

Chapter 2

Background

Although in principle all attention models serve the same purpose, i.e. to highlight potentially relevant and thus interesting—that is to say “salient”—data, attention models can differ substantially in which parts of the signal they mark as being of interest. This is to a great extent due to the varying research questions and interests in relevant fields such as, most importantly, neuroscience, psychophysics, psychology, and computer science. However, it is also caused by the vagueness as well as application- and task-dependence of the underlying problem description, i.e. what is interesting?

The purpose of this chapter is to provide an introduction to visual and auditory attention (Sect. 2.1) and its applications (Sect. 2.2) that serves as background information for the remainder of this book.

2.1 Attention Models

In general, it is possible to distinguish three types of attention models by the respective research field: First, neurobiological models try to understand and model in which part of the brain attentional mechanisms reside and how they operate and interact on a neurobiological level. Second, psychological models try to model, explain, and better understand aspects of human perception and not the brain’s neural system and layout. Third, computational models implement principles of neurobiological and psychological models, but they are also often subject to an engineering objective. Such an engineering objective is less to model the human brain or perception, but to be part of and improve artificial systems such as, e.g., vision systems or complex robots.

For visual attention, the following text focuses on computational and to a lesser extent psychological models, because well-studied, elaborated psychological and computational models exist. Furthermore, a deep understanding of neurobiological aspects of the human brain’s neural visual system is of minor relevance for the

remainder of this book. An interesting complementary lecture to this section is the excellent survey by Frintrop et al. [FRC10], which specifically tries to explain attention related concepts and ideas across the related fields of neurobiology, psychology, and computer science. For auditory attention, it is necessary to address neurobiological aspects of the human auditory system, because concise elaborated psychological and computational do not exist and a basic understanding of the human auditory system is important to understand the motivation of proposed computational models. Here, Fritz et al. and Hafer et al. provide very good neurobiological overviews of auditory attention [FEDS07, HSL07].

2.1.1 Visual Attention

Psychological Models

The objective of psychological attention models is to explain and better understand human perception, not to model the brain's neural structure. Among the psychological models, the feature integration theory (FIT) by Treisman et al. [TG80] and Wolfe et al.'s guided search model (GSM) [Wol94] are probably by far the most influential models. Aspects of both models are still present in modern models and both models have constantly been adapted to incorporate later research findings. A deeper discussion of psychological models can be found in the review by Bundesen and Habekost [BH05].

Treisman's feature integration theory [TG80], see Fig. 2.1, assumes that "different features are registered early, automatically, and in parallel across the visual fields, while objects are identified separately and only at a later stage, which requires focused attention" [TG80]. This simple description includes various aspects that are still fundamental for psychological and computational attention models. Conspicuity in a feature channel are represented in topological "conspicuity" or "feature maps". The information from the feature map is integrated in a "master map of location". A concept that is nowadays most widely known as "saliency map" [KU85]. This master map of location encodes "where" things are in an image, but not "what" they are, which reflects the "where" and "what" pathways in the human brain [FRC10]. Attention is serially focused on the highlighted locations in the master map and the image data around the attended location is passed as data to higher perception tasks such as, most importantly, object recognition to answer "what" is shown at that location.

Although Treisman's early model primarily focused on bottom-up perceptual saliency, Treisman also considered how attention is affected during visual search, i.e. when looking for specific target objects. A target is easier—i.e., faster—to find during visual search the more distinctive features it exhibits that differentiate it from the distractors. To implement visual search mechanism in FIT, Treisman proposed to inhibit the feature maps that encode the features of distractors, i.e. non-target features.

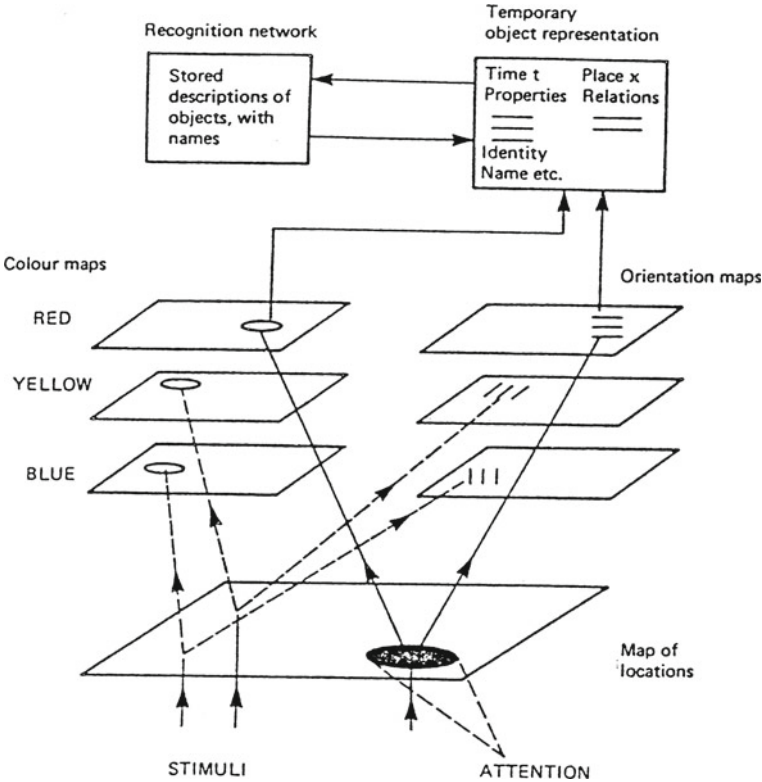


Fig. 2.1 Treisman’s feature integration theory model. Image from [TG88] reprinted with permission from APA

Treisman et al. also introduced the concept of object files as “temporary episodic representations of objects”. An object file “collects the sensory information that has so far been received about the object. This information can be matched to stored descriptions to identify or classify the object” [KTG92].

Wolfe et al. [WCF89, Wol94] introduced the initial guided search model to address shortcomings of early versions of Treisman’s FIT model, see Fig. 2.2. As its name suggests, Wolfe’s GSM focuses on modeling and predicting the results of visual search experiments. Accordingly, it explicitly integrates the influence of top-down information to highlight potential target objects during visual search. For this purpose, it uses the top-down information to select the feature type that best distinguishes between target and distractors.

Computational Models—Traditional Structure

Most computational attention models follow a similar structure, see Fig. 2.3, which is adopted from Treisman’s feature integration theory [TG80] and Wolfe’s guided search model [WCF89, Wol94] (see Figs. 2.1 and 2.2, respectively). The first com-

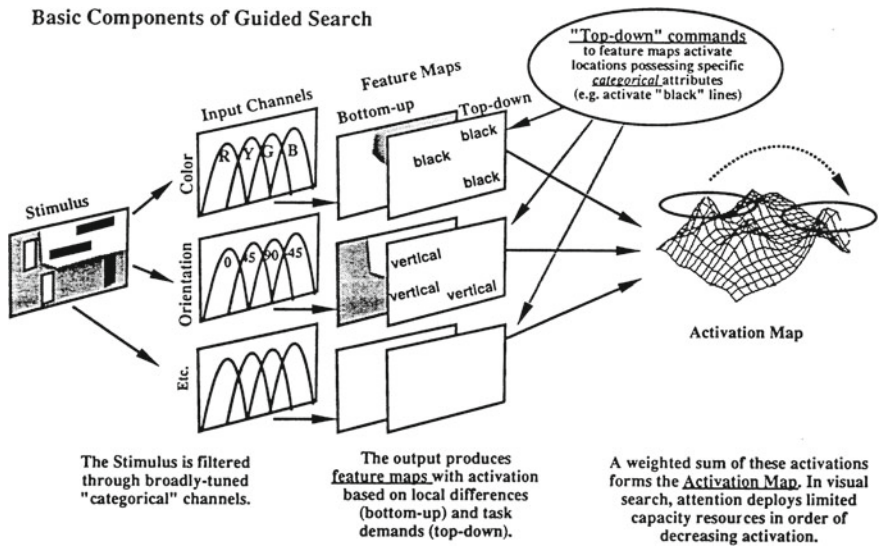


Fig. 2.2 Wolfe’s guided search model. Image from [Wol94] reprinted with permission from Springer

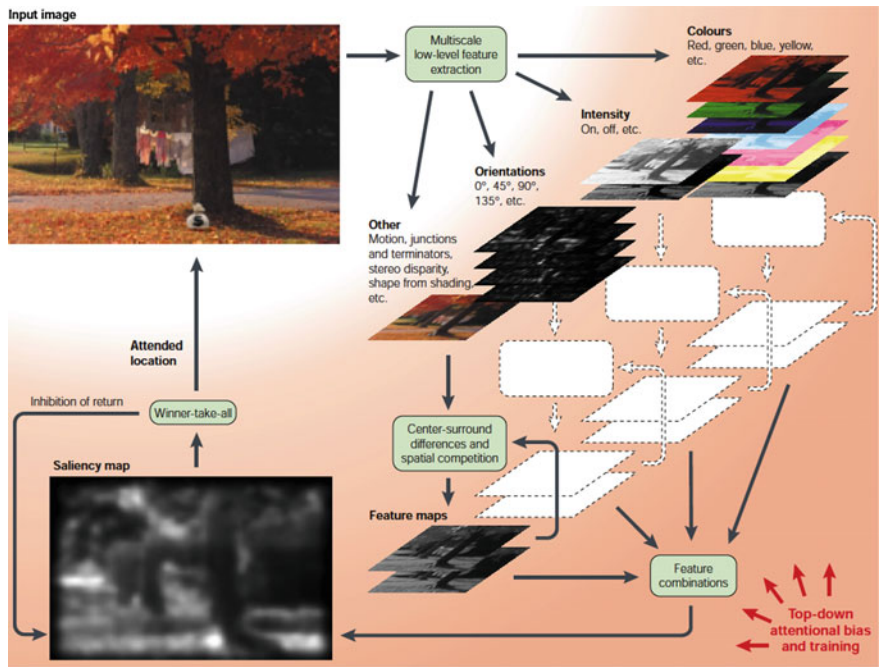


Fig. 2.3 The traditional structure of feature-based computational visual attention models, which is the basis of Itti and Koch’s [IKN98] traditional visual attention model. Image from [IK01a] reprinted with permission from NPG

putational implementation of this model was proposed by Koch and Ullman [KU85], who also coined the term “saliency map” that is identical to the concept of Treisman’s “master map of location”. The general idea is to compute several features in parallel that are fused to form the final saliency map. This traditional structure consists of several processing steps to calculate the saliency map and the different computational models differ in how they implement these steps. For example, Frintrop’s visual object detection with computational attention system (VOCUS) uses integral images to calculate the center-surround differences [Fri06], Harel et al.’s graph-based visual saliency model [HKP07] implements Itti and Koch’s model [IKN98], which is depicted in Fig. 2.3, in a consistent graph-based framework.

In this model, one or several image pyramids are computed to facilitate the subsequent computation features are computed on different scales. Then, image features are computed, which typically are based on local contrast operations such as, most importantly, “center-surround differences” that compare the average value of a center region with the average value in the surrounding region [Mar82]. The most common low-level feature channels are intensity, color, orientation, and motion. Each feature channel is subdivided into several feature types such as, for example, red, green, blue, and yellow feature maps for color. The features are commonly represented in so-called “feature maps”, which are also known as “conspicuity maps”. These feature maps are then normalized and fused to form a single “saliency map”.

How the conspicuity maps are fused is a very important aspect of attention models. It is important that image regions that stand out in one feature map are not suppressed by the other feature maps. Furthermore, the feature calculation can be non-linear, leading to strong variations in the value range across and even within feature channels. Typical normalizations not just try to normalize the value range but also try to highlight local maxima and suppress the often considerable noise in the feature maps [IK01a, IKN98, Fri06]. The feature maps can be weighted, for example, bottom-up by their uniqueness or top-down to incorporate task knowledge when fused into the final saliency map.

Although the saliency map can serve as input to subsequent processing operations, e.g. as a relevance map for image regions, many applications require a trajectory of image regions similar to human saccades. Saccadic movement of the human eye is an essential part of the human visual system and critical to focus and resolve objects. By moving the eye, the small part of the scene that is fixated can be sensed with greater resolution, because it is projected on the central part of the retina, i.e. the fovea, which is responsible for highly resolved, sharp, non-peripheral vision. To serially attend image regions, the saliency map’s local maxima are determined and sequentially attended, typically in the order of descending saliency. A major contribution of Koch and Ullman [KU85] was to show that serially extracting the local maxima can be implemented with biologically-motivated winner-take-all (WTA) neural networks. To serially shift the focus of attention, the saliency of an attended region is suppressed so that the return of the focus of attention to previously attended regions is inhibited.

The computational model as described so far mostly reflects bottom-up attention, i.e. it does not explicitly handle task-specific top-down information (e.g. as is given by a sentence that describes a searched object such as “search for the red ball”). The

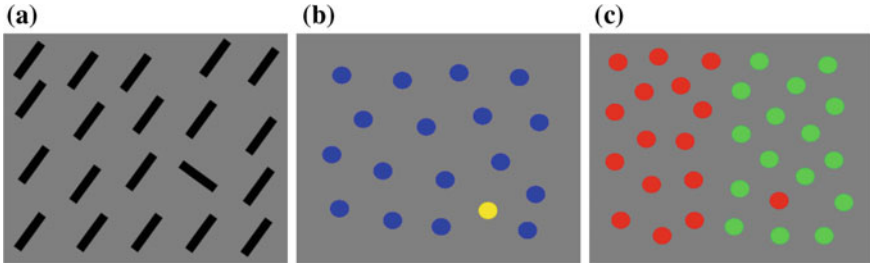


Fig. 2.4 Psychologically motivated test patterns that haven been and are still used to assess the capabilities of visual attention models [KF11]. The goal is to highlight the irregularities in the patterns. Images from and used with permission from Simone Frintrap. **a** Orientation. **b** Color. **c** Locality

most common approach to integrate top-down information is control the influence of the feature maps during the fusion and adapt the weights in such way that feature maps that are likely to highlight distractors are suppressed [WoI94, NI07]. The weights can either be static or dynamic to adapt the model to specific scenarios [XZKB10]. Additionally, it is possible to integrate further, more specialized feature maps that encode, for example, faces, persons, or even cars [JEDT09, CFK09].

Computational Models—Non-traditional Approaches

Since human eye movements are controlled by visual attention, which can easily be observed, gaze trajectories have long served as basis to study visual attention and aspects of human cognition in general. For example, in 1967, Yarbus showed that eye movements depend on the task that is given to a person [Yar67]. Consequently, the main goal of psychological models is to explain and predict eye movements that are recorded during eye tracking experiments. However, due to the lack of modern computerized eye tracking equipment, the abilities of visual attention models where for a long time assessed by testing whether or not they were able to replicate effects that have been observed on psychological test patterns, see Fig. 2.4. In the last five years, several eye tracking datasets have been made publicly available to evaluate visual attention models (e.g., [KNd08, BT09, CFK09, JEDT09]; Winkler and Subramanian provide an up-to-date overview of eye tracking datasets [WS13]). Among other aspects, such easily accessible datasets and the resulting quantitative comparability of test results has lead to a plethora of novel algorithms such as, for example, attention by information maximization [BT09], saliency using natural statistics [ZTM+08], graph-based visual saliency [HKP07], context-aware saliency [GZMT12, GZMT10], and Judd et al.’s machine learning model [JEDT09]. Interestingly, Borji et al. recently evaluated many proposed visual saliency algorithms on eye tracking data [BI13, BSI13b].

However, although being often evaluated on eye tracking data, most recently proposed models do not try to implement or explain any psychological or neurobiological models (e.g., [HHK12, HZ07]). However, a biological plausibility can sometimes be discovered later (e.g., [BZ09]). One such recent trend are spectral saliency models

that were first proposed by Hou et al. [HZ07]. These models operate in the image’s frequency spectrum and exploit the well-known effect that spectral whitening of signals will “accentuate lines, edges and other narrow events without modifying their position” [OL81]. Since these models are based on the fast Fourier transform (FFT), they combine state-of-the-art results in predicting where people look with the computational efficiency inherited from the FFT. Please note that spectral saliency models are discussed in detail in Sect. 3.1.1.

Another recent trend is to use machine learning techniques to learn to predict where humans look, which was first proposed by Judd et al. [JEDT09]. Most saliency models that rely on machine learning are either pixel- or patch-based. Pixel-based approaches have in common with the traditional structure of computational models that they calculate a collection of feature maps. Then, classification or regression methods such as, for example, support vector machines [JEDT09] or boosting [Bor12] can be trained to learn how to optimally fuse the individual feature maps into the final saliency map. Patch-based approaches compare image patches against each other to calculate the saliency of each patch. For example, it is possible to rank the image patches by their uniqueness and assign a high saliency to patches that contain features that are rarely seen across the image [LXG12]. However, all approaches that rely on machine learning have the disadvantage that they require enough training data, which can be problematic, because most datasets consist of a very limited number of eye tracked images.

Computational Models—Salient Object Detection

Recently, Liu et al. adapted the traditional definition of visual saliency by incorporating the high level concept of a salient object into the process of visual attention computation [LSZ+07]. A “salient object” is defined as being the object in an image that attracts most of the user’s interest such as, for example, the man, the cross, the baseball players and the flowers that are shown in Fig. 2.5. Accordingly, Liu et al.



Fig. 2.5 Example images from Achanta et al.’s and Liu et al.’s salient object detection dataset [AS10, LSZ+07]

[LSZ+07] defined the task of “salient object detection” as the binary labeling problem of separating the salient object from the background. Here, it is important to note that the selection of a salient object happens consciously by the user whereas the gaze trajectories, which are recorded with eye trackers, are the result of mostly unconscious processes. Consequently, considering that salient objects naturally attract human gaze [ESP08], salient object detection and predicting where people look are very closely related yet different tasks with different evaluation measures and characteristics.

Since the ties of salient object detection to psychology and neurobiology are relatively loose, a wide variety of models has been proposed in recent years that are even less restricted by biological principles than traditional visual saliency algorithms. Initially, Liu et al. [LSZ+07] combined multi-scale contrast, center-surround histograms, and color spatial-distributions with conditional random fields. Liu et al.’s ideas—a combination of histograms, segmentation, and machine learning—can still be found in most salient object detection algorithms. Alexe et al. [ADF10] combine traditional bottom-up saliency, color contrast, edge density, and superpixels in a Bayesian framework. Closely related to Bayesian surprise [IB06], Klein et al. [KF11] use the Kullback-Leibler divergence of the center and surround image patch histograms to calculate the saliency map, whereas Lu and Lim [LL12] calculate and invert the whole image’s color histogram to predict the salient object. Achanta et al. [AHES09, AS10] rely on the difference of pixels to the average color and intensity value of an image patch or even the whole image. Cheng et al. [CZM+11] use segmentation and define each segments saliency based on the color difference and spatial distance to all other segments.

2.1.2 Auditory Attention

Auditory attention is an important, complex system of bottom-up—i.e., sound-based salience—and top-down—i.e., task-dependent—aspects. Among other aspects, auditory attention assists in the computation of early auditory features and acoustic scene analysis,¹ the identification and recognition of salient acoustic objects, enhancement of signal processing for the attended features or objects, and the planning of actions in response to incoming auditory information [FEDS07]. Moreover, auditory attention can be directed to a rich set of acoustic features including, among others, spatial location, auditory pitch, frequency or intensity, tone duration, timbre, speech versus non-speech, and characteristics of individual voices [FEDS07]. The best example for these abilities is the “cocktail party effect” [Che08], which illustrates that we are able to attend and selectively listen to different speakers in a crowded room that is filled with a multitude of ongoing conversations. Consequently, auditory attention influences many levels of auditory processing; ranging from processing in the cochlea to the association cortex. Not unlike the “what” and “where” pathways in

¹Auditory scene analysis describes the process of segregating and grouping sounds from a mixture of sources to determine and represent relevant auditory streams or objects [Bre90].

the human brain's visual system, there seem to be auditory “what” and “where” pathways, whose activation depends on whether an auditory task requires attending to an auditory feature or object or to a spatial location [ABGA04, DSCM07].

However, since auditory attention is an active research field in neurobiology, psychophysics, and psychology, it is only possible to provide a brief overview of selected aspects in the following. There exist however two detailed literature overviews: First, Hafter et al.'s review [HSL07] focuses on bottom-up aspects of auditory attention. Second, Fritz et al.'s survey [FEDS07] nicely presents aspects of top-down auditory and crossmodal attention. However, although there exists a large body of existing work, it is important to say that there are still many open research questions [FEDS07]. Some of these questions are directly related to the work presented in this book such as, for example: How much of the brain's acoustic novelty detection mechanisms can be explained by simple habituation mechanisms? What are the differences and similarities between visual and auditory attention? What is an appropriate computational model of auditory attention?

How Humans Perceive Sound

As shown in Fig. 2.6b, the cochlea is a coiled system of three ducts: the vestibular duct (scala vestibuli, upper gallery), the tympanic duct (scala tympani, lower gallery), and the cochlear duct (scala media, middle gallery). All of which are filled with lymphatic fluid. The cochlea contains a partition which is known as the “basilar membrane”, see Fig. 2.7. The basilar membrane is essential for our sense of hearing and consists of, most importantly, the scala media, the organ of Corti, the tectorial membrane, and the basilar membrane.

Sound waves that reach the ear lead to oscillatory motions of the auditory ossicles. The oval window allows the transmission of this stimulus into the cochlea. In the cochlea, this stimulus sets the basilar membrane as well as the fluids in the scalae vestibuli and tympani in motion. The location of the maximal amplitude of the travelling wave that moves the basilar membrane depends on the frequency of the incoming sound signal. In other words, the basilar membrane performs a frequency analysis of the incoming sound wave. The motion along the basilar membrane stimulates nerve cells that are located in the organ of Corti, see Fig. 2.7. These nerve cells send electrical signals to the brain, which are finally perceived as sound.

Bottom-Up Auditory Attention

As mentioned before, attentional effects in the human auditory system can occur at various levels of auditory processing. Interestingly, the earliest, mostly bottom-up attentional mechanisms can be observed already in the cochlea [FEDS07, DEHR07, HPSJ56].

The ability to detect “novel”, “odd”, or “deviant” sounds amidst the environmental background noise is an important survival skill of humans and animals. Accordingly, the brain has evolved a sophisticated system to detect novel, odd, and deviant sounds. This system includes an automatic, pre-attentive component that analyzes stability and novelty of the acoustic streams within the acoustic scene, even for task-irrelevant acoustic streams [FEDS07, WTSH+03, WCS+05, Sus05].

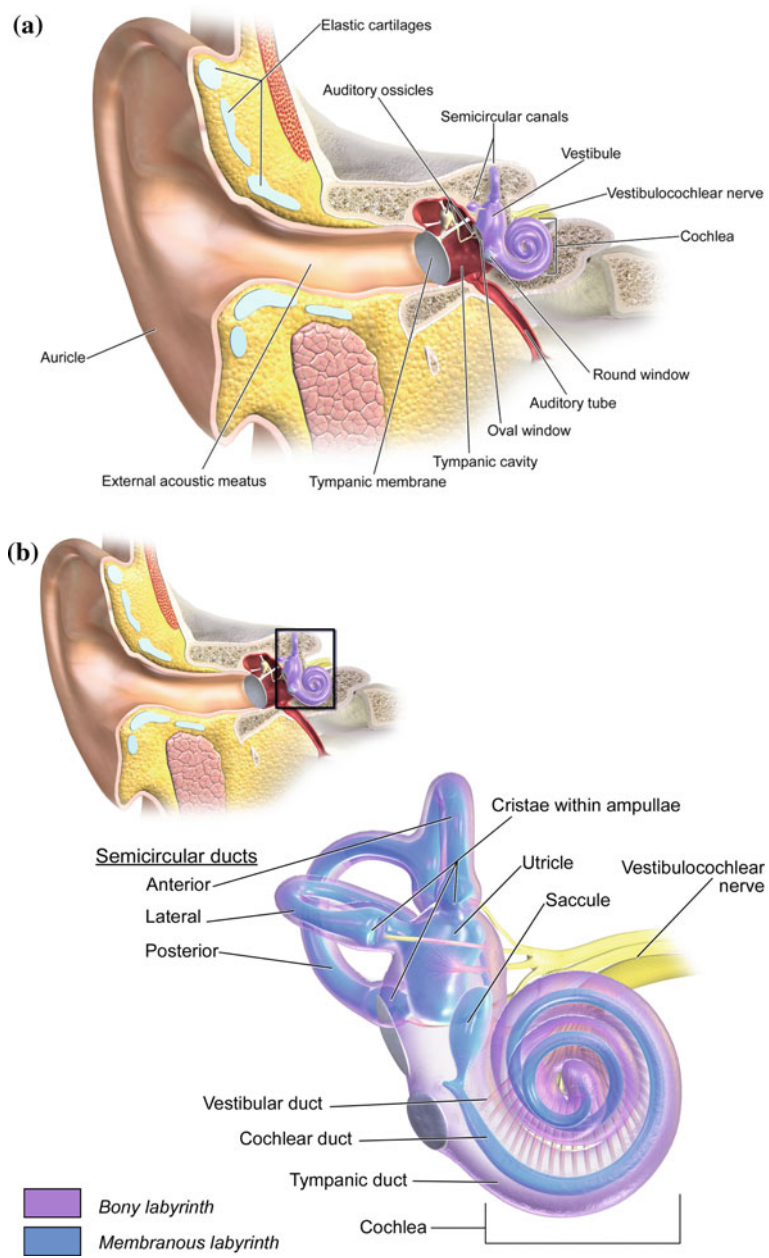


Fig. 2.6 Images illustrating the structure and transmission of sounds in the human ear. Images from [Bla14]. **a** Structure of the human ear [Wika]. **b** Structure of the human inner ear [Wikb]

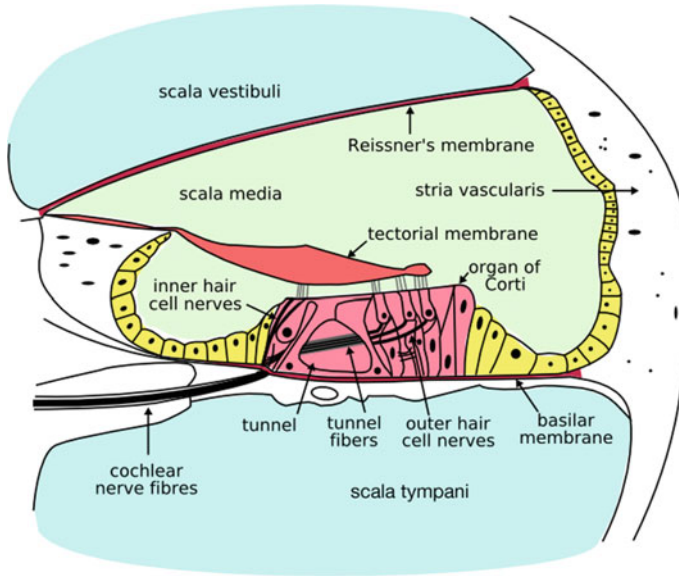


Fig. 2.7 Crosssection of the cochlea. Image from [Wikd].

The brain's acoustic novelty detection system consists of an interconnected set of mechanisms, which includes "adaptive" neurons and a specialization of so-called "novelty" detection neurons. Here, novelty detection neurons specifically encode deviations from the pattern of preceding stimuli. There exist two alternative views on this "change detection" within the auditory scene, depending on where the triggered novelty responses arise in the brain. According to the first view, novelty signals can occur very early in the human auditory system [PGMC05] and suggest the possibility of subcortical pathways for change detection [FEDS07]. However, most research focuses on projections of current neural sound representations that are matched against incoming sounds [FEDS07]. In this view, the change detection system continuously monitors the auditory environment, tracks changes, and updates its representation of the acoustic scene [SW01]. Here, the matching and the novelty response is a largely pre-attentive mechanism, which however can be influenced by top-down mechanisms. It has been shown that this kind of signal mismatch detection can be triggered by deviations in stimulus frequency, intensity, duration or spatial location, or by irregularities in spectrotemporal sequences (over periods of up to 20s), or even in patterns of complex sounds such as speech and music [FEDS07]. Once such a novel or odd stimulus is detected and marked, it can be analyzed by the auditory system to decide whether it should receive further attention or even trigger a behavioral response. Unfortunately, the exact neural basis of this impressive fast, pre-attentive change detection system has not conclusively been found so far.

Computational Models

In contrast to visual attention, hardly any computational auditory attention models exist (cf. [Kal09]). Most closely related to the work presented in this book is the model by Kayser et al. [KPLL05] and Kalinli and Narayanan [KN07]. In both models, Itti et al.'s [IKN98] visual saliency model, see Fig. 2.3, is applied on a map representation of the acoustic signal's frequency spectrum, see Fig. 2.8, which is equivalent or very similar to the signal's spectrogram. This model has been successfully applied and extended for speech processing by Kalinli and Narayanan [KN09, Kal09, KN07], where Kalinli and Narayanan focus on integrating top-down influences in the auditory attention model. Using the auditory spectrum of incoming sound as the basis for bottom-up auditory attention mimics “the process from basilar membrane to the cochlear nucleus in the human auditory system” [Kal09]. Transferring Itti et al.'s visual model to auditory signals is a radical implementation of the idea that the human visual and auditory systems have many similarities. But, it is not clear whether there exists an accessible time-frequency memory in early audition as is implied by the model's time-frequency map, see Fig. 2.8.

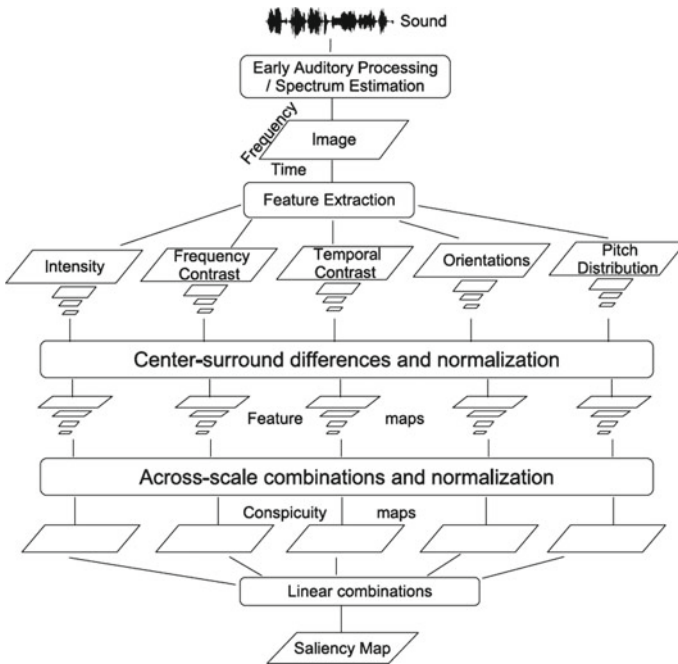


Fig. 2.8 Itti and Koch’s visual saliency model [IKN98] transferred to auditory saliency detection as has been proposed by Kayser et al. [KPLL05] and similarly by Kalinli and Narayanan [KN07]. Image from [KN07] with permission from ISCA

2.1.3 *Multimodal Attention*

Crossmodal Integration

There exist substantial similarities between the visual and auditory attention systems: Most importantly, both consist of bottom-up and top-down components and there appear to be specialized “what” and “where” processes. Since a few years, there is an increasing amount of experimental results that show that all sensory processing in the human brain is in fact multisensory [GS06]. For example, it has been shown that lipreading [CBB+97] or the observation of piano playing without hearing the sound [HEA+05] can activate areas in the auditory cortex.

Several studies have shown that the presence of a visual stimulus or attending a visual task can draw away attention from an auditory stimulus, which is indicated by a decreased activity in the auditory cortex (e.g., [LBW+02, WBB+96]). Similarly, auditory attention can negatively influence visual attention. In fact, it was shown that there exists a reciprocal inverse relationship between auditory and visual activation, which means that increases in visual activation correlate with a decrease in auditory activation and vice versa. A very interesting study was performed by Weissman et al. [WWW04]. Weissman et al. created a conflict between auditory and visual target stimuli, and crossmodal distractors. They observed that when the “distracting stimulus in the task-irrelevant sensory channel is increased, there was a compensatory increase in selective attention to the target in the relevant channel and a corresponding increase in activation in the relevant sensory cortex” [FEDS07]. This suggests that it is likely that there exists a top-down mechanism that regulates the relative strengths of the sensory channels.

How auditory and visual sensory information interact for the control of overt attention, i.e. directing the sensory organs toward interesting stimuli, has recently been investigated by Onat et al. [OLK07]. Onat et al. performed eye tracking studies in which the participants were listening to sounds coming from different directions. It was shown that eye fixation probabilities increase toward the location where the sound originates, which means—unsurprisingly—that the selection of fixation points depends on auditory and visually salient stimuli. Furthermore, Onat et al. used the data to test several biologically plausible crossmodal integration mechanisms and found “that a linear combination of both unimodal saliencies provides a good model for this integration process” [OLK07]. Interestingly, such a linear combination is not just optimal in an information theoretic sense (see [OLK07]), but it also allows to adjust the relative strength of the sensory channels. However, this model assumes the existence of a 2-dimensional auditory saliency map that encodes where salient stimuli occur in the scene and how salient these stimuli are, which can be directly fused with the visual saliency map to form a joint audio-visual saliency map.

High-Level Influences

Not just crossmodal effects can influence what is interesting in one modality. Instead, there exist many top-down signals that can direct attention toward specific targets (e.g., [Ban04, Hob05, CC03, TTDC06, STET01, WHK+04, NI07]).

Verbal descriptions of object properties can directly influence what is perceived as being perceptually salient (e.g., [STET01, WHK+04, NI07]). For example, it has been shown that knowledge about an object’s visual appearance can influence the perceptual saliency to highlight an object that we are actively searching, i.e. in a so-called visual search task. But, only specific information that refers to primitive preattentive features allows such attentional guidance [WHK+04]. Accordingly, if we have a good visual impression or memory of the target that we are looking for (e.g., we have just seen it a few moments ago) or if we at least know the target’s color, then we can find the target faster. In these cases, the visual saliency will be guided in such way that it stronger highlights image regions that exhibit the target’s preattentive features. In contrast, for example, categorical information about the search target (e.g., search for an animal) typically does not provide such top-down guidance (see [WHK+04]).

Interestingly, certain features that would typically be associated with high-level vision tasks can attract our low-level attention independent of task. Most importantly, it has been shown that faces and face-like patterns attract the gaze of infants as young as 6 weeks, i.e. before they can consciously perceive the category of faces [SS06]. The fact that the gaze and, consequently, interest of infants is attracted by face-like patterns seems to be an important aspect of early infant development, especially for social signals and processes (see, e.g., [KJS+02]). Interestingly, infants show the ability to follow the observed gaze direction of caregivers at an age of 6 months [Hob05]. If people talk about objects that are part of the environment, where and at what people are looking at is related to the object that is being talked about. Consequently, the ability to follow the caregiver’s gaze makes it possible for an infant to associate what it sees with the words it hears, an important ability to learn a language. Similar to gaze but more direct and less subtle, infants also soon develop the ability to interpret pointing gestures (see [LT09]). Accordingly, pointing gestures and gaze are both non-verbal signals that direct the attention toward a spatial region of interest (see, e.g., [Ban04, LB05, LT09]). This is an essential aspect in natural interaction, because it makes it possible to direct and coordinate the attention of interacting persons and, thus, helps to establish a joint focus of attention. In other words, such non-verbal signals are used to influence where an interaction partner is looking in order to direct his gaze toward a specific object that is or will become the subject of the conversation. Consequently, the generation and interpretation of such signals is fundamental for “learning, language, and sophisticated social competencies” [MN07a].

2.2 Applications of Attention Models

Knowing in advance what people might find interesting and attend to is an important information that can be integrated into many applications. Images and videos can be compressed better, street signs can be designed to immediately grab the attention, and advertisement can put stronger emphasis on the intended message. Furthermore, having an estimate of what is probably a relevant signal in a data stream allows us to focus computational algorithms. This way, machine learning can learn better models

from less data, class-independent object detection as well as object recognition can be improved, and robots are able to process incoming sensory information in real-time despite limited computational power.

2.2.1 *Image Processing and Computer Vision*

Image and video compression algorithms can improve the perceptual quality of compressed images and videos by allocating more bits to code image regions that exhibit a high perceptual saliency [GZ10, OBH+01]. This way, image regions that are likely to attract the viewers' interest are less compressed and thus show fewer disturbing alterations such as compression artifacts. Ouerhani et al. [OBH+01] implement such an adaptive coding scheme that favors the allocation of a higher number of bits to those image regions that are more conspicuous to the human visual system. The compressed image files are fully compatible with the JPEG standard. An alternative approach was recently proposed by Hadizadeh and Bajic [HB13]. Their method uses saliency to automatically reduce potentially attention-grabbing coding artifacts in regions of interest.

Visual attention and object recognition are tightly linked processes in human perception. Accordingly, although most models of visual attention and object recognition are separated, there is an increasing interest in integrating both processes to increase the performance of computer vision systems. Initial approaches tried to use attention as a front-end to detect salient objects or keypoint locations. Miao et al. [MPI01] use an attentional front-end with the biologically motivated object recognition system HMAX [RP99]. Walther and Koch [WK06] combine an attention system with a SIFT-based object recognition [Low04] and demonstrate that they are able to improve object recognition performance. Going a step further, Walter and Koch [WK06] suggest a unifying attention and object recognition framework. In this framework, the HMAX object recognition is modulated to suppress or enhance image locations and features depending on the spatial attention.

Related to such attentional front-ends for object recognition, principles of visual attention have recently been integrated into approaches for general, class independent object detection [ADF10]. This way, sampling windows can be distributed according to the "objectness" distribution and used as location priors for class-specific object detectors. This can greatly reduce the necessary number of windows evaluated by class-specific object detectors as has been shown in the PASCAL Visual Object Classes (VOC) challenge 2007 [EVGW+]. Interestingly, going in the other direction, high-level object detectors are being integrated into saliency models to model, for example, that the human visual attention is attracted by faces and face-like patterns. For this purpose, some models integrate detectors for faces, the horizon, persons and even cars [CHEK07, JEDT09]. This shows that attention and object recognition might grow together in the future.

Saliency has also been employed as a spatial prior to learn object attributes, categories, or classes from weakly labeled images. For example, Fei-Fei Li et al.'s

[[FFFFP03](#)] approach to “one-shot learning” uses Kadir and Brady’s saliency detector [[KB01](#)] to sample features at highly salient locations. The most salient regions are clustered over location and scale to give a reasonable number of distinctive features per image.

2.2.2 *Audio Processing*

In contrast to applications in computer vision, only few applications of acoustic or auditory saliency models have been explored so far. Coensel and Botteldooren [[CB10](#)] propose to use an auditory attention model in soundscape design to assess how specific sounds can mask unwanted environmental sounds. Lin et al. [[LZG+12](#)] as well as, in principle, Kalinli and Narayanan [[KN07](#), [KN09](#)] use Itti’s classic visual saliency model [[IKN98](#)] to highlight visually salient patterns in the spectrogram. Lin et al. [[LZG+12](#)] fuse the spectrogram’s saliency map with the original spectrogram and use the resulting saliency-maximized audio spectrogram to enable faster than real-time detection of audio events by human audio analysts. Kalinli and Narayanan [[KN07](#)] use the spectrogram’s saliency map to detect prominent syllable and word locations in speech, achieving close to human performance. The task of syllable detection was chosen by the authors to investigate low-level auditory saliency models, because during speech perception, a particular phoneme or syllable can be perceived to be more salient than the others due to the coarticulation between phonemes, and other factors such as the accent, and physical and emotional state of the talker [[KN07](#)].

2.2.3 *Robotics*

In addition to reducing computational requirements by focusing on the most salient stimuli, robots can benefit from attentional mechanisms at several conceptual levels [[Fri11](#)]. On a low level, attention can be used for salient landmark detection and subsequent scene recognition and localization. On a mid level, attention can serve as a pre-processing step for object recognition. On a high level, attention can be implemented in a human-like fashion to guide actions and mimic human behavior, for example, during object manipulation or human-robot interaction.

Salient landmarks are excellent candidates for localization, because they are visually outstanding and distinctive, often having unique features. This makes them easy to (re-)detect and allows for a very sparse set of localization landmarks that can easily be detected, accessed in memory, and matched in real-time. The ARK project [[NJW+98](#)] is one of the earliest projects that investigated the use of salient landmarks for localization. The localization was based on manually generated maps of static obstacles and natural visual landmarks. Siagian and Itti [[SI09](#)] presented an integrated system for coarse global localization based on the “gist” of the scene and fine localization within a scene using salient landmarks. Frintrop and Jensfelt [[FJ08](#)] combined attention and salient landmark detection with simultaneous localization

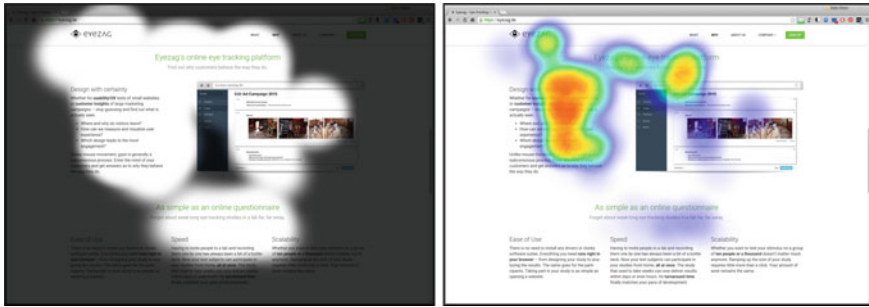


Fig. 2.9 Eye tracking experiments are used to optimize the layout of websites. The opacity map (*left*) illustrates what people see and the heat map (*right*) represents the distribution of interest on the website. Images generated with Eyezag’s online eye tracking platform [Eye]

and mapping (SLAM). The attention system VOCUS [Fri06] detects salient regions. These regions are tracked and matched to all entries in a database of previously seen landmarks to estimate a 3D position.

The main difference between robotic applications and, for example, image processing is that a robot can move its body parts to interact with its environment and influence what it perceives. This way, robots can control their geometric parameters, e.g. where it looks, and manipulate the environment to improve the perception quality of specific stimuli [AWB88]. This can be implemented with an attentive two-step object detection and recognition mechanism: First, regions of interest are detected in a peripheral vision system based on visual saliency and a coarse view of the scene. Second, the robot then investigates each region of interest by focusing its sensors on the target object, which provides high-resolution images for object recognition (e.g., [MFL+08, GAK+07]). It is noteworthy to say that using this strategy Meger et al.’s robot “Curious George” [MFL+08] won the 2007 and 2008 Semantic Robotic Vision Challenge [Uni] (Fig. 2.9).

A common assumption in the field of socially interactive robots is that “humans prefer to interact with machines in the same way that they interact with other people” [FND03]. This is based on the observation that humans tend to treat robots like people and, as a consequence, tend to expect human-like behavior from robots [FND03, NM00]. According to this assumption, a computational attention system that mimics how humans direct their attention can facilitate human-robot interaction. For example, this idea has been implemented in the social robot Kismet, whose gaze is controlled by a visual attention system [BS99].

2.2.4 Computer Graphics

Naturally, knowing what attracts the viewer’s attention is important when automatically generating or manipulating images. For example, it is possible to automatically

crop an image to only present the most relevant content to a user and/or act as a thumbnail [SLBJ03, SAD+06, CXF+03]. Similarly, content-aware media retargeting automatically changes the aspect ratio of images and videos to optimize the presentation of visual content across platforms and screen sizes [SLNG07, AS07, RSA08, GZMT10]. For this purpose, saliency models are used to automatically determine image regions that are likely to contain relevant information. Depending on their estimated importance, image regions are then deleted or morphed so that the resized image best portrays the most relevant information (see, e.g., [SLNG07]).

2.2.5 Design, Marketing, and Advertisement

There exist several companies such as, for example, SMIvision [SMI] and Eyezag [Eye] that offer eye tracking experiments as a service. This enables companies to analyze how people view their webpage, advertisement, or image and video footage, see Figs. 2.9 and 2.10. Other companies such as Google have their own in-house laboratories and solutions to perform eye tracking experiments and research [Goo].

With increasingly powerful computational attention models that predict human fixations, it becomes possible to reduce the need for expensive and intrusive eye tracking experiments. In 2013, 3M has started to offer its visual attention service [AMR3M] that uses a computational attention model as a cheaper and faster alternative to eye tracking experiments. Potential usage scenarios as proposed by 3M are in-store merchandising, packaging, advertising, web and banner advertisement, and video analysis [AMR3M].

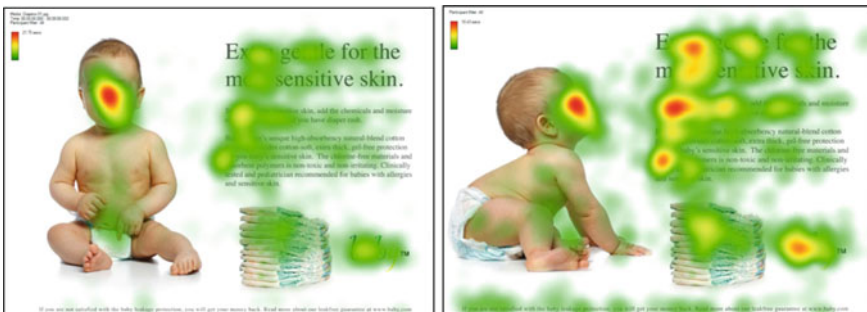


Fig. 2.10 Eye tracking experiments are used to optimize the layout of advertisement. Images from and used with permission from James Breeze

References

- [AHES09] Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: Proceedings of the International Conference on Computer Vision Pattern Recognition (2009)
- [AS10] Achanta, R., Süsstrunk, S.: Saliency detection using maximum symmetric surround. In: Proceedings of the International Conference on Image Processing (2010)
- [ADF10] Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: Proceedings of the International Conference on Computer Vision Pattern Recognition, pp. 73–80 (2010)
- [AWB88] Aloimonos, Y., Weiss, I., Bandopadhyay, A.: Active vision. *Int. J. Comput. Vis.* **1**(4), 333–356 (1988)
- [ABGA04] Arnott, S.R., Binns, M.A., Grady, C.L., Alain, C.: Assessing the auditory dual-pathway model in humans. *Neuroimage* **22**, 401–408 (2004)
- [AS07] Avidan, S., Shamir, A.: Seam carving for content-aware image resizing. *ACM Trans. Graph.* **26**(3) (2007)
- [AMR3M] 3SMM: 3M visual attention service. http://solutions.3m.com/wps/portal/3M/en_US/VAS-NA?MDR=true
- [Ban04] Bangerter, A.: Using pointing and describing to achieve joint focus of attention in dialogue. *Psychol. Sci.* **15**(6), 415–419 (2004)
- [BZ09] Bian, P., Zhang, L.: Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In: Proceedings of the Annual Conference on Neural Information Processing Systems (2009)
- [Bla14] Blausen.com staff: Blausen gallery 2014. Wikiversity J. Med. (2014)
- [Bor12] Borji, A.: Boosting bottom-up and top-down visual features for saliency estimation. In: Proceedings of the International Conference on Computer Vision Pattern Recognition (2012)
- [BI13] Borji, A., Itti, L.: State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 185–207 (2013)
- [BSI13b] Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *IEEE Trans. Image Process.* **22**(1), 55–69 (2013)
- [BH05] Breazeal, C., Scassellati, B.: A context-dependent attention system for a social robot. In: Proceedings of the International Joint Conference on Artificial Intelligence (1999)
- [BS99] Bregman, A.S.: Auditory Scene Analysis: The Perceptual Organization of Sounds. MIT Press (1990)
- [Bre90] Bruce, N., Tsotsos, J.: Saliency, attention, and visual search: an information theoretic approach. *J. Vis.* **9**(3), 1–24 (2009)
- [BT09] Bundesen, C., Habekost, T.: Handbook of Cognition. Sage Publications, Chap. Attention (2005)
- [CBB+97] Calvert, G.A., Bullmore, E., Brammer, M., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., David, A.S.: Activation of auditory cortex during silent lipreading. *Science* **276**, 593–596 (1997)
- [CC03] Cashon, C., Cohen, L.: The Construction, Deconstruction, and Reconstruction of Infant Face Perception. Chapter The development of face processing in infancy and early childhood, Current perspectives, pp. 55–68. NOVA Science Publishers (2003)
- [CHEK07] Cerf, M., Harel, J., Einhäuser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. In: Proceedings of the Annual Conference on Neural Information Processing Systems (2007)
- [CFK09] Cerf, M., Frady, E.P., Koch, C.: Faces and text attract gaze independent of the task: experimental data and computer model. *J. Vis.* **9** (2009)
- [CXF+03] Chen, L.-Q., Xie, X., Fan, X., Ma, W.-Y., Zhang, H.-J., Zhou, H.-Q.: A visual attention model for adapting images on small displays. *Multim. Syst.* **9**(4), 353–364 (2003)

- [CZM+11] Cheng, M.-M., Zhang, G.-X., Mitra, N.J., Huang, X., Hu, S.-M.: Global contrast based salient region detection. In: Proceedings of the International Conference on Computer Vision Pattern Recognition (2011)
- [Che08] Cherry, E.C.: Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* **25**, 975–979 (2008)
- [CB10] Coensel, B.D., Botteldooren, D.: A model of saliency-based auditory attention to environmental sound. In: Proceedings of the International Congress on Acoustics (2010)
- [DEHR07] Delano, P.H., Elgueda, D., Hamame, C.M., Robles, L.: Selective attention to visual stimuli reduces cochlear sensitivity in chinchillas. *J. Neurosci.* **27**, 4146–4153 (2007)
- [DSCM07] De Santis, L., Clarke, S., Murray, M.M.: Automatic and intrinsic auditory what and where processing in humans revealed by electrical neuroimaging. *Cereb Cortex* **17**, 9–17 (2007)
- [ESP08] Einhäuser, W., Spain, M., Perona, P.: Objects predict fixations better than early saliency. *J. Vis.* **8**(14) (2008)
- [EVGW+] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>
- [Eye] Eyezag: Eyezag—eye tracking in your hands. <http://www.eyezag.com/>
- [FFFP03] Fei-Fei, L., Fergus, R., Perona, P.: A bayesian approach to unsupervised one-shot learning of object categories. In: Proceedings of the International Conference on Computer Vision (2003)
- [FND03] Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. *Robot. Auton. Syst.* **42**(3–4), 143–166 (2003)
- [Fri06] Frintrop, S.: VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search, ser. Springer, Lecture Notes in Computer Science (2006)
- [FJ08] Frintrop, S., Jensfelt, P.: Attentional landmarks and active gaze control for visual slam. *IEEE Trans. Robot.* **24**(5), 1054–1065 (2008)
- [FRC10] Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundation: a survey. *ACM Trans. Applied Percept.* **7**(1), 6:1–6:39 (2010)
- [FEDS07] Fritz, J.B., Elhilali, M., David, S.V., Shamma, S.A.: Auditory attention-focusing the searchlight on sound. *Curr. Opin. Neurobiol.* **17**(4), 437–455 (2007)
- [GS06] Ghazanfar, A.A., Schroeder, C.E.: Is neocortex essentially multisensory? *Trends Cogn. Sci.* **10**, 278–285 (2006)
- [GZMT10] Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. In: Proceedings of the International Conference on Computer Vision Pattern Recognition (2010)
- [Goo] Google: Eye-tracking studies: more than meets the eye. <http://googleblog.blogspot.de/2009/02/eye-tracking-studies-more-than-meets.html>
- [GAK+07] Gould, S., Arfvidsson, J., Kaehler, A., Sapp, B., Messner, M., Bradski, G., Baumstarck, P., Chung, S., Ng, A.Y.: Peripheral-foveal vision for real-time object recognition and tracking in video. In: Proceedings of the International Joint Conference on Artificial Intelligence (2007)
- [GZ10] Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans. Image Process.* **19**, 185–198 (2010)
- [HB13] Hadizadeh, H., Bajic, I.: Saliency-aware video compression. *IEEE Trans. Image Process* (99) (2013)
- [HSL07] Hafer, E.R., Sarampalis, A., Loui, P.: Auditory Perception of Sound Sources. Springer (2007) (ch. Auditory attention and filters (review))
- [HEA+05] Haslinger, B., Erhard, P., Altenmüller, E., Schroeder, U., Boecker, H., Ceballos-Baumann, A.O.: Transmodal sensorimotor networks during action observation in professional pianists. *J. Cogn. Neurosci.* **17**, 282–293 (2005)

- [HKP07] Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Proceedings of the Annual Conference on Neural Information Processing Systems (2007)
- [HPSJ56] Hernandez-Peon, R., Scherrer, H., Jouvett, M.: Modification of electric activity in cochlear nucleus during attention in unanesthetized cats. *Science* **123**, 331–332 (1956)
- [Hob05] Hobson, R.: Joint attention: Communication and other minds. Oxford University Press (2005) (Chap.. What puts the jointness in joint attention?), pp. 185–204
- [HHK12] Hou, X., Harel, J., Koch, C.: Image signature: highlighting sparse salient regions. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(1), 194–201 (2012)
- [HZ07] Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: Proceedings of the International Conference on Computer Vision Pattern Recognition (2007)
- [IB06] Itti, L., Baldi, P.: Bayesian surprise attracts human attention. In: Proceedings of the Annual Conference on Neural Information Processing Systems (2006)
- [IKN98] Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), 1254–1259 (1998)
- [IK01a] Itti, L., Koch, C., Niebur, E.: Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2**(3), 194–203 (2001)
- [JSF12] Jaspers, H., Schauerte, B., Fink, G.A.: Sift-based camera localization using reference objects for application in multi-camera environments and robotics. In: Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods (ICPRAM), Vilamoura, Algarve, Portugal (2012)
- [JEDT09] Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: Proceedings of the International Conference on Computer Vision (2009)
- [KB01] Kadir, T., Brady, M.: Saliency, scale and image description. *Int. J. Comput. Vis.* **45**(2), 83–105 (2001)
- [KTG92] Kahneman, D., Treisman, A., Gibbs, B.J.: The reviewing of object files: object-specific integration of information. *Cogn. Psychol.* **24**(2), 175–219 (1992)
- [Kal09] Kalinli, O.: Biologically inspired auditory attention models with applications in speech and audio processing. Ph.D. dissertation, University of Southern California, Los Angeles, CA, USA (2009)
- [KPLL05] Kayser, C., Petkov, C.I., Lippert, M., Logothetis, N.K.: Mechanisms for allocating auditory attention: an auditory saliency map. *Curr. Biol.* **15**(21), 1943–1947 (2005)
- [KN09] Kalinli, O.: Prominence detection using auditory attention cues and task-dependent high level information. *IEEE Trans. Audio Speech Lang Proc.* **17**(5), 1009–1024 (2009)
- [KN07] Kalinli, O., Narayanan, S.: A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In: Proceedings of the Annual Conference on International Speech Communication Association (2007)
- [KF11] Klein, D.A., Frintrop, S.: Center-surround divergence of feature statistics for salient object detection. In: Proceedings of the International Conference on Computer Vision (2011)
- [KJS+02] Klin, A., Jones, W., Schultz, R., Volkmar, F., Cohen, D.: Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Arch. Gen. Psychiatry* **59**(9), 809–816 (2002)
- [KU85] Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* **4**, 219–227 (1985)
- [KSS13] Koester, D., Schauerte, B., Stiefelhagen, R.: Accessible section detection for visual guidance. In: IEEE/NSF Workshop on Multimodal and Alternative Perception for Visually Impaired People (2013)
- [KNd08] Kootstra, G., Nederveen, A., de Boer, B.: Paying attention to symmetry. In: Proceedings of the British Conference on Computer Vision (2008)
- [KSSK12] Kühn, B., Schauerte, B., Stiefelhagen, R., Kroschel, K.: A modular audio-visual scene analysis and attention system for humanoid robots. In: Proceedings of the 43rd International Symposium on Robotics (ISR), Taipei, Taiwan (2012)

- [KSKS12] Kühn, B., Schauerte, B., Kroschel, K., Stiefelbogen, R.: Multimodal saliency-based attention: A lazy robot's approach. In: *Proceedings of the 25th International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ, Vilamoura, Algarve, Portugal (2012)
- [SS13b] Kühn, B., Schauerte, B., Kroschel, K., Stiefelbogen, R.: Wow! Bayesian surprise for salient acoustic event detection. In: *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, Vancouver, Canada (2013)
- [LBW+02] Laurienti, P., Burdette, J.H., Wallace, M.T., Yen, Y.F., Field, A.S., Stein, B.E.: Deactivation of sensory-specific cortex by cross-modal stimuli. *J. Cogn. Neurosci.* **14**, 420–429 (2002)
- [LXG12] Li, J., Xu, D., Gao, W.: Removing label ambiguity in learning-based visual saliency estimation. *IEEE Trans. Image Process.* **21**(4), 1513–1525 (2012)
- [LT09] Liebal, K., Tomasello, M.: Infants appreciate the social intention behind a pointing gesture: commentary on "children's understanding of communicative intentions in the middle of the second year of life" by T. Aureli, P. Perucchini and M. Genco. *Cogn. Dev.* **24**(1), 13–15 (2009)
- [LZG+12] Lin, K.-H., Zhuang, X., Goudeseune, C., King, S., Hasegawa-Johnson, M., Huang, T.S.: Improving faster-than-real-time human acoustic event detection by saliency-maximized audio visualization. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (2012)
- [Fri11] Lin, K.-H., Zhuang, X., Goudeseune, C., King, S., Hasegawa-Johnson, M., Huang, T.S.: Towards attentive robots. *Paladyn* **2**(2), 64–70 (2011)
- [LSZ+07] Liu, T., Sun, J., Zheng, N.-N., Tang, X., Shum, H.-Y.: Learning to detect a salient object. In: *Proceedings of the International Conference on Computer Vision Pattern Recognition* (2007)
- [LL12] Lu, S., Lim, J.-H.: Saliency modeling from image histograms. In: *Proceedings of the European Conference on Computer Vision* (2012)
- [LB05] Louwerse, M., Bangerter, A.: Focusing attention with deictic gestures and linguistic expressions. In: *Proceedings of the Annual Conference on Cognitive Science Society* (2005)
- [Low04] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**, 91–110 (2004)
- [Mar82] Marr, D.: *VISION—A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H Freeman and Company (1982)
- [NI07] Marr, D.: Search goal tunes visual features optimally. *Neuron* **53**(4), 605–617 (2007)
- [MCS+14] Martinez, M., Constantinescu, A., Schauerte, B., Koester, D., Stiefelbogen, R.: Cognitive evaluation of haptic and audio feedback in short range navigation tasks. In: *Proceedings of the 14th Int. Conf. Computers Helping People with Special Needs (ICCHP)*. Springer, Paris, France (2014)
- [MSS13] Martinez, M., Schauerte, B., Stiefelbogen, R.: BAM! Depth-based body analysis in critical care. In: *Proceedings of the 15th International Conference on Computer Analysis of Images and Patterns (CAIP)*. Springer, York, UK (2013)
- [SS13a] Martinez, M., Schauerte, B., Stiefelbogen, R.: How the distribution of salient objects in images influences salient object detection. In: *Proceedings of the 20th International Conference on Image Processing (ICIP)*. IEEE, Melbourne, Australia (2013)
- [MFL+08] Meger, D., Forssén, P.-E., Lai, K., Helmar, S., McCann, S., Southey, T., Baumann, M., Little, J.J., Lowe, D.J.: Curious george: an attentive semantic robot. *Robot. Auton. Syst.* **56**(6), 503–511 (2008)
- [MPI01] Miao, F., Papageorgiou, C., Itti, L.: Neuromorphic algorithms for computer vision and attention. In: Bosacchi, B., Fogel, D.B., Bezdek, J.C. (eds.) *Proceedings of the SPIE 46 Annual International Symposium on Optical Science and Technology*, vol. 4479, pp. 12–23 (2001)

- [MN07a] Mundy, P., Newell, L.: Attention, joint attention, and social cognition. *Curr. Dir. Psychol. Sci.* **16**(5), 269–274 (2007)
- [NM00] Nass, C., Moon, Y.: Machines and mindlessness: social responses to computers. *J. Soc. Issues* **56**(1), 81–103 (2000)
- [NJW+98] Nickerson, S.B., Jasiobedzki, P., Wilkes, D., Jenkin, M., Milios, E., Tsotsos, J.K., Jepson, A., Bains, O.N.: The ark project: autonomous mobile robots for known industrial environments. *Robot. Auton. Syst.* **25**, 83–104 (1998)
- [OLK07] Onat, S., Libertus, K., König, P.: Integrating audiovisual information for the control of overt attention. *J. Vis.* **7**(10) (2007)
- [OL81] Oppenheim, A., Lim, J.: The importance of phase in signals. *Proc. IEEE* **69**(5), 529–541 (1981)
- [OBH+01] Ouerhani, N., Bracamonte, J., Hugli, H., Ansorge, M., Pellandini, F.: Adaptive color image compression based on visual attention. In: *Proceedings of the International Conference on Image Analysis and Processing*, pp. 416–421 (2001)
- [PGMC05] Perez-Gonzalez, D., Malmierca, M.S., Covey, E.: Novelty detector neurons in the mammalian auditory midbrain. *Eur. J. Neurosci.* **22**, 2879–2885 (2005)
- [RP99] Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999)
- [RSA08] Rubinstein, M., Shamir, A., Avidan, S.: Improved seam carving for video retargeting. In: *Proceedings of the Annual Conference on Special Interest Group on Graphics and Interactive Techniques* (2008)
- [RSAHS14] Rybok, L., Schauerte, B., Al-Halah, Z., Stiefelhausen, R.: Important stuff, everywhere! Activity recognition with salient proto-objects as context. In: *Proceedings of the 14th IEEE Winter Conference on Applications of Computer Vision (WACV)*, Steamboat Springs, CO, USA (2014)
- [SAD+06] Santella, A., Agrawala, M., DeCarlo, D., Salesin, D., Cohen, M.: Gaze-based interaction for semi-automatic photo cropping. In: *Proceedings of the International Conference on Human Factors Computing Systems (CHI)* (2006)
- [SF10a] Schauerte, B., Fink, G.A.: Focusing computational visual attention in multi-modal human-robot interaction. In: *Proceedings of the 12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI)*. ACM, Beijing, China (2010)
- [SF10b] Schauerte, B., Fink, G.A.: Web-based learning of naturalized color models for human-machine interaction. In: *Proceedings of the 12th International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, Sydney, Australia (2010)
- [Sch14] Schauerte, B.: Multimodal computational attention for scene understanding. Ph.D. dissertation, Karlsruhe Institute of Technology (2014)
- [SKMS14] Schauerte, B., Koester, D., Martinez, M., Stiefelhausen, R.: Way to Go! Detecting open areas ahead of a walking person. In: *ECCV Workshop on Assistive Computer Vision and Robotics (ACVR)*. Springer (2014)
- [SWS15b] Schauerte, B., Koester, D., Martinez, M., Stiefelhausen, R.: A web-based platform for interactive image sonification. In: *Accessible Interaction for Visually Impaired People (AI4VIP)* (2015)
- [SS14] Schauerte, B., Koester, D., Martinez, M., Stiefelhausen, R.: Look at this! Learning to guide visual saliency in human-robot interaction. In: *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ (2014)
- [SS15] Schauerte, B., Koester, D., Martinez, M., Stiefelhausen, R.: On the distribution of salient objects in web images and its influence on salient object detection. *PLoS ONE* **10**, 07 (2015)
- [SKKS11] Schauerte, B., Kühn, B., Kroschel, K., Stiefelhausen, R.: Multimodal saliency-based attention for object-based scene analysis. In: *Proceedings of the 24th International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ, San Francisco, CA, USA (2011)

- [SMCS12] Schauerte, B., Martinez, M., Constantinescu, A., Stiefelhagen, R.: An assistive vision system for the blind that helps find lost things. In: Proceedings of the 13th International Conference on Computers Helping People with Special Needs (ICHP). Springer, Linz, Austria (2012)
- [SPF09] Schauerte, B., Plötz, T., Fink, G.A.: ‘A multi-modal attention system for smart environments. In: Proceedings of the 7th International Conference on Computer Vision Systems (ICVS). Lecture Notes in Computer Science, vol. 5815. Springer, Liège (2009)
- [SRF10] Schauerte, B., Richarz, J., Fink, G.A.: Saliency-based identification and recognition of pointed-at objects. In: Proceedings of the 23rd International Conference on Intelligent Robots and Systems (IROS). IEEE/RSJ, Taipei, Taiwan (2010)
- [SRP+09] Schauerte, B., Richarz, J., Plötz, T., Thureau, C., Fink, G.A.: Multi-modal and multi-camera attention in smart environments. In: Proceedings of the 11th International Conference on Multimodal Interfaces (ICMI). ACM, Cambridge (2009)
- [SS12a] Schauerte, B., Stiefelhagen, R.: Learning robust color name models from web images. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR). IEEE, Tsukuba, Japan (2012)
- [SS12b] Schauerte, B., Stiefelhagen, R.: Predicting human gaze using quaternion DCT image signature saliency and face detection. In: Proceedings of the IEEE Workshop on the Applications of Computer Vision (WACV). IEEE, Breckenridge, CO, USA (2012)
- [SS12c] Schauerte, B., Stiefelhagen, R.: Quaternion-based spectral saliency detection for eye fixation prediction. In: Proceedings of the 12th European Conference on Computer Vision (ECCV). Springer, Firenze, Italy (2012)
- [SSS14] Schneider, T., Schauerte, B., Stiefelhagen, R.: Manifold alignment for person independent appearance-based gaze estimation. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR). IEEE, Stockholm, Sweden (2014)
- [SWS15a] Schauerte, B., Wörtwein, T., Stiefelhagen, R.: Color decorrelation helps visual saliency detection. In: Proceedings of the 22nd International Conference on Image Processing (ICIP). IEEE (2015)
- [SZ14] Schauerte, B., Zamfirescu, C.T.: Small k-pyramids and the complexity of determining k. *J. Discrete Algorithms (JDA)* (2014)
- [SLNG07] Setlur, V., Lechner, T., Nienhaus, M., Gooch, B.: Retargeting images and video for preserving information saliency. *IEEE Comput. Graph. Appl.* **27**(5), 80–88 (2007)
- [SI09] Siagian, C., Itti, L.: Biologically inspired mobile robot vision localization. *IEEE Trans. Robot.* **25**(4), 861–873 (2009)
- [SS06] Simion, C., Shimojo, S.: Early interactions between orienting, visual sampling and decision making in facial preference. *Vis. Res.* **46**(20), 3331–3335 (2006)
- [SMI] SMIvision: Sensomotoric instruments gmbh. <http://www.smivision.com/>
- [STET01] Spivey, M.J., Tyler, M.J., Eberhard, K.M., Tanenhaus, M.K.: Linguistically mediated visual search. *Psychol. Sci.* **12**, 282–286 (2001)
- [SLBJ03] Suh, B., Ling, H., Bederson, B.B., Jacobs, D.W.: Automatic thumbnail cropping and its effectiveness. In: ACM Symposium on User interface Software and Technology (2003)
- [SW01] Sussman, E.S., Winkler, I.: Dynamic sensory updating in the auditory system. *Cogn. Brain Res.* **12**, 431–439 (2001)
- [Sus05] Sussman, E.S.: Integration and segregation in auditory scene analysis. *J. Acoust. Soc. Am.* **117**, 1285–1298 (2005)
- [TG88] Treisman, A.M., Gormican, S.: Feature analysis in early vision: evidence from search asymmetries. *Psychol. Rev.* **95**(1), 15–48 (1988)
- [TG80] Treisman, A.M., Gelade, G.: A feature-integration theory of attention. *Cogn. Psychol.* **12**(1), 97–136 (1980)
- [TTDC06] Triesch, J., Teuscher, C., Deák, G.O., Carlson, E.: Gaze following: why (not) learn it? *Dev. Sci.* **9**(2), 125–147 (2006)

- [Uni] University of British Columbia: Curious George Project. https://www.cs.ubc.ca/labs/lci/curious_george/. Accessed 3 April 2014
- [WK06] Walther, D., Koch, C.: Modeling attention to salient proto-objects. *Neural Netw.* **19**(9), 1395–1407 (2006)
- [WWW04] Weissman, D.H., Warner, L.M., Woldorff, M.G.: The neural mechanisms for minimizing cross-modal distraction. *J. Neurosci.* **24**, 10 941–10 949 (2004)
- [Wika] Wikimedia Common (Blausen.com staff): Blausen gallery 2014, ear anatomy. http://commons.wikimedia.org/wiki/File:Blausen_0328_EarAnatomy.png. 23 Feb 2015 (License CC BY 3.0)
- [Wikb] Wikimedia Common (Blausen.com staff): Blausen gallery 2014, the internal ear. http://commons.wikimedia.org/wiki/File:Blausen_0329_EarAnatomy_InternalEar.png. 23 Feb 2015 (License CC BY 3.0)
- [Wikd] Wikimedia Common (Oarih): Cochlea-crosssection. <http://commons.wikimedia.org/wiki/File:Cochlea-crosssection.png>. 23 Feb 2015 (License CC BY-SA 3.0)
- [WTSH+03] Winkler, I., Teder-Salejarvi, W.A., Horvath, J., Naatanen, R., Sussman, E.: Human auditory cortex tracks task-irrelevant sound sources. *Neuroreport* **14**, 2053–2056 (2003)
- [WCS+05] Winkler, I., Czigler, I., Sussman, E., Horvath, J., Balazs, L.: Preattentive binding of auditory and visual stimulus features. *J. Cogn. Neurosci.* **17**, 320–339 (2005)
- [WS13] Winkler, S., Subramanian, R.: Overview of eye tracking datasets. In: *International Workshop on Quality of Multimedia Experience* (2013)
- [WCS+15] Woertwein, T., Chollet, M., Schauerte, B., Stiefelwagen, R., Morency, L.-P., Scherer, S.: Multimodal public speaking performance assessment. In: *Proceedings of the 17th International Conference on Multimodal Interaction (ICMI)*. ACM (2015)
- [WSMS15] Woertwein, T., Schauerte, B., Mueller, K., Stiefelwagen, R.: Interactive web-based image sonification for the blind. In: *Proceedings of the 17th International Conference on Multimodal Interaction (ICMI)*. ACM (2015)
- [WHK+04] Wolfe, J.M., Horowitz, T.S., Kenner, N., Hyle, M., Vasan, N.: How fast can you change your mind? the speed of top-down guidance in visual search. *Vis. Res.* **44**, 1411–1426 (2004)
- [Wol94] Wolfe, J.M.: Guided search 2.0: a revised model of visual search. *Psychon. Bull. Rev.* **1**, 202–238 (1994)
- [WCF89] Wolfe, J.M., Cave, K., Franzel, S.: Guided search: an alternative to the feature integration model for visual search. *J. Exp. Psychol.: Hum. Percept. Perform.* **15**, 419–433 (1989)
- [WBB+96] Woodruff, P.W., Benson, R.R., Bandettini, P.A., Kwong, K.K., Howard, R.J., Talavage, T., Belliveau, J., Rosen, B.R.: Modulation of auditory and visual cortex by selective attention is modality-dependent. *Neuroreport* **7**, 1909–1913 (1996)
- [XZKB10] Xu, T., Zhang, T., Kühnlenz, K., Buss, M.: Attentional object detection of an active multi-vocal vision system. *Int. J. Humanoid.* **7**(2) (2010)
- [Yar67] Yarbus, A.L.: *Eye Movements and Vision*. Plenum Press (1967)
- [ZTM+08] Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: a bayesian framework for saliency using natural statistics. *J. Vis.* **8**(7) (2008)
- [GZMT12] Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Context-aware saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell* (2012)

Multimodal Computational Attention for Scene
Understanding and Robotics

Schauerte, B.

2016, XXIV, 203 p. 55 illus., 51 illus. in color., Hardcover

ISBN: 978-3-319-33794-4