001.

[55] A. Olmos and F. Kingdom. A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception*, 33(12):1463–1473, 2003.

[56] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[57] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.

[58] O. Pele and M. Werman. A linear time histogram metric for improved sift matching. In *10th European Conference on Computer Vision (ECCV)*, pages 495–508. Springer, October 2008.

[59] O. Pele and M. Werman. Fast and robust earth mover's distances. In *12th International Conference on Computer Vision (ICCV)*, pages 460–467, September 2009.

[60] U. Rajashekar, I. Van der Linde, A. C. Bovik, and L. K. Cormack. GAFFE: a Gaze-Attentive fixation finding engine. *IEEE Transactions on Image Processing*, 17(4):564–573, April 2008.

[61] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T. Chua. An eye fixation database for saliency detection in images. In *European Conference on Computer Vision*, pages 30–43. Springer, 2010.

[62] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context-free attentional operators: the generalized symmetry transform. *International Journal of Computer Vision*, 14(2):119–130, 1995.

[63] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1153–1160, December 2013.

[64] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit. Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. volume 28, pages 642–658. Elsevier, 2013.

[65] R. Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, 39(19):3157–3163, 1999.

[66] Y. I. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000.

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

[68] R. Salakhutdinov and H. Larochelle. Efficient learning of deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 693–700, 2010.

[69] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15, November 2009.

[70] H. R Sheikh, M. F Sabir, and A. C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, November 2006.

[71] C. Shen and Q. Zhao. Learning to predict eye fixations for semantic contents using multi-layer sparse network. *Neurocomputing*, 138:61–68, 2014.

[72] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45:643–659, 2005.

[73] H. R. Tavakoli, E. Rahtu, and J. Heikkila. Fast and efficient saliency detection using sparse sampling and kernel density estimation. In *Scandinavian Conference on Image Analysis*, pages 666–675. Springer, 2011.

[74] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[75] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *European Conference on Computer Vision*, pages 589–600. Springer, 2006.

[76] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2798–2805, 2014.

[77] T. N. Vikram, M. Tscherepanow, and B. Wrede. A saliency map based on sampling an image into random rectangular regions of interest. *Pattern Recognition*, 45(9):3114–3124, 2012.

[78] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[79] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *European Conference on Computer Vision*, pages 29–42. Springer, 2012.

[80] S. Wen, J. Han, D. Zhang, and L. Guo. Saliency detection based on feature learning using deep boltzmann machines. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014.

[81] J. M. Wolfe, K. R. Cave, and S. L. Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance*, 15(3):419, 1989.

[82] V. Yanulevskaya, J. Uijlings, J. Geusebroek, N. Sebe, and A. Smeulders. A proto-object-based computational model for visual saliency. *Journal of Vision*, 13(13):27, 2013.

[83] A. L. Yarbus. Eye movements during perception of complex objects. In *Eye Movements and Vision*, pages 171–211. Springer, 1967.

[84] J. Zhang and S. Sclaroff. Saliency detection: A boolean map approach. In *IEEE International Conference on Computer Vision (ICCV)*, pages 153–160, 2013.

[85] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32, 2008.

[86] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11:1–15, 2011.

[87] Q. Zhao and C. Koch. Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost. *Journal of Vision*, 12(6):22, January 2012.