Chapter 14
Advances in Learning Visual Saliency: From Image Primitives to Semantic Contents

Qi Zhao and Christof Koch

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. While previous research focused on low-level image features in saliency, the problem of "semantic gap" has recently attracted attention from vision researchers, and higher-level features have been proposed to fill the gap. Based on various features, machine learning has become a popular computational tool to mine human data in the exploration of how people direct their gaze when inspecting a visual scene. While learning saliency consistently boosts the performance of a saliency model, insights of what is learned inside the black box is also of great interest to both the human vision and computer vision communities. This chapter introduces recent advances in features that determine saliency, reviews related learning methods and insights drawn from learning outcomes, and discusses resources and metrics in saliency prediction.

14.1 Introduction

Besides understanding the mechanism that drives the 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

Q. Zhao
National University of Singapore, Singapore, Singapore

C. Koch (✉)
California Institute of Technology, Pasadena, CA, USA

Allen Institute for Brain Science, Seattle, WA, USA
e-mail: koch.christof@gmail.com

Z. Yang (ed.), *Neural Computation, Neural Devices, and Neural Prosthesis*,
DOI 10.1007/978-1-4614-8151-5_14,

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 [1–9] 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.
   The saliency literature has focused on low-level image features, where commonly used ones include contrast [10], edge content [11], intensity bispectra [12], color [13], and symmetry [14]. To better predict attention in visual scenes with semantic contents, higher-level features such as faces and people are also included in several recent models [15–19].
2. Compute individual feature maps to quantify saliency in that particular feature dimension.
   Low-level feature maps can be derived from biologically plausible filters such as Gabor or Difference of Gaussian filters, or more sophisticated inference algorithms—for example, Itti and Baldi [20] hypothesize that the information-theoretical concept of spatio-temporal surprise is central to saliency, and compute saliency using Bayesian statistics. Vasconcelos et al. [9, 21] quantify saliency based on a discriminant center-surround hypothesis. Raj et al. [22] derive an entropy minimization algorithm to select fixations. Seo and Milanfer [23] 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 [24] present a model based on "self-information" after Independent Component Analysis (ICA) decomposition [25] that is in line with the sparseness of the response of cortical cells to visual input [26]. Wang et al. [27] calculate the Site Entropy Rate to quantify saliency also based on ICA decomposition. Avraham and Lindenbaum [28] use a stochastic model to estimate the probability that an image part is of interest. In Harel et al.'s work [29], an activation map within each feature channel is generated based on graph computations. In [30], Carbone and Pirri propose a Bernouli mixture model to capture context dependency. Recently, high-level information has been incorporated into the saliency models where a high-level feature map is usually generated by an object detector such as a face detector [31] and a person detector [32].
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) [33–36] or "max" [37, 38] type of integration, where the former one has been commonly employed in computational modeling [1, 15]. Later, under the linear assumption, Itti and Koch [39] suggest various ways to normalize the feature maps based on map distributions. Hu et al. [40] compute feature contributions to saliency using "composite saliency indicator," a measure based on spatial compactness and saliency density. In a recent work by

Zhao and Koch [18, 19], it is suggested that feature integration is nonlinear, which raises the question of the extent to which the primate brain takes advantages of such nonlinear integration strategies. Biological neurons are highly nonlinear devices [41]. Thus, implementing the type of nonlinearities is not particularly problematic for the brain. Future psychophysical and neurophysiological research is needed to untangle this question.

This conventional structure of a computational saliency map requires many design parameters such as the number and type of features, the shape and size of the filters, and the choice of feature weights and normalization schemes. 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 much to be explored in the design of an effective saliency model, 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. In addition, learning provides a unified framework for analyzing data and making comparisons under different conditions.

Before discussions on main areas that bear on the topic of learning saliency, as will be elaborated in Sects. 14.2–14.5, we would like to bring up a recently aware issue of "semantic gap" in the saliency community, and the following discussions in each of the sections would also describe approaches to fill the gap.

14.1.1 Semantic Gap in Saliency

The semantic gap refers to the gap between the predictive power of computational saliency models and human behavior. That is, while existing research focuses on low-level image features, such features fail to encode object and/or social semantic information that is also important to saliency, many times more important than low-level information. Recent neurophysiological studies [42, 43] suggest that primates use a more powerful representation in which raw sensory input is perceptually grouped by dedicated neuronal circuitry. Psychophysical experiments [6, 44] show that humans frequently allocate their gaze to interesting objects in a scene and a large portion of fixations are close to the center of objects. Further, on top of the object-level information that attracts attention, social semantic information also contributes much to the final saliency: for example, a face tends to attract attention more than other objects [15]. It is also known that survival-related features (e.g., food, sex, danger, pleasure, and pain) possess an innate saliency which is determined by the activity of evolutionarily selected value systems in the brain [45, 46].

To fill the gap, [12] and [47] suggested the incorporation of higher order statistics. Computation models [15–17, 19] have also been developed to improve the prediction of attentional selection by adding object detectors, using linear or nonlinear methods to integrate all feature maps together and formulate the final saliency map. While boosting performance to some extent, adding object detectors does not scale well to the many object categories in real world as each object requires a particular detector. The question then is: what and to which extent high-level cues predict gaze allocation—we show that face is salient and we add a face channel into the saliency model. How about others? Are animals salient? Cars? Text? How many detectors should we add? And what is their relative importance? The conventional object/event detection-based method does not work here as each category requires building a particular detector which does not scale; yet saliency, on the other hand, requires a generally applicable mechanism to interpret the natural scenes for attention allocation. In fact, human brain does not work in a way that each object or event category has a region or a pathway for processing, therefore the current approach is not neutrally plausible. What, then, are inherent to the object/event categories that make them salient? We hypothesize that this type of saliency relates to prior knowledge that is hard-wired in neural ensembles, either from genetic propagation or from neuronal synaptic modifications through task training; yet, not much is known about the underlying mechanisms of semantic saliency.

In a recent effort [48], we make a first step to the exploration of learning higher-level saliency backed on human behavioral data. Briefly, based on observations of where humans look at in natural images, we propose an attribute-based framework. Besides low-level attributes that have been intensively researched in the literature, we also summarize higher-level attributes to encode semantic contents. A unique and important feature of the attribute-based framework is that unlike object detector-based methods, each attribute captures inherent object- or semantic-level information that is important to saliency and the combination of a limited set of attributes is able to describe a much larger set of object categories, in theory an infinite number of categories. Learning from human data, this work aims to better understand how various (i.e., low-level and higher-level) factors contribute to saliency, e.g., what attributes are more important, and how are they combined to fill the semantic gap. We believe that implications derived from this work are of great interest to both neuroscientists and psychophysicists, as well as serving as a useful guideline for computational modeling.

14.1.2 Challenges in Learning Saliency

There are several challenges particular to learning visual saliency 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 large effort. Examples of such image databases are LableMe [49] and

ImageNet [50]. 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. For example, crowdsourcing the task with Amazon Mechanic Turk is difficult with existing eye tracking technology. As will be introduced in Sect. 14.5.1, 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, photographers and subjects tend to look at the center of the image. This is largely due to the experimental setup [16, 51–53] and the feature distributions of the image sets [2, 6, 10, 16, 54]. 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 [55–57].
c. The problem is loosely defined: unlike typical computer vision tasks such as image segmentation or object detection where the objective is clearly specified, for predicting where people look at, the paradigm is more ambiguous. Some studies [1, 2, 13, 29, 58–60] focus on stimulus-dependent factors while others [8, 61] 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 [1, 17–19, 62], or proto-objects [63], or objects [6]—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 [57]. Thus, it depends upon the readers to identify relevant design assumptions and paradigms. For example, for a free-viewing model, task-dependent data are not applicable.

There are four main areas that bear on the topic of learning saliency: feature representation, learning techniques, data, and metrics. In the following sections, we will be describing recent advances in each of them: Sect. 14.2 introduces features to capture saliency, especially recent progress in encoding semantic saliency, Sect. 14.3 reviews methods in learning visual saliency, Sect. 14.4 discusses insights regarding the human visual system that are derived from learning outcomes, Sect. 14.5 reviews public datasets and performance metrics for sharing and comparisons in the saliency community, and Sect. 14.6 concludes the chapter.

14.2 Feature Representation

14.2.1 Low-Level Image Features

There is a vast literature on low-level features for saliency, and this chapter does not aim to exhaust all of them but provides a brief overview of this category.

Starting from the early proposal by Koch and Ullman [64], and later implemented by Itti et al. [1], a series of works (e.g., [2, 13, 15, 17, 19]) extract early visual features (e.g., color, intensity, and orientation) by computing a set of linear "center-surround" operations akin to visual receptive fields, where the feature maps are typically implemented as the difference between fine and coarse scales. Such mechanisms follow the general computational principles in the retina, lateral geniculate nucleus, and primary visual cortex [65] and are good at detecting regions that stand out from their neighbors. Typical low-level features are designed based on various visual characteristics using different region statistics or spatial frequencies (e.g., [10–12, 66, 67]). In theory, whether biologically plausible or not, the rich body of image features from the computer vision or image processing communities can be potentially incorporated for saliency models—the problem then is how to select from the vast pool the most relevant features that are inherent to saliency.

While low-level features are indicative of saliency to some extent, pushing along this line shows limitations and gains only marginal improvements on the predictive power of computational models, especially for scenes with semantic contents. To fill the semantic gap, higher-level features at the object- and semantic-levels are crucial, and recent progress in the design of such higher-level features is introduced in the next sections.

14.2.2 Object-Level Features

Attributes at this level describe object properties at non-semantic/social level. Based on psychophysical and neurophysiological evidence [6, 42–44], we hypothesize that any object, despite its semantic meanings, attracts attention more than non-object regions.

Gestalt psychologists have found many perceptual organization rules like convexity, surroundedness, orientation, symmetry, parallelism, and object familiarity [68]. In this model [48], we introduce five measures at this level that are simple and effective in predicting saliency: size, convexity, solidity, complexity, and eccentricity. Our observations show that these object-level features describing object shapes from different angles are strongly correlated with saliency.

Before the introduction of the object-level features, we first define several relevant notations for objects and the convex hull of the objects (illustrations are

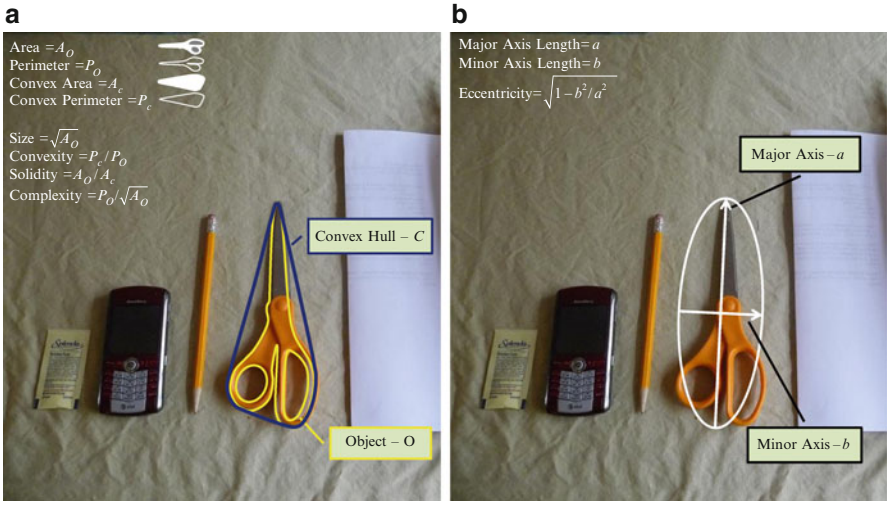

Fig. 14.1 Illustration of object-level attributes: (**a**) size, convexity, solidity, complexity, and (**b**) eccentricity

shown in Fig. 14.1). Particularly an object is denoted as O, and the convex hull of an object as C. Thus the area and perimeter of an object are represented as A_O and P_O, and the area and perimeter of the convex hull of an object are denoted as A_C and P_C.

Size Size is an important object-level feature; yet, it is not clear how it affects saliency—whether large or small objects tend to attract attention. Generally, a larger object might have more attractive details, but will probably be ignored for being a background as well. This feature is denoted as $\sqrt{A_O}$ where A_O represents the object's area.

Convexity The convexity of an object is denoted as P_C/P_O, where P_C represents the perimeter of the object's convex hull, and P_O represents the perimeter of the object's outer contour. Thus, a convex object has a convexity value of 1.

Solidity The solidity feature is intuitively similar to convexity, but it also measures holes in objects. Formally, solidity is denoted as A_O/A_C where A_O and A_C are the areas of the object and its convex hull, respectively. If an object is convex and without holes in it, it has a solidity value of 1.

Complexity Complexity is denoted as $P_O\sqrt{A_O}$. With the area of the object fixed, the complexity is higher if the contour is longer. A circle has the minimum complexity.

Eccentricity Eccentricity is computed as $\sqrt{1-b^2a^2}$, where a and b are the major and minor axes of the region. It describes how much a region's length differs in different directions. A circle's eccentricity is 0, while a line segment's eccentricity is 1.

Table 14.1 Semantic attributes

Name	Description
Face	Back, profile and frontal faces are labelled with this attribute
Emotion	Faces with obvious emotions
Touched	Objects touched by a human or animal in the scene
Gazed	Objects gazed by a human or animal in the scene
Motion	Moving/flying objects, including humans/animals with meaningful gestures
Sound	Objects producing sound (e.g. a talking person, a musical instrument)
Smell	Objects smelling good or bad (e.g. a flower, a fish, a glass of wine)
Taste	Food, drink and anything that can be tasted
Touch	Objects with a strong tactile feeling (e.g. a sharp knife, a fire, a soft pillow, a glass of cold drink)
Text	Digits, letters, words and sentences are all labelled as text
Watchability	Man-made objects designed to be watched (e.g. a picture, a display screen, a traffic sign)
Operability	Natural or man-made tools used by holding or touching with hands

Fig. 14.2 Sample figures illustrating semantic attributes. Each column is a list of sample objects with each semantic attribute and the last column shows sample objects without any defined semantic attributes

14.2.3 Semantic-Level Features

On top of the object-level attributes, humans tend to allocate attention to important semantic/social entities. Many cognitive psychological, neuropsychological, and computational approaches [69–71] have been proposed to organize semantic concepts in terms of their fine-grained features. Inspired by these works, we construct a semantic vocabulary [48], that broadly covers the following three categories: (1) Directly relating to humans (i.e., face, emotion, touched, gazed, motion). (2) Relating to other (non-visual) senses of humans (i.e., sound, smell, taste, touch). Observing whether objects relating to non-visual senses attract visual attention allows an analysis of cross-modality interaction [35]. (3) Designed to attract attention or for interaction with humans (i.e., text, watchability, operability). For each attribute, each object is either scored 1 to address the existence of the corresponding attribute, or a 0 to represent the absence of the attribute. In Table 14.1 we briefly list the annotation (with examples) for each attribute. Some objects may have all-zero scores if none of these attributes are apparent. Figure 14.2 demonstrates sample objects with or without semantic attributes.

14.3 Learning Visual Saliency

The problem of saliency learning is formulated as a classification problem [16–19, 62]. 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 non-fixated (or background) regions together with a label of "−1." A typical saliency learning algorithm including a training stage and a testing stage as illustrated in Fig. 14.3.

14.3.1 Features for Learning

Most existing learning-based saliency models use raw image data or low-level features (several works with a few object detectors) to represent positive and negative samples.

For example, Kienzle et al. [62] 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

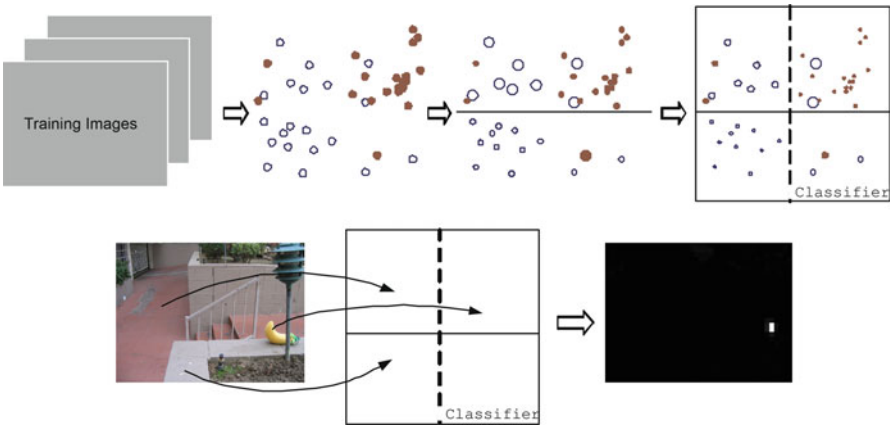

Fig. 14.3 Illustration of learning visual saliency. (**a**) Training stage: a saliency predictor (classifier) is trained using samples from training images in which observers fixated within the scene. The dimension of the feature vector of each sample is usually much higher than 2. We use 2 here for 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

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 [62] show a comparable performance with the conventional saliency model by Itti et al. [1] (i.e., 0.63 [62] vs. ~ 0.65 [1], using an Receiver Operating Characteristics (ROC)-based analysis as discussed in Sect. 14.5.2), although not any design prior is used in [62].

Given the very large input vectors if using raw image data, an alternative is to perform a feature extraction step before learning. This way, positive samples correspond to extracted features at fixated locations and negative samples to extracted features at non-fixated locations. For example, Zhao and Koch [17, 19] extract biologically plausible features that include low-level ones [1, 2] as well as faces [15] for learning. Judd et al. [16] use low-level image features (e.g., [72]), a mid-level horizon detector [73], and two high-level object detectors [31, 32] in their model. With feature extraction, feature dimensions are substantially reduced [16–18] compared with training on raw image data, and better performance is achieved [16–18] by learning in the lower-dimensional feature space.

Extraction and selection of good features for saliency, however, are not trivial. Besides constant efforts made on designing low-level image features, high-level detectors have recently shown to be effective in improving performance and have been added into saliency models. The problem is that adding detectors does not scale well in practice. We expect that with a vocabulary of higher-level attributes as described in Sect. 14.2, a much larger set of object categories can be considered in a saliency model. Further research and engineering efforts in both the human vision and computer vision communities would be needed for this incorporation. With designed features at various levels, a feature selection step picks up the most relevant ones to build a saliency model. For example, Zhao and Koch [18] propose an AdaBoost-based framework for saliency learning which automatically selects from a large feature pool the most informative features that nevertheless have significant variety. This framework could easily incorporate any candidate features, including both low-level and high-level ones, and naturally select the best ones in a greedy manner.

14.3.2 Learning Visual Saliency

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 Sect. 14.1) could 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. [62] aim to learn a completely parameter-free model directly from raw data using support vector machine (SVM) [74] 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 computational techniques have been employed to make saliency learning computationally more tractable. For example, feature extraction [16–18] largely reduces feature dimension. Besides, in the work by Zhao and Koch [17], the linear integration assumption is used and feature weights are learned using constraint linear regression. The simple structure makes the results applicable to numerous studies in psychophysics and physiology and leads to an extremely easy implementation for real-world applications. Similarly, using a set of predefined features, Judd et al. [16] learn the saliency model with liblinear SVM [75] which is used to achieve performance no worse than models with RBF kernels as proposed by Kienzle et al. [62]. Later, Zhao and Koch [18, 19] propose an AdaBoost [76–79] 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 [80]. In some of the methods [16–18], parameters of the spatial prior 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 [81, 82]. To make the object detection step robust and consistent, pixel neighborhood information is included. Thus, Khuwuthyakorn et al. [81] extend generic image descriptors of [1, 81] to a neighbourhood-based descriptor setting by considering the interaction of image pixels with neighboring pixels. In other efforts, Conditional Random Field (CRF) [63] that encodes interaction of neighboring pixels effectively detects salient objects in images [82] and videos [84], although CRF learning and inference are quite slow. In a recent work, we built an object-based model upon a variety of geometric and color features that are generic to "objectness." Superpixels are used as the basic representation unit which, on the one hand, saves computation compared with pixel-based methods; while, on the other hand, retains geometric and color information of the basic forming components of objects.

14.4 Insights from Learning Saliency

Besides being a powerful computational tool to build saliency models and boost performance of saliency detection, it is also of interest to draw inspirations as of what is "learned" from learning saliency. This section summarizes several recent efforts in interpreting the learning outcomes regarding the functioning of the human visual system.

Table 14.2 Optimal weights learned from four datasets [15, 16, 24, 85]

	Color	Intensity	Orientation	Face
FIFA [15]	0.027	0.024	0.222	0.727
Toronto [24]	0.403	0.067	0.530	0
MIT [16]	0.123	0.071	0.276	0.530
NUSEF [85]	0.054	0.049	0.256	0.641

Face is the most important (except the Toronto dataset [24] that includes fewfrontal faces), followed by orientation, color, and intensity

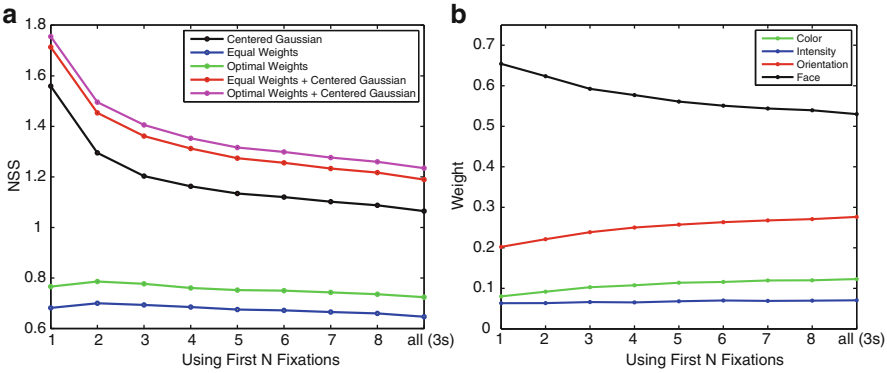

Fig. 14.4 (**a**) Illustration of model performance—using the Normalized Scanpath Saliency (Sect. 14.5.2)—with respect to viewing time using the MIT dataset [16]. The performance of all these bottom-up saliency models degrades with viewing time, as more top-down factors come into play. (**b**) Optimal weights with respect to viewing time using the MIT dataset [16]. The weight of face decreases while the weights for other channels increase, indicating that face attracts attention faster than the other channels

14.4.1 Faces Attract Attention Strongly and Rapidly

Using linear regression with constraints, Zhao and Koch [17, 18] learned on four published datasets (i.e., the FIFA [15], Toronto [24], MIT [16], and NUSEF [85] datasets, as detailed in Sect. 14.5.1) that (1) people rely on certain features more than others in deciding where to look at and setting proper weights to different features improves model performance and (2) faces attract attention the most strongly, independent of tasks. The learned weights of the face, color, intensity, and orientation channels are shown in Table 14.2 [18].

With the same learning techniques on fixations at different time instances (i.e., computing weights using the first N fixations), it is further shown (Fig. 14.4a [17]) that saliency decreases with time, consistent with the findings [85] that initial fixations are more driven by stimulus-dependent saliency compared to later ones. Besides, by making comparisons over time [17], observations show that face attracts attention faster than other visual features (Fig. 14.4b [17]).

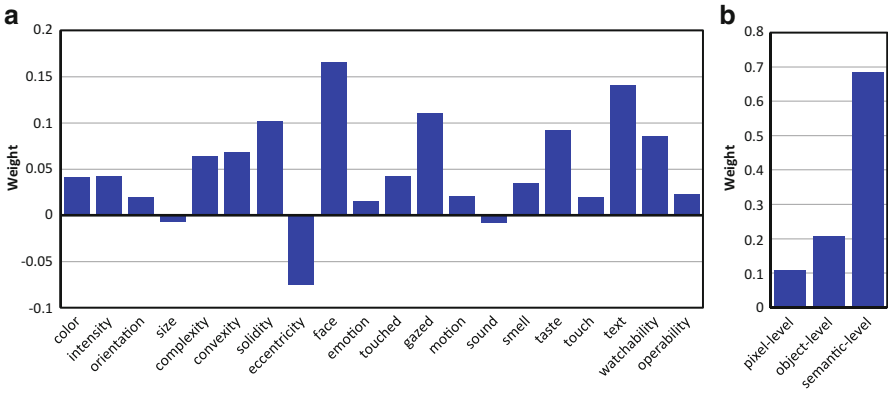

Fig. 14.5 (**a**) The learnt weights of different features. Face outweighs other semantic features, followed by text, gazed, and taste. (**b**) The importance of three levels of features

14.4.2 Semantic Contents are Important

It has been shown both qualitatively [15] and quantitatively [17] that face plays an important role in gaze allocation, which leads to the question of whether other semantic/social categories have similar properties, and if so, to what extent do they have. To approach this big question, we recently build a large eye tracking dataset with object and semantic saliency ground truth (see Sect. 14.5.1 for details), and learn using linear SVM [75] the weights of low-, object-, and semantic-level features in determining their importance in attention allocation [48]. The learned weight of each feature is shown in Fig. 14.5a. For semantic attributes, in consistent with the previous finding [15], face and text outweigh other attributes, followed by gazed, taste, and watchability. The high weight of "gazed" channel shows the effect of a joint attention. Viewers readily detect the focus of attention from other people's eye gaze, and orient their own to the same location [87, 88]. The weights of object-level features also agree with previous finding in figure-ground perception, that smaller, more convex regions tend to be foreground [89]. A complex shape contains more information, so it is also more salient than a simple one. The negative weight of eccentricity shows that longer shapes are less salient than round blob-like ones. We further compare the overall weights of the low-, object- and semantic-levels, by combining feature maps within each level into an intermediate saliency map of that particular level using the previously learned weights, and performing a second pass learning using the three intermediate maps. The learned weights of each level are 0. 11, 0. 21, and 0. 68 for low-, object-, and semantic-information, respectively, suggesting that semantic-level attributes attract attention most strongly, followed by object-level ones.

To further investigate the nature of semantic attributes in driving gaze, attribute weights as a function of fixation were calculated and compared over time.

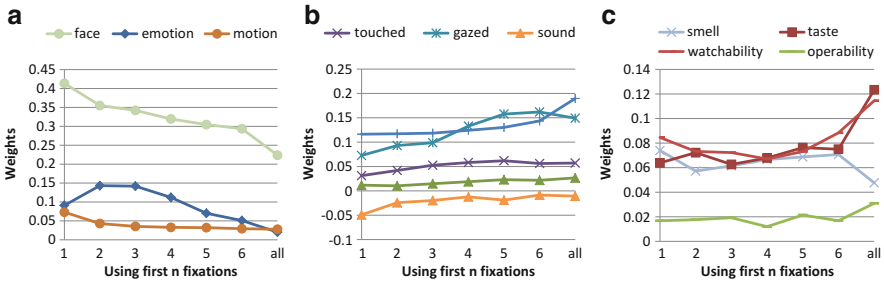

Fig. 14.6 Optimal weights with respect to viewing time for semantic features. (**a**) Features whose weights decrease over time attract attention rapidly. This is particular to face-related information, in consistent with the fact that face has its dedicated processing region and pathway in human brains. (**b**) Features whose weights increase over time attract attention not as rapidly. (**c**) Attributes whose weights do not show an obvious trend over time

As shown in Fig. 14.6, three types of trends are observed: (1) the weight decreases over time—when the training data include only the first fixations from all subjects, the weights of face, emotion, and motion are the largest, and they decrease monotonically as more fixations per image per subject are used (as shown in Fig. 14.6a). Attributes in this category attract attention rapidly, especially for the face and emotion channels—which may be due to the fact that humans have a dedicated face region and pathway to process face-related information. (2) As shown in Fig. 14.6b, the weights of text, sound, touch, touched, and gazed increase as viewing proceeds, indicating that although some of the attributes attract attention, they are not as rapid. (3) The weights of other semantic attributes including smell, taste, operability, and watchability do not show apparent trend over time, as illustrated in Fig. 14.6c.

14.5 Resources and Metrics

14.5.1 Public Eye Tracking Datasets

There is a growing interest in the neurosciences as well as in the computer science disciplines to understand how humans and other animals interact with visual scenes and to build artificial visual models. Thus, several eye tracking datasets have recently 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 these stimuli. A typical image set contains on the order of hundreds or a thousand of images. Different from the conventional laboratory psychophysics/eye tracking experiments based on highly simplified synthetic stimuli (e.g., a bunch of gratings or colored, singleton letters), natural stimuli reflect realistic visual input

and offer a better platform for the study of vision and cognition under ecological relevant conditions. On the other hand, natural stimuli are less controlled and therefore require more sophisticated computational techniques for analysis. Usually tens of subjects are asked to view the stimuli while the locations of their eyes in image coordinates are tracked over time (typically at rates between 32 and 1,000 Hz). A critical consideration is the task subjects had to perform when looking at the images, as it is known that the nature of the task can strongly influence fixation patterns [57]. Most common is a so-called *free-viewing* task with instructions such as "simply look at the image," "look at the image as you'll be asked to later on recognize it," or "look at the image and judge how interesting this image is compared to all other images" [15, 16, 24, 85]. In some datasets, Matlab codes are also available for basic operations such as calculating fixations and visualizing eye traces. Furthermore human labeling such as object bounding boxes, contours, and social attributes are available in certain datasets as ground truth data for learning and analysis of particular problems.

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

FIFA Dataset In the FIFA dataset from Cerf et al. [15], fixation data are collected from 8 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 through 10, how interesting each image is. Scenes are indoor and outdoor still images in color. Most of the images include faces of different skin colors, age groups, gender, positions, and sizes.

Toronto Dataset The dataset from Bruce and Tsotsos [24] 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 ($32^\circ \times 24^\circ$), 4s each. One distinction between this dataset and that of the FIFA [15] is that a large portion of images here do not contain particular regions of interest, while in the FIFA dataset typically contain very salient regions (e.g., faces or noticeable non-face objects).

MIT Dataset The eye tracking dataset from Judd et al. [16] includes 1, 003 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 3s. 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.

NUSEF Dataset The NUSEF database was published by Subramanian et al. [85]. An important feature of this dataset compared to others is that its 758 images contain many semantically affective objects/scenes such as expressive faces, nudes, aversive images of accidents, trauma and violence, and interactive actions, thus providing a good source to study social and emotion-related topics.

Images are from *Flickr, Photo.net, Google*, and from the emotion-evoking standard psychology database, *IAPS* [90]. In total, 75 subjects free-view ($26^{\circ} \times 19^{\circ}$) part of the image set for 5s each (each image is viewed by an average of 25 subjects).

OSIE Dataset The Object and Semantic Images and Eye-tracking (OSIE) dataset was created to facilitate research relating to object and semantic saliency. The dataset contains eye tracking data from 15 participants for a set of 700 images. Each image is manually segmented into a collection of objects on which semantic attributes are manually labelled. The images, eye tracking data, labels, and Matlab codes are publicly available [48]. Two main contributions of the dataset are: first, the image set is novel in that (a) it contains a large number of object categories, including a sufficient number of objects with semantic meanings and (b) most images contain multiple dominant objects in each image. Second, this dataset for the first time provides large-scale ground truth data of (a) 5, 551 objects segmentation with fine contours and (b) semantic attribute scores of these objects. The image contents and the labels allow quantitative analysis of object- and semantic-level attributes in driving gaze deployment.

Figure 14.7 illustrates sample images of each dataset and Table 14.3 summarizes a comparison between several recent eye tracking datasets.

Tracking the eye movements of subjects over many seconds while they are visually inspecting a static images allows researchers to evaluate how internal measures of saliency evolve in time given a constant input (Figs. 14.4 and 14.6). This mimics the situation naturally encountered when free-viewing a photograph or a webpage on a computer monitor. Dynamic scenes, such as those obtained from video or film sequence, lose this aspect yet correspond to the more ecologically relevant situation of a constantly changing visual environment. Unfortunately, there are less dynamic scenes available in which subjects' eye movements have been tracked under standardized conditions, and the following provides one example.

USC Video Dataset The USC video dataset from Itti's laboratory 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 8 distinct subjects watching 50 different video clips ($\sim$ 25 min of total playtime; [91, 92]), and from another 8 subjects watching the same set of video clips after scrambling them into randomly re-ordered sets of 1–3s clippets [93, 94].

Besides being valuable recourses for the saliency research community, these public datasets allow a fair comparison of different computational models. For example, the Toronto dataset [24] has been used as a benchmark for several recent saliency algorithms (e.g., [17, 19, 21, 51, 95]). Given the different nature and size of the datasets, researchers could either select specific ones to study particular problems (e.g., using the FIFA dataset to study face fixations, the NUSEF dataset for emotion-related topics, or the OSIE dataset to study object and semantic saliency) or carry out a comprehensive comparisons across all datasets for general issues that arise with any type of image.

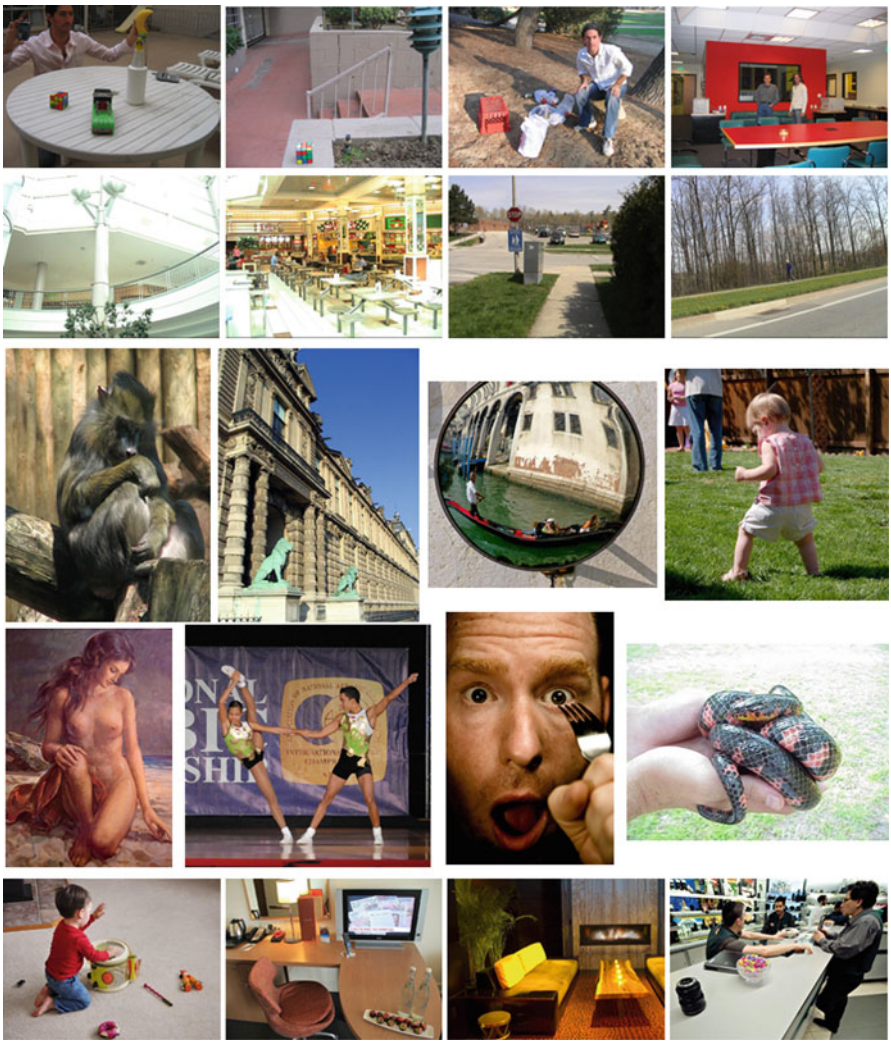

Fig. 14.7 Sample images of the four sets used in [17, 18]. *Top row*: FIFA dataset [15]. *Second row*: Toronto Dataset [24]. *Third row*: MIT Dataset [16]. *Fourth row*: NUSEF Dataset [85]. *Bottom row*: OSIE Dataset [48]

14.5.2 Performance Evaluation

Similarity measures are important to quantitatively evaluate the performance of saliency models. However, the question of how to define similarity in the saliency context is still open. In a number of parameters describing eye movements including fixation locations, fixation orders, fixation numbers, fixation durations, and saccade magnitude, how to scale each of them and how to integrate them in

Table 14.3 Comparisons of recent eye tracking datasets

Database	MIT [16]	FIFA [15]	Toronto [24]	NUSEF [85]	OSIE [48]
# Images	1,003	200	120	758	700
Resolution	1,024×(405−1,024)	1,024×768	681×511	1,024 × 728	800 × 600
# Viewers per image	15	8	11	25.3 (75 subjects each viewing a random set of 400 images)	15
Viewing time per image	3 s	2 s	4 s	5 s	3 s
Theme / distinguishing features	Everyday scenes	Images with faces	Indoor and outdoor scenes	Affective objects, e.g., expressive faces, nudes, unpleasant concepts, and interactive actions	Everyday scenes, many object categories with semantic meanings, multiple dominant objects per image
Ground truth annotation	None	Location of faces	None	ROIs, foreground segmentation for some objects (1 object per image and 54 images), valence and arousal scores, text captions	Object segmentation with fine contours for all objects (5, 551) and semantic attribute labels for all objects

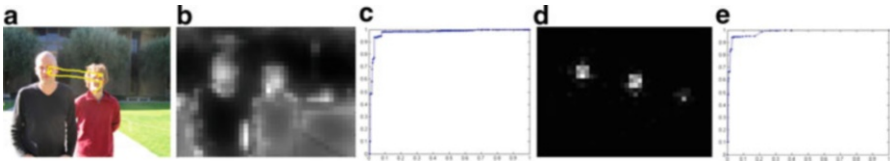

Fig. 14.8 Illustration of ROC limitations. (**a**) Original image with eye movements of one subject (fixations denoted as *red circles*). (**b**) Saliency map from linear combination with equal weights. (**c**) ROC of (**b**), with AUC = 0.973. (**d**) A saliency map with higher predictability power. (**e**) ROC of (**d**), with AUC = 0.975. Although (**b**) has a much larger false alarm rate, its AUC score is almost the same as that of (**d**). It could be observed that the ROC plot in (**c**) has a large number of points with high false alarm rate, but they do not affect the AUC score much as long as the hit rates at corresponding thresholds are high. In comparison, the NSS [58] of (**b**) and (**d**) are 1.50 and 4.47, and the EMD [105] between the fixation map and (**b**) and (**d**) are 5.38 and 2.93, respectively. Color figure online

quantifying similarity? For example, is the difference linearly proportional to the fixation distance in the image coordinate? Or is it a sigmoid type of function, or a step one? How about the difference in fixation durations? Is a shorter fixation less weighted than a longer one? And how to quantify the same fixations with different orders? Is it more different than two fixation sequences with the same order but certain translation in location? The definition of similarity between a saliency map and eye movement data, or between two sequences of eye movement data, is itself an intricate problem in a high-dimensional space. In practice, it is difficult to address all of the above issues in a single measure and most of the current measures, as we will discuss below, identify the most discriminative, or the most basic factors in viewing images while keeping the measures computationally tractable.

The commonly used similarity measures in the literature include the ROC [96], the Normalized Scanpath Saliency (NSS) [58], correlation-based measures [13, 97], Kullback-Leibler (KL) Divergence-based distances [54, 98], the least square index [99, 100], and the "string-edit" distance [101–103].

The ROC [97] is the most popular method in the community [21, 24, 29, 62]. In signal detection theory, an ROC curve plots the true positive rate vs. false positive rate for a binary system as its discrimination threshold is varied. In assessing saliency maps, the saliency map is considered as a binary classifier on every pixel in the image and human fixations are used as ground truth. By plotting the ROC curve and calculating the Area Under the ROC Curve (AUC), how well the saliency map matches human performance is quantified. An important characteristic of ROC is that it only depends on the ordering of the fixations (ordinality) and does not capture the metric amplitude differences. The desirable aspect of this feature is its transformation invariance; on the other hand, however, it loses magnitude information. In practice, as long as the hit rates are high, the AUC is always high regardless of the false alarm rate, as illustrated in Fig. 14.8 [17].

Another commonly used measure for saliency models is the NSS [58]. By definition, NSS evaluates salience values at fixated locations. It works by first linearly normalizing the saliency map to have zero mean and unit

standard deviation. Next, it extracts from each point corresponding to the fixation locations along a subject's scanpath its computed saliency and averages these values to compute the NSS that is compared against the saliency distribution of the entire image (which is, by definition, zero mean). The NSS is the average distance between the fixation saliency and zero. A larger NSS implies a greater correspondence between fixation locations and the saliency predictions. A value of 0 indicates no such correspondence. The NSS is intuitive in notion and simple in computation, and has been extended into the spatio-temporal domain [105]. One limitation of this measure is that it only captures information at fixated locations while completely ignoring the rest.

Mannan et al. [99] develop the index of similarity measure to compare two sets of fixations by summing up the distances between each fixation in one set and its nearest one from the other set. The overall spatial variability in the distribution of fixations over an image is not well accounted in this measure.

A large category of measures compare differences between two maps: the predicted saliency map and the fixation map that is usually the recorded fixations convolved with an isotropic Gaussian kernel, assuming that each fixation gives rise to a Gaussian-distributed activity [17, 29, 98]. An intuitive measure in this class is the correlation-based measure [13, 97] that calculates the correlation coefficient of the two maps. By definition, the coefficient lies in the $[-1, 1]$ interval. A value of 1 indicates that both maps are exactly similar, and a value of 0 indicates that both maps are totally different. Besides the correlation-based measures, theoretically all similarity measures for distributions can be applied to compare the saliency map and the fixation map that are essentially two distributions. Among them, the Kullback-Leibler (KL) Divergence and its extension that makes it a real metric [106] have been used in several saliency works [54, 98]. This measure is based on information theory and specifies the information one distribution provides given knowledge of the second distribution. One limitation of most distribution measures including the KL divergence is the bin-by-bin nature, meaning that they only capture the difference between corresponding bins in the distributions. For example, if a predicted salient location does not match a real fixation and as long as they do not fall into the same bin (i.e., a certain region in the maps), bin-by-bin distances return a large value no matter what the real distance of the predicted and the real fixations. Although bin-by-bin measures are relatively simple to compute, they do not catch global discrepancy well. Recently Zhao and Koch [17, 18] employ the Earth Mover's Distance (EMD) [104] that encodes cross-bin differences. Intuitively, given two distributions, EMD measures the least amount of work needed to move one distribution to map onto the other one. It is computed through linear programming and accommodates distribution alignments well. Compared with the other measures, a common weakness of both the KL divergence and the EMD is a lack of an intuitive interpretation for the closeness of prediction to actual eye fixation—identical distributions have a KL Divergence/an EMD of 0, but the interpretation of the (theoretically unbounded) result for non-identical distributions is not straightforward.

The "string-edit" algorithm [101–103] maps a list of fixation locations to a string of letters based on a predefined table and reduces the location sequence comparison problem to a string comparison problem where costs are defined for insertion, deletion, and substitution of letters. The minimum costs of this transformation are usually computed using dynamic programming. Drawbacks of this method are the division of stimuli to make the table and the indistinguishability of fixation durations. Differently from all the above methods, the order of fixations in the temporal dimension is accounted in this measure [107].

Different measures have different attributes, e.g., informative versus simple, ordering (i.e., the difference in order of fixation) versus magnitude (i.e., the measured difference in value), local versus global, and so on. While a single measure may not suffice in certain cases, a complementary combination can be a good candidate. For example, Zhao and Koch [17, 18] combine the AUC, NSS, and EMD for performance evaluation—while AUC captures only ordinality, NSS and EMD 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. Such a complementary combination enables a more objective assessment of saliency models. Further, given the extant variability among different subjects looking at the same image, no saliency algorithm can perform better (on average) than the measures dictated by inter-subject variability. Several previous works [15, 17–19, 29] 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. Particularly for the four published eye tracking datasets with color images [15, 16, 24, 85], these AUC values are 78. 6 % for the FIFA [15], 87. 8 % for the Toronto [24], 90. 8 % for the MIT [16], and 85. 7 % for the NUSEF [85] datasets. The performance of saliency algorithms taking into account such inter-subject variability is expressed in terms of normalized AUC (nAUC) values, which is the AUC using the saliency algorithm normalized by the ideal AUC.

14.6 Summary

This chapter reviews several issues relating to advances in learning visual saliency. 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. Further, biological interpretations can be derived from the learning outcomes and associated with the human visual system, which is of great interest to vision researchers.

Besides low-level features that have been intensively studied, recent findings in both the neuroscience and computational domains have found the importance of higher-level (i.e., object/semantic-level) features in saliency. Integration of features at various levels successfully fill the "semantic gap" and lead to models that are more consistent with human behaviors.

Lastly, as an important component in the data-driven approaches, a steady progress is also 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.

References

1. L. Itti, C. Koch, E. Niebur, A model for saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 1254–1259 (1998)
2. D. Parkhurst, K. Law, E. Niebur, Modeling the role of salience in the allocation of overt visual attention. Vision Res. **42**, 107–123 (2002)
3. A. Oliva, A. Torralba, M. Castelhano, J. Henderson, Top-down control of visual attention in object detection. In: *International Conference on Image Processing*, vol I, 2003, pp. 253–256
4. D. Walther, T. Serre, T. Poggio, C. Koch, Modeling feature sharing between object detection and top-down attention. J. Vis. **5**, 1041–1041 (2005)
5. T. Foulsham, G. Underwood, What can saliency models predict about eye movements spatial and sequential aspects of fixations during encoding and recognition. J. Vis. **8**, 601–617 (2008)
6. W. Einhauser, M. Spain, P. Perona, Objects predict fixations better than early saliency. J. Vis. **8**(18), 1–26(2008)
7. C. Masciocchi, S. Mihalas, D. Parkhurst, E. Niebur, Everyone knows what is interesting: Salient locations which should be fixated. J. Vis. **9**(25), 1–22 (2009)
8. S. Chikkerur, T. Serre, C. Tan, T. Poggio, What and where: a bayesian inference theory of attention. Vision Res. **50**, 2233–2247 (2010)
9. V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in highly dynamic scenes. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 171–177 (2010)
10. P. Reinagel, A. Zador, Natural scene statistics at the center of gaze. Network Comput. Neural Syst. **10**, 341–350 (1999)
11. R. Baddeley, B. Tatler, High frequency edges (but not contrast) predict where we fixate: a bayesian system identification analysis. Vision Res. **46**, 2824–2833 (2006)
12. G. Krieger, I. Rentschler, G. Hauske, K. Schill, C. Zetzsche, Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. Spat. Vis. **13**, 201–214 (2000)
13. T. Jost, N. Ouerhani, R. von Wartburg, R. Muri, H. Hugli, Assessing the contribution of color in visual attention. Comput. Vis. Image Und. **100**, 107–123 (2005)
14. C. Privitera, L. Stark, Algorithms for defining visual regions-of-interest: comparison with eye fixations. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 970–982 (2000)
15. M. Cerf, E. Frady, C. Koch, Faces and text attract gaze independent of the task: experimental data and computer model. J. Vis. **9**(10), :1–15 (2009)
16. T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look. In: IEEE International Conference on Computer Vision (2009)
17. Q. Zhao, C. Koch, Learning a saliency map using fixated locations in natural scenes. J. Vis. **11**(9), :1–15 (2011)
18. Q. Zhao, C. Koch, Learning visual saliency. In: *Conference on Information Sciences and Systems*, 2011, pp. 1–6

19. Q. Zhao, C. Koch, Learning visual saliency by combining feature maps in a nonlinear manner using adaboost. J. Vis. **12**(22), 1–15 (2012)
20. L. Itti, P. Baldi, Bayesian surprise attracts human attention. Adv. Neural Inform. Process. Syst. **19**, 547–554 (2006)
21. 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
22. R. Raj, W. Geisler, R. Frazor, A. Bovik, Contrast statistics for foveated visual systems: fixation selection by minimizing contrast entropy. J. Opt. Soc. Am. A **22**, 2039–2049 (2005)
23. H. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance. J. Vis. **9**(15), 1–27 (2009)
24. N. Bruce, J. Tsotsos, Saliency, attention, and visual search: an information theoretic approach. J. Vis. **9**, 1–24 (2009)
25. A. Hyvarinen, E. Oja, Independent component analysis: algorithms and applications. Neural Netw. **13**, 411–430 (2000)
26. D. Field, What is the goal of sensory coding Neural Comput. **6**, 559–601 (1994)
27. 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
28. T. Avraham, M. Lindenbaum, Esaliency (extended saliency): meaningful attention using stochastic image modeling. IEEE Trans. Pattern Anal. Mach. Intell. **99**, 693–708 (2009)
29. J. Harel, C. Koch, P. Perona, Graph-based visual saliency. In: *Advances in Neural Information Processing Systems*, 2007, pp. 545–552
30. A. Carbone, F. Pirri, Learning saliency. an ica based model using bernoulli mixtures. In *Proceedings of Brain Ispired Cognitive Systems*, 2010
31. P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol I, 2001, pp. 511–518
32. 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
33. A. Treisman, G. Gelade, A feature-integration theory of attention. Cognit. Psychol. **12**, 97–136 (1980)
34. H. Nothdurft, Salience from feature contrast: additivity across dimensions. Vision Res. **40**, 1183–1201 (2000)
35. S. Onat, K. Libertus, P. König, Integrating audiovisual information for the control of overt attention. J. Vis. **7**(11), 1–6 (2007)
36. 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. Atten. Percept. Psychophys. **71**, 1337–1352 (2009)
37. Z. Li, A saliency map in primary visual cortex. Trends Cogn. Sci. **6**, 9–16 (2002)
38. A. Koene, L. Zhaoping, Feature-specific interactions in salience from combined feature contrasts: evidence for a bottom-up saliency map in v1. J. Vis. **7**(6), 1–14 (2007)
39. L. Itti, C. Koch, Comparison of feature combination strategies for saliency-based visual attention systems. In: *Proceedings of SPIE Human Vision and Electronic Imaging*, vol 3644, 1999, pp. 473–482
40. 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
41. C. Koch, *Biophysics of Computation: Information Processing in Single Neurons* (Oxford University Press, New York, 1999)
42. E. Craft, H. Schütze, E. Niebur, R. von der Heydt, A neural model of figure–ground organization. J. Neurophysiol. **97**, 4310–4326 (2007)
43. S. Mihalas, Y. Dong, R. von der Heydt, E. Niebur, Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects. J. Vis. **10**, 979–979 (2010)

44. A. Nuthmann, J. Henderson, Object-based attentional selection in scene viewing. J. Vis. **10**(8), 20, 1–19 (2010)
45. G. Edelman, *Neural Darwinism: The Theory of Neuronal Group Selection* (Basic Books, New York, 1987)
46. K. Friston, G. Tononi, G. Reeke, O. Sporns, G. Edelman, et al. Value-dependent selection in the brain: simulation in a synthetic neural model. Neuroscience **59**, 229–243 (1994)
47. W. Einhauser, U. Rutishauser, E. Frady, S. Nadler, P. Konig, C. Koch, The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. J. Vis. **6**(1), 1148–1158 (2006)
48. J. Xu, M. Jiang, S. Wang, M. Kankanhalli, Q. Zhao, Predicting human gaze beyond pixels. J. Vis. **14**(1), 1–20, Article 28 (2014)
49. B. Russell, A. Torralba, K. Murphy, W. Freeman, Labelme: a database and web-based tool for image annotation. Int. J. Comput. Vis. **77**, 157–173 (2008)
50. 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
51. B. Tatler, The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. J. Vis. **7**, 1–17 (2007)
52. L. Zhang, M. Tong, T. Marks, H. Shan, G. Cottrell, Sun: a bayesian framework for saliency using natural statistics. J. Vis. **8**, 1–20 (2008)
53. 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
54. B. Tatler, R. Baddeley, I. Gilchrist, Visual correlates of fixation selection: effects of scale and time. Vision Res. **45**, 643–659 (2005)
55. F. Schumann, W. Einhauser, J. Vockeroth, K. Bartl, E. Schneider, P. Konig, Salient features in gaze-aligned recordings of human visual input during free exploratoin of natural environments. J. Vis. **8**(12), 1–17 (2008)
56. F. Cristino, R. Baddeley, The nature of the visual representations involved in eye movements when walking down the street. Vis Cogn. **17**, 880–903 (2009)
57. B. Tatler, M. Hayhoe, M. Land, D. Ballard, Eye guidance in natural vision: reinterpreting salience. J. Vis. **11**(5), 1–23 (2011)
58. R. Peters, A. Iyer, L. Itti, C. Koch, Components of bottom-up gaze allocation in natural images. Vision Res. **45**, 2397–2416 (2005)
59. J. Xu, Z. Yang, J. Tsien, Emergence of visual saliency from natural scenes via contextmediated probability distributions coding. PLoS One **5**, e15796 (2010)
60. V. Yanulevskaya, J. Marsman, F. Cornelissen, J. Geusebroek, An image statistics-based model for fixation prediction. Cogn. Comput. **3**, 94–104 (2010)
61. V. Navalpakkam, L. Itti, Modeling the influence of task on attention. Vision Res. **45**, 205–231 (2005)
62. 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
63. S. Mihalas, Y. Dong, R. von der Heydt, E. Niebur, Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects. Proc. Natl. Acad. Sci. **108**, 75–83 (2011)
64. C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. **4**, 219–227 (1985)
65. A. Leventhal, *The Neural Basis of Visual Function: Vision and Visual Dysfunction* (CRC Press, Boca Raton, 1991)
66. J. Elder, R. Goldberg, Ecological statistics of gestalt laws for the perceptual organization of contours. J. Vis. **2**(5), 324–353 (2002)
67. N. Bruce, J. Tsotsos, Saliency based on information maximization. Adv. Neural Inform. Process. Syst. **18**, 155 (2006)

68. S. Palmer, *Vision Science: Photons to Phenomenology*, vol. 1 (MIT Press, Cambridge, 1999)
69. P. Garrard, M. Ralph, J. Hodges, K. Patterson, Prototypicality, distinctiveness, and intercorrelation: analyses of the semantic attributes of living and nonliving concepts. Cogn. Neuropsychol. **18**, 125–174 (2001)
70. G. Cree, K. McRae, Analyzing the factors underlying the structure and computation of the meaning of *chipmunk*, *cherry*, *chisel*, *cheese*, and *cello* (and many other such concrete nouns). J. Exp. Psychol. Gen. **132**, 163 (2003)
71. A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes. In: *IEEE Conference on Computer Vision and Pattern Recognition, 2009 (CVPR 2009)*. IEEE (2009), pp. 1778–1785
72. E. Simoncelli, W. Freeman, The steerable pyramid: a flexible architecture for multi-scale derivative computation. In: *International Conference on Image Processing*, vol III, 1995 pp. 444–447
73. A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**, 145–175 (2001)
74. C. Burges, A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. **2**, 121–167 (1998)
75. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
76. Y. Freund, R. Schapire, Game theory, on-line prediction and boosting. In: *Conference on Computational Learning Theory*, 1996, pp. 325–332
77. R. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions. Mach. Learn. **37**, 297–336 (1999)
78. J. Friedman, T. Hastle, R. Tibshirani, Additive logistic regression: a statistical view of boosting. Ann. Stat. **38**, 337–374 (2000)
79. A. Vezhnevets, V. Vezhnevets, Modest adaboost - teaching adaboost to generalize better. In: *Graphicon*. (2005)
80. R. Jin, Y. Liu, L. Si, J. Carbonell, A.G. Hauptmann, A new boosting algorithm using input-dependent regularizer. In: *International Conference on Machine Learning*, 2003
81. P. Khuwuthyakorn, A. Robles-Kelly, J. Zhou, Object of interest detection by saliency learning. In: *European Conference on Computer Vision*, vol 6312, 2010, pp. 636–649
82. T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H. Shum, Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 353–367 (2011)
83. 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
84. 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
85. 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*, vol 6314, 2010, pp. 30–43
86. S. Mannan, C. Kennard, M. Husain, The role of visual salience in directing eye movements in visual object agnosia. Curr. Biol. **19**, 247–248 (2009)
87. L. Nummenmaa, A. Calder, Neural mechanisms of social attention. Trends Cogn. Sci. **13**, 135–143 (2009)
88. C. Friesen, A. Kingstone, The eyes have it! reflexive orienting is triggered by nonpredictive gaze. Psychon. Bull. Rev. **5**, 490–495 (1998)
89. C. Fowlkes, D. Martin, J. Malik, Local figure–ground cues are valid for natural images. J. Vis. **7**(8), 2, 1–9 (2007)
90. P. Lang, M. Bradley, B. Cuthbert, (IAPS): Affective ratings of pictures and instruction manual. Technical Report, University of Florida. (2008)
91. L. Itti, Automatic foveation for video compression using a neurobiological model of visual attention. IEEE Trans. Image Process. **13**, 1304–1318 (2004)

92. L. Itti, Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Vis. Cogn. **12**, 1093–1123 (2005)
93. R. Carmi, L. Itti, The role of memory in guiding attention during natural vision. J. Vis. **6**, 898–914 (2006)
94. R. Carmi, L. Itti, Visual causes versus correlates of attentional selection in dynamic scenes. Vision Res. **46**, 4333–4345 (2006)
95. X. Hou, L. Zhang, Dynamic visual attention: searching for coding length increments. In: *Advances in Neural Information Processing Systems*, 2008
96. D. Green, J. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966)
97. U. Rajashekar, I. van der Linde, A. Bovik, L. Cormack, Gaffe: a gaze-attentive fixation finding engine. IEEE Trans. Image Process. **17**, 564–573 (2008)
98. U. Rajashekar, L. Cormack, A. Bovik, Point of gaze analysis reveals visual search strategies. In: *Proceedings of SPIE Human Vision and Electronic Imaging IX*, vol 5292, 2004, pp. 296–306
99. S. Mannan, K. Ruddock, D. Wooding, The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. Spat. Vis. **10**, 165–188 (1996)
100. J. Henderson, J. Brockmole, M. Castelhano, M. Mack, Visual saliency does not account for eye movements during visual search in real-world scenes, in *Eye Movements: A Window on Mind and Brain*, ed. by R. van Gompel, M. Fischer, W. Murray, R. Hill (Elsevier, Amsterdam, 2007), pp. 537–562
101. S. Hacisalihzade, J. Allen, L. Stark, Visual perception and sequences of eye movement fixations: a stochastic modelling approach. IEEE Trans. Syst. Man Cybern. **22**, 474–481 (1992)
102. Y. Choi, A. Mosley, L. Stark, String editing analysis of human visual search. Optom. Vis. Sci. **72**, 439–451 (1995)
103. S.A. Brandt, L.W. Stark, Spontaneous eye movements during visual imagery reflect the content of the visual scene. J. Cogn. Neurosci. **9**, 27–38 (1997)
104. Y. Rubner, C. Tomasi, L. Guibas, The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. **40**, 99–121 (2000)
105. M. Dorr, T. Martinetz, K. Gegenfurtner, E. Barth, Variability of eye movements when viewing dynamic natural scenes. J. Vis. **10**(28), 28:1–17 (2010)
106. D. Johnson, S. Sinanovic, Symmetrizing the kullback-leibler distance. Technical Report, Rice University. (2001)
107. M. Clauss, P. Bayerl, H. Neumann, A statistical measure for evaluating regions-of-interest based attention algorithms. In: *Pattern Recognition: Lecture Notes in Computer Science*, vol 3175, 2004, pp. 383–390