IEICE TRANS. INF. & SYST., VOL.E96–D, NO.3 MARCH 2013

INVITED SURVEY PAPER

Computational Models of Human Visual Attention and Their Implementations: A Survey

Akisato KIMURA[†a)], ***Senior Member***, **Ryo YONETANI**[††b)], ***Student Member***, ***and*** **Takatsugu HIRAYAMA**[†††c)], ***Member***

SUMMARY We humans are easily able to instantaneously detect the regions in a visual scene that are most likely to contain something of interest. Exploiting this pre-selection mechanism called *visual attention* for image and video processing systems would make them more sophisticated and therefore more useful. This paper briefly describes various computational models of human visual attention and their development, as well as related psychophysical findings. In particular, our objective is to carefully distinguish several types of studies related to human visual attention and *saliency* as a measure of attentiveness, and to provide a taxonomy from several viewpoints such as the main objective, the use of additional cues and mathematical principles. This survey finally discusses possible future directions for research into human visual attention and saliency computation.

key words: *human visual attention, computational model, saliency, bottom-up, top-down*

1. Motivation

Developing sophisticated algorithms for detecting and recognizing something like objects from a given image and video has been a long distance challenge in pattern recognition and computer vision research fields. In fact, a huge number of studies, techniques and theories related to object detection and recognition have already been developed. In particular, several methods for detecting certain specific categories of objects such as human bodies and human faces have already been put to practical use in for example surveillance, authentication and the human-centric enhancement of image quality, with the best possible use of the prior knowledge of target objects (human bodies and faces) [1], [2]. However, generic object detection and recognition without any constraints as regards the target objects has remained major challenge, because (1) various kinds of objects might constitute the targets and (2) target objects in the same category might have different appearances due to variations of instances in a specific category, illumination changes and so on.

Fig. 1 One promising way of understanding human visual attention based on visual saliency. The model assumes that for a given input image (left) a gray-scale image called a saliency map (right) is automatically calculated in the brain, and we would first pay attention to the position yielding the maximum pixel value in the saliency map.

On the other hand, human beings seem to be able to detect various kinds of objects without any thought or effort. For example, from Fig. 1 left, we can easily and instantly detect a red car, a blue traffic sign and a broad white line. *Visual attention* [3] is considered to play an important role in achieving this function. Visual attention is one of the built-in mechanisms of the human visual system that quickly selects regions in a visual scene, which are most likely to contain items of interest. Such a pre-selection mechanism focusing only on relevant data would be essential in enabling computers to undertake subsequent processing such as generic object recognition and sentence generation from images.

With the above background, mimicking visual attention and computing *saliency* as a measure of attentiveness have attracted much attention in relation to both biological and artificial systems, especially in the last couple of years. A lot of researches on this issue are under way in several fields including psychophysics, neuroscience and computer vision. These researches take a *bottom-up* approach, meaning that a given image is the only resource for computing visual attention and saliency. In contrast, recent studies have attempted to incorporate additional cues such as prior knowledge about search targets, human intention and cognitive states, which are generally called a *top-down* approach. Moreover, the research dealing with regions-of-interest (ROI) extraction or *salient region extraction* has also been actively pursued in various fields such as image processing, computer vision and machine learning. This research is mainly aimed at real-world applications including medical imaging, intelligent cars, surveillance, image segmentation, object detection, generic object recognition and content-based image retrieval. Consequently, the word "saliency" is becoming a buzzword with certain undesirable

Manuscript received August 3, 2012.
Manuscript revised November 7, 2012.

[†]The author is with NTT Communication Science Laboratories, NTT Corporation, Kyoto-fu, 619–0237 Japan.
[††]The author is with the Graduate School of Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan.
[†††]The author is with the Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8603 Japan.
a) E-mail: akisato@ieee.org
b) E-mail: yonetani@vision.kuee.kyoto-u.ac.jp
c) E-mail: hirayama@is.nagoya-u.ac.jp
DOI: 10.1587/transinf.E96.D.562

side effects:

1. we are facing too many different types of saliency, each of which depends on the policies of researchers, background theories and applications,
2. this fact makes the definition of saliency ambiguous, fragile, and application-driven,
3. and therefore we cannot find any common procedures, measures and benchmark data for appropriately evaluating their validity.

Based on the above discussions, this paper systematically reviews previously reported studies related to human visual attention modeling and saliency computation, and provides a taxonomy from several viewpoints. We mainly focus on computational models of human visual attention, especially models that can be implemented on computers. First, Sect. 2 presents several scientific theories and findings, which provide useful basic knowledge for understanding computational models. With the help of these theories and findings, Sect. 3 offers a general overview of related research issues, and categorizes them from several points of view. Section 4 reviews representative or epoch-making methods based on the bottom-up approach, namely those methods that do not rely on any specific tasks, targets, intentions and cognitive states. Section 5 describes several computational models based on top-down approaches that consider the prior knowledge, intentions and cognitive states of humans, which build on the theories and findings presented in Sect. 2. Section 6 introduces public resources, especially datasets and source codes, all for evaluating computational models of human visual attention and methods of salient region detection. Section 7 summarizes this paper and discusses promising future work related to human visual attention, not limited to computational models.

2. Scientific Theories and Findings

This section presents some psychophysical theories and findings that lead to understanding of visual attention models.

2.1 Visual Search

Visual search is a task that involves detecting a specific visual target from various other stimuli (the distractors), which is widely accepted as clarifying human visual perception. The visual search can be classified into the following two types based on the relationship between the target and distractors: *feature search* that employs a target which can be distinguished from the distractors by a unique feature (e.g., intensity, color and edge orientation), and *conjunction search* with a target involving several features different from the distractors.

Many studies including the pioneering work undertaken by Neisser [4] have examined visual perception mechanisms via the visual search with various relationships of targets and distractors. In following sections, we introduce

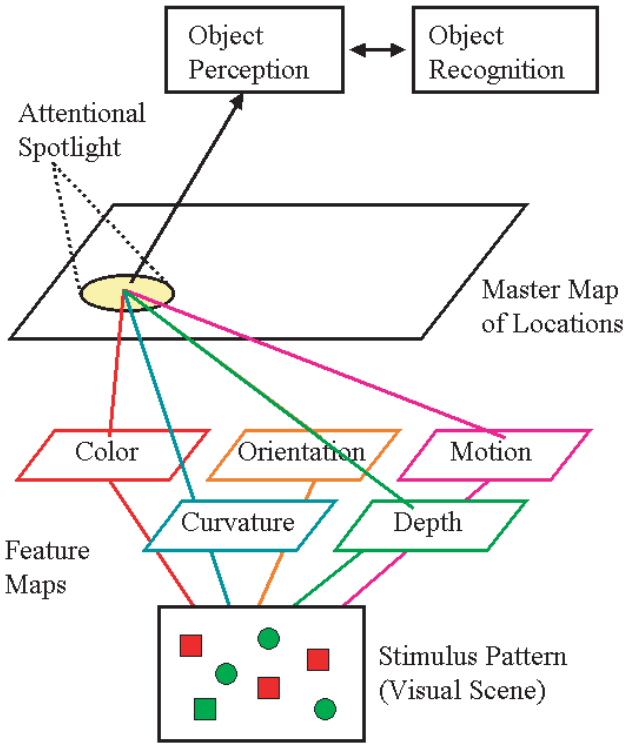

Fig. 2 Feature integration theory.

the two major theories: *the feature integration theory* [5] and *the guided search model* [6], [7].

2.2 Feature Integration Theory

The basic methodology in a study on visual search is to measure a reaction time to detect a target from several different numbers of distractors under the condition of the feature and conjunction searches. If the reaction time is constant independently of the number of distractors, the search is assumed to be conducted in parallel under such conditions. On the other hand, if the reaction time increases according to the number of distractors, the search is assumed to be sequential[†].

The feature integration theory (FIT), which has been proposed by Treisman and Gelade [5], has argued that the feature search and the conjunction search are conducted in parallel and sequentially, respectively (see Fig. 2). The FIT consists of the following points:

- Each single feature (e.g., color, edge orientation, size, motion and orientation) is processed individually by the corresponding specific module, spatially in parallel. Thus, humans are not required to pay attention to a specific location to search targets (the bottom of Fig. 2). This claim explains that feature search is processed in parallel.
- In the case of conjunction search, humans have to localize and integrate the results from multiple modules

[†]It is often referred to as a *set-size effect* [8].

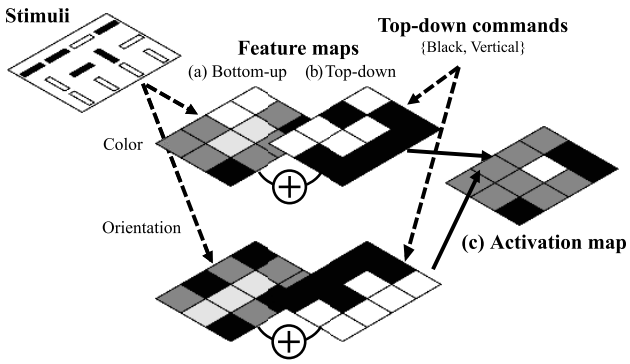

Fig. 3 Concept figure of the guided search. (a) bottom-up map, (b) top-down map, and (c) activation map, where lighter regions obtain higher activation.

of a unique feature. This integration requires an attentional shift to a specific location. Because of this attentional shift, the conjunction search is processed sequentially.

The FIT has received broad attention thanks to the clear explanation on parallel processing for feature search and sequential processing for conjunction search. On the other hand, several following studies have conducted experiments under various conditions and have revealed many phenomena, which are impossible to be explained by the FIT:

1. Reaction times can differ greatly according to the given tasks despite of the same sets of visual stimuli [9], [10]. This phenomenon is interpreted in terms of trade-off between the spatial coverage and resolution of attention.
2. Several studies have reported the results indicating that both feature and conjunction search have a difficulty depending on the similarity between targets and distractors as well as the similarity between distractors [11]–[13].

Moreover, there are some findings on "search asymmetry" that occurs when a search for stimulus A among stimulus B produces different results from a search for B among A [14], [15].

2.3 Guided Search Model

The model of guided search, proposed by Wolfe *et al.* [6], [7], has introduced top-down knowledge on characteristics of target stimuli in visual search.

The guided search model employs an *activation map* computed from both bottom-up activation based on the existing feature-search process and top-down activation based on the characteristics of target stimuli (see Fig. 3). The bottom-up activation in the guided search comes from saliency in each feature channel (e.g., color, orientation) in input stimuli, and it is obtained thanks to the "guide" by feature maps obtained from feature search (Fig. 3 (a)). On the other hand, the top-down activation is obtained based on the correlation between input and target stimuli with regard to each of the features. For instance, if the target stimulus has features consisting of "black" and "vertical line", the black regions in the top-down color map and the vertical-line regions in the top-down orientation map are activated (Fig. 3 (b)). The activation map is achieved by summing up the top-down and bottom-up activation maps (Fig. 3 (c)). Consequently, the focus of attention is oriented in the order of activation levels.

Guided search is a model that explicitly implements the characteristics of target stimuli in visual search, and it has now become important as a basis for recent top-down computation models. The literatures above [6], [7] only conducted some primal examinations using artificial images. In addition, recent studies tackle visual search and target characteristics using more general images and videos. Those studies are specifically presented in Sect. 5.

2.4 Relationship between Visual Attention, Human Intentions and Cognitive States

Visual attention is closely related to human internal states such as his/her intentions, cognitive states, and given tasks. Together with target characteristics, these human states can be regarded as top-down aspects as shown in the related survey [16]. Here, we present several psychophysical findings on the relationship between visual attention and the human states.

Yarbus has revealed that visual attention highly depends on human intentions [17]. His experiments measuring eye movements toward "The Unexpected Visitor" have demonstrated that different tendencies of gaze distributions can be observed according to the given tasks (e.g., free examination, give the ages of the people). The same results have been reported in the tutorial by Tatler *et al.* [18]. Moreover, The tutorial by Hayhoe *et al.* [19] presents some examples that human attention can be oriented to task-relevant locations rather than salient locations in the case humans are given some tasks.

Another important finding provided by Yarbus's experiments is that eye movements of several subjects toward the same images and those in several examinations by individual subjects are similar but not identical. This finding indicates the possibility to realize an estimation of the human states based on statistical learning of gaze information.

Several studies aim to clarify the relationship between visual attention and various cognitive states. For instance, Balkenius *et al.* [20] regard orienting of attention as an "action", and apply the mechanism in psychology of learning such as habituation and conditioning. Moreover, Taylor *et al.* [21] have discussed the possibility to associate endogenous/exogenous attention with emotion in view of neurophysiology.

3. Brief Classification of Related Research

This section offers a general taxonomy of human attention models in a broad sense from several viewpoints, which

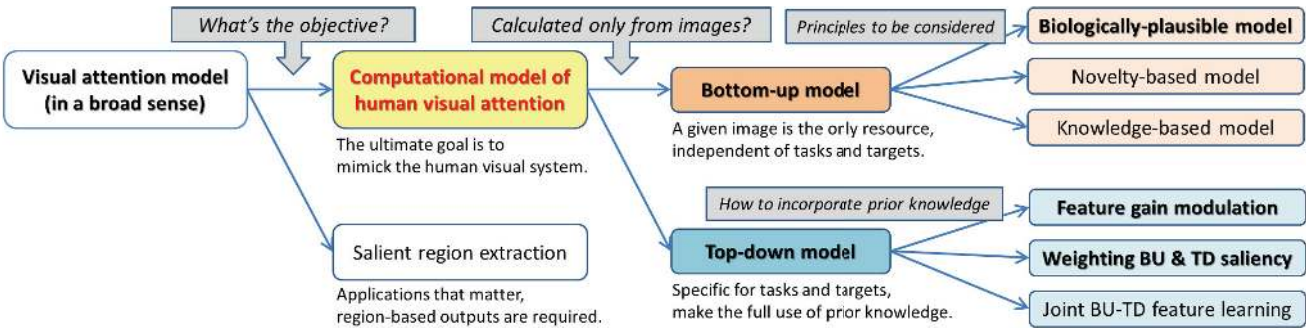

Fig. 4 Classifying human attention models (in a broad sense) from 4 standpoints: the main objective, usage of additional cues, fundamental policy and principle for building models, and way of incorporating additional cues.

would help us to understand individual computational models and their implementations. Note that the models and implementations introduced in the following sections do not always have a biological background such as the feature integration theory and the guided search model shown in Sect. 2.

Figure 4 depicts an overview of the taxonomy. The first standpoint is the main objective of individual computational models. Studies of "saliency" include not only scientific studies that aim to implement the psychophysical findings of human visual attention systems but also engineering studies simply designed to extract meaningful objects. These different objectives inevitably lead to different approaches, different outputs and different evaluations. More specifically, the first approach computes how much attention a pixel in a given image or video attracts, and compares outputs of methods with actual human gaze data. These outputs are often referred to as *saliency maps*. The second approach aims to estimate regions in a given image that contain meaningful objects, and utilizes ground-truth region labels for evaluation. This approach is often referred to as *salient region extraction*. Recently, the term "saliency" often appears indistinguishably in both two types of studies, which may sometimes result in inappropriate evaluations. One of the main goals of this paper is to carefully distinguish between these two approaches and provide a taxonomy of human visual attention models.

All the computational models of human visual attention can be separated into two categories from the second standpoint, namely the existence and availability of specific tasks, targets or intentions. As introduced in Sect. 4, studies on modeling visual attention have started with the implementation of feature integration theory and the shift of selective attention, which play a fundamental role in bottom-up attention. However, it has often been pointed out that such bottom-up models cannot completely explain entire visual attention systems, and some top-down concepts have been proposed (e.g., the guided search presented in Sect. 2.3). To this end, we introduce the second standpoint to classify the visual attention models into two categories, meaning the existence of specific tasks, targets or intentions. If some specific tasks or targets are given in advance, it is natural for computational models to utilize knowledge or side information related to the targets or tasks. This type of computational models are called *top-down models*. In contrast, without any specific tasks and targets, the signal is the only available resource for activating computational models. This type of computational models are called *bottom-up models*. Note that the bottom-up attention dealt with in this paper is mainly location-based attention [22], [23] that depends on only image stimuli at the location, and we do not take particular note of object-based attention [24]–[26] that relies on only high-level knowledge of objects to be focused rather than image features at the location.

The third standpoint relates to their background, mathematical principles and specific approaches to their implementation. Both bottom-up and top-down models have this standpoint. For instance, the center-surround differences, which play a central role in modeling bottom-up attention, were implemented by Itti *et al.* [27] as faithful as psychophysical theories and findings related to human vision, i.e., FIT and selective attention. By contrast, several models find irregularity or non-stationarity in a given image with the help of information theory and signal processing techniques, and other models employ general knowledge and pre-trained features that characterize visual attention from actual gaze data in a machine-learning manner. Note that the knowledge-based methods are essentially different from top-down ones, where the former categorization approach forms scientific findings and basic principles, and the latter relies on the use of knowledge obtained from targets or tasks.

Moreover, top-down models have three ways of introducing additional cues. Two of them relate closely to the guided search shown in Sect. 2.3, which directly modifies bottom-up saliency maps or combines them with top-down attention maps derived from the additional cues. The last model extracts both bottom-up and top-down related features and learns their importance from the gaze data using machine learning techniques.

4. Bottom-Up Models

4.1 FIT-Based Computational Models

As described in Sects. 2.2 and 2.3, the feature integration theory (FIT) took the central role in the development of bottom-up visual attention models for two decades after its proposal. Several conceptual models modifying FIT were proposed during this process [3], [9], [28]–[30]. In addition to these conceptual models, computational models have also been developed that enable us to clarify the process of visual search and to verify their performance by implementing them on computers. A significant work regarding computational models of visual attention is the saliency map model proposed by Itti, Koch and Niebur [27]. The following briefly describes this computational model, which we call the *Itti saliency map model* for simplicity.

The Itti saliency map model is based on the bottom-up architecture of visual attention proposed by Koch and Ullman [3]. This architecture introduces a multi-resolution structure to solve the first problem of FIT shown in Sect. 2.2, and serves as a foundation for later computational models. The Itti saliency map model can be regarded as a representative implementation of the Koch-Ullman architecture.

Figure 5 is a sketch of the Itti saliency map model. The details can be seen in the original paper [27] and some other reviews [31], [32]. In the first stage of the procedure, several fundamental features such as intensity, color opponents and edge directions are extracted from an input image, and a Gaussian pyramid is constructed for each fundamental feature. Taking pixel-wise differences across the center (= fine) and surround (= coarse) scales of the Gaussian pyramid, we can compute multi-scale spatial contrasts for combinations of three center scales and two center-surround scale differences. These pixel-wise differences can be viewed as an approximation of the convolutions of difference-of-Gaussian (DoG) filters. Each center-surround difference is normalized to obtain a sparse representation, called *a feature map*, so that only outlier locations have larger pixel values than those of their surroundings. Feature maps with the same feature channel (intensity, color opponent and edge directions) are integrated into a single map and again normalized to obtain *a conspicuity map*. All the conspicuity maps finally contribute to a unique *saliency map* representing the uniqueness of each location in the visual field. The Itti saliency map model introduces a mechanism, *winner-takes-all (WTA)*, which selects the location where the pixel value of the saliency map is greater than at any other locations. From this viewpoint, the concept behind the Itti saliency map model is somewhat similar to that behind guided search [6], [7].

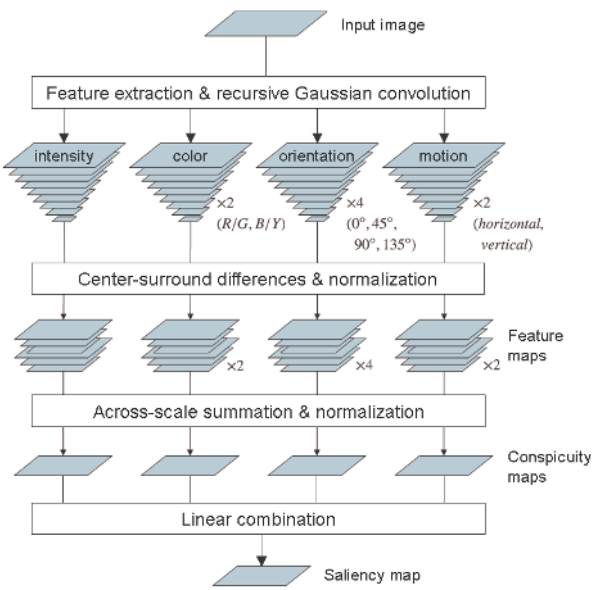

Fig. 5 Computational model of saliency maps by Itti *et al.*

The Itti saliency map model is very simple, easy to implement and provides reasonable outputs for various kinds of input images. Therefore, this model has had a considerable impact on broader research areas such as image processing, pattern recognition, computer vision, robotics and neuroscience [33]–[37]. Several extended models have also been developed in parallel: Leung *et al.* [38] were inspired by neural adaptation [39], which describes the habituation caused by a continuously displayed visual stimulus, and proposed a neural adaptation mechanism that employs standard image processing. Maki *et al.* [40] took particular note of the feature whereby regions that are closer in terms of depth are more salient [41], and they proposed a model for estimating human visual attention that integrates several features obtained from binocular cameras such as motion and disparity. Ouerhani and Hugli [42] directly incorporated a depth feature taken from a range finder into the Itti saliency map model. Jeong *et al.* [43] proposed a model for computing saliency maps from every image taken from binocular cameras and correcting the maps with the help of disparity information in highly salient regions.

4.2 Introducing Stochastic Ambiguity

Although the Koch-Ullman architecture and Itti saliency map model provide good explanation of the bottom-up human visual attention, they pose one crucial problem: A saliency map is extracted from an input image in a deterministic way, which implies that all the subjects would focus on the same location for the same input image. However, according to the finding described in Sect. 2.4, humans may focus on different locations in the same input image.

As shown in Sect. 2.4, conventional theories state that this inconsistent visual attention is mainly caused by top-down intention, knowledge and preferences, and a lot of psychophysical studies supported this hypothesis (See e.g. [44]).

On the other hand, another theory to understand human visual attention was introduced, called the *signal detection theory*, which has widely been employed in the field of com-

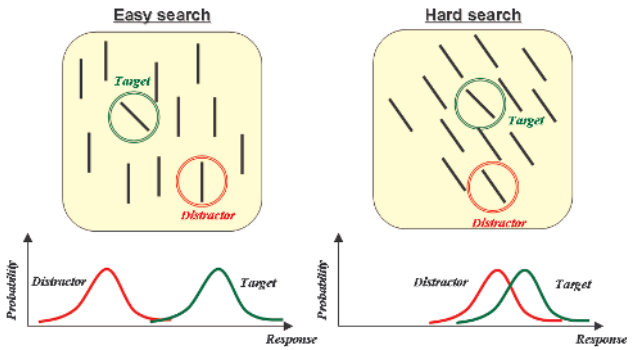

Fig. 6 Intuitive interpretation of the signal detection theory applied to the understanding of human visual attention.

munication theory and psychology [45], [46] for about 50 years. Eckstein *et al.* [47], [48] applied the signal detection theory to visual attention modeling, and described a mechanism that causes attention ambiguity using only bottom-up processing. The idea behind the signal detection theory can be intuitively explained with the help of Fig. 6, where a visual search task is considered with a single 45° target among a lot of distractors[†]. The essential difference between conventional theories and the signal detection theory lies in whether or not distractors can be recognized as an (incorrect) target. The feature integration theory never allows us to detect a distractor as a target, and thinks of humans as capable of making mistakes. In contrast, the signal detection theory assumes that distractors might be recognized as a target due to internal noise of a visual cortex.

Based on the Itti saliency map model, a saliency map can be uniquely obtained for a given input image. Here we assume that people may observe a different map from the saliency map because of the internal noise of the visual cortex, where every internal noise is emitted from an independent Gaussian distribution [47]. The observed saliency map is called a stochastic saliency map, and each of its pixel values is called a stochastic saliency. The stochastic saliency at every position can be represented as a Gaussian random variable with the same mean value as the saliency value at this position, as shown in Fig. 6 below. Namely, the response of a distractor tuned to the target orientation[††] is represented as a Gaussian density with a lower mean than that of the target.

In Fig. 6 left with a 45° target and vertical distractors, these densities barely overlap, which implies that we can immediately detect the target. On the other hand, in Fig. 6 right with a 45° target and very similar distractors, the target density is identical to the easy search case, while the distractor density is shifted to the right, so that the two densities overlap considerably. This implies that the probability that we first focus on the distractors becomes high, and therefore it takes a lot of time to detect the target. This is the mechanism that causes attention ambiguity with only bottom-up processing.

Several findings [11]–[13] have already indicated some relationships between the complexity of a visual search and the similarity between a target and distractors, as shown in subsection 2.2. The major contribution of Eckstein *et al.* [47] is that they were the first to clarify the computational mechanism of those relationships.

Several stochastic models of human visual attention have been proposed based on the finding derived from signal detection theory. Koike and Saiki [49] first introduced a stochastic mechanism of human visual attention into a computational model, and verified it with psychophysical settings. Pang *et al.* [50], [51] extended this model to video inputs, and constructed a dynamic Bayesian network that considered the stochastic ambiguity and temporal smoothness of visual saliency simultaneously. Miyazato *et al.* [52] achieved real-time computing of this model with the full use of parallel processors in GPUs.

4.3 Temporal Aspects in Saliency

The computational models presented so far in this paper have focused only on spatial aspects of visual attention and saliency. Namely, they were interested in where is salient in a given still image, and saliency values were computed based on the spatial contrasts of image features. However, when dealing with a video as a saliency calculation target, certain temporal events such as sudden changes in (parts of) image frames and the motions of objects would be significant cues for visual attention, which implies that the temporal dynamics of image features should be considered when modeling human visual attention.

Itti and Baldi [53], [54] first incorporated the temporal dynamics of image features into computational models of human visual attention, and proposed the Bayesian Surprise model that regards the difference between the visual features that are expected to be obtained and those that are actually obtained as indicating saliency. Figure 7 outlines the Bayesian Surprise model. Its basic idea involves the parametric modeling of the distributions of visual features. First assume that a prior distribution of visual features at time t can be modeled by a Poisson distribution. For a given set of visual features at time t, a posterior distribution of the visual features at time t can also be obtained as a Poisson distribution via a state-space model. From the definition of the state-space model, the posterior at time t can be re-used as the prior at time $t+1$. The main contribution of the Bayesian Surprise model is that it adopts a Kullback-Leibler divergence from the prior to the posterior as a saliency measure. More specifically, a sequence of similar visual features continuously gives low saliency values, while unexpected visual features such as sudden scene changes provides high saliency values.

[†]This example is task-driven and includes top-down factors even if a subject does not know the target stimuli. However, we have to note that the principle of the signal detection theory does not rely on any top-down factors.

[††]Note that responses from any other filters would be useless in this example, and therefore the mechanism for filter selection does not necessarily depend on the target information.

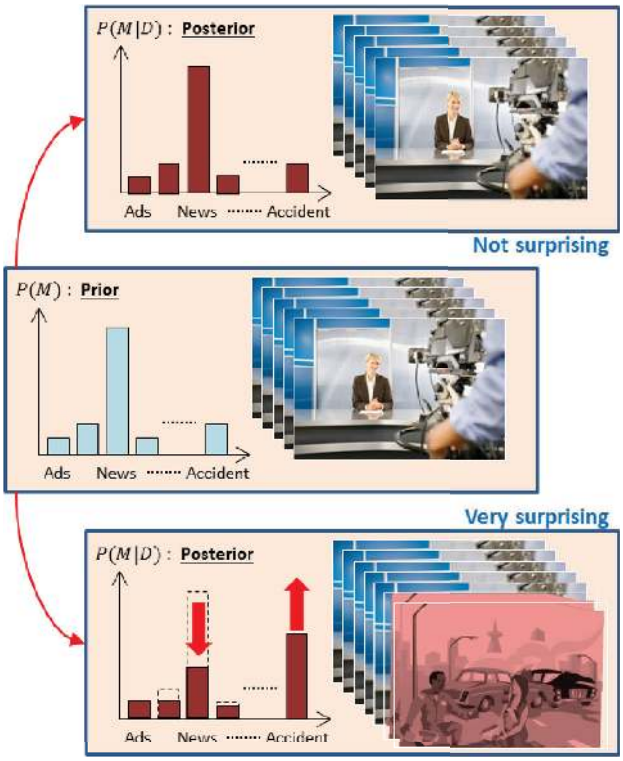

Fig. 7 Model of Bayesian surprise proposed by Itti and Baldi [53].

When incorporating temporal aspects into computational models, dynamic visual features such as motion and flicker would be significant. Studies by Maki *et al.* [40] and Marat *et al.* [55] are typical examples of the exploitation of motion features for computational models of human visual attention. According to the psychophysical findings reported by Reichardt [56], several computational models utilize the spatial contrasts of motion and flicker [35], [57].

4.4 Novelty-Based Models

In addition to computational models based on psychophysical theories and findings, novelty-based models have also been developed that try to find spatial irregularities and temporal non-stationarity as saliency in a given image and video. Most of these models would not rely on any kinds of physiological or psychophysical theories and findings. Instead, their priority is mainly placed on the performance of extracting salient regions that would be useful for further applications. As a result, novelty-based approaches are now becoming mainstream techniques, especially in the field of image processing and computer vision.

Hou and Zhang [58] took particular note of $1/f$ noise phenomenon [59] in a frequency spectrum, where the power spectral density is inversely proportional to the frequency. If we assume that images with natural scenes also obey this system, the log power spectrum is proportional to the frequency. Hou and Zhang employed a differential of the log power spectrum with respect to the frequency as a measure of saliency, and approximated it by subtracting the log spectrum at the current position from the averaged log spectrum around this position. This model is called the *spectral residual model*. The saliency map can be obtained by applying an inverse Fourier transform to the spectral residual. A variant of spectral residual model has also been developed by employing a phase spectral density [60].

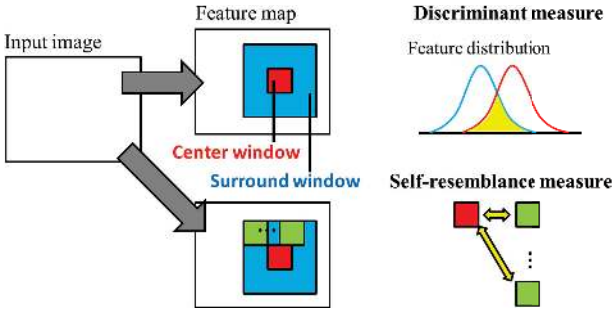

Fig. 8 Center-surround measures proposed by Gao and Vasconcelos [64], [65] (top; discriminant measure) and Seo *et* al. [66] (bottom; self-resemblance measure).

Achanta *et al.* [61] slightly modified the Itti saliency map model where a Lab color space is used instead of an RGB space. Later, they further simplified the procedure for computing multi-scale contrasts, and eventually exploited the difference between smoothed and averaged input images as a basis for salient region extraction [62]. This model is called the *frequency-tuned method*. The simplification of the frequency-tuned method can also be derived from the spectral residual model [58].

Avraham *et al.* proposed Esaliency (extended saliency) [63], which detects salient regions based on the similarity of regions in a whole image. It begins with the segmentation of an image into small regions, and some features are extracted from them. A Bayesian network is introduced to describe the co-occurrence of region indices that can be a salient region, while considering the similarity of region features. Saliency of the regions is finally computed based on the joint probability of the Bayesian network. Esaliency can incorporate top-down knowledge about target locations, and it is applied to natural scenes by learning target locations from several datasets (see Section 5.3).

Several models evaluate a local novelty at each location against its neighbors as a measure to evaluate the center-surround differences. The following two models introduce two windows of different sizes, the center and surround window, to compute saliency on a certain location.

Gao and Vasconcelos [64], [65] proposed a discriminant measure based on the decision theory. The basic concept behind the discriminant measure is that the center-surround difference at a location is evaluated using expected error probabilities when discriminating the feature distributions of the center and surround windows (top of Fig. 8). That is, saliency values become high when the two feature distributions are easy to discriminate. The discriminant measure is applied to each feature map (such as intensity, color and orientation), and a saliency map is finally obtained

by summing up the results. The neurophysiological plausibility of the measure is discussed in [65].

Moreover, Seo *et al.* [66] introduced a self-resemblance measure. Unlike the discriminant measure, the self-resemblance measure slides a window within the surround window, and computes the similarity of features between the center window and the cropped window (the bottom of Fig. 8). A collection of resemblance scores finally indicates how rarely (how salient) the features are at the center location.

4.5 Knowledge-Based Models

When visual features that would generally contribute to saliency calculation or salient region extraction are available in advance, the problem might become easier with the full use of this generic prior knowledge. Also in some cases, we may be able to obtain a number of pairs of images and their ROIs as ground-truth, which would be promising for capturing useful visual features via machine learning techniques. This section briefly reviews several models and methods based on generic prior knowledge independent of or unspecified by targets to be focused on or tasks to be executed.

One typical example is the *self-information* model by Bruce *et al.* [67], where the concept of self information is incorporated into computational models of visual attention. It is well known that a process for learning a sparse code for natural image statistics has been achieved through the emergence of simple-cell receptive fields in the primary visual cortex of primates [68], [69]. Based on this finding, the self-information model first derives the bases of image patches by performing independent component analysis (ICA) on a lot of samples of small image patches in advance. For a given image, the probability distribution of each basis coefficient is estimated across the entire image by employing non-parametric kernel density estimation. The product of these distributions over local neighborhoods yields the joint distribution of the entire set of basis coefficients. The level of saliency is finally computed as the negative log likelihood (self information) of the basis coefficient. Several similar models have also been presented by Renninger *et al.* [70] and Li *et al.* [71]. Furthermore, as an extended variant of those approaches, Sun *et al.* [72], [73] recently proposed a framework for modeling saccadic eye movements via on-the-fly FastICA [74]–[76].

Kienzle *et al.* [77] proposed a framework to discover relevant visual features for saliency calculation using human gaze data. They assume that there are local image patterns (perceptive fields) that guide human gaze. The discovery of such patterns begins with non-linear mapping of image patch textures to target/non-target labels, where the labels are derived from the gaze data. The perceptive fields that excite and inhibit visual attention are then obtained as local patches that maximize and minimize the mapping function, respectively. They found that the excitatory perceptive fields exhibit center-surround structures and the inhibitory fields exhibit a flat, ramp-like structure. A saliency map is finally computed based on a feed-forward network with the excitatory and inhibitory perceptive fields.

Ma and Zhang [78], [79] employed various cues that have the possibility of being related to human visual attention from videos, and combined them to compute saliency. The cues include not only visual features but also acoustic and linguistic features: motions, contrasts, faces, camera motions, audio loudness peaks, and how similar the sound is to speech and music.

5. Top-Down Models

Many researchers have proposed computational models involving top-down processes, including Wolfe's guided search model [6], [7]. The models fall into the following two classes based on what kind of top-down knowledge manages the computational processes: (1) computational models based on prior knowledge of the search target in a visual search task, (2) computational models based on contextual knowledge in a specific task situation other than a visual search task. The former deals with the situation where the humans determine a visual object to which they turn their attention before beginning the task, while the latter never pre-specifies an object but the subjects focus on an object linked to the controlled cognitive state under the situation such as where they memorize something or play a game. The former has been proposed more than the latter because psychophysics has long focused on visual search tasks.

Each also falls into the following three classes based on how a model computes top-down saliency.

(a) Weight modulation of bottom-up features, which learns each weight of feature channels (e.g., color, orientation) by reference to psychophysical findings on top-down processes so that some regions related to top-down knowledge have higher saliency values (Fig. 9 (a), Fig. 3 (b)).
(b) Weighted combination of outputs from bottom-up and top-down models[†], which combines the bottom-up saliency map with visual similarity map between an appearance model of object/scene and an image region or the target-/task-dependent saliency map (Fig. 9 (b), Fig. 3 (c)). The weight parameters to combine them are modulated by top-down knowledge.
(c) Joint learning of bottom-up and top-down features, which learns a function from a pair consisting of feature vectors including top-down factors and supervisory signals reflecting top-down knowledge (e.g., the corresponding eye positions) to a saliency value using machine learning techniques (Fig. 9 (c)).

Classes (a) and (c) modulate the relationships among the features, while class (b) modulates the relationship between the maps using a weight.

[†]Overall saliency maps computed using the process (class (b)) are called "activation map" in the guided search model [6], [80], [81] or "priority map" in cognitive science [82].

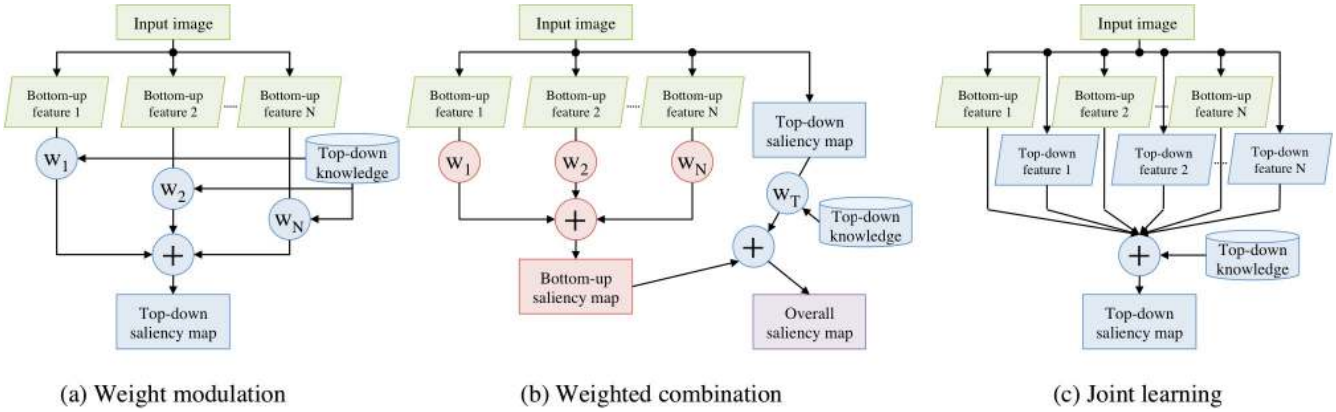

Fig. 9 (a) Weight modulation of bottom-up features, (b) weighted combination of bottom-up and top-down saliency maps, (c) joint learning of bottom-up and top-down features.

Table 1 Classification of top-down computational models.

	(a) Weight modulation of BU features	(b) Weighted combination of BU and TD saliency maps	(c) Joint learning of BU and TD features
(1) Target search	Guided search [80], [81], Navalpakkam *et al.* [83], VOCUS [84], SalBayes [86]	Guided search [6], [80], [81], VOCUS [84], Zelinsky *et al.* [85], Cerf *et al.* [87], Rao *et al.* [88], Contextual guidance [89], SUN [90], [91]	
(2) Other tasks		Peters *et al.* [92], Borji *et al.* [93], Yamada *et al.* [96], [97]	Judd *et al.* [94], Borji [95], Esaliency [63], Ozeki *et al.* [98]

Some typical computational models are classified according to the above two categories as shown in Table 1. To the best of our knowledge there are no models in the two blank cells. This shows that the biologically plausible approaches have been useful for computing saliency in target search tasks, while the psychophysics have not fully reported the findings needed to compute visual attention in the other tasks. We present an overview of these approaches in the next section.

5.1 Weight Modulation of Bottom-Up Features

The human performance of a visual target search depends on both the visual features of the target and those of the distractors [12], [15], [99]–[101]. Most weight modulation models learn each feature channel weight based on the differences between bottom-up features representing the target and the distractors (or background scenes). As regards the guided search model, the second version [80] gives more weight to feature channels that uniquely represent the target. The weighted response of each channel to the target is compared with its average response to the distractors. The channel with the greatest positive difference is selected to compute the top-down saliency map.

The signal-to-noise ratio (SNR), i.e., the ratio between target salience and distractor salience, is effective information for controlling the weights of feature channels. Navalpakkam *et al.* improved the Itti saliency map model by exploiting SNR maximization as an objective function of the weight modulation [83], [102] (Fig. 10). Frintrop *et al.* also proposed VOCUS [84], which directly applies SNRs computed from the feature channels of training images to their weights. The top-down saliency map results from the difference between the excitation map, which consists of the

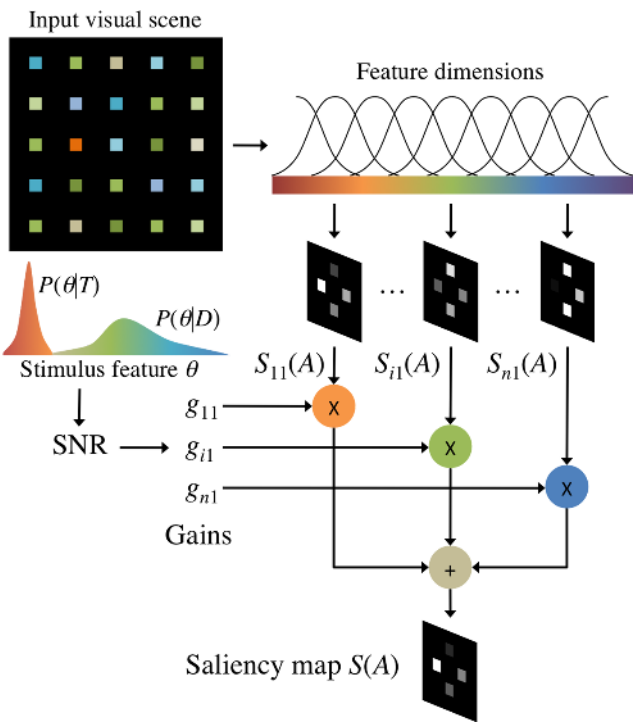

Fig. 10 Weight modulation of feature channels realized by maximization of SNR.

weighted responses of channels indicating SNR > 1, and the inhibition map, which is computed by SNR < 1. However, they never evaluated quantitatively the advantage of using the distractor knowledge to modulate the weight. Also, they do not confirm that the model acquires sufficient knowledge about every distractor, especially the background scene, to learn of the weight.

The above-mentioned computational models, which learn the optimum weight of each feature channel, multiply the learned value by every response of the channel. The uniform bias does not provide a greater ability to express visual saliency. One solution is to learn the likelihood distribution of the channel for the target. It can also deal with the uncertainty of an object's appearance. Elazary *et al.* proposed a Bayesian model called SalBayes, which regards the posterior probability that the visual features extracted from an image region belong to an object class as saliency [86]. The model recognizes the object and detects the target by means of maximum a posteriori probability estimation. It is worth remarking that SalBayes is a novel top-down and simple machine learning approach for computing visual attention in conjunction with object classification.

5.2 Weighted Combination of Bottom-Up and Top-Down Saliency Maps

(1) In visual target search

The guided search model, which was covered in the previous sections, is a pioneering model of the combination of bottom-up and top-down saliency [6], [80]. Bottom-up saliency is combined using equal weight with top-down saliency calculated using the selected feature channels shown in Sect. 5.1. The model has evolved through several versions. The latest version [81] has an additional top-down mechanism to handle a scene context and the inhibition tagging mechanism [103] to simulate visual attention more accurately. As a feature of the guided search, the guidance module of attention is separated from the main pathway to object recognition. VOCUS [84] also computes a global saliency map as the weighted sum of the top-down saliency map and the Itti saliency map. Giving more weight for the top-down saliency provided high performance in a complex scene that was even difficult for subjects to search.

Zelinsky *et al.* analyzed the relationships between the weighted parameters and the computed saliency maps [85]. The computation process first convolves a filter that simulates the spatial distortion of the retina [104] with an input image, and computes the bottom-up saliency based on several basic features (intensity, color, and orientation) and the top-down saliency based on the correlation between the model features of a target and an image region. They varied the weights of the combination of bottom-up and top-down saliency. The analysis revealed that the eye movements predicted using only the top-down saliency, that is with the bottom-up saliency completely-suppressed, closely matched human search behaviors. The finding did not agree with the observation by Navalpakkam *et al.* [83] that the visual search performance of a purely top-down model is extremely good beyond human ability. In contrast to this, Zelinsky *et al.* insisted that some traits of human vision such as central visual field, saccade distance limitation, and the inhibition of return (IOR) [105][†] contributed to improve performance.

A computational model specialized for one specific search target may optimize the visual search performance. Cerf *et al.* focused on the high-level feature of faces that fires a specific neuron in a nervous system [87]. The proposed model detects face regions using the Viola & Jones face detector [1] and computes the top-down saliency around the regions by convolving a Gaussian filter according to the size of the regions and the graph-based visual saliency [106] as the bottom-up saliency, and combines them using equal weight[††]. The model provided a good fit to human fixations not only in a face search task but also during free viewing. Its accuracy was better than that of the purely bottom-up saliency model. In addition, Walther *et al.* proposed a specific face search model that assigns a higher weight to the feature channel detecting skin hue [107].

Rao *et al.* revealed that human search performance improves by having subjects observe a scene before informing a target [88]. The prior knowledge of the scene includes visual information about both the target and the distractor. That is, the knowledge helps to promote a visual search as a top-down component. Rao's computational model simulates an initial saccade toward the center of the scene (center-of-gravity) by computing a weighted average of the top-down saliency map. The weighting is based on a weight function with a sharper peak near the center of a higher spatial resolution filter used to extract the basic features. It provided a very small error of 0.7 deg. between the model and the human gaze position.

Torralba *et al.* focused on the role of global features in a target search and proposed a Bayesian model of attentional guidance called the contextual guidance [89]. The model computes the probability of the presence of the target object at a location by integrating a pure bottom-up prior not depended on the target and a context-based prior on the location of the target as top-down knowledge of global features. In the experimental evaluation, they gave the subjects the task of counting target objects within an image and compared the model and human visual search behavior. The consistency when the model predicted the location of the target was superior to that of a purely bottom-up saliency model. The use of contextual knowledge improved the prediction of the first few fixations of the early stage of the search. However, it was poorer than the consistency of the fixations of the subjects.

Based on a Bayesian framework as well as the con-

[†]In psychology, this mechanism in visual search is not called inhibition of return but inhibitory tagging [103].

[††]The model does not employ only bottom-up processes to compute a higher saliency of face. We therefore grouped it under the top-down models.

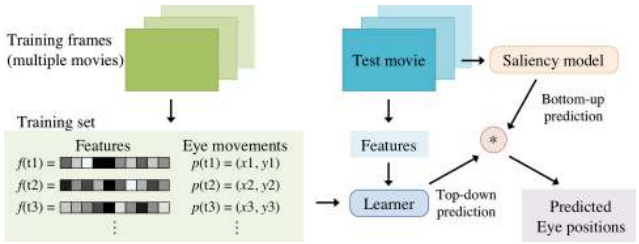

Fig. 11 Combination of the Itti bottom-up saliency map and the top-down prediction map learned from a series of visual features and the corresponding eye positions.

textual guidance, Zhang *et al.* proposed SUN (saliency using natural statistics) [90], [91] which incorporated top-down knowledge of target appearances, rather than the scene context. SUN computes a bottom-up prior and likelihood that denotes local features consistent with prior knowledge of the target appearance to derive the posterior probability that the local features belong to the target class. It can be interpreted in an information theoretic way by looking at the log salience. The probability can be computed using the probabilistic support vector machine for the ICA features presented in Sect. 4.5. The accuracy of the predicted fixations based on SUN was better than that achieved with the contextual guidance. Zhang *et al.* also showed that SUN can be regarded as a natural statistics-based saliency model simulating the visual search asymmetry [14], [15].

(2) In a situation other than a visual target search

Peters *et al.* addressed visual attention in an interactive visual task that consisted of playing a video game [92]. Their model learns to pair the basic visual features from a series of video clips with the corresponding eye positions that reflect the top-down factor; once trained, it generates a top-down eye position prediction map of previously unseen video frames. Finally, the Itti bottom-up saliency map and the top-down map are combined via point-wise multiplication (Fig. 11). The individual and combined maps were compared against the observed eye positions. The test results suggested that the predictions by the proposed model were better than those of the purely bottom-up model and top-down models. A concern is for over-training resulting from tasks undertaken by subjects with individual differences and skill improvement.

Borji *et al.* learned a Bayesian network that has some feature variables (scene gists [108], bottom-up saliency maps, game controllers, game events) connected to the corresponding eye positions and incorporated MEP (mean of the distribution of all training eye positions) as a prior distribution [109]. The approach obtained higher prediction of eye fixations than classical discriminative classifiers, including regression, support vector machine, and k nearest neighbor. They also proposed a framework that introduces a hidden Markov model to predict time-varying visual attention maps using a previous gaze point, subjects' inputs from game controllers, and the scene gist using features shared with a visual attention model [93].

Yamada *et al.* focused on human egocentric visual attention during walking [96], [97]. They generated an egomotion-based attention map by integrating motion maps using egomotion information and a purely bottom-up saliency map. The computation consists of the following steps: 1) estimating camera motion (rotation and translation) from an egocentric video that includes visual motions caused by human head motions, 2) computing angular velocity and generating a rotation-based attention map, 3) computing the focus of expansion (FOE) of a moving scene and generating a translation-based attention map, 4) combining the maps and the graph-based visual saliency [106] using equal weights. They demonstrated that the combination of the bottom-up saliency map and the rotation-based attention map could achieve the most accurate predictions of human attention in egocentric scenes [97], whereas they revealed that saliency maps using typical dynamic features (motion and flicker) reduced a prediction accuracy [96].

5.3 Joint Learning of Bottom-Up and Top-Down Features

Judd *et al.* measured human eye movements during a landscape and portrait image memory task [94]. They used the top-down controlled eye movement data as training and testing examples to learn a saliency model based on a large set of image features (low-level: 33 local features, middle-level: a horizontal line, and high-level: face, human, car region [110]). Each weight for a linear combination of the features was learned from the eye movement data. They demonstrated the importance of the center prior feature, which indicates the distance to the center for each pixel. A similar approach was recently proposed by Borji [95]. This approach describes saliency as being binary (i.e., salient or not salient), and it obtains the weights using classification algorithms such as support vector machine and Adaboost [111].

Avraham *et al.* asked their experimental subjects to mark the interesting objects in each scene and evaluated their proposed model Esaliency [63] (for details, see Section 4.4). In a human-robot interaction, the user gives the robot some tasks. To realize a natural interaction, the robot needs to control its attention according to the user's command. Ozeki *et al.* employed bottom-up features and face pose detection, and simulated the dynamic variation of visual attention using a particle filter [98]. The set of particles that approximates the spatial probability density distribution of the attention is distributed with weight placed on a saliency region close to the command such as "pay attention to the red object" and "establish joint attention with human partner".

6. Evaluating Computational Models

6.1 Evaluation Measures

As surveyed in the previous sections, there have been pro-

posed a huge variety of computational models and their implementations for human visual attention. This section introduces several evaluation measures commonly used in the state-of-the-art studies.

Given gaze data while examining targets, the strength of saliency at gaze locations is often evaluated. The normalized scan path saliency (NSS) is a measure of comparing the strength of saliency at gaze locations with the average strength of saliency in input images, which is employed in [50]–[52], [92], [106]. Moreover, the Kullback-Leibler divergence between saliency distributions sampled from gaze locations and those sampled at random is regarded as a measure to evaluate saliency map from videos [53].

Studies that assume search tasks including visual search can employ an evaluation measure that counts the number of shifts of gaze locations to find targets, by simulating such gaze shifts based on obtained saliency maps. This measure is employed not only in the pioneer work by Itti *et al.* [27] but in several other studies [49], [84], [85], [90].

Especially for the still input images, gaze data are often regarded as gaze-point distributions. Some studies estimated human attention map resulted from convolving a Gaussian kernel to the gaze-point distributions and measure the correlation between the human attention map and computed saliency maps [67], [112].

Other studies binalize a saliency map based on a threshold and evaluate how often human gazes locate at high salient regions (region of interests; ROI) [63], [86], [89]. Furthermore, the studies which aim at detection of ROI often compare the ROI obtained from their saliency map with those annotated manually [58], [61], [113]. In such cases no subjects nor gaze data are required for the evaluation. With this, salient region extraction techniques such as [114], [115] also employ this measure. Those techniques have different goals from the saliency computation as shown in Sect. 3. Consequently, it is important to evaluate methods appropriately based on their goals.

Several computational models have been quantitatively evaluated by Toet [116]. Qualitative evaluations are also released in https://sites.google.com/site/saliencyevaluation/home. In addition, Borji *e*t al. have presented a quantitative evaluation of several salient object detection (i.e., salient region extraction) techniques in [117].

6.2 Dataset for Model Evaluation

When evaluating visual attention models, datasets are usually required consisting of image or video for gaze target as well as the corresponding gaze data. CRCNS eye-1[†] (eye-1) is a dataset containing all of those data, which is mainly used in [53]. Eye-1 contains totally 50 videos consisting of TV programs, video games as well as artificial visual stimuli, and the corresponding gaze data obtained from 4 to 6 subjects. In addition, several research groups publish gaze datasets associated with static images or videos, as shown in the following list. Notice that this list and the list in the next section include not only the models introduced in this survey paper but several other models excluded.

- Bruce et al. [67] (http://www-sop.inria.fr/members/Neil.Bruce/)
- Torralba et al. [89] (http://people.csail.mit.edu/torralba/GlobalFeaturesAndAttention/)
- Judd et al. [94] (http://people.csail.mit.edu/tjudd/WherePeopleLook/)
- Cerf et al. [87] (http://www.fifadb.com/)
- Le Meur et al. [118] (http://www.irisa.fr/temics/staff/lemeur/visualAttention/)
- Ehinger et al. [119] (http://cvcl.mit.edu/searchmodels/)
- Rajashekar et al. [120] (http://live.ece.utexas.edu/research/doves/)
- The DIEM (Dynamic Images and Eye Movements) Project (http://thediemproject.wordpress.com/)
- NUSEF: The National University of Singapore Eye-Fixation database (http://mmas.comp.nus.edu.sg/NUSEF.html)

6.3 Open Source Codes

Finally, here is a partial list of open source codes which implement existing computational models. See the links to access the detail of the codes.

- Biologically-plausible models
  - iLab Neuromorphic Vision C++ Toolkit [27], [53] (http://ilab.usc.edu/toolkit/)
  - Saliency Toolbox [27], [121] (http://www.saliencytoolbox.net)
  - Graph-based visual saliency [106] (http://www.klab.caltech.edu/~harel/share/gbvs.php)
  - Implementation of Itti saliency map with OpenCV [27] (http://pub.ne.jp/akisato/?entry_id=4437100)
  - The bottom-up visual saliency of Itti et al. [27] to run on the Nokia N810 internet tablet (http://maemo.org/downloads/product/OS2008/saliency/)
- Novelty-based models
  - Spectral residual [58] (http://www.klab.caltech.edu/~xhou/projects/spectralResidual/spectralresidual.html)
  - Frequency-tuned salient region detection [61], [62] (http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/)
  - Esaliency [63] (http://isl.cs.technion.ac.il/index.php/research/research-projects/30)
  - The incremental coding length [122] (http://www.klab.caltech.edu/~xhou/projects/dva/dva.html)

[†]http://crcns.org/data-sets/eye/eye-1

- Knowledge-based models
  - Saliency based on information maximization [67] (http://www-sop.inria.fr/members/Neil.Bruce/)
- Top-down models
  - Learning to predict where to look [94] (http://people.csail.mit.edu/tjudd/WherePeopleLook/)
  - Predicting human gaze using low-level saliency combined with face detection [87] (http://www.fifadb.com/)
  - A combined source model of eye guidance [119] (http://live.ece.utexas.edu/research/doves/)

A list of model implementations as well as performances is also provided at http://people.csail.mit.edu/tjudd/SaliencyBenchmark/.

7. Concluding Remarks

This paper overviewed computational models of human visual attention and their implementations, which consists of bottom-up models that compute a saliency map from images (or videos) and top-down models that exploit some knowledge on human cognitive states, tasks, or prior knowledge of the images. Both bottom-up and top-down models were individually classified into several types based on their mathematical backgrounds or the formulations. We conclude the paper with foresight of in models of human visual attention, which includes not only a pure computational model of visual attention but salient region extraction.

In computer science communities, traditional biologically-plausible models of bottom-up saliency, which include Itti's saliency map [27], are basically utilized to serve a baseline, and now novelty-based and knowledge-based models seem to be dominant. Whereas novelty-based models are well studied for these 5 years, knowledge-based models do not mature so much. The knowledge-based models have a potential to apply various machine learning techniques. For instance, Li *et al.* [123] introduced multi-task learning to simulate the conjunction search (cf. Sect. 2.1), where each task corresponds to a simple function of feature search such as color, intensity or edge orientation.

On top-down models, many weighted combination and joint-learning models have been introduced recently. Top-down information on target characteristics is easy to be introduced because of its familiarity with various pattern recognition and computer vision techniques such as Viola-Jones face detector [1] and deformable part models [110].

On the other hand, it still remains undiscovered to build practical models and implementation with human cognitive states because of the difficulties in experimental setup and validation. As shown in Sects. 2.4 and 5, many psychophysical findings and conceptual models on such human-states-related aspects have been already reported. When experimental setups including the introductions of eye trackers become much easier, top-down models with human states will be one of the important topics in the future.

Finally, visual attention models have many applications. Recently the models have been used to boost some computer vision and pattern recognition techniques such as object detection [113], [124], [125], object recognition [126]–[131], action recognition [132], [133], segmentation [37], [114], [115], [134], [135] and background subtraction [136]. Besides, specific applications include video summarization [137] and compression [138], scene understanding [139]–[141], computer-human interaction [98], [142]–[147], robotics [132], [148]–[150], and driver assistance [151], [152]. The potential of the visual attention models that are capable of extracting important regions promises their contributions to many other domains.

Acknowledgements

We thank the IEICE Editorial Committee for the opportunity to write this paper. We also thank Prof. Tomohiro Shibara of Nara Advanced Institute of Science and Technology (NAIST), Mr. Kazuhiko Kojima of Sanyo Electric Co.Ltd, and anonymous associate editors and reviewers for their valuable discussions and useful comments, which helped to improve of this work.

References

[1] P. Viola and M. Jones, "Robust real-time face detection," Int. J. Comput. Vis., vol.57, no.2, pp.137–154, 2004.
[2] M. Enzweiler and D.M. Gavrila, "Monocular pedestrian detection: Survey and experiments," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, pp.2179–2195, 2009.
[3] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," Human Neurobiology, vol.4, pp.219–227, 1985.
[4] U. Neisser and H. Beller, "Searching through word lists," British Journal of Psychology, vol.56, pp.349–358, 1965.
[5] A. Treisman and G. Gelade, "A feature-integration theory of attention," Cognitive Psychology, vol.12, pp.97–136, 1980.
[6] J. Wolfe, K. Cave, and S. Franzel, "Guided search: An alternative to the feature integration model for visual search," J. Experimental Psychology: Human Perception and Performance, vol.15, no.3, pp.419–433, 1989.
[7] K. Cave and J. Wolfe, "Modeling the role of parallel processing in visual search," Cognitive Psychology, vol.22, no.2, pp.225–271, 1990.
[8] J. Palmer, "Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks," Vision Research, vol.34, no.13, pp.1703–1721, 1994.
[9] K. Nakayama, "The iconic bottleneck and the tenuous link between early visual processing and perception," in Vision: coding and efficiency, Cambridge University Press, 1990.
[10] M. Bravo and K. Nakayama, "The role of attention in different visual-search tasks," Perception & Psychophysics, vol.51, pp.465–472, 1992.
[11] J. Duncan, "Boundary conditions on parallel processing in human vision," Perception, vol.18, pp.457–469, 1989.
[12] J. Duncan and G. Humphreys, "Visual search and stimulus similarity," Psychological Review, vol.96, pp.433–458, 1989.
[13] J. Duncan and G. Humphreys, "Beyond the search surface: Visual search and attentional engagement," J. Experimental Psychology: Human Perception and Performance, vol.18, pp.578–588, 1992.
[14] A. Treisman and S. Gormican, "Feature analysis in early vision: Evidence from search asymmetries," Psychol Review, vol.95, no.1,

pp.15–48, Jan 1988.
[15] J. Wolfe, "Asymmetries in visual search: An introduction," Attention, Perception, & Psychophysics, vol.63, no.3, pp.381–389, 2001.
[16] S. Frintrop, E. Rome, and H. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," ACM Trans. Applied Perception, vol.7, no.1, pp.1–39, 2010.
[17] A. Yarbus, "Eye movements and vision," Plenum, 1967.
[18] B. Tatler, N. Wade, H. Kwan, J. Findlay, and B. Velichkovsky, "Yarbus, eye movements, and vision," i-Perception, vol.1, no.1, pp.7–27, 2010.
[19] M. Hayhoe and D. Ballard, "Eye movements in natural behavior.," Trends in Cognitive Sciences, vol.9, no.4, pp.188–194, 2005.
[20] C. Balkenius, "Attention, habituation and conditioning: Toward a computational model," Cognitive Science Quarterly, vol.1, pp.171–214, 2000.
[21] J. Taylor and N. Fragopanagos, "Modelling the interaction of attention and emotion," Proc. International Joint Conference on Neural Networks (IJCNN), pp.1663–1668, 2005.
[22] M.I. Posner, "Orienting of attention," The Quarterly Journal of Experimental Psychology, vol.32, no.1, pp.3–25, 1980.
[23] J.W. Bisley and M.E. Goldberg, "Neuronal activity in the lateral intraparietal area and spatial attention," Science, vol.299, no.5603, pp.81–86, 2003.
[24] J. Duncan, "Selective attention and the organization of visual information," J. Experimental Psychology: General, vol.113, no.4, pp.501–517, Dec. 1984.
[25] B.J. Scholl, "Objects and attention: The state of the art," Cognition, vol.80, no.1-2, pp.1–46, June 2001.
[26] W. Einhauser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," J. Vision, vol.8, no.14, pp.1–26, 2008.
[27] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol.20, no.11, pp.1254–1259, 1998.
[28] H. Müller, G. Humphreys, and N. Donnelly, "SEarch via Recursive Rejection (SERR): Visual search for single and dual form-conjunction targets," J. Experimental Psychology-human Perception and Performance, vol.20, no.2, pp.235–258, 1994.
[29] S. Peter, "Simulating visual attention," J. Cognitive Neuroscience, vol.2, pp.213–231, 1990.
[30] O. Bruno, A. Charles, and V. David, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," J. Neuroscience, vol.13, pp.4700–4719, 1993.
[31] S. Frintrop, "Computational visual attention," in Computer Analysis of Human Behavior, pp.69–101, Springer, 2011.
[32] A. Kimura, "A stochastic model of human visual attention: Latest approach based on dynamic Bayesian networks (in Japanese)," Proc. IEICE Signal Processing Symposium, 2010.
[33] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," Vision Research, vol.40, no.10-12, pp.1489–506, 2000.
[34] L. Itti, "Real-time high-performance attention focusing in outdoors color video streams," SPIE Human Vision and Electronic Imaging (HVEI), pp.235–243, 2002.
[35] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," SPIE International Symposium on Optical Science and Technology, pp.64–78, 2003.
[36] S. Li and M. Lee, "An efficient spatiotemporal attention model and its application to shot matching," IEEE Trans. Circuits Syst. Video Technol., vol.17, no.10, pp.1383–1387, 2007.
[37] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," Proc. IEEE International Conference on Multimedia and Expo (ICME), pp.638–641, June 2009.
[38] C. Leung, A. Kimura, T. Takeuchi, and K. Kashino, "A computational model of saliency depletion/recovery phenomena for the salient region extraction of videos," Proc. IEEE International Conference on Multimedia and Expo (ICME), pp.300–303, 2007.
[39] H. Hartline, "The nerve messages in the fibers of the visual pathway," J. Optics Society of America, vol.30, pp.239–247, 1940.
[40] A. Maki, P. Nordlund, and J. Eklundh, "A computational model of depth-based attention," Proc. IAPR International Conference on Pattern Recognition (ICPR), pp.734–739, Aug. 1996.
[41] K. Nakayama and G.H. Silverman, "Serial and parallel processing of visual feature conjunctions," Nature, vol.320, no.6059, pp.264–265, March 1986.
[42] N. Ouerhani and H. Hugli, "Computing visual attention from scene depth," Proc. IAPR International Conference on Pattern Recognition (ICPR), pp.375–378, 2000.
[43] S. Jeong, S.W. Ban, and M. Lee, "Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment," Neural Netw., vol.21, no.10, pp.1420–1430, 2008.
[44] B. Scholl, "Objects and attention: The state of the art," Cognition, vol.80, no.1-2, pp.1–46, 2001.
[45] W. Peterson, T. Birdsall, and W. Fox, "The theory of signal detectability," IRE Trans. Inf. Theory, vol.4, pp.171–212, 1954.
[46] W. Tanner and J. Swets, "A decision-making theory of visual detection," Psychological Review, vol.61, pp.401–409, 1954.
[47] M. Eckstein, J. Thomas, J. Palmer, and S. Shimozaki, "A signal detection model predicts effects of set size on visual search accuracy for feature, conjunction, triple conjunction and disjunction displays," Perception and Psychophysics, vol.62, pp.425–451, 2000.
[48] P. Verghese, "Visual search and attention: A signal detection theory approach," Neuron, vol.31, pp.525–535, Aug. 2001.
[49] T. Koike and J. Saiki, "Stochastic saliency-based search model for search asymmetry with uncertain targets," Neurocomputing, vol.69, pp.2112–2126, 2006.
[50] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino, "A stochastic model of selective visual attention with a dynamic Bayesian network," Proc. IEEE International Conference on Multimedia and Expo (ICME), pp.1073–1076, 2008.
[51] A. Kimura, D. Pang, T. Takeuchi, K. Miyazato, J. Yamato, and K. Kashino, "A stochastic model of human visual attention with a dynamic Bayesian network," arxiv.org, pp.1–13, 2010.
[52] K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Real-time estimation of human visual attention with dynamic Bayesian network and MCMC-based particle filter," Proc. IEEE International Conference on Multimedia and Expo (ICME), pp.250–257, 2009.
[53] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," Vision Research, vol.49, no.10, pp.1295–306, 2009.
[54] P. Baldi and L. Itti, "Of bits and wows: A Bayesian theory of surprise with applications to attention," Neural Netw., vol.23, no.5, pp.649–666, 2010.
[55] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," Int. J. Comput. Vis., vol.82, no.3, pp.231–243, May 2009.
[56] W. Reichardt, "Evaluation of optical motion information by movement detectors," J. Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology, vol.161, pp.533–547, 1987.
[57] R. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," Vision Research, vol.46, no.26, pp.4333–4345, 2006.
[58] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2007.
[59] P. Bak, C. Tang, and K. Wiesenfeld, "Self-organized criticality: An explanation of the 1/*f* noise," Phys. Rev. Lett., vol.59, pp.381–384, July 1987.
[60] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection

using phase spectrum of quaternion Fourier transform," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2008.

[61] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," Proc. International Conference on Computer Vision Systems (ICVS), pp.66–75, 2008.

[62] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1597–1604, 2009.

[63] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): Meaningful attention using stochastic image modeling," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.4, pp.693–708, 2010.

[64] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," Proc. Conference on Neural Information Processing Systems (NIPS), 2007.

[65] D. Gao and N. Vasconcelos, "Decision-theoretic saliency: Computational principles, biological plausibility, and implications for neurophysiology and psychophysics," Neural Comput., vol.21, pp.239–271, 2009.

[66] H.J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," J. Vision, vol.9, no.12, pp.1–27, 2009.

[67] N. Bruce and J. Tsotsos, "Saliency based on information maximization," Proc. Conference on Neural Information Processing Systems (NIPS), vol.18, p.155, 2006.

[68] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature, vol.381, no.6583, pp.607–609, 1996.

[69] A.J. Bell and T.J. Sejnowski, "The "independent components" of natural scenes are edge filters," Vision Research, vol.37, no.23, pp.3327–3338, Dec. 1997.

[70] L. Renninger, J. Coughlan, P. Verghese, and J. Malik, "An information maximization model of eye movements," Proc. Conference on Neural Information Processing Systems (NIPS), pp.1121–1128, 2005.

[71] Y. Li, Y. Zhou, J. Yan, Z. Niu, and J. Yang, "Visual saliency based on conditional entropy," Proc. Asian Conference on Computer Vision (ACCV), pp.246–257, 2010.

[72] X. Sun, H. Yao, and R. Ji, "What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1552–1559, 2012.

[73] X. Sun, H. Yao, R. Ji, P. Xu, X. Liu, and S. Liu, "Visual saliency as sequential eye fixation probability," Proc. IEEE International Conference on Image Processing (ICIP), pp.1093–1096, Sept. 2010.

[74] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, Adaptive and Learning Systems for Signal Processing, Communications, and Control, John Wiley & Sons, 2001.

[75] S. Amari, A. Cichocki, and H.H. Yang, "A new learning algorithm for blind signal separation," Proc. Conference on Neural Information Processing Systems (NIPS), pp.757–763, 1995.

[76] A. Cichocki and S. Amari, Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications, John Wiley & Sons, 2002.

[77] W. Kienzle, M.O. Franz, B. Schölkopf, and F.A. Wichmann, "Center-surround patterns emerge as optimal predictors for human saccade targets," J. Vision, vol.9, no.5, pp.1–15, 2009.

[78] Y.F. Ma and H.J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," Proc. ACM International Conference on Multimedia (ACMMM), pp.374–381, 2003.

[79] Y.F. Ma, X. Hua, L. Lu, and H.J. Zhang, "A generic framework of user attention model and its application in video summarization," IEEE Trans. Multimed., vol.7, no.5, pp.907–919, 2005.

[80] J. Wolfe, "Guided search 2.0: A revised model of visual search," Psychonomic Bulletin & Review, vol.1, no.2, pp.202–238, 1994.

[81] J. Wolfe, "Guided search 4.0: Current progress with a model of visual search," Integrated Models of Cognitive Systems, vol.1, no.3, pp.99–119, 2007.

[82] J.H. Fecteau and D.P. Munoz, "Salience, relevance, and firing: A priority map for target selection," Trends in Cognitive Sciences, vol.10, no.8, pp.382–390, 2006.

[83] V. Navalpakkam and L. Itti, "Search goal tunes visual features optimally," Neuron, vol.53, pp.605–617, 2007.

[84] S. Frintrop, VOCUS: A visual attention system for object detection and goal-directed search, Springer, 2006.

[85] G. Zelinsky, W. Zhang, B. Yu, and X. Chen, "The role of top-down and bottom-up processes in guiding eye movements during visual search," Proc. Conference on Neural Information Processing Systems (NIPS), pp.1569–1576, 2006.

[86] L. Elazary and L. Itti, "A Bayesian model for efficient visual search and recognition," Vision Research, vol.50, no.14, pp.1338–1352, 2010.

[87] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," Proc. Conference on Neural Information Processing Systems (NIPS), pp.1–8, 2007.

[88] R. Rao, G. Zelinsky, M. Hayhoe, and D. Ballard, "Eye movements in iconic visual search," Vision Research, vol.42, no.11, pp.1447–1463, 2002.

[89] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," Psychological Review, vol.113, no.4, pp.766–789, 2006.

[90] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," J. Vision, vol.8, no.7, pp.32.1–20, 2008.

[91] C. Kanan, M. Tong, L. Zhang, and G. Cottrell, "SUN: Top-down saliency using natural statistics," Visual Cognition, vol.17, no.6-7, pp.979–1003, 2009.

[92] R. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2007.

[93] A. Borji, D. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2012.

[94] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," Proc. International Conference on Computer Vision (ICCV), pp.2106–2113, 2009.

[95] A. Borji, "Boosting bottom-up and top-down visual features for saliency detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2012.

[96] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki, "Can saliency map models predict human egocentric visual attention?," Proc. Asian Conference on Computer Vision (ACCV), pp.420–429, 2010.

[97] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki, "Attention prediction in egocentric video using motion and visual saliency," Proc. Pacific-Rim Symposium on Image and Video Technology (PSIVT), pp.277–288, 2011.

[98] M. Ozeki, Y. Kashiwagi, M. Inoue, and N. Oka, "Top-down visual attention control based on a particle filter for human-interactive robots," Proc. International Conference on Human System Interaction (ICHSI), pp.188–194, 2011.

[99] V. Navalpakkam and L. Itti, "Top-down attention selection is fine grained," J. Vision, vol.6, no.11, pp.1180–1193, 2006.

[100] J. Hodsoll and G. Humphreys, "Driving attention with the top down: The relative contribution of target templates to the linear separability effect in the size dimension," Perception & psychophysics, vol.63, no.5, pp.918–926, 2001.

[101] J. Wolfe and T. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?," Nature Reviews Neuroscience, vol.5, no.6, pp.495–501, 2004.

[102] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2049–2056, 2006.
[103] R.M. Klein, "Inhibitory tagging system facilitates visual search," Nature, vol.334, pp.430–431, 1988.
[104] J. Perry and W. Geisler, "Gaze-contingent real-time simulation of arbitrary visual fields," SPIE Human Vision and Electronic Imaging, pp.57–69, 2002.
[105] M. Posner and Y. Cohen, "Components of visual orienting," in Attention and Performance, pp.531–556, Erlbaum, 1984.
[106] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," Proc. Conference on Neural Information Processing Systems (NIPS), pp.545–552, 2007.
[107] D. Walther, "Interactions of visual attention and object recognition: Computational modeling, algorithms, and psychophysics," Ph.D. Thesis, California Institute of Technology, 2006.
[108] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," IEEE Trans. Pattern Anal. Mach. Intell., vol.29, no.2, pp.300–312, 2007.
[109] A. Borji, D. Sihite, and L. Itti, "Computational modeling of top-down visual attention in interactive environments," Proc. British Machine Vision Conference (BMVC), pp.85.1–85.12, Sept. 2011.
[110] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2008.
[111] Y. Freund and R. Schapire, "A decicion-theoretic generalization of on-line learning and an application to boosting," in Computational Learning Theory, ed. P. Vitányi, Lect. Notes Comput. Sci., vol.904, pp.23–37, Springer, 1995.
[112] N. Ouerhani, R. von Wartburg, H. Hugli, and R. Muri, "Empirical validation of the saliency-based model of visual attention," Electronic Letters on Computer Vision and Image Analysis (ELCVIA), vol.3, no.1, pp.13–24, 2003.
[113] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.6, pp.989–1005, 2009.
[114] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.409–416, 2011.
[115] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng, "Automatic salient object segmentation based on context and shape prior," Proc. British Machine Vision Conference (BMVC), pp.110.1–110.12, 2011.
[116] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," IEEE Trans. Pattern Anal. Mach. Intell., vol.33, no.11, pp.1–18, 2011.
[117] A. Borji, D.N. Sihite, and L. Itti, "Salient object detection: A benchmark," Proc. European Conference on Computer Vision (ECCV), 2012.
[118] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," IEEE Trans. Pattern Anal. Mach. Intell., vol.28, no.5, pp.802–817, 2006.
[119] K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modeling search for people in 900 scenes: A combined source model of eye guidance," Visual Cognition, vol.18, no.6-7, pp.945–978, 2009.
[120] U. Rajashekar, I. van der Linde, A. Bovik, and L. Cormack, "Gaffe: A gaze-attentive fixation finding engine," IEEE Trans. Pattern Anal. Mach. Intell., vol.17, no.4, pp.564–573, 2008.
[121] D. Walther and C. Koch, "Modeling attention to salient proto-objects," Proc. International Joint Conference on Neural Networks (IJCNN), pp.1395–1407, 2006.
[122] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," Proc. Conference on Neural Information Processing Systems (NIPS), pp.681–688, 2008.
[123] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," Int. J. Comput. Vis., vol.90, no.2, pp.150–165, Nov. 2010.
[124] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li, "Salient object detection for searched web images via global saliency," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2012.
[125] J. Yang and M.H. Yang, "Top-down visual saliency via joint CRF and dictionary learning," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2012.
[126] N. Ouerhani and H. Hügli, "A model of dynamic visual attention for object tracking in natural image sequences," Proc. International Work-Conference on Artificial Neural Networks (IWANN), pp.702–709, 2003.
[127] N. Ouerhani, H. Hügli, G. Gruener, and A. Codourey, "A visual attention-based approach for automatic landmark selection and recognition," Proc. International Workshop on Attention and Performance on Computational Vision (WAPCV), pp.183–195, 2004.
[128] S. Frintrop, A. Nüchter, H. Surmann, and J. Hertzberg, "Saliency-based object recognition in 3D data," Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.2167–2172, 2004.
[129] D. Walther and C. Koch, "Attention in hierarchical models of object recognition.," Progress in Brain Research, vol.165, no.6, pp.57–78, 2007.
[130] J.Y. Zhu, J. Wu, Y. Wei, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2012.
[131] G. Sharma, F. Jurie, and C. Schmid, "Discriminative spatial saliency for image classification," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2012.
[132] Y. Nagai, "From bottom-up visual attention to robot action learning," IEEE International Conference on Development and Learning (ICDL), pp.5198–5203, 2009.
[133] Y. Nagai, C. Mühl, and K. Rohlfing, "Bottom-up attention improves action recognition using histograms of oriented gradients," Proc. IAPR Conference on Machine Vision Applications (MVA), pp.467–470, 2011.
[134] Y. Fu, J. Cheng, Z. Li, and H. Lu, "Saliency cuts: An automatic approach to object segmentation," Proc. IAPR International Conference on Pattern Recognition (ICPR), pp.1–4, 2008.
[135] K. Fukuda, T. Takiguchi, and Y. Ariki, "Automatic segmentation of object region using graph cuts based on saliency maps and adaBoost," IEEE International Symposium on Consumer Electronics (ISCE), pp.36–37, 2009.
[136] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.1, pp.171–177, Jan. 2010.
[137] Y.F. Ma, X.S. Hua, L. Lu, and H.J. Zhang, "A generic framework of user attention model and its application in video summarization," IEEE Trans. Multimed., vol.7, no.5, pp.907–919, 2005.
[138] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," IEEE Trans. Image Process., vol.13, no.10, pp.1304–1318, 2004.
[139] T. Jost, N. Ouerhani, R. von Wartburg, R. Müri, and H. Hügli, "Assessing the contribution of color in visual attention," Computer Vision and Image Understanding (CVIU), vol.100, no.1, pp.107–123, 2005.
[140] Y. Nakashima, N. Babaguchi, and J. Fan, "Automatically protecting privacy in consumer generated videos using intended human object detector," Proc. ACM International Conference on Multimedia (ACMMM), pp.1135–1138, 2010.
[141] A. Hagiwara, A. Sugimoto, and K. Kawamoto, "Saliency-based

image editing for guiding visual attention," Proc. International Workshop Pervasive Eye Tracking and Mobile Eye-based Interaction (PETMEI), pp.43–48, 2011.
[142] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, "Overt visual attention for a humanoid robot," Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.2332–2337, 2001.
[143] G. Heidemann, R. Rae, H. Bekel, I. Bax, and H. Ritter, "Integrating context-free and context-dependent attentional mechanisms for gestural object reference," Proc. International Conference on Computer Vision Systems (ICVS), pp.22–33, 2003.
[144] M. Hikita, S. Fuke, M. Ogino, T. Minato, and M. Asada, "Visual attention by saliency leads cross-modal body representation," Proc. IEEE International Conference on Development and Learning (ICDL), pp.157–162, 2008.
[145] Y. Sugano, Y. Matsushita, and Y. Sato, "Calibration-free gaze sensing using saliency maps," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2667–2674, 2010.
[146] R. Yonetani, H. Kawashima, and T. Matsuyama, "Multi-mode saliency dynamics model for analyzing gaze and attention," Proc. ACM Symposium on Eye Tracking Research & Applications (ETRA), 2012.
[147] H. Kubota, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki, "Incorporating visual field characteristics into a saliency map," Proc. ACM Symposium on Eye Tracking Research and Applications (ETRA), pp.1507–1514, 2012.
[148] S. Frintrop and P. Jensfelt, "Active gaze control for attentional visual SLAM," Proc. IEEE International Conference on Robotics and Automation (ICRA), pp.3690–3697, 2008.
[149] Y. Nagai and K. Rohlfing, "Computational analysis of motionese toward scaffolding robot action learning," IEEE Trans. Autonomous Mental Development, vol.1, no.1, pp.44–54, 2009.
[150] Y. Nagai, C. Mühl, and K. Rohlfing, "Toward designing a robot that learns actions from parental demonstrations," Proc. IEEE International Conference on Robotics and Automation (ICRA), pp.3545–3550, 2008.
[151] A. Doshi and M. Trivedi, "Investigating the relationships between gaze patterns, dynamic vehicle surround analysis, and driver intentions," Proc. IEEE Intelligent Vehicles Symposium (IVS), pp.887–892, 2009.
[152] K. Doman, D. Deguchi, T. Takahashi, Y. Mekada, I. Ide, H. Murase, and Y. Tamatsu, "Estimation of traffic sign visibility considering temporal environmental changes for smart driver assistance," Proc. IEEE Intelligent Vehicles Symposium (IVS), pp.667–672, 2011.

Akisato Kimura received his B.E., M.E. and D.E. degrees in Communications and Integrated Systems from Tokyo Institute of Technology, Japan in 1998, 2000 and 2007, respectively. Since 2000, he has been with NTT Communication Science Laboratories, NTT Corporation, where he is currently a senior research scientist in Innovative Communication Laboratory. He has been engaged in work on multimedia content identification, computational models of human visual attention, automatic image/audio/video annotation, and multimedia mining. His research interests include pattern recognition, computer vision, image processing, human visual perception, machine learning and social media. He is a senior member of IEEE and a member of ACM SIGMM/SIGKDD.

Ryo Yonetani received his B.E. degree in electrical and electronic engineering and M.S. degree in informatics from Kyoto University, Japan in 2009 and 2011, respectively. He is currently a Ph.D. candidate at the Graduate School of Informatics, Kyoto University. His research interests include pattern recognition, computer vision and human-computer interaction. He received the IBM Best Student Paper Award at the International Conference on Pattern Recognition '10. He is a student member of IPSJ.

Takatsugu Hirayama received his M.E. and D.E. degrees in engineering science from Osaka University, Japan in 2002 and 2005, respectively. From 2005 to 2011, he was a research assistant professor in the Graduate School of Informatics, Kyoto University. He is currently an assistant professor in the Graduate School of Information Science, Nagoya University. His research interests include computer vision, human vision, human communication, and human-computer interaction. He is a member of IPSJ, the Human Interface Society of Japan, and ACM.