Springer Series in Cognitive and Neural Systems 10

Matei Mancas
Vincent P. Ferrera
Nicolas Riche
John G. Taylor† *Editors*

From Human Attention to Computational Attention

A Multidisciplinary Approach

Springer

Springer Series in Cognitive and Neural Systems

Volume 10

More information about this series at http://www.springer.com/series/8572

Matei Mancas • Vincent P. Ferrera • Nicolas Riche
John G. Taylor[†]
Editors

From Human Attention to Computational Attention

A Multidisciplinary Approach

Springer

Editors

Matei Mancas
Numediart Institute
University of Mons (UMONS)
Mons, Belgium

Nicolas Riche
Numediart Institute
University of Mons (UMONS)
Mons, Belgium

Vincent P. Ferrera
Department of Neuroscience
Columbia University
New York, NY, USA

John G. Taylor (deceased)
King's College
London, UK

ISSN 2363-9105   ISSN 2363-9113 (electronic)
Springer Series in Cognitive and Neural Systems
ISBN 978-1-4939-3433-1   ISBN 978-1-4939-3435-5 (eBook)
DOI 10.1007/978-1-4939-3435-5

Library of Congress Control Number: 2016934707

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media LLC New York

Foreword

We all know what attention is. Attention is so obvious and apparent that up until recently nobody really took notice of it. According to William James (1890), attention “is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. It implies withdrawal from some things in order to deal effectively with others.” Thus, attention is the first step towards perceiving the world. The late John G. Taylor fiercely argued that attention and consciousness are inseparable. Conscious awareness of the environment cannot occur without attention.

Mancas, Ferrera, and Riche have masterfully put together this book of attention from a multi-disciplinary perspective. To them, like James, attention is the first step to perception. Attention analyzes the world and creates inner representations of it. Thus to them, like Taylor, attention is a gate of conscious awareness at the interface between the inner and outer and it is the key to survival of the organism. Mancas, Ferrera, and Riche go one step further in detailing how modeling attention can improve current artificial intelligence and what advantages these attentive systems will have over others. Aside from being faster and more efficient in terms of memory storage, attentive AI systems will be able to detect novel patterns in their input streams of information and react appropriately to potentially dangerous situations.

Undoubtedly, this book is an enthusiastic applause of attention and it will prove highly valuable as a resource to engineers, computer scientists, and neuroscientists, as it will allow each community to see what the others do, what is left to do, and what needs to done.

Foundation for Research & Techn. Hellas
Inst. of Molecular Bio. & Biotechn.,
N. Plastira 100, 70013 Heraklion, Crete, Greece
September 11, 2015

Vassilis Cutsuridis

Contents

1 Why Do Computers Need Attention? 1
Matei Mancas, Vincent P. Ferrera, and Nicolas Riche

Part I Foundations

2 What Is Attention? 9
Matei Mancas

3 How to Measure Attention? 21
Matei Mancas and Vincent P. Ferrera

4 Where: Human Attention Networks and Their Dysfunctions After Brain Damage 39
Tal Seidel Malkinson and Paolo Bartolomeo

Part II Modeling

5 Attention and Signal Detection: A Practical Guide 63
Vincent P. Ferrera

6 Effects of Attention in Visual Cortex: Linking Single Neuron Physiology to Visual Detection and Discrimination 87
Vincent P. Ferrera

7 Modeling Attention in Engineering 105
Matei Mancas

8 Bottom-Up Visual Attention for Still Images: A Global View 123
Fred Stentiford

9 Bottom-Up Saliency Models for Still Images: A Practical Review 141
Nicolas Riche and Matei Mancas

10 Bottom-Up Saliency Models for Videos: A Practical Review 177
Nicolas Riche and Matei Mancas

11 Databases for Saliency Model Evaluation 191
Nicolas Riche

12 Metrics for Saliency Model Validation 209
Nicolas Riche

13 Study of Parameters Affecting Visual Saliency Assessment 227
Nicolas Riche

14 Saliency Model Evaluation 245
Nicolas Riche

Part III Evolution and Applications

15 Object-Based Attention: Cognitive and Computational Perspectives 271
Anna Belardinelli

16 Multimodal Saliency Models for Videos 291
Antoine Coutrot and Nathalie Guyader

17 Toward 3D Visual Saliency Modeling 305
Leroy Julien and Nicolas Riche

18 Applications of Saliency Models 331
Matei Mancas and Olivier Le Meur

19 Attentive Content-Based Image Retrieval 379
Dounia Awad, Vincent Courboulay, and Arnaud Revel

20 Saliency and Attention for Video Quality Assessment............ 399
Dubravko Culibrk

21 Attentive Robots 413
Simone Frintrop

Part IV Conclusion

22 The Future of Attention Models: Information Seeking and Self-awareness 447
Matei Mancas, Vincent P. Ferrera, and Nicolas Riche

Index............ 461

Chapter 1
Why Do Computers Need Attention?

Matei Mancas, Vincent P. Ferrera, and Nicolas Riche

The focus of this book is to present a multidisciplinary perspective on modelling of attention. In this introductory chapter, we first address the question of why one should care about modelling attention, and then we detail the structure of this book and explain who are the targeted readers.

1.1 Why Care About Attention and Attention Modelling?

1.1.1 First Step in Perception of Living Beings . . .

Any animal [1] from the tiniest insect [2] to humans is perfectly able to "pay attention". Attention is the first step of perception: it analyses the outer real world and turns it into an inner conscious representation. Even during dreams and REM sleep (Rapid Eye Movements), eye movement activity suggests that attentional mechanism is at work. But, in this case, it analyses a virtual world coming from the inner subconscious and turns it into an inner conscious representation. Attention seems to be not only the first step of perception but also the gate to conscious awareness.

M. Mancas (✉) • N. Riche
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: matei.mancas@umons.ac.be; nicolas.riche@umons.ac.be

V.P. Ferrera
Department of Neuroscience, Columbia University, 1051 Riverside Drive, Unit 87, New York, NY 10032, USA
e-mail: vpf3@cumc.columbia.edu; vincent.ferrera@gmail.com

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_1

1.1.2 ... From Foetus to Death, Awake and During Dreams ...

Attention probably arises during embryonic development in parallel with sensory systems. The development of attention may correlate with the first REM dreams beginning around the sixth month of foetal development [3]. This mechanism is one of the first cognitive processes to be set up, and factors like smoking, drugs, alcohol or even stress during pregnancy may lead to later attention disorders and even a higher chance of developing psychopathologies [4, 5]. In cognitive disorders like in autism or schizophrenia, attentive processes are highly affected, as suggested by studying eye-tracking traces which can be very different between patients and the control groups [6, 7]. The attentive process is set up as early as the prenatal period when it already begins to operate during babies' dreams. Until death, it occurs in every single moment of the day when people are awake but also during dreams. This shows the importance of attention: it cannot be dissociated from perception and consciousness. Even when a person is sleeping without dreaming and the eyes are not moving, a person can be awakened by important stimuli. Attention is never turned off; it can only be reduced to a standby mode (excepting drug-induced states when consciousness is altered or eliminated as in coma). It is thus safe to say that if there is conscious life in a body, there is attention.

1.1.3 ... Attention Is the Gate to Consciousness

As a gateway to conscious awareness at the interface between the external world and internal experience, attention can be both conscious (attentive) and unconscious (pre-attentive), and it is the key to survival. Attention is also a sign of limited computation capabilities. Vision, audition, touch, smell or taste, all provide the brain with a huge amount of information. Gigabits of rough sensorial data flow every second into the brain, which overloads the capacity to think and respond coherently. Attention provides the brain with the capacity of selecting relevant information and prioritizing tasks. While there are a lot of definitions and views of attention, the one core idea which justifies attention regardless of the discipline, methodology, or intuition is "information reduction" [8].

Attention only began to be scientifically studied from the nineteenth century with the arrival of modern experimental psychology. Some thoughts and concepts related to attention may be found in Descartes and Malebranche, but no rigorous and intensive scientific study was done until psychologists developed the tools to quantify perceptual and motor performance. How did philosophers since antiquity miss such a key concept as attention for so long? Part of the answer is given by William James, the father of psychology, in his famous definition of attention: "Everybody knows what attention is". Attention is so natural and self-evident, so linked to life and partly unconscious, so obvious that ... nobody really noticed it until recently.

1.1.4 Attention in Computers Might Be a First Step . . .

However, little by little, a new transversal research field has coalesced around the concept of “attention”, gathering first psychologists, then neuroscientists and, since the end of the1990s, engineers and computer scientists. While covering the whole research on attention would require a series of books, the topic is here narrowed to focus on attention modelling, a crucial step towards wider artificial intelligence.

Indeed, this key process of attention is currently rarely used within computers. As with the brain, a computer is a processing unit. As with the brain it has limited computation capabilities and memory. As with the brain, computers are required to analyse a surfeit of data. But unlike the brain they do not pay attention. While a classical computer will be more precise in quantifying the whole input data, an attentive computer will autonomously focus on the most “interesting” data which has several advantages:

- It will be faster and more efficient in terms of memory storage due to its ability to process only part of the input data.
- It will be able to find regularities and irregularities in the input signal and thus be able to detect and react to unexpected or abnormal events.
- It will be able to optimize data prediction by describing novel patterns, and depending on the information reduction result (how efficient the information reduction was), it will be capable of being curious, bored or annoyed. This curiosity which constantly pushes to the discovery of more and more complex patterns to better reduce information is a first step towards creativity.

1.1.5 . . . To Real Artificial Intelligence

As in humans attention is the gate to awareness and consciousness; in computers attention can lead to novel emergent computational paradigms beyond classical preprogrammed machines. To perform tasks autonomously, machines must be able to select and prioritize information. While the path towards self-programming computers is still very long, computational attention is developing at an exponentially increasing pace, letting more and more applications benefit from it.

1.2 Who Should Read This Book and Why?

The first point in this book is that we had a multidisciplinary approach of attention modelling in a world with little communication between those disciplines. This is especially the case for engineering and cognitive psychology/neuroscience. Engineers are at least aware of the fact that attention is studied in psychology and neuroscience because the first computational model [9] was based on the Koch and

Ullman architecture [10]. From that point new models emerged, and some of them are very far from the biological considerations of Koch and Ullman. Despite this diversity, engineers and computer scientists like the "cognitive" or "biologically inspired" labels even if they do not really know what a "cognitive model" should be. Despite this fact, few engineers take the time to read and understand papers on attention modelling in neuroscience. The other way around, neuroscientists are also aware about the existence of attention models in the engineering domain but often do not follow the rapid evolutions in this area. One of the main goals of the book is to show to each community some insights on what the others do and what they achieve because we think that having different views on the same issues can help improve knowledge and progress in both communities.

The second point of the book is that chapters are of a mixed complexity so they can be interesting both for students and specialists. Following the same idea, there is also a balance between theory and practical approaches, leading to both deeper understanding of attention and fast ability to test and improve existing models. This book intends to be accessible by a wide range of people. Students can easily read some of the chapters and can progressively go deeper in the topic with others. Specialists can directly focus on more complex chapters, but they can also benefit from practical reviews of others.

A third point of this book is an exhaustive application review and future research avenues that can help the reader to orient his research or application development efficiently. People from industry or researchers focusing on applications related to human perception can improve their applications by incorporating attention-related algorithms. Sometimes we realise that some applications could be improved by using attention or saliency models, but the literature is very scarce because people working in this community are not yet aware about what attention models can bring to them.

If you are a student in engineering but also in neuroscience or even psychology interested in researching the field of attention modelling, this book is everything you need to start quickly and efficiently. You can quickly acquire the state of the art in attention modelling but also see practical and exhaustive reviews.

If you already work in the field as an engineer, you will find a quick introduction to psychological and biological approaches to attention, and you will be able to go deeper in the concepts linked to attention modelling and the brain.

If you already work in the field as a neuroscientist, you will find engineering approaches to exponentially improve attention models and implement them into real-life applications. Some of the concepts used by engineers are clearly inspired from biological facts, but other much less. The latter models are also interesting because if they achieve good results in predicting human gaze, maybe part of the concepts they use might be found as relevant in the brain.

If you work in industry and focus on perception, images or sound, you might find here your next innovation. From video surveillance to ads optimisation passing by compression, robotics and computer graphics, many domains can benefit from attention models.

1.3 Book Structure

In this book, a synthesis of what attention is, how it can be measured and modelled and an overview of current and emerging applications is presented. The structure is organized around three parts.

The first one focuses on fundamentals and is a comprehensive introduction to attention modelling. These chapters attempt to answer basic questions one may have before modelling attention: why model attention in computers, what is attention or more precisely what are attentions, how to measure attention and where it is localized in the brain.

The second part deals with attention modelling itself. It begins with practical guides on signal detection and neurophysiology from the study of a single neuron to visual performance. Afterwards, attention modelling in engineering and computer science is introduced. After two chapters on the bottom-up attention models for still images which are the most common in computer science, another chapter presents attention modelling for video sequences. The set of four chapters which follow describe anything which needs to be known about model validation in computer science to assess how well the attention models can predict human eye fixations: the datasets which are used as ground truth, the metrics used to compute the similarity between the ground truth and the saliency models output, the influence of several parameters on the validation results and the validation itself on a set of state of the art models for still images and videos.

The third part discusses current developments in attention modelling in computer science with chapters on 3D saliency, multimodal saliency and the link between saliency and proto-objects. Finally, this part presents an exhaustive review of attention modelling applications followed by more in deep chapters on some of the possible applications in object recognition, video quality and robotics.

Finally, new research directions and foreseeable evolution of the field are discussed in the conclusion.

1.4 Summary

- Attention is of utmost importance: first step of perception, it is the gate to consciousness. It is active from before birth until death and during sleep and waking.
- Attention is so fundamental, and perhaps obvious, that it was not recognized as a legitimate object of inquiry until relatively recently.
- The study of attention has spread from philosophy and psychology to neuroscience and computer science.
- Attentive computers can benefit substantially from an implementation of attentive mechanism in their quest for artificial intelligence. This book focuses on the computational aspects of attention.

- The multidisciplinary approach presented here targets students and researchers (from both engineering and neuroscience communities) and developers from industry who work in applications on perception, video or sound. The latter might find here their next innovation.

References

1. Zentall, T. R. (2005). Selective and divided attention in animals. *Behavioural Processes, 69*(1), 1–15.
2. Hoy, R. R. (1989). Startle, categorical response, and attention in acoustic behavior of insects. *Annual Review of Neuroscience, 12*(1), 355–375.
3. Hopson, J. L. (1998). Fetal psychology. *Psychology Today, 31*(5), 44.
4. Mick, E., et al. (2002). Case–control study of attention-deficit hyperactivity disorder and maternal smoking, alcohol use, and drug use during pregnancy. *Journal of the American Academy of Child and Adolescent Psychiatry, 41*(4), 378–385.
5. Linnet, K. M., et al. (2003). Maternal lifestyle factors in pregnancy risk of attention deficit hyperactivity disorder and associated behaviors: Review of the current evidence. *American Journal of Psychiatry, 160*(6), 1028–1040.
6. Holzman, P. S., et al. (1974). Eye-tracking dysfunctions in schizophrenic patients and their relatives. *Archives of General Psychiatry, 31*(2), 143–151.
7. Klin, A., et al. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry, 59*(9), 809–816.
8. Itti, L., Rees, G., & Tsotsos, J. K. (Eds.). (2005). *Neurobiology of attention*. Boston: Academic Press.
9. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.
10. Koch, C., & Ullman, S. (1987). *Shifts in selective visual attention: Towards the underlying neural circuitry. Matters of intelligence* (pp. 115–141). Amsterdam: Springer.

Part I
Foundations

Chapter 2
What Is Attention?

Matei Mancas

2.1 The Study of Attention: A Transversal Approach

Human attention is a self-evident mental phenomenon that is active during every single moment of awareness. It was studied first in philosophy, followed by experimental psychology, cognitive psychology, cognitive neuroscience, and finally computer science for modelling in humans and machines. These studies emerged sequentially but they added one on top of the others as the layers of an "attention onion" (Fig. 2.1).

Due to the highly diverse applications of attention, a precise and general definition is not easy to find. Moreover, views on attention have evolved over time and research domains. This chapter is structured into two parts. In the first part, we briefly survey the long history of related research from philosophy to cognitive psychology, to which were added cognitive neuroscience and computer science. The second part of the chapter covers different aspects of attention in an attempt to arrive at a working definition.

2.2 A Short History of Attention

Attention seems almost absent from the writings until the modern age. How did most of the philosophers miss such a key concept from the ancient times to the Enlightenment? Part of the answer is probably that attention is such a self-evident part of life that very few noticed it until recently.

M. Mancas (✉)
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: matei.mancas@umons.ac.be

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_2

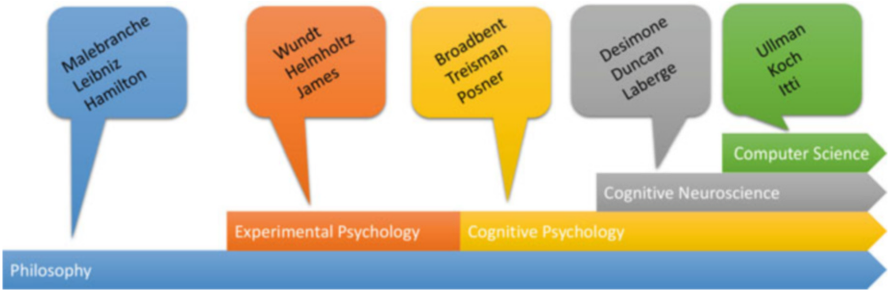


Fig. 2.1 Attention history and the attention onion, an accumulation of research domains

2.2.1 Conceptual Findings: Attention in Philosophy

Selective attention was briefly treated by Greek philosophers like Aristotle in relation to the spirit or psyche which is an early link between attention and awareness. In the fourth century, St. Augustine talked about objects of cognitive interest which can automatically tug at one's attention inferring the existence of involuntary attention. Descartes in the seventeenth century added more on the distinction between voluntary and involuntary attention. He called the first one "attention" and the latter "admiration." The concept of admiration that he linked to the notion of "wonder" is close to the idea of "surprise" which is used by some modern computational attention models.

An early and important inquiry into human attention was that of Nicolas Malebranche, a French Oratorian priest who was also a philosopher and follower of Descartes. In his "De la Recherche de la Verité" (Concerning the Search after Truth) published in 1675, Malebranche focused on the role of attention in providing structure in scene understanding and thought organization. He also saw attention as the basis of free will, writing that "the occasional cause of the presence of ideas is attention ... and it is easy to recognize, that this is the principle of our freedom" [1]. Thus, from the very beginning attention was seen as linked to volition and consciousness.

In the eighteenth century, G. W. Leibniz introduced the concept of "apperception" which refers to the assimilation of new and past experience into a current view of the world [2]. Leibniz' intuited an involuntary form of attention (known today as "bottom up" or "stimulus driven"), which is *needed for a perceived event to become conscious*. Here attention is viewed as a reflexive and involuntary gate to consciousness.

In the nineteenth century, Sir W. Hamilton, a Scottish metaphysician, challenged the previous view on attention, which consisted in thinking that humans can only focus on a single stimulus at once. Hamilton noted that when people throw marbles, the placement of about seven of the marbles could be remembered [3]. This finding opened the way to the notion of "*divided attention.*" The limited span of divided attention led about one century later to the famous paper of G.A. Miller, "The Magical Number Seven, Plus or Minus Two" in 1956 [4].

2.2.2 Attention in Experimental Psychology

After the first philosophical investigations, attention entered a scientific phase when approached by the emergence of experimental psychology in the nineteenth century. By studying individual differences in the ability of trained astronomers to judge the transit of a celestial body through a telescope, W. Wundt introduced the study of consciousness and attention to the field of psychology [5]. He interpreted this observation error as the time needed to switch voluntarily one's attention from one stimulus to another and initiated a series of studies on the speed of mental processing. This was made possible by new measuring methods proposed by F. Donders [6]. Here attention comes to be related to reflection and not reflex alone.

In the second half of the nineteenth century, H. Von Helmholtz, in his "Treatise on Physiological Optics" [7], noted that despite the illusion of seeing our entire visual environment at the same spatial resolution, humans need to move their eyes around the whole visual field "because that is the only way we can see as distinctly as possible all the individual parts of the field in turn." Although his experimental work mainly involved the analysis of the eye movement scanpath (overt attention), he also noted the existence of a covert attention, which is the ability to focus on different parts of a scene without moving the eyes. Von Helmholtz focused on the role of attention as an answer to the question *"where" the objects of interest are*. Adding to the concepts of reflex attention and divided attention, the notion of parallel versus serial processing was born.

In 1890, W. James published his textbook "The Principles of Psychology" [8] and remarked again that attention is closely related to consciousness and structure. According to James, attention makes people perceive, conceive, distinguish, remember, and shortens reaction time. He indeed linked attention to the notion of data compression and memory. He also developed a taxonomy of attention that distinguished between "passive" and "voluntary" attention. Contrary to Von Helmholtz, James was more focused on the fact that attention should answer the question of *"what" are the objects of interest*.

2.2.3 Attention in Cognitive Psychology

Between the very beginning of the twentieth century and 1949, the mainstream approach in psychology was behaviorism, which focused almost exclusively on the external causes of behavior. During this period, the study of mind was considered as barely scientific and few important advances were achieved in the field of attention. Despite this "hole" in the study of attention, important work was done on so-called interference effects. One of the most famous examples, the "Stroop effect," was reported by J. R. Stroop [9], who showed that reaction times are considerably lengthened when a single stimulus affords two conflicting responses, for example, reading a red-printed word such as “GREEN” as opposed to reporting the color of the ink in which the word was printed. Attention was invoked as a means to resolve the response conflict.

After the Second World War, a vastly more technological world emerged. Advances in information theory, statistical decision theory, and, perhaps most importantly, digital computing gave rise to the information age. Human performance in complex environments ranging from battlefields to factory floors became a central concern. The study of attention made a tremendous comeback. To the behaviorist view, which states that the organism’s behavior is controlled by stimulus–response-outcome associations, cognitive psychology showed that behavior can be modulated by attention. The resurgence of attention begun with the work of C. Cherry in 1953 on the “cocktail-party” paradigm [10]. This approach models how people select the conversation that they are listening to and ignore the rest. This problem was called “*focused attention,*” as opposed to “*divided attention.*”

In the late 1950s, D. Broadbent [11] proposed a “bottleneck” model in which he described the selective properties of attention. His idea was that *attention acts like a filter (selector) of relevant information* based on basic features, such as color or orientation for images. If the incoming information matches the filter, it can reach awareness (conscious state); otherwise it will be discarded. At that time, the study of attention seemed to become very coherent and was called “early selection.” Nevertheless, after this short positive period, most of the findings summarized by Broadbent proved to be conflicting.

The first “attack” came from the alternative model of Deutsch and Deutsch [12] who used some properties of the cocktail-party paradigm to introduce a “*late selection*” *model*, where attentional selection is basically a matter of memory processing and response selection. The idea is that all information is acquired, but only that which fits semantic or memory-related objects is selected to reach awareness. This is an opposite view to Broadbent who professed an early selection of the features before they reach any further processing.

New models were introduced like the attenuated filter model of A. Treisman [13] which is a softer version than Broadbent’s bottleneck and which let stimuli with a *response higher than a given threshold through the filter*, thus determining the focus of the selective attention.

Later, in 1980, Treisman and Gelade [14] proposed a new “feature integration” theory, where attention occurs in two distinct steps. First, *a preattentive parallel effortless step* analyzes objects and extracts features from those objects. In a second step, those *features are combined to obtain a hierarchy of focus* attention which pushes information towards awareness.

Despite its importance, the feature integration theory was also highly disputed. Other theories emerged as M. Posner [15] *spotlight supporting a spatial selection* approach or D. Kahneman [16] and his theory of capacity supporting the idea of *mental effort*.

In the late 1980s, a plethora of theories on attention flourished, and none of them was capable of accounting for all previous findings. According to H. Pashler [17], after several decades of research in cognitive psychology, more questions were raised than answers given. As a provocative rejoinder to the famous “Everyone knows what attention is” proposed by James a century before, Pashler declared that “No one knows what attention is.”

2.2.4 The Need for New Approaches: After the Late 1980s “Crisis”

Attention deals with the allocation of cognitive resources to prioritize incoming information in order to bring them to a conscious state, update the scene model and memory, and influence behavior. Between consciousness, memory, and behavior, attention was revealed to be much more complex than initially expected, and some people even questioned whether attention was a single concept or, rather, that there are several different forms of attentions. The number of issues and the complexity of the nature of attention led to an interesting move in splitting attention studies from one single community into two different communities.

The cognitive neuroscience community has the goal of getting further into the theoretical and biological nature of attention using simple stimuli. The arrival of advanced tools such as functional imaging, EEG, MEG, or single-cell recordings in awake, behaving subjects allows them to make huge steps towards relating neural recordings with behavioral correlates of attention.

The segment of the computer science community working in the field of attention has a goal of making the concept work with real data such as images, videos, audio, or 3D models. From the late 1990s and the first computational models of visual attention, the cognitive neuroscience and computer science approaches have developed in parallel, one trying to get more insight on the biological brain and the other trying to get results which can predict eye movements and other behavior for real-life stimuli and environments. Even if the computational attention community led to some models very different from what is known to happen in the brain, the engineers’ creativity is impressive, and the results on real-life data begin to be significant and the applications endless.

2.2.5 Attention in Cognitive Neuroscience

Cognitive neuroscience arrived with a whole set of new tools and methods. If some of them were already used in cognitive psychology (e.g., EEG, eye-tracking devices), others are new tools providing new insights on brain behavior:

- Psychophysiological methods: scalp recording of EEG (electroencephalography: measures the large-scale electric activity of the neurons) and MEG (magnetoencephalography: measures the magnetic fields produced by electrical currents in the brain) which are complementary in terms of sensitivity on different brain areas of interest.

- Neuroimaging methods: functional MRI and PET scan images which both measure the areas in the brain which have intense activity given a task that the subject executes (visual, audio, etc.). Magnetic resonance spectroscopy can provide information about specific neurotransmitters.
- Electrophysiological methods: single-cell recordings, which measure the electrophysiological responses of a single neuron using a microelectrode system. While this system is much more precise, it is also more invasive.
- Other methods: TMS and TDCS (transcranial magnetic stimulation and transcranial direct current stimulation, which can be used to stimulate a region of the brain and to measure the activity of specific brain circuits in humans) and multielectrodes technology which allows the study of the activity of many neurons simultaneously showing how different neuron populations interact and collaborate.

Using those techniques two main families of theories have been constructed.

The first and most well-known model is the biased competition model of Desimone and Duncan on [18]. The central idea is that at any given moment, there is more information in the environment than can be processed. Relevant information always competes with irrelevant information to influence behavior. Attention biases this competition, increasing the influence of behaviorally relevant information and decreasing the influence of irrelevant information. Desimone explicitly suggests a physiologically plausible neural basis that mediates this competition for the visual system. A receptive field of the neuron is a window to the outside world. The neuron reacts only to stimuli in this window and is insensitive to stimulation in other areas. The authors assume that the competition between stimuli takes place if more than one stimulus shares the same receptive field. This approach is very interesting as each neuron can be seen as a filter by itself and the neurons receptive field can vary from small and precise (like in the primary visual cortex V1) to large enough to focus on entire objects (like higher visual areas in the temporal and parietal lobes). This basic idea suggests different domains of attention (location-based, feature-based, object-based, attentional bottleneck) in a very natural and elegant way. Moreover, a link is achieved with memory based on the notion of attentional templates in working memory which enhance neuronal responses depending on previously acquired data. This idea is embodied in the selective tuning model of Tsotsos in 1995 [19].

The second family of models was developed by Laberge in the late 1990s [20]. It is a structural model based on neuropsychological findings and data from neuroimaging studies. Laberge conjectures that at least three brain regions are concurrently involved in the control of attention: frontal areas, especially the prefrontal cortex and thalamic nuclei, especially the pulvinar and posterior sites, the posterior parietal cortex, and the interparietal sulcus. Laberge proposes that these regions are necessary for attention, and all these regions presumably give rise to attentional control together. While cognitive neuroscience brought a lot of new methods and information to cognitive psychology, attention is still far from being fully understood, and a lot of work is undergoing in the field.

2.2.6 Attention in Computer Science

While cognitive neuroscience focuses on researching the biological nature of attention, a different angle arose in the 1980s with the improvements in computational power. Building on the feature integration theory of Treisman and Gelade [14], C. Koch and S. Ullman [21] proposed that the different visual features that contribute to attentive selection of a stimulus (color, orientation, movement, etc.) are combined into one single topographic map, called the "saliency map." The saliency map integrates the normalized information from the individual feature maps into one global measure. Bottom-up saliency is determined by how different a stimulus is from its surround at several scales. The saliency map provides the probability for each region in the visual field to be attended. This saliency map concept is close to that of the "master map" postulated in the feature integration theory by Treisman and Gelade.

The first computational implementation of Koch and Ullman architecture was achieved by Laurent Itti in his seminal work [22]. This very first computational implementation of an attention system takes as an input of any image and outputs of a saliency map of this image and also the winner-take-all-based mechanism, simulating the eye fixations during scene analysis. From that point, hundreds of models developed first for images, then for videos, and some for audio or even 3D data very recently.

From the initial biologically inspired models, a number of models based on mathematics, statistics, or information theory arrived on the "saliency market," making better and better predictions about human attention. These models are all based on features extracted from the signal (most of the time low-level features but not always), such as luminance, color, orientation, texture, motion, objects' relative position, or even simply neighborhoods or patches from the signal. Once those features are extracted, all the existing methods are essentially based on the same principle: looking for "contrasted, rare, surprising, novel, worthy-to-learn, less compressible, or information maximizing" areas. All those terms are actually synonyms, and they all amount to searching for some unusual features in a given context. This context can be local (typically center–surround spatial or temporal contrasts) and global (whole image or very long temporal history), or it can be a model of normality (the image average, the image frequency content). Very recently learning is more and more involved in computing saliency: first it was mainly about adjusting model coefficients given a precise task; now complex classifiers like deep neural networks are beginning to be used to both extract the features from the signal and train the most salient features based on ground truth obtained with eye-tracking or mouse-tracking data.

2.3 So ... What Is Attention?

The transdisciplinary nature of attention naturally leads to a lot of different definitions. Attention deals with the allocation of cognitive resources to prioritize incoming information in order to bring them to a conscious state, update a scene model, update memory, and influence behavior. But several attention mechanisms were highlighted especially from Cherry's cocktail-party phenomenon. A dichotomy appeared between divided attention and selective attention. From there, clinical observations led to a model of attention divided into five different "kinds" appeared. One can also talk about different kinds of attention that rely on gaze or not or that use only image features vs. memory and emotions ... While its purpose seems to be the relation between the outer world and inner consciousness, memory, and emotions, the clinical manifestation of attention tends to show that there might be several attentions.

2.3.1 Overt Versus Covert: The Eye

Overt versus covert attention is a distinction that was noted at the very beginning of psychological studies on attention. Overt attention is manifested by changes in posture that prepare sensory receptors for expected input. Eye movements, head movements, external ear (pinna) movements, changes in pupil size, and so forth are all examples of overt attention. Covert attention does not induce eye movements or other postural changes: it is the ability to catch (and thus be able to bring to consciousness) regions of a scene which are not fixated by the eyes. The eye achieves mainly three types of movements which are dues to the nonuniform distribution of receptive cells (cones and rods) on the retina. The cones which provide a high resolution and color are mainly concentrated in the middle of the retina in a region called "fovea." This means that in order to acquire a good spatial resolution of an image, the eye must gaze towards this precise area to align it on the fovea. This constraint led to mainly three types of eye movements:

1. Fixations: the gaze stays a minimal time period on approximately the same spatial area. The eye gaze is never still. Even when gazing a specific location, micro-saccades can be detected. The micro-saccades are very small movements of the eye during area fixations.
2. Saccades: the eyes have a ballistic movement between two fixations. They disengage from one fixation and they are very rapidly shifted to the second fixation. Between the two fixations, no visual data is acquired.
3. Smooth pursuit: a smooth pursuit is like fixation on a moving object. The eye will follow a moving object to maintain it in the fovea (central part of the retina). During smooth pursuits, more rapid small corrections can be done to correct position errors.

Modelling overt attention attempts to predict human fixation locations and the dynamical path of the eye (called the eye "scanpath").

2.3.2 Serial Versus Parallel: The Cognitive Load

While focused, sustained, and selective attention deal with a serial processing of information, alternating and divided attention deal with parallel processing of several tasks. These distinctions show that attention can deal with information both serially and in parallel. While there is a limit to the number of tasks which are processed in parallel during divided attention (around five tasks), in the case of preattentive processing, massively parallel computation can be done. Some notions such as the "gist" [23] seem to be very fast and able to process the entire visual field to get a first and very rough idea about the context of the environment. The five kinds of attention follow a hierarchy based on the degree of focus, thus the cognitive load which is needed to achieve the attentive task. This approach is sometimes called the clinical model of attention:

1. Focused attention: respond to specific stimuli (focus on a precise task).
2. Sustained attention: maintain a consistent response during longer continuous activity (stay attentive a long period of time and follow the same topic).
3. Selective attention: selectively maintain the cognitive resource on specific stimuli (focus only on a given object while ignoring distractors).
4. Alternating attention: switch between multiple tasks (stop reading to watch something).
5. Divided attention: deal simultaneously with multiple tasks (talking while driving).

2.3.3 Bottom Up Versus Top Down: Memory and Actions

Another fundamental property of attention needs to be taken into account: attention is a mix of two components referred to as bottom–up (or exogenous) and top-down (or endogenous) components. The bottom-up component is reflex-based and is driven by the acquired signal. Attention is attracted by the novelty of some features in a given context (spatial local, a contrasted region; spatial global, a red dot, while all the others are blue; temporal, a slow motion, while before motion was fast). Its main purpose is to alert in the case of unexpected or rare situations, and it is tightly related to survival. This first component of attention is the one which is the best modeled in computer science as the signal features are objective cues which can be easily extracted in a computational way.

The second component of attention (top-down) deals with individual subjective feelings. It is related to memory, emotions, and individual goals. This component of attention is less easy to model by computers as it is more subjective and it requires information about internal states, goals, a priori knowledge, or emotions. Top-down attention can be itself divided into two subcomponents:

1. Goal/action related: Depending on an individual current goal, certain features or locations are inhibited and others receive more weight. The same individual with the same prior knowledge responds differently to the same stimuli when the task in hand is different. This component is also called “volitional.”
2. Memory/emotion related: This process is related to experience and prior knowledge (and the emotions related to them). In this category one can find the scene context (experience from previously viewed scenes with similar spatial layouts or similar motion behavior) or object recognition (you see your grandmother first in the middle of other unknown people). This component of attention is more “automatic,” it does not need an important cognitive load, and it can come along with volitional attention. The other way around, the volitional top-down attention, cannot inhibit the memory-related attention which will still work even if a goal is present or not. More generally, bottom-up attention cannot be inhibited if there is a strong and unusual signal acquired. If someone searches for his keys (volitional top-down), he will not take care about a car passing by. But if he hears a strange sound (bottom-up) and then recognizes a lion (memory-related top-down attention), he will stop searching for the keys and run away. Volitional top-down attention is able to inhibit the other components of attention only if they are not very intense.

2.3.4 Attention Versus Attentions: A Summary

The study of attention is an accumulation of disciplines ranging from philosophy to computer science and passing by psychology and neuroscience. Those disciplines study sometimes different aspects or views of attention, which leads to a situation where a single and precise definition of attention is simply not feasible.

To sum up the different approaches, attention is about:

- Eye/neck mechanics and outside world information acquisition: the attentional “embodiment” leads to parallel and serial attention (overt vs. covert attention).
- Allocation of cognitive resources to important incoming information: the attentional “filtering” is the first step towards data structuring (degree of focus and clinical model of attention).
- Mutual influence on memory and emotions: passing of important information to a conscious state and get feedback from memory and emotions (bottom-up and memory-related top-down attention).
- Behavior update: react to novel situations but also manage the goals and actions (bottom-up and volitional top-down attention).

Attention plays a crucial role, partly conscious and partly unconscious, from signal acquisition to action planning going through the main cognitive steps, or maybe there are simply several attentions and not only one. At this point in time, this question still has no final answer.

References

1. Greenberg, S. (2008). Things that undermine each other: Occasionalism, freedom, and attention in Malebranche. In D. Garber & S. Nadler (Eds.), *Oxford studies in early modern philosophy* (Vol. 4, pp. 113–140). Oxford: Clarendon.
2. Runes, D. D. (Ed.). (2001). *The dictionary of philosophy*. New York: Citadel.
3. Hamilton, W. (1859). *Lectures on metaphysics and logic* (Vol. 1). Boston: Gould and Lincoln.
4. Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81.
5. Wundt, W. M. (1904). *Principles of physiological psychology* (Vol. 1). London: Sonnenschein.
6. Goldstein, E. (2014). *Cognitive psychology: Connecting mind, research and everyday experience*. Stamford: Cengage Learning.
7. von Helmholtz, H. (2005). *Treatise on physiological optics* (Vol. 3). Courier Corporation: Dover Phoenix Editions.
8. James, W. (1913). *The principles of psychology* (Vol. II). New York: Henry Holt and Co, vi, 708 pp.
9. Jensen, A. R., & Rohwer, W. D. (1966). The Stroop color-word test: A review. *Acta Psychologica, 25*, 36–93.
10. Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*(5), 975–979.
11. Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review, 64*(3), 205.
12. Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review, 70*(1), 80.
13. Treisman, A. M. (1968). *Contemporary theory and research in visual perception* (pp. 258–266). New York: Holt, Rinehart and Winston.
14. Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.
15. Gazzaniga, M. S. (Ed.). (1995). *The cognitive neurosciences* (pp. 615–624). Cambridge: The MIT Press, xiv, 1447 pp.
16. Friedenberg, J., & Silverman, G. (2011). *Cognitive science: An introduction to the study of mind*. London: Sage.
17. Pashler, H. E., & Sutherland, S. (1998). *The psychology of attention* (Vol. 15). Cambridge, MA: MIT Press.
18. Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*(1), 193–222.
19. Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence, 78*(1), 507–545.
20. Laberge. (1999). Networks of attention. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences*. Cambridge, MA: MIT Press, 2004.
21. Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. In L. M. Vaina (Ed.), *Matters of intelligence* (pp. 115–141). Amsterdam: Springer.

22. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.
23. Torralba, A., et al. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*(4), 766.

Chapter 3
How to Measure Attention?

Matei Mancas and Vincent P. Ferrera

Researchers who are interested in attention generally have one or more of the following goals: (1) to identify sources of information in the environment that are selected and prioritized by the observer, (2) to quantify the effect of attention on task performance, and (3) to identify neural correlates of attention. When considering methods to measure attention, it is important to distinguish between overt and covert orienting mechanisms. Overt attention is expressed by movements of the body and can be measured directly by determining the position and velocity of the relevant effectors – primarily the eyes, head, and hands. Covert orienting refers to the ability to direct attention without body movement and is primarily measured by differences in task performance (e.g., reaction time) that cannot be attributed to changes in the external stimulus.

In this chapter, we will focus on quantitative techniques that provide fine-grain spatial and temporal information about attentive responses at a macroscale. We do not discuss the many psychophysical paradigms that have been used to infer attention based on the speed and accuracy of observer judgments. Micro-measurements of single neuron or several neurons using microelectrodes are not described here. However, in the Chap. 6, the use of microelectrodes to measure single neuron responses is described.

At a macroscale, the attentive response can be either measured directly in the brain or indirectly through participants' behavior. Only one of the techniques that

M. Mancas (✉)
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: matei.mancas@umons.ac.be

V.P. Ferrera
Department of Neuroscience, Columbia University, 1051 Riverside Drive, Unit 87, New York, NY 10032, USA
e-mail: vpf3@cumc.columbia.edu; vincent.ferrera@gmail.com

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_3

are described here is based on participant active feedback: mouse tracking. This is because the mouse tracking feedback is very close to eye tracking, and this is an emerging approach of interest for the future: it requires less time and less money to be conducted and provides more data than classical eye tracking. All the other methods are direct or indirect and provide objective measures of attention. In a first part, the indirect methods are described, while direct methods are mainly dealt with in a second stage.

3.1 Indirect Measures of Attention

3.1.1 Eye Tracking: A Gold Standard for Overt Attention

If “the eyes are windows to the soul”, eye tracking consists of taking a look to it. Indeed, eye tracking is probably the most widely used tool for measuring visual attention. Although attention can be directed without moving the eyes, it is generally the case that humans look where they attend, and vice versa. There is ample neurophysiological support for this proposition as several structures that are involved in attention – in prefrontal cortex, parietal cortex, and the midbrain – are also involved in guiding voluntary eye movements.

Eye trackers are devices that determine the orientation of the eye relative to the head (eye in head) or to an external frame of reference (eye in space.) If head position is known, then the orbital position of the eye (eye in head) is sufficient to determine gaze direction (eye in space.)

Eye-tracking technology has evolved over time. Different technologies are described in [1]. One of the earliest techniques to be widely used is EOG (electrooculography). The eye itself generates an electric dipole oriented along the corneo-retinal axis. This potential can be measured by placing electrodes on the skin around the eye. From these electrodes, the eye orientation relative to the head can be reconstructed. To determine the orientation of the eye in space, either the head must be attached to a fixed system (chin rest or bite bar) or a head tracking system must be used in addition to the EOG. EOG signals are noisy and confounded by skin conductance or the activity of facial muscles. Reliable measurements typically require averaging over trials.

A more precise method was developed in the 1960s [2, 3] using the scleral search coil. Here, a loop of wire is embedded in an annular contact lens placed around the cornea. A small electric current is passed through the wire, generating a magnetic dipole which orientation moves with the eye. The subject sits with their head inside an oscillating magnetic field generated by a pair of large (roughly 0.6–0.9 ms in diameter) field coils. Electronics are used to sense the orientation of the scleral coil and hence the orientation of the eye. This system measures eye orientation relative to the field coils, which are fixed in space. The head generally needs to be stabilized to avoid confusing the rotation of the eye with translations due to head movement. A separate head coil can be used to record head movement. Binocular search coil systems allow experimenters to reconstruct vergence angle. Torsional eye movements can also be recorded. Scleral search coil systems provide continuous temporal resolution, limited only by front end filtering and the sampling rate of the recording device use to convert the analog signal to digital samples. Spatial resolution is typically 0.1° of visual angle or better, and noise is extremely low. Contact lens search coils can only be worn for a short time (<30 min) as they cause an increase in intraocular pressure during the time that they are in contact with the sclera. This method should be used only under the supervision of a trained clinical ophthalmologist.

The technique that most of the current commercial and research solutions use is video-oculography (VOG), based on a video camera to detect the pupil and corneal reflection. An infrared light source illuminates the eyes. The light is either reflected (bright pupil) or absorbed (dark pupil) by the pupil, and image processing software (usually embedded in dedicated hardware) is used to detect the edges of the pupil either by filling in or fitting an ellipse to the edge of the iris (Fig. 3.1). This processing also provides an estimate of pupil size. Crosshairs identify the horizontal and vertical position of the center of the pupil. Some light is also reflected from the cornea and is called the corneal reflection (CR). The position of the pupil and corneal reflection is sensitive to head movement. However, the difference (pupil – CR) discounts the influence of head motion and gives a robust estimate of eye orientation in space. Nevertheless, for precise measurements, it is more appropriate to stabilize the head with a chin rest or bite bar.

It must be kept in mind that VOG trackers operate on a two-dimensional image of the eye. To obtain eye orientation, the appropriate transformation must be done considering the geometry of the camera relative to the eye and the projection of a 3D sphere onto a 2D image. Alternatively, a look-up table matching eye position

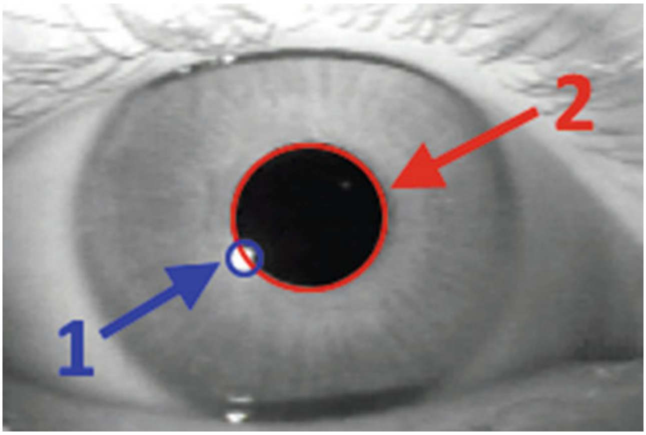


Fig. 3.1 The relative position of the pupil (*arrow 2*) and the corneal reflection (*arrow 1*) are used to compute the gaze direction

to tracker output can be generated by having subjects fixate on targets at known positions. A grid of at least nine positions should be used for this calibration. VOG systems work best when the optics of the camera are aligned with the optical axis of the eye when the subject is looking straight ahead (primary position). An infrared or “hot” mirror placed in front of the eye can be used to achieve this alignment. The infrared mirror is transparent to visible light. This way, the subject can directly view the visual display or scene through the mirror, while the camera is placed off to the side.

The temporal resolution of VOG systems is limited by the frame rate of the camera and the speed of the image processing algorithm that identifies the pupil and corneal reflection. Commercially available systems range from 30 Hz to over 1000 Hz. Spatial resolution is limited by the resolution of the camera. Typically, this is enhanced by using telephoto and close-up lenses to magnify the image of the eye. Many systems provide spatial resolution comparable to search coils (0.1° of visual angle or less). Drawbacks of VOG systems include sensitivity to stray light, which may cause large apparent changes in eye position. These systems can also be sensitive to eye color and might not work with subjects wearing glasses due to uncontrolled reflection. Furthermore, these systems are unable to function when the subject blinks and typically set their output to a default value whenever this happens.

While the fundamental technique is most of the time the same, the embodiment of the eye tracker can be very different. The main eye-tracking manufacturers propose the system under different forms [4–6].

Some eye trackers are directly incorporated into the screen (Fig. 3.2) which is used to present the data. This setup has the advantage of a very short calibration, but it can only be used with its own screen.

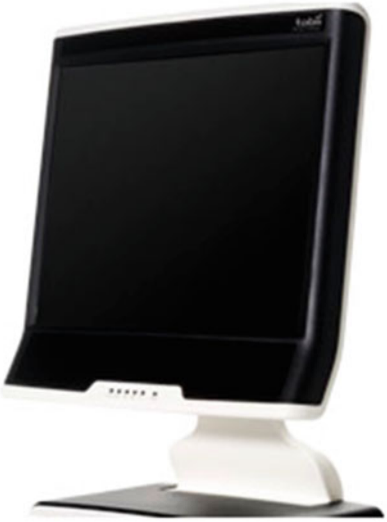


Fig. 3.2 Example of eye-tracking device included in a high-resolution screen (here a Tobii system)

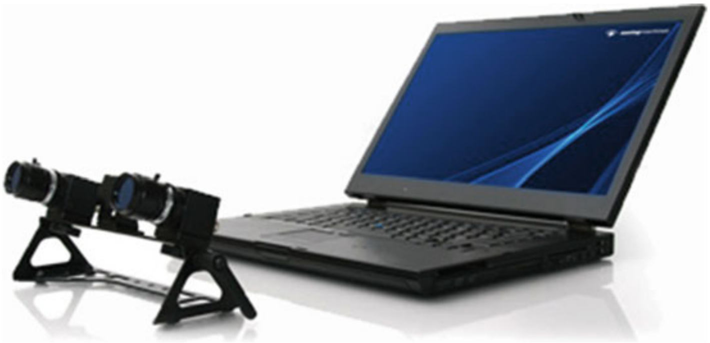


Fig. 3.3 Binocular eye-tracking system independent from the screen (here a Facelab system)

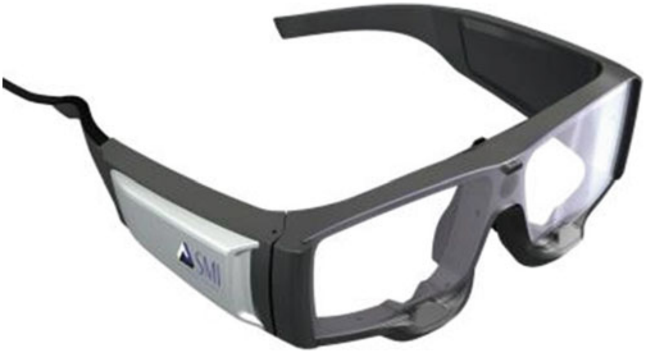


Fig. 3.4 Eye tracking embedded in wireless (here a SMI system)

Separate cameras need some additional calibration time, but the tests can be done on any screen and even in real scenes by using a scene camera to which the system needs to be calibrated (Fig. 3.3).

The eye-tracking glasses (Fig. 3.4) can be used in very ecological setups, even outside on real-life scenes. An issue of those systems is that it is not easy to aggregate the data from several viewers as the scene which is viewed is not the same. The aggregation needs a nontrivial registration of the scenes which might imply to install markers before the experiment.

Cheap devices (Fig. 3.5) come to market, and quite precise cameras are sold less than 100 EUR [7] which is a fraction of the price of a professional eye tracker. An issue with these eye trackers is that they are packaged with minimal software and it is often difficult to synchronize the stimuli and the related eye movement data. These eye trackers are mostly used as real-time human-machine interaction devices in gaming applications. Nevertheless, there are open-source projects which allow recording of data from low-cost eye trackers like Ogama [8], but mainly on still images and not moving stimuli.

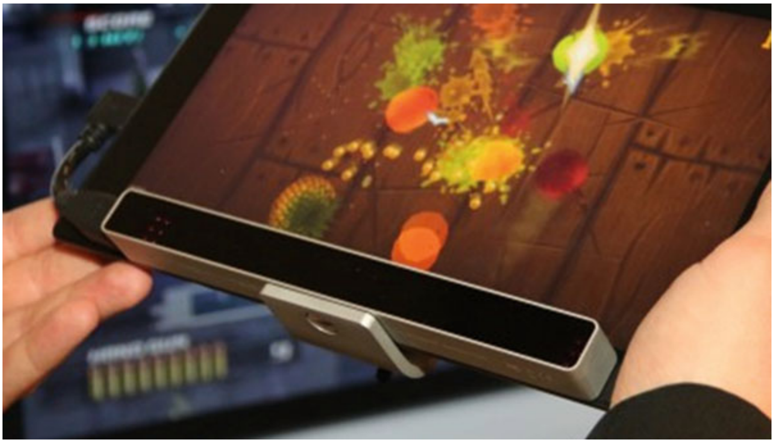


Fig. 3.5 Low-cost eye-tracking device here attached to a tablet (here the Eye-tribe system)

Finally, webcam-based software is freely available [9]. They are able to provide good quality data and to be used remotely with existing webcams [10].

Eye movement behavior has a rich variety of features that are indicative of attention. In primates, voluntary eye movements consist of saccades (rapid changes in position with peak velocity >>100°/s), vergence (changes in the alignment of the two eyes), and smooth pursuit (slow movements, generally under 100°/s, that track small moving targets). Between these movements are periods of fixation, though microscopic movements (drift, tremor, and microsaccades) may still occur even when the eye is relatively still. Fixations can be detected using clustering algorithms [11] or simply by using a double threshold: a time threshold and a spatial threshold to be sure that the gaze focused a small region. Fixation duration can be a measure of attention [12]. Fixations can be used to generate scanpaths (Fig. 3.6) or heatmaps (Fig. 3.7). A heatmap is a low-pass filtered accumulation of scan paths, and it indicates the average attention attraction of each pixel. Usually for a result to be significant, there is a need of a minimum of ten participants per stimulus.

During fixations, subjects often make very small eye movements called microsaccades [13, 47]. These are saccades with amplitudes of less than 2° of visual angle. Spontaneous microsaccades are often correlated with attention [14].

When viewing static scenes at a fixed depth, the most common eye movements are saccades, which normally occur roughly 2–3 times/s. The onset of a saccade can be detected to within a few milliseconds using algorithms based on eye velocity or acceleration. The latency of saccades relative to the sudden appearance of a target is generally 150–300 ms. Variations in saccade latency may be related to attention [15]. Attention may alter saccade direction [16], or may result in curved saccade trajectories [17].

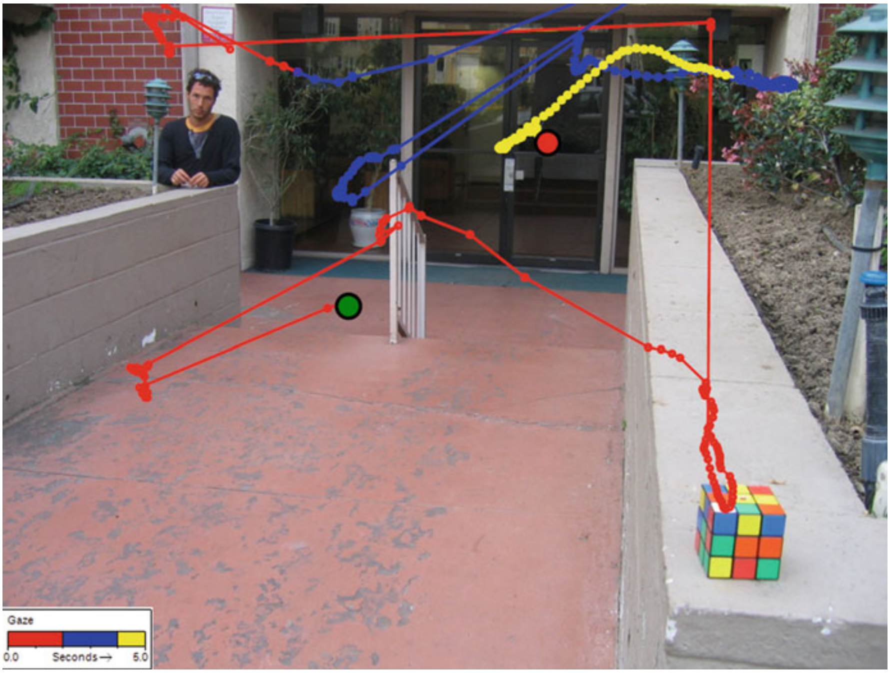


Fig. 3.6 Example of eye scan path provided by eye-tracking systems

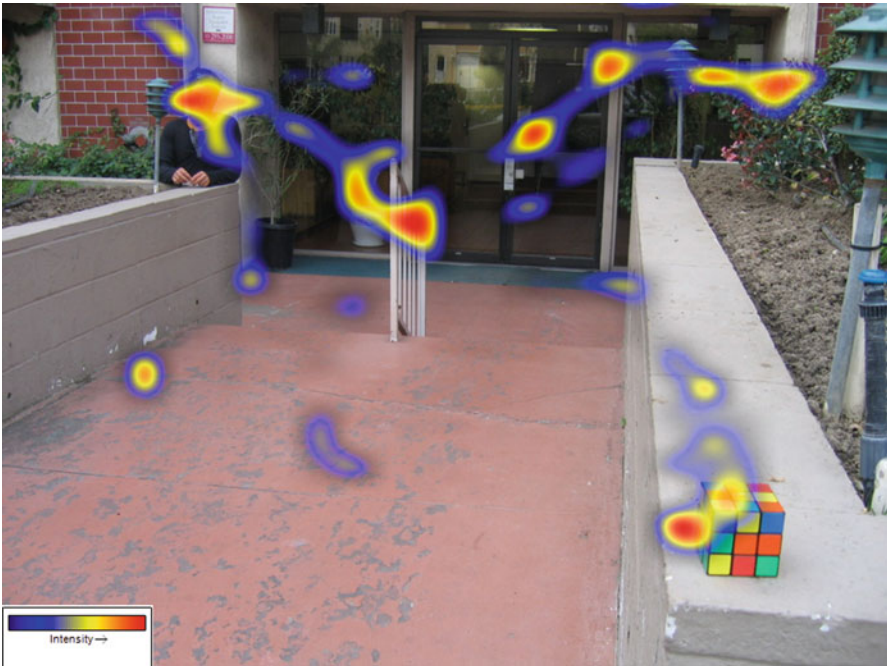


Fig. 3.7 Example of attention heatmap averaged over the participants

3.1.2 Mouse Tracking: The Low-Cost Eye Tracking

If eye tracking is the most reliable ground truth in the study of overt visual attention, it has several drawbacks in addition to the high cost of the professional devices:

- It needs minimal practice for the operator.
- The user head might need to be stabilized.
- The calibration process might be long.
- The infrared light pointing the eyes might induce eye fatigue especially during long tests.
- The system might work much less well depending on the user eye color or if she/he wears glasses.

A much simpler way to acquire data about visual attention may be the use of mouse tracking. The mouse can be precisely followed while an Internet browser is open by using a client-side language like JavaScript. The mouse precise position on the screen can be either captured using homemade code or existing libraries like [18, 19]. This technique may appear as not very reliable; however, its accuracy depends on the context of the experiment.

The first case is the one where the participant is unaware of the fact that the mouse motion is recorded. In this case, mouse motion is not accurate enough. Indeed, there is no automatic following of the eye gaze by the hand even if a tendency of the hand (and consequently the mouse) to follow the gaze is visible. Sometimes, the mouse is only used to scroll a page, and the eyes are very far from the mouse pointer, for example.

The second case is the one where the participant is aware of the experiment and has a task to follow. This can go from a simple "point the mouse where you look" instruction as in [20] where mouse tracking was used for the first time for saliency evaluation to more recent approaches as the one of SALICON in [21] where multi-resolution interactive cursor mimicking the fovea resolution is used to encourage people to point the mouse curser where they look. Indeed, as the image resolution is decreased far from the cursor, people tend to point at the locations they are interested in to have a full-resolution view of those regions.

In this second case where the participant is aware about his mouse motion tracking, the results of mouse tracking are very close to eye tracking as shown by Egner and Scheier (Fig. 3.8) on their website [22]. However, small or unconscious eye movements may be missed.

The main advantages of mouse tracking are low price and the complete transparency for the users (they just move a mouse pointer). The output can be the same as in eye tracking. It can either be a heatmap (Fig. 3.9), but also scan paths, raw data, etc.

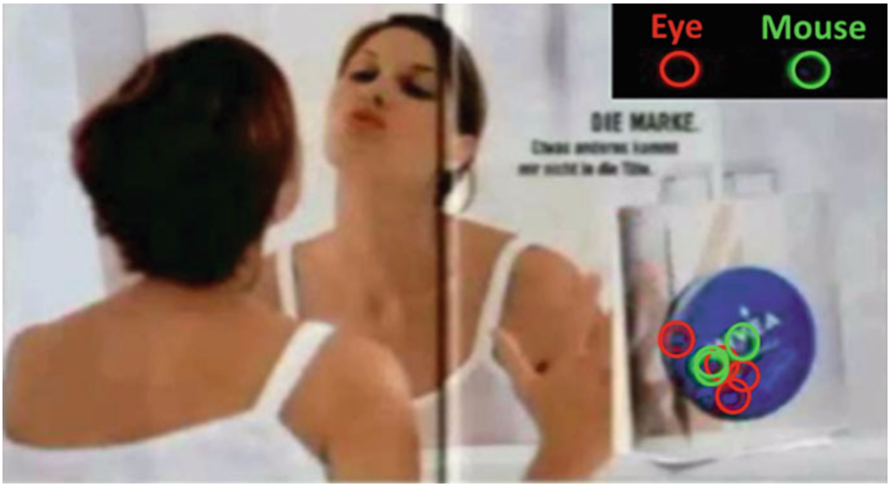


Fig. 3.8 Eye-tracking and mouse-tracking correlation (Adapted from Ref. [13])

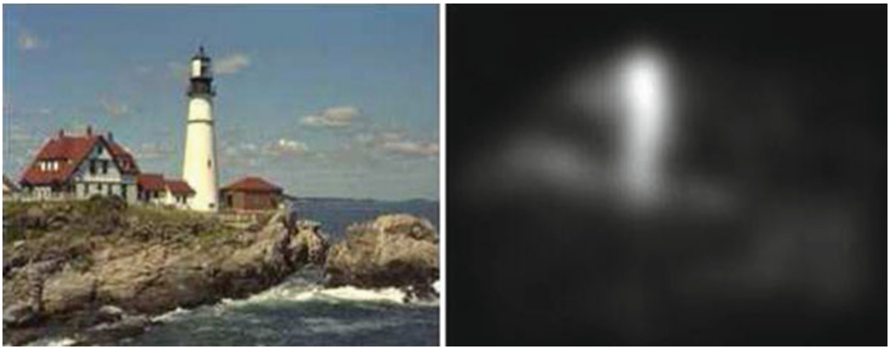


Fig. 3.9 *Left*: initial presented image. *Right*: mouse-tracking heatmap after averaging across participants

However, mouse tracking has also several drawbacks:

- The first place where the mouse pointer is located is quite important as the observer may look for the pointer. Should it be located outside the image or in the center of the image? Ideally, the pointer should initially appear randomly in the image to avoid introducing a bias of the initial position of the pointer.
- Mouse tracking only highlights areas that are consciously important for the observer. This is more a theoretical drawback than a practical one as one should try to predict the overtly interesting regions.
- The pointer hides the image region it overlaps; thus the pointer position is never on the important areas but very close to them. This drawback may be partially

eliminated by the low-pass filter step performed after the mean of the whole observer set. It is also possible to make a transparent pointer as in [21].

Mouse tracking was neglected with few publications since [20] and somehow considered as a "poor man's eye tracking." However, the rise of learning-based computational models using deep neural networks, which need huge datasets to provide correct results, has changed the situation. Mouse tracking can be done online by a virtually unlimited number of participants allowing the generation of big datasets of mouse tracking data. As eye tracking can only provide datasets with a limited number of stimuli and users per stimulus, even if they are more precise, the development of mouse tracking has certain advantages that complement eye tracking. Moreover, the combined use of eye and hand tracking can also provide insight into the deployment of attention in natural tasks [22].

3.2 Direct Measures of Attention

3.2.1 EEG: Get the Electric Activity from the Brain

The EEG technique (electroencephalography) uses electrodes placed on the participant's scalp. Those electrodes amplify the electrical potentials originating in the brain. An issue of this technique is that the skull and scalp attenuate those electrical signals.

While classical research setups have a high number of electrodes (Fig. 3.10) with manufacturers like [23, 24], some low-cost commercial systems like Emotiv [25] are more compact and easier to install and calibrate (Fig. 3.11). While the latter are easier to use, they are obviously less precise.

EEG studies provided interesting results as the modulation of the gamma band [26] during selective visual attention. Other papers [27] also provide cues about the alpha band modification during attentional shifts.

One very important cue about attention which can be measured using EEG is the P300 event-related potential (ERP).

The work of Näätänen et al. [28] in 1978 on auditory attention provided evidence that the evoked potential has an enhanced negative response when the subject was presented with rare stimuli compared to frequent ones. This negative component is called the mismatch negativity (MMN), and it was observed in several experiments. The MMN occurs 100–200 ms after the stimulus, a time that is perfectly in the range of the preattentive attention phase.

Depending on the experiments, different auditory features were isolated: audio frequency [29], audio intensity [30–32], spatial origin [33], duration [34], and phonetic changes [35]. All these features were not salient alone, but saliency was induced by the rarity of each one of these features.

The study of the MMN signal for visual attention has been investigated several times in conjunction with audio attention [36–38]. But a few experiments were made

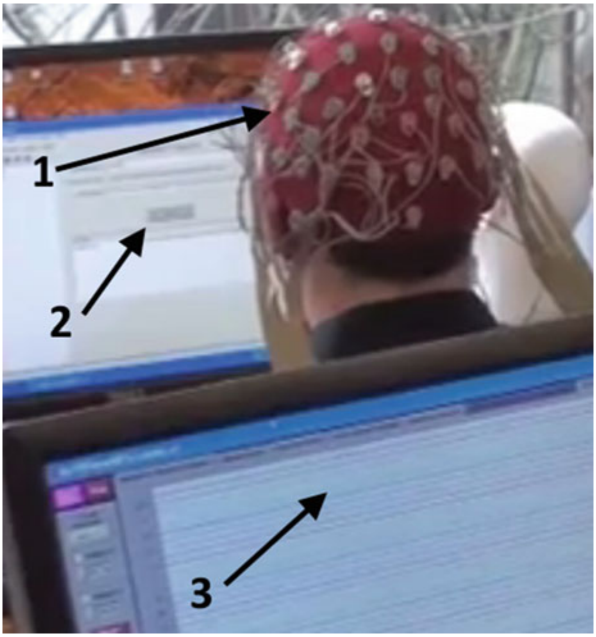


Fig. 3.10 Example of a research EEG device with a lot of electrodes (*1*), screen for the participant to visualize stimuli and tasks (*2*), screen for the operator to visualize the signals (*3*)

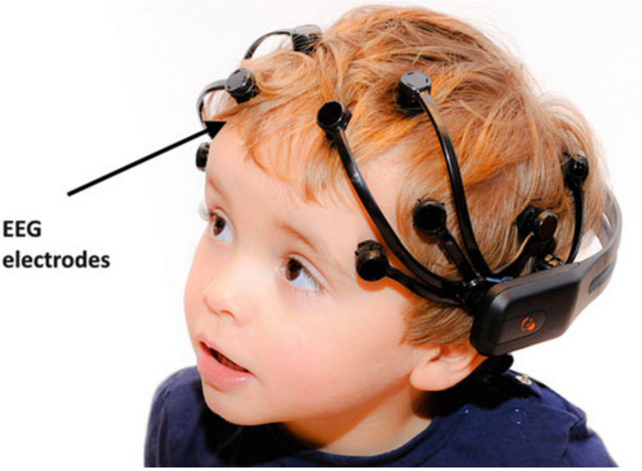


Fig. 3.11 A low-cost commercial EEG (here the Emotiv EEG system)

using only visual stimuli. Crottaz-Herbette in her thesis [39] conducted a visual experiment in the same conditions as for auditory MMN, and she has shown a high increase of the negativity of the evoked potential when seeing rare stimuli compared with the evoked potential when seeing frequent stimuli. The visual MMN occurs from 120 to 200 ms after the stimulus. The 200 ms frontier approximately matches the 200 ms needed to initiate a first eye movement, thus to engage the "attentive" serial attentional mechanism. As for the audio MMN detection, no specific task was asked to the subject who only had to hear the stimuli; this MMN component is thus preattentive, unconscious, and automatic. This study and others [40] also suggest the presence of a MMN response for the somatosensory modality (touch, taste, etc.). The MMN seems to be a universal component of the brain response reflecting an unconscious preattentive process. Any unknown stimulus (novel, rare) will be very salient as measured by P300. Rarity or novelty is a major driver of the attentional mechanism for visual, auditory, and all the other senses.

3.2.2 Functional Imaging: fMRI

MRI stands for magnetic resonance imaging. The main idea behind this kind of imaging system is that human body is mainly made of water which is itself composed of hydrogen atoms that have a single proton. Those protons have a magnetic moment (spin) which is randomly oriented most of the time. The MRI device uses a very high magnetic field (B0) to align the magnetic moment of a small fraction of protons in the patient's body. Radio frequency (RF) pulses are used to drive the proton spins into a plane orthogonal to B0. As the spins reorient or "relax" parallel to the orientation of B0, RF emissions are produced. Those emissions are captured, and an inverse Fourier transform is used to construct an image where clear gray levels mean that there are more protons; therefore, more water in the body parts (like in fat) and a darker gray levels reveal regions with less water (like bones).

MRI was initially an anatomical imaging technique, but it was soon discovered that the susceptibility artifact created by iron in the blood could be used to measure blood volume and oxygenation. Since blood volume and oxygenation respond to the metabolic demands of neural tissue, they can be used as a proxy for neuronal activity. In that way, when a region in the brain, for example, is activated, then the blood may have an increased flow. The hemodynamic response has multiple components that bear a complicated relationship to the metabolic and electrical activity of the neural tissue. Nevertheless, fMRI imaging is capable of detecting the areas in the brain which are more or less active and has become a great tool for neuroscientists to visualize which area in the brain responds during an attention-related patient exercise (Fig. 3.12).

3.2.3 Functional Imaging: MEG

MEG stands for magnetoencephalography. The idea is simple: while the EEG detects the electrical field which is heavily distorted when traversing the skull and skin, MEG detects the magnetic field induced by this electrical activity. The magnetic field has the advantage of not being influenced by the skin or the skull. While the idea is simple, in practice the magnetic field is very weak which makes

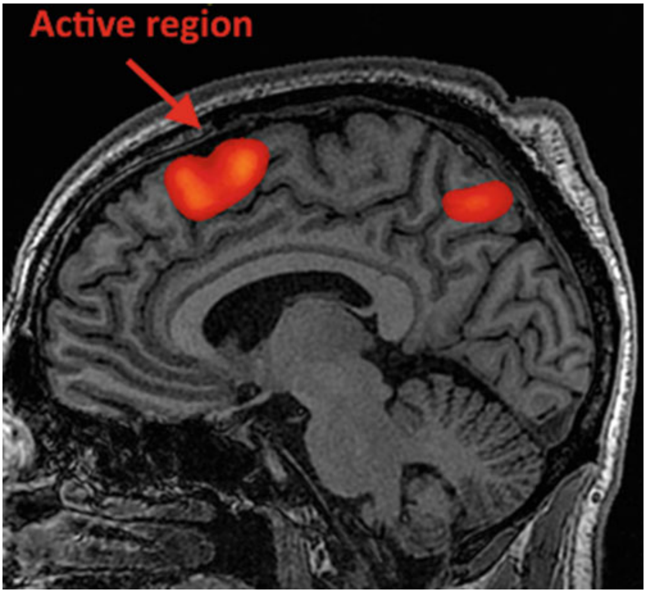


Fig. 3.12 Example of fMRI output: *red* active regions superimposed on an anatomical MRI sagittal image (Adapted from Ref. [41])

it very difficult to measure. This is why the MEG imaging is relatively new: the technological advances that allow MEG to be effective are based on SQUID (superconducting quantum interference devices). The magnetic field of the brain can induce electricity in a superconducting device which can be precisely measured. Modern devices have spatial resolutions of 2 mm and temporal resolutions of some milliseconds. Moreover, MEG images can be superimposed on MRI anatomic images which help to rapidly localize the main active areas. Finally, participants in MEG imaging can have an upright seated position (Fig. 3.13) which is more natural during testing than the horizontal position of fMRI or PET scan.

3.2.4 Functional Imaging: PET Scan

As for fMRI, PET scanning (positron electron tomography) is also a functional imaging tool, and it can thus produce also a higher signal in case of brain activity. The main idea of PET scan is that a mildly radioactive substance which is injected to the patient releases positrons (antielectrons which are particles of the same properties as an electron but with positive charges). Those positrons will almost instantaneously meet an electron and have a very exo-energetic reaction (called annihilation). This annihilation will transform the whole mass of the two particles into energy and release gamma photons in two opposite directions which will be detected by the scanner sensors. The substance which is injected will go and fixate

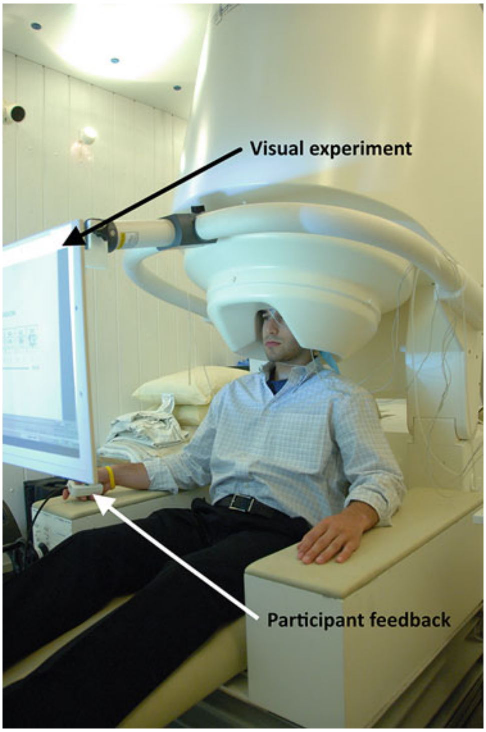


Fig. 3.13 A participant set into the MEG device and a visual experiment (Adapted from Ref. [42])

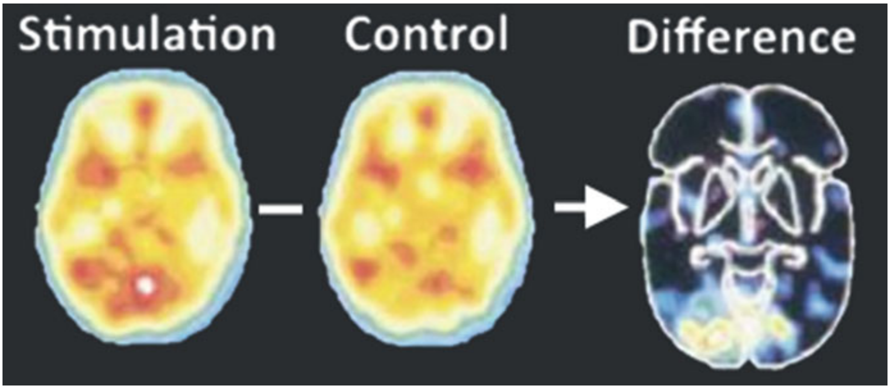


Fig. 3.14 Example of output in case of a repetitive visual pattern (flickering). The difference let us see the areas activated by the stimulus (Adapted from Ref. [43])

on the areas of the brain which are the most active, which means that those areas will exhibit a high number of annihilations. As for fMRI, the PET scan let the neuroscientists know which areas of the brain are activated when the patient is performing an attention task. Figure 3.14 shows an example of the use of PET scan to see the influence of a flickering visual pattern in the brain.

3.2.5 Complementary Techniques to Manipulate Brain Activity: TMS or tDCS

TMS stands for transcranial magnetic stimulation, and it uses electromagnetic induction to stimulate a precise region of cortex. A current passing through a coil of wire generates a magnetic field. Rapid variations of this magnetic field induce a transient electric field which in turn influences the membrane potential of nearby neurons.

Beginning with 1980s, TMS has been used first for clinical diagnostic and then in psychiatric therapy. It is now also used in conjunction with other imaging modalities such as fMRI, PET scans, and even with EEG devices.

Indeed, imaging techniques allow to find the active areas of the brain for a given task. However, they cannot say which part of those regions and when exactly they are really necessary to solve the task. By interfering with the normal functioning of a brain area, TMS, which has a very good spatiotemporal resolution, provides cues about when and where exactly a brain area is making its critical contribution to behavior.

Figure 3.15 shows a TMS which influences EEG signals (top-right), fMRI images (bottom-left), and PET scan (bottom-right).

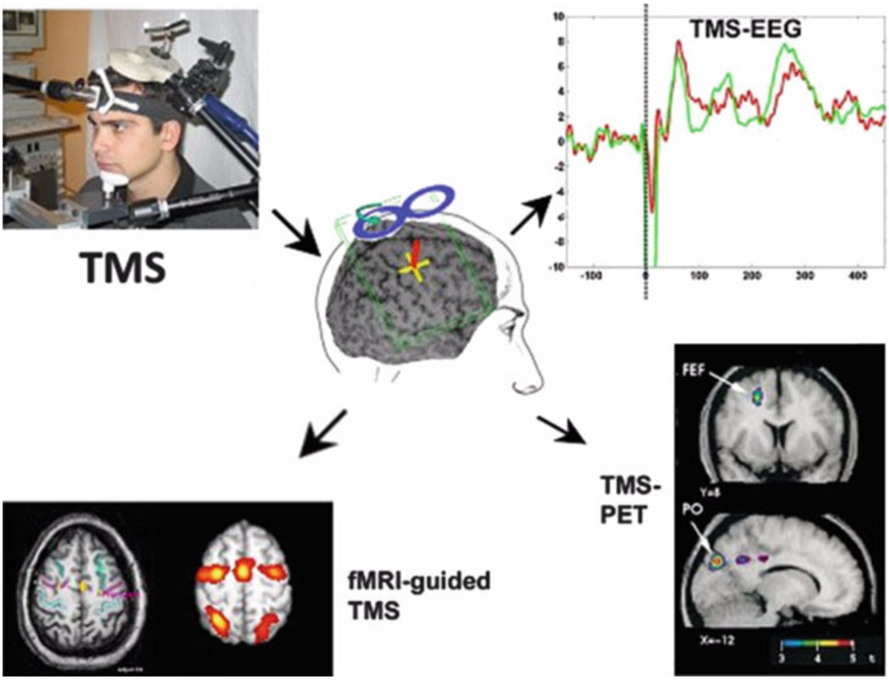


Fig. 3.15 *Top left*: a TMS setup. *Top right*: EEG modification following a TMS. *Bottom left*: fMRI images response after the TMS. *Bottom right*: PET scan response after the TMS (Adapted from Ref. [44])

Transcranial direct current stimulation (tDCS) is another method which aims in providing neurostimulation. The difference with the TMS is that it uses constant current delivered to the brain area of interest via electrodes on the scalp.

3.2.6 Functional Imaging and Attention

Positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) have been extensively used to explore the functional neuroanatomy of cognitive functions. MEG imaging becomes to be used in the field as in [45]. In [46], a review of 275 PET and fMRI studies of attention type, perception, visual attention, memory, language, etc. is described. Depending on the setup and task, a large variety of brain regions seem to be involved in attention and related functions (language, memory). This findings support again the idea that at the brain level, there are several attentions and their activity is largely distributed across almost all the brain. Attention goes from low-level to high-level processing, from reflexes to memory and emotions, and across all the human senses.

3.3 Summary

- Eye tracking remains a gold standard mainly in engineering and computer science even if it is used also in psychology.
- Mouse tracking can be more and more used with the need to build very large stimuli datasets to model attention in computer science.
- In neuroscience, fMRI has the best spatial resolution and EEG/ERP and MEG the best temporal resolution.
- fMRI has become one of the most used methods in neuroscience.
- The use of TMS or tDCS in conjunction with other imaging techniques provides precise cues about when and where exactly a brain area is making its critical contribution to behavior.

References

1. Duchowski, A. (2007). *Eye tracking methodology: Theory and practice* (Vol. 373). London: Springer Science & Business Media.
2. Rashbass, C. (1960). New method for recording eye movements. *Journal of the Optical Society of America, 50*, 642–644.
3. Robinson, D. A. (1963). A method of measuring eye movement using a scleral search coil in a magnetic field. *IEEE Transactions on Biomedical Engineering, 10*, 137–145.
4. Tobii eye tracking technology, http://www.tobii.com/
5. SMI eye tracking technology, http://www.smivision.com
6. SR-Research eye tracking technology, http://www.sr-research.com/

7. Eyetribe low cost eye-trackers, https://theeyetribe.com/
8. Open source recording from several eye trackers, http://www.ogama.net/
9. Open source eye-tracking for webcams, http://sourceforge.net/projects/haytham/
10. Xu, P., et al. (2015). TurkerGaze: Crowdsourcing saliency with webcam based eye tracking. arXiv preprint arXiv:1504.06755.
11. Konig, S. D., & Buffalo, S. D. (2014). A nonparametric method for detecting fixations and saccades using cluster analysis: Removing the need for arbitrary thresholds. *Journal of Neuroscience Methods, 227*, 21–131.
12. Findlay, J. M., & Kapoula, Z. (1992). Scrutinization, spatial attention, and the spatial programming of saccadic eye movements. *The Quarterly Journal of Experimental Psychology. A, 45*(4), 633–647.
13. Poletti, M., & Rucci, M. (2015). A compact field guide to the study of microsaccades: Challenges and functions. *Vision Research*. Epub. (2015), 83–97.
14. Yuval-Greenberg, S., Merriam, E. P., & Heeger, D. J. (2014). Spontaneous microsaccades reflect shifts in covert attention. *Journal of Neuroscience, 34*(41), 13693–13700.
15. Braun, D., & Breitmeyer, B. G. (1988). Relationship between directed visual attention and saccadic reaction times. *Experimental Brain Research, 73*(3), 546–552.
16. Kustov, A. A., & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. *Nature, 384*(6604), 74–77.
17. Doyle, M., & Walker, R. (2001). Curved saccade trajectories: Voluntary and reflexive saccades curve away from irrelevant distractors. *Experimental Brain Research, 139*(3), 333–344.
18. Heatmapjs, javascript API, http://www.patrick-wied.at/static/heatmapjs/
19. Simple mouse tracker, http://smt.speedzinemedia.com/
20. Mancas, M. (2009). Relative influence of bottom-up and top-down attention. In *Attention in cognitive systems* (pp. 212–226). Berlin/Heidelberg: Springer.
21. Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). SALICON: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, pp. 1072–1080.
22. Ballard, D. H., Hayhoe, M. M., Li, F., & Whitehead, S. D. (1992). Hand-eye coordination during sequential tasks. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 337*(1281), 331–338.
23. Mediaanlyzer web site (2015) http://www.mediaanalyzer.net
24. Cadwell EEG, http://www.cadwell.com/
25. Natus EEG, www.natus.com
26. Emotiv EEG, https://emotiv.com/
27. Müller, M. M., Gruber, T., & Keil, A. (2000). Modulation of induced gamma band activity in the human EEG by attention and visual information processing. *International Journal of Psychophysiology, 38*(3), 283–299.
28. Sauseng, P., et al. (2005). A shift of visual spatial attention is selectively associated with human EEG alpha activity. *European Journal of Neuroscience, 22*(11), 2917–2926.
29. Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica, 42*, 313–329.
30. Sams, H., Paavilainen, P., Alho, K., & Näätänen, R. (1985). Auditory frequency discrimination and event-related potentials. *Electroencephalography and Clinical Neurophysiology, 62*, 437–448.
31. Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and analysis of the component structure. *Psychophysiology, 24*, 375–425.
32. Paavilainen, P., Alho, K., Reinikainen, K., Sams, M., & Näätänen, R. (1991). Right hemisphere dominance of different mismatch negativities. *Electroencephalography and Clinical Neurophysiology, 78*, 466–479.
33. Paavilainen, P., Karlsson, M. L., Reinikainen, K., & Näätänen, R. (1989). Mismatch negativity to change in spatial location of an auditory stimulus. *Electroencephalography and Clinical Neurophysiology, 73*, 129–141.

34. Paavilainen, P., Jiang, D., Lavikainen, J., & Näätänen, R. (1993). Stimulus duration and the sensory memory trace: An event-related potential study. *Biological Psychology, 35*(2), 139–152.
35. Aaltonen, O., Niemi, P., Nyrke, T., & Tuhkahnen, J. M. (1987). Event-related brain potentials and the perception of a phonetic continuum. *Biological Psychology, 24*, 197–207.
36. Neville, H. J., & Lawson, D. (1987). Attention to central and peripheral visual space in a movement detection task: An event-related potential and behavioral study. I. Normal hearing adults. *Brain Research, 405*, 253–267.
37. Czigler, I., & Csibra, G. (1990). Event-related potentials in a visual discrimination task: Negative waves related to detection and attention. *Psychophysiology, 27*(6), 669–676.
38. Alho, K., Woods, D. L., Alagazi, A., & Näätänen, R. (1992). Intermodal selective attention. II. Effects of attentional load on processing of auditory and visual stimuli in central space. *Electroencephalography and Clinical Neurophysiology, 82*, 356–368.
39. Crottaz-Herbette, S. (2001). *Attention spatiale auditive et visuelle chez des patients héminégligents et des sujets normaux: Étude clinique, comportementale et électrophysiologique*. PhD thesis, University of Geneva, Geneva.
40. Desmedt, J. E., & Tomberg, C. (1989). Mapping early somatosensory evoked potentials in selective attention: Critical evaluation of control conditions used for titrating by difference the cognitive P30, P40, P100 and N140. *Electroencephalography and Clinical Neurophysiology, 74*, 321–346.
41. http://www.neuroscientificallychallenged.com/glossary/functional-magnetic-resonance-imaging-fmri/
42. NIMH Image library, http://infocenter.nimh.nih.gov/il/public_il/image_details.cfm?id=80
43. David Heeger courses, http://www.cns.nyu.edu/~david/courses/perception/lecturenotes/neuroimaging/neuroimaging.html
44. Centre for Cognitive Neuroimaging (CCNi), University of Glasgow, http://www.ccni.gla.ac.uk/index.php/component/content/article/46-facilities/labs/16-ranscranial-magnetic-stimulation
45. Downing, P., Liu, J., & Kanwisher, N. (2001). Testing cognitive models of visual attention with fMRI and MEG. *Neuropsychologia, 39*(12), 1329–1342.
46. Cabeza, R., & Nyberg, L. (2000). Imaging cognition II: An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience, 12*(1), 1–47.
47. Otero-Millan, J., Castro, J. L., Macknick, S. L., & Martinez-Conde, S. (2014). Unsupervised clustering method to detect microsaccades. *Journal of Vision, 14*(2), 18.

Chapter 4
Where: Human Attention Networks and Their Dysfunctions After Brain Damage

Tal Seidel Malkinson and Paolo Bartolomeo

4.1 Taxonomies of Human Attention

To behave in a coherent way in a changing environment, we need to select stimuli appropriate to our goals. On the other hand, because of capacity limitations, we must be capable of ignoring other, less important objects, which also compete to become the focus of our subsequent behavior. Neural mechanisms of attention resolve this competition by integrating the information relative to the agent's goals and to the salience of the sensorial stimuli [1]. Thus, attention to external information can help the agent select locations in space, points in time, or modality-specific input [2]. Other attention processes select, modulate, and maintain internally generated information, such as task rules, responses, long-term memory, or working memory [2].

T. Seidel Malkinson (✉)
Inserm U 1127, Hôpital Pitié Salpêtrière, ICM building, 47, bd de l'hôpital, 75013 PARIS, France

Sorbonne Universités, UPMC Univ Paris 06, F-75013 Paris, France

CNRS, UMR 7225, F-75013 Paris, France

Institut du Cerveau et de la Moelle épinière, ICM, F-75013 Paris, France
e-mail: tal.seidel@mail.huji.ac.il

P. Bartolomeo
Inserm U 1127, Hôpital Pitié Salpêtrière, ICM building, 47, bd de l'hôpital, 75013 PARIS, France

Sorbonne Universités, UPMC Univ Paris 06, F-75013 Paris, France

CNRS, UMR 7225, F-75013 Paris, France

Institut du Cerveau et de la Moelle épinière, ICM, F-75013 Paris, France
e-mail: paolo.bartolomeo@gmail.com

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_4

Attention and its neural correlates are not unitary phenomena; they can be better understood as a heterogeneous, if interacting, set of processes. According to traditional theories, attention is broadly divided into two domains: a *selectivity* aspect and an *intensity* aspect [3]. On the other hand, Parasuraman [4] identified at least three independent but interacting components of attention: (1) selection, that is, mechanisms determining more extensive processing of some input rather than others; (2) vigilance, the capacity of sustaining attention over time; and (3) control, the ability of planning and coordinating different activities (Table 4.1). Some authors have distinguished between vigilance and sustained attention as being two extremes of a continuum within the intensity domain. Thus, *vigilance* has been considered "a state of readiness to detect and respond to *small* changes occurring at random time intervals in the environment" [7] and is studied primarily through long, tedious tasks – vigils – requiring individuals to continuously monitor the environment for rare events, e.g., the detection of an infrequent blip on a radar screen. On the other hand, *sustained attention* would intervene when the flow of information is more rapid, requiring continuous active processing and monitoring [8]. For example, an interpreter giving an "online" translation of a speech would be considered to be actively sustaining attention to the words of the speaker. In our view, both ends of this intensity spectrum require holding current goal or task

Table 4.1 A schematic taxonomy of attention processes and of their anatomical bases

Type of attention	Function	Anatomy
Spatial selective attention	Orienting of attention to spatial locations and to objects in space	Bilateral DAN (superior parietal lobule, intraparietal sulcus, and dorsolateral prefrontal cortex)
Stimulus-driven attentional capture	Processing of unexpected events	Right-hemisphere VAN (inferior parietal lobule, temporoparietal junction, and ventrolateral prefrontal cortex)
Sustained (vigilant) attention or tonic alertness	Rapid responses to external stimuli (independent of their spatial position)	Right-hemisphere VAN, thalamic and brain stem nuclei (esp. locus coeruleus), anterior cingulate cortex, anterior insula
Phasic alertness	Alertness externally generated by a warning signal	Vigilant attention networks + left prefrontal cortex and thalamus
Arousal	General wakefulness and responsiveness	Diffuse cortical projections from brain stem nuclei (basal forebrain, locus coeruleus, medial forebrain bundle, dorsal raphe nucleus)
Executive control	Monitoring and conflict solving	Dorsal anterior cingulate cortex, dorsolateral prefrontal cortex, right ventrolateral anterior cingulate cortex

Note the functional and neural overlap between attentional capture (exogenous attention) and sustained (vigilant) attention. *DAN* dorsal attention network, and *VAN* ventral attention network [5]. From Bartolomeo [6]. See also Figs. 4.1 and 4.3 for an illustration of the anatomy of these brain regions.

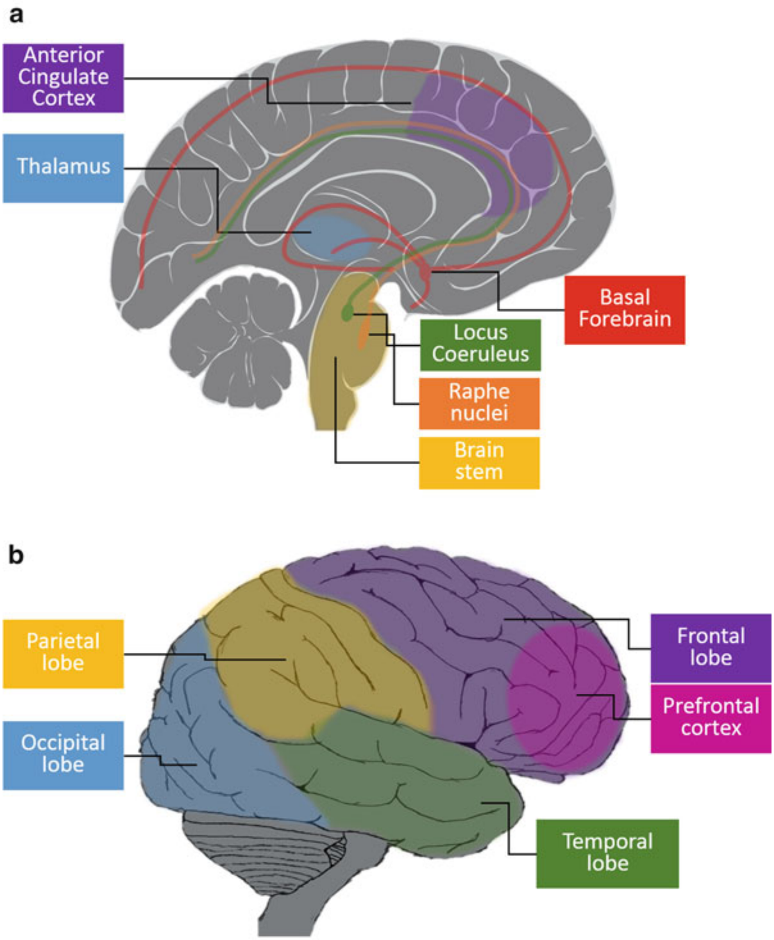


Fig. 4.1 (**a**) An illustration of the midsagittal surface of the brain, depicting the thalamus (*blue*), the anterior cingulate cortex (*purple*), the brain stem (*yellow*), the locus coeruleus and its projections (*green*), the raphe nuclei and their projections (*orange*), and the basal forebrain and its projections. (**b**) An illustration of the lateral view of the brain, depicting the occipital lobe (*blue*), the parietal lobe (*yellow*), the temporal lobe (*green*), the frontal lobe (*purple*), and the prefrontal cortex (*magenta*) within it

instructions in mind in order to monitor incoming information from the environment and produce (motor) outputs that satisfy the goal/task demands. In this sense, both vigilance and sustained attention require processes that are often termed as being “top-down” in current parlance [9].

In sum, attention must allow an organism to successfully cope with a continuously changing external and internal environment while maintaining its goals. This flexibility calls for mechanisms that (a) allow for the processing of novel, unexpected events that could be either advantageous or dangerous, in order to respond appropriately with either approaching or avoidance behavior, and (b) allow for the maintenance of finalized behavior despite distracting events [10]. For example, attention can be directed at an object in space either in a relatively reflexive way (e.g., when a honking car attracts the attention of a pedestrian) or in a more controlled mode (e.g., when the pedestrian monitors the traffic light waiting for the "go" signal to appear). It is thereby plausible that different attention processes serve these two partially conflicting goals [11]. A traditional distinction in experimental psychology refers to more exogenous (or stimulus-dependent, bottom-up) processes for orienting attention to novel events [12, 13], as opposed to more endogenous (or strategy-driven, top-down) orienting processes, which would be responsible for directing the organism's attention toward relevant targets despite the presence of distractors in the environment [14].

4.1.1 Spatial Selective Attention

The concept of spatial selective attention refers operationally to the advantage in speed and accuracy of processing for objects lying in attended regions of space as compared to objects located in non-attended regions [15].

When several events compete for limited processing capacity and control of behavior, attention selection may resolve the competition. In their influential neurocognitive model of selective attention, Desimone and Duncan [16] proposed that competition is biased toward some stimuli over others by neural attention processes on the basis of the organisms' goals and of the sensory properties of the objects, thereby giving priority to some objects over others.

A subset of these selective attention processes deals with objects in space. In ecological settings outside the laboratory, agents usually orient toward important stimuli by turning their gaze, head, and trunk toward the spatial location of the attended stimulus [17]. This is done in order to align the stimulus with the part of the sensory surface with highest resolution (e.g., the retinal fovea). This allows further perceptual processing of the detected stimulus, for example, its classification as a useful or as a dangerous object. Even very simple artificial organisms display orienting behavior when their processing resources are insufficient to process the whole visual scene in parallel [18]. However, attention can also be oriented in space without eye movements, via so-called "covert" orienting [15].

4.1.2 Cued Detection Tasks

Posner and his co-workers developed a manual response time (RT) paradigm to study the covert orienting of attention. Subjects are presented with three horizontally arranged boxes (Fig. 4.2). They fixate the central box and respond by pressing a

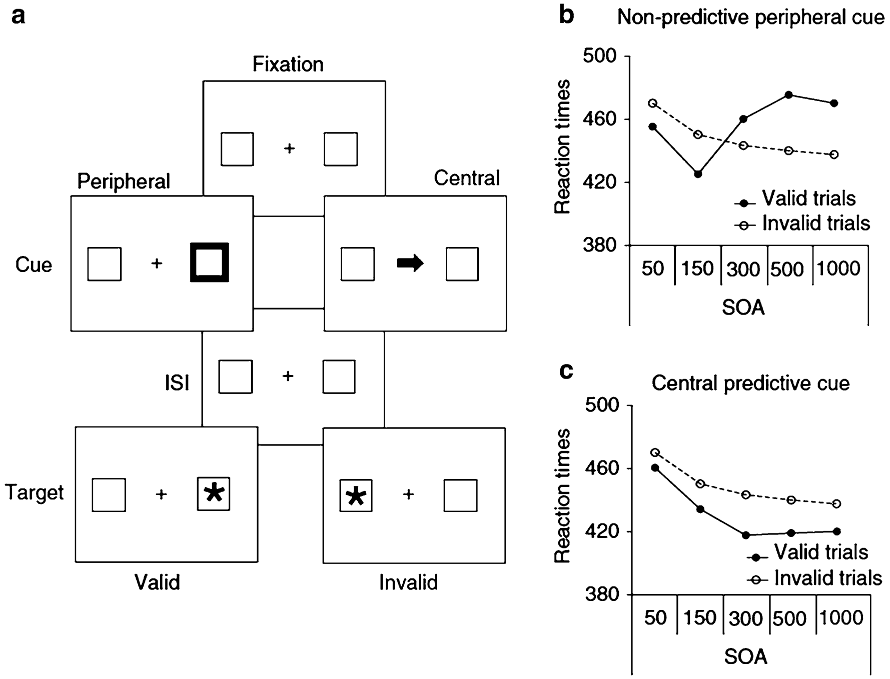


Fig. 4.2 Frontoparietal attention networks in the monkey and in the human brain. *SLF* superior longitudinal fasciculus, *DAN* dorsal attention network (intraparietal sulcus/superior parietal lobule and frontal eye field/dorsolateral prefrontal cortex), *VAN* ventral attention network (temporoparietal junction, TPJ, and inferior/middle frontal gyri). The DAN is often considered to be bilateral and symmetric; the VAN is lateralized to the right hemisphere (From Bartolomeo et al. [19])

key to a target (an asterisk) appearing in one of two lateral boxes. The target is preceded by a cue indicating one of the two lateral boxes. Cues can be either central (an arrow or another symbol presented in the central box) or peripheral (a brief brightening of one peripheral box). Valid cues correctly predict the box in which the target will appear, whereas invalid cues indicate the wrong box. Normal subjects usually show a cue validity effect consisting in faster RTs and increased accuracy for valid cue-target trials than for invalid trials (but see the phenomenon of inhibition of return described below). This suggests that the cue prompts an orienting of attention toward the cued location, which speeds up the processing of targets appearing in that region and slows down responses to targets appearing in other locations.

In this paradigm, it is often the case that a large majority (e.g., 80 %) of cues are valid; in this case, cues are said to be informative of the future position of the target. Alternatively, cues may be non-informative, when targets can appear with equal probabilities in the cued or in the uncued location. Peripheral non-informative cues attract attention automatically or exogenously. This exogenous attention shift (revealed by a cue validity effect) is typically observed only for short stimulus-onset asynchronies (SOAs) between cue and target. For SOAs longer than 300 ms, uncued targets evoke faster responses than cued targets [20–22]. This phenomenon is known as inhibition of return (IOR) [23, 24] and is often interpreted as reflecting a mechanism which promotes the exploration of the visual scene by inhibiting repeated orientations toward the same locations (but see [21, 25, 26]). Exogenous, or stimulus-dependent, and endogenous, or strategy-driven, mechanisms of attention orienting are thus qualitatively different, though highly interactive, processes [11]. An interesting property of exogenous orienting of attention is that it does not remain focused on the stimulated spatial position, but tends to spread to the whole perceptual object presented in that region [27, 28].

4.2 Networks of Human Attention

4.2.1 Sustaining Attention in Time

An important component of attention, which does not necessarily involve selection, is the capacity to rapidly respond to external stimuli, whether or not accompanied by distractors. This aspect is often referred to as alertness, vigilant or sustained attention, with a typical time span measured in seconds [29].

The alerting system is believed to produce a general alert state that would be responsible for spreading attention over a broad area of space and is believed to be modulated by the locus coeruleus (Fig. 4.1a) [30], a collection of neurons in the pons (part of the brain stem) that secrete the neurotransmitter norepinephrine and whose axons project throughout nearly the entire central nervous system. Release of norepinephrine increases alertness. A higher alert state allows for faster processing of information, independently of its spatial location [31]. We can voluntarily maintain our level of alertness over time, a function known as sustained attention, which involves the right prefrontal cortex (PFC, Fig. 4.1b); [32], the inferior parietal lobule (IPL), and the subcortical structures [33]. Right frontoparietal systems (Fig. 4.1b) can be important for modulating alertness, especially when alertness is to be generated in the absence of suitable external stimuli [29]. Thus, brain networks important for sustained attention include the PFC and PPC (posterior parietal cortex) primarily in the right hemisphere [34], with additional contribution from thalamic and brain stem nuclei [35].

A "salience networks" comprising the dorsal anterior cingulate cortex (dACC; Fig. 4.1a) in the medial wall of the frontal lobe, the anterior insula, the thalamus, and the anterior PFC may be important to maintain tonic (sustained) alertness and facilitate stimulus detection [36]. The ACC might thus constitute an important interface between the right frontoparietal cortical system and subcortical arousal mechanisms [29].

In particular, the ACC could assume a key role in the modulation of alertness depending on task demands [35, 37–39]. Neuroimaging studies (review in [40]) showed that task difficulty was strongly correlated with activation peaks, especially in the supracallosal part of the ACC. More difficult tasks possibly call for an increased level of alertness and a higher activation of the brain stem catecholaminergic (i.e., norepinephrine and dopamine) systems. Consistent with these notions, the ACC is densely connected to the noradrenergic [41] and cholinergic [42] subcortical systems involved in the regulation of alertness (see also [43]).

The alertness level can also be modulated experimentally by presenting warning signals that carry information about when, but not where, targets will appear. This is so-called phasic alertness. In addition to the (mainly right-lateralized) neural structures involved in sustained attention, phasic alertness is associated with activity in the left PFC and thalamus [33].

Although sometimes used interchangeably with alertness, *arousal* should be referred to general wakefulness and responsiveness and related to slow circadian rhythms. Of particular importance for arousal are systems projecting to the cortex from the brain stem [44], the cholinergic basal forebrain, the noradrenergic locus coeruleus (also implicated in alertness [45]), the dopaminergic medial forebrain bundle, and the serotoninergic dorsal raphe nucleus [29].

4.2.2 Orienting and Reorienting to Objects in Space

Today, we know a fair amount of detailed information about the anatomy, functions, dynamics, and pathological dysfunctions of the brain networks that subserve the orienting of gaze and attention in the human brain. Here, we describe some of the observations using neurophysiological techniques in the monkey or functional magnetic resonance imaging (fMRI) in humans to pinpoint the anatomical structures and networks which are activated during the performance of attention-related functions. Important components of these networks include the dorsolateral prefrontal cortex (PFC) and the posterior parietal cortex (PPC) (Fig. 4.3).

Physiological studies indicate that these two structures show interdependence of neural activity and thus compose a functional frontoparietal networks. In the monkey, analogous PPC and PFC areas show coordinated activity when the animal selects a visual stimulus as a saccade target [46].

Functional MRI studies in healthy human participants (reviewed by [5]) indicate the existence of multiple frontoparietal networks for spatial attention (Fig. 4.3, right panel).

A dorsal attention network (DAN), composed of the intraparietal sulcus (IPS)/superior parietal lobule and the frontal eye field (FEF)/dorsolateral PFC, shows increased blood oxygenation level-dependent (BOLD) responses during the orienting period. Functional MRI also demonstrated a ventral attention network (VAN), which includes the temporoparietal junction (TPJ) and the ventral PFC (inferior and middle frontal gyri), and shows increased BOLD responses when participants have to respond to targets presented in unexpected locations.

Thus, the VAN is considered important for detecting unexpected but behaviorally relevant events. Importantly, the DAN is considered to be bilateral and symmetric, whereas the VAN is strongly lateralized to the right hemisphere. According to

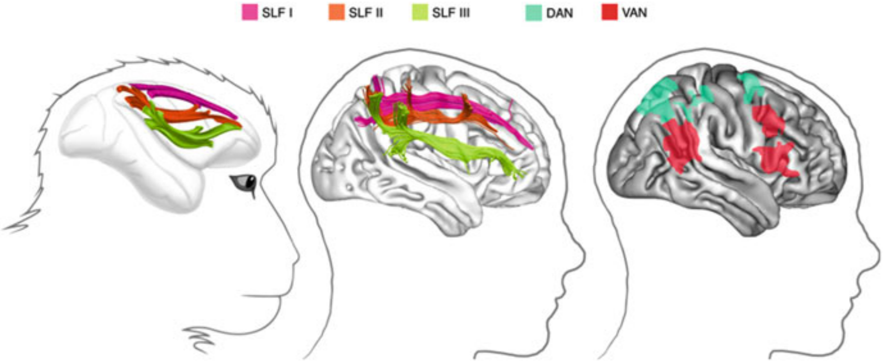


Fig. 4.3 (**a**) Illustration of a typical Posner paradigm (From Ref. [11]). Targets can be preceded by either peripheral cues (*left*) or central cues (*right*). (**b**) Typical response time results (in milliseconds) observed when peripheral non-predictive cues precede targets at different SOAs (stimulus-onset asynchronies, the time intervals between the onset of the cue and the onset of the target). Reaction times are faster for valid versus invalid trials at short SOAs, but the effect reverses at SOAs longer than 300 ms, demonstrating an IOR effect. (**c**) Typical response time results observed when central predictive cues precede targets at different SOAs. Reaction times are faster for valid versus invalid trials, and the effect is sustained even at the longest SOA. ISI (interstimulus interval) is the time interval between the end of the cue and the beginning of the target (Reproduced from Chica et al. [11]. © 2013, with permission from Elsevier)

Singh-Curry and Husain [9], the VAN is not only dedicated to salience detection in a stimulus-driven way but is also responsible for maintaining attention on goals or task demands, which is a top-down process. In support of this proposal, functional MRI has suggested a role for the inferior frontal junction (parts of Brodmann areas 9, 44, 6) in mediating interactions between bottom-up and top-down attention [47]. Furthermore, TPJ, the caudal node of the VAN, demonstrates increased BOLD response for behaviorally relevant distractors, but not for nonrelevant but highly salient ones [48].

Importantly, despite some resemblance, human and monkeys differ fundamentally in the structure and function of these two networks. A study directly comparing the brain activity in humans and monkeys during the performance of the same attention-demanding task found that the VAN is unique to humans and thus has probably developed after the evolutionary divergence of humans from monkeys [49]. Moreover, the DAN, which exists in monkeys, exhibits fundamental differences in its structure and organization between the two species. In humans, it encompasses more brain areas, and its potentially homologous areas present major differences in their basic organization, such as in their receptive field distribution. These results suggest that the human and macaque attention systems have separately evolved to meet the unique challenges each species faces [49].

Studies using noninvasive brain stimulation with transcranial magnetic stimulation (TMS) have further specified the hemispheric functions and asymmetries of the attention networks. Double pulses of TMS on the right TPJ interfered with IOR when delivered between cue and target [50], thus indicating that not only the DAN, but also the VAN does play a role during the orienting period [6]. Repetitive TMS over IPS or TPJ in the right hemisphere lastingly interfered with manual IOR for ipsilateral right-sided targets [51], thus mimicking the effects of brain lesions [52]. In sharp contrast, repetitive TMS over the homolog regions in the left hemisphere had no measurable effect on IOR [53]. Thus, there is a clear hemispheric asymmetry favoring the right hemisphere in the cortical control of IOR, which not only concerns the VAN, but also the DAN.

Importantly, and not surprisingly given the functional neuroimaging evidence of frontoparietal attention networks, PFC and PPC are directly and extensively interconnected by anatomical white matter tracts. In particular, studies in the monkey brain have identified three distinct frontoparietal long-range branches of the superior longitudinal fasciculus (SLF) on the basis of cortical terminations and course [54, 55] (see Fig. 4.3, left panel). Recent evidence from advanced in vivo tractography techniques and postmortem dissections suggests that a similar architecture exists in the human brain [56] (Fig. 4.3, middle panel). In humans, the most dorsal branch (SLF I) originates from Brodmann areas (BA) 5 and 7 and projects to BA 8, 9, and 32. The middle pathway (SLF II) originates in BA 39 and 40 within the inferior parietal lobule (IPL) and reaches prefrontal BA 8 and 9. The most ventral pathway (SLF III) originates in BA 40 and terminates in BA 44, 45, and 47. These results are consistent with the functional MRI evidence on attention networks mentioned above. In particular, the SLF III connects the cortical nodes of the VAN, whereas the DAN is connected by the human homolog of SLF I. The SLF II connects the parietal component of the VAN to the prefrontal component of the DAN, thus allowing direct communication between ventral and dorsal attention networks.

Anatomical evidence is in good agreement with asymmetries of BOLD response during functional MRI, because the SLF III (connecting the VAN) is anatomically larger in the right hemisphere than in the left hemisphere, whereas the SLF I (connecting the DAN) is more symmetrically organized [56]. SLF II also tends to be right lateralized but with substantial interindividual differences. The lateralization of SLF II is strongly correlated to behavioral signs of right-hemisphere specialization for visuospatial attention such as pseudo-neglect on line bisection, i.e., small leftward deviations of the subjective midline produced by normal individuals [57–59], and asymmetries in the speed of detection of events presented in the right or in the left hemifield [56].

4.2.2.1 Attention and Visual Perception

Cortical Streams of Visual Processing

According to an influential model [60], visual information processed in the primary visual cortex (or striate cortex, see Fig. 4.4 below) follows two major pathways in the macaque brain. A dorsal cortical visual stream, concerned with visually guided movements in space [62], but also overlapping in part with the dorsal attention

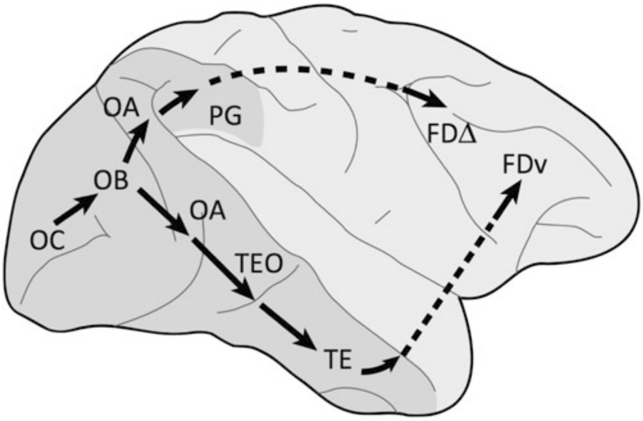


Fig. 4.4 The ventral and dorsal cortical visual streams in the macaque monkey. In the original description [60], the ventral stream is a multisynaptic pathway projecting from the striate cortex [cytoarchitectonic area (*OC*)] to area *TE* in the inferior temporal (*IT*) cortex, with a further projection from *TE* to the ventrolateral prefrontal region *FDv*. The dorsal pathway was described as a multisynaptic pathway projecting from the striate cortex to area *PG* in the inferior parietal lobule, with a further projection from *PG* to the dorsolateral prefrontal region *FD*Δ. The behavioral effects of lesions in monkeys suggested that the ventral pathway subserves object vision ("what"), whereas the dorsal pathway was characterized as supporting spatial vision ("where") (Reproduced from Ref. [61]. © 2013, with permission from Elsevier)

systems, reaches the IPL and the dorsolateral PFC. The dorsal stream is often referred to as the "where" or "how" pathway as it is concerned with where objects are located and with guidance of movements toward them. A ventral cortical visual stream, important for perceptual identification, projects from the occipital striate cortex to the inferior temporal cortex, with a further projection from the inferior temporal cortex to the ventral prefrontal cortex (Fig. 4.4). The ventral stream is often referred to as the "what" pathway, as it is involved in identifying objects.

More recently, the concept of the dorsal visual stream has been refined by the identification of several pathways emerging from the dorsal stream that consist of projections to the prefrontal and premotor cortices [63] and a further projection to the medial temporal lobe [64]. Also the ventral visual stream has recently been subdivided into several components, and the original hypothesis of a serial mode of processing from V1 to the inferior temporal cortex has now been revised to include more complex interactions, both feed-forward and feedback [61].

Indeed, the anatomy of long-range white matter tracts in these regions does suggest that both the dorsal and ventral streams can be further divided into distinct components. As mentioned before, there are at least three major subdivisions of the frontoparietal superior longitudinal fasciculus (SLF), both in the monkey [55] and in the human brain [56]. Concerning the occipitotemporal pathway, several functional systems are starting to emerge in the monkey [61]. Anatomically, two major systems

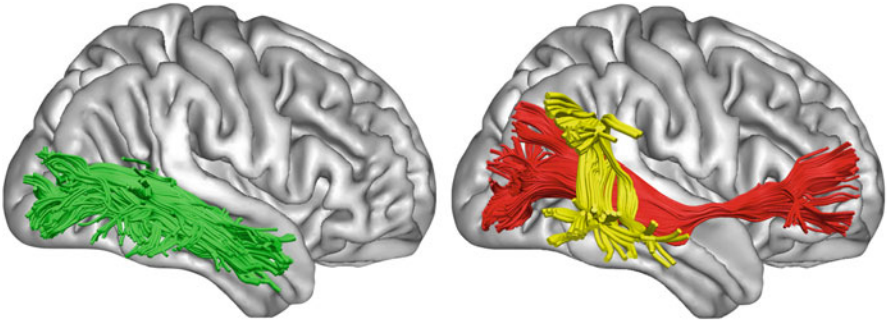


Fig. 4.5 Virtual in vivo dissection of the ILF (in *green*), the IFOF (in *red*), and the posterior segment of the superior longitudinal fasciculus (in *yellow*) (Reproduced from Thiebaut de Schotten et al. [56]. © 2012, with permission from Oxford University Press)

have been identified in the human brain. They run along the inferior longitudinal fasciculus (ILF) and the inferior fronto-occipital fasciculus (IFOF) [65] (Fig. 4.5).

Attentional Modulations of Visual Perception

Attention influences in important ways not only the perception of near-threshold visual targets [66], but also the subjective perception of suprathreshold visual stimuli, for example, by increasing spatial resolution, i.e., the ability to discriminate between two nearby points in space [67].

Thus, neural activity in the ventral visual pathways is modulated by attentional processes [68,69]. In particular, attention increases the neuronal responses and alters the profile and position of the receptive fields of ventral stream neurons near the attended location [70]. Although attention effects are seen almost all through the visual cortex, attention modulatory power follows a clear gradient. When moving up the visual processing hierarchy, the strength of attentional effects dramatically increases [71]. Attentional modulation in humans can be seen as early in the visual processing hierarchy as the LGN [72]. Moreover, the attentional modulation of population receptive fields, i.e., the "attention field," was recently studied using single-voxel modeling of fMRI time courses [73]. Attention fields were found to scale with eccentricity and varied across visual areas. In addition, voxels in multiple visual areas exhibited suppressive attentional effects such that they had an enhancing Gaussian center with a suppressive surround. This study suggests that large-scale brain networks, including frontoparietal attention networks and more ventral occipitotemporal streams of processing, are involved in conscious visual perception.

4.2.3 Target Salience

Stimuli that stand out from their surroundings are more likely to capture selective attention. This feature-based attention is influenced both by bottom-up processes, which compare the difference between a stimulus and its surroundings over different visual features like contrast, color, etc., as well as top-down processes, assessing the behavioral relevance of the stimulus. An influential computational framework explaining how salience may be computed is based on the concept of saliency maps [74, 75]. According to this framework, the visual information is first processed by early visual neurons, which are sensitive to the basic visual features of the stimulus. Locations, which significantly differ from their neighbors, are then highlighted. All highlighted locations from all feature maps are combined into a single saliency map which represents a pure salience signal that is independent of visual features [74]. The resulting sparse representation of the visual environment reflects the system's best guess as to the most relevant information [71].

Based on primate neurophysiological studies, two main saliency-related cortical regions were identified. The neuronal responses in the primate FEF (a part of DAN in the human brain) were found to be linked both to bottom-up aspects of stimulus saliency and to top-down contextual factors, suggesting it may be involved in the generation of saliency maps [5, 71]. Additionally, Bisley and Goldberg [76] proposed that area LIP acts as a priority map in which objects are represented by activity proportional to their behavioral priority, combining bottom-up inputs with an array of top-down signals. These regions seem to be tightly linked to those areas responsible for the planning and execution of eye movements, which is in agreement with the frequent need to foveate salient regions of the visual environment for a more detailed analysis [71].

In humans, target salience is often assessed using simple behavioral tasks like the oddball paradigm, in which infrequently occurring target stimuli (to which the subject must respond) are presented among a stream of frequently occurring nontarget stimuli, to which responses must be withheld [9]. The neurophysiological signature of the detection of salient events in this paradigm is a positive event-related response (ERP) centered over the parietal lobe occurring approximately 300–500 ms after target presentation but not after familiar nontargets, known as the P3 or P300 [9]. Pathological alterations of the P3 were found following lesions to the TPJ [77] and the prefrontal cortex [78] and in patients with visual neglect [79]. The cortical regions most consistently activated during target detection in functional imaging studies are the right-sided IPL, IPS, TPJ, and frontal regions, with substantial overlap with the VAN [9].

4.3 Visual Neglect

A lot can be learned about the cognitive and neuroanatomical aspects of human attention from the case of visual neglect.

This common and severely disabling neurological condition typically affects the left side of the patient's space and results from right-hemisphere damage, usually centered on the inferior parietal lobule [80] or on the superior temporal lobe [81]. Neglect patients ignore events occurring on their left, sometimes to the dramatic extent of “forgetting” to eat from the left part of their dish or of bumping into obstacles situated on their left. Patients with left neglect also display a tendency to look to right-sided details as soon as a visual scene deploys, as if their attention was “magnetically” attracted by these details [82]. They are usually unaware of their deficits (anosognosia) and often obstinately deny being hemiplegic. Neglect is a substantial source of handicap and disability for patients and entails a poor functional outcome. Unilateral neglect negatively affects patients’ motor recovery [83] and social rehabilitation. Deficits at different levels of impairment may be at work in different patients; however, the frequency and severity of attentional problems in neglect patients have been repeatedly underlined [84]. Patients with left-brain damage may also show signs of right-sided neglect, albeit more rarely and usually in a less severe form [85, 86]. For example, using a neglect battery, Bartolomeo et al. [87] found signs of contralesional neglect in 17 of 30 right-brain-damaged patients (57 %), but only in two of 30 left-brain-damaged patients (7 %). Right visual neglect seems to result from extensive left-hemisphere lesions concomitant with a (partial) right-hemisphere impairment [88]. Hence, right visual neglect might be more common with neurodegenerative conditions than with focal brain lesions ([89, 90]; but see [91]).

Neglect patients present an abnormal behavioral pattern that can be readily seen using the Posner location-cueing paradigm. In general, endogenous orienting is relatively spared, if slowed, in visual neglect, whereas exogenous orienting appears heavily biased toward the right side [84]. Specifically, in exogenous attention orientation, the patients’ RTs are much slower on both the affected and the intact side [84]. In addition, patients typically show prolonged RTs on invalid trials when the target is presented on the left, suggesting difficulties to disengage from the preceding right-sided cue and to transfer their attention [84, 92, 93]. Moreover, even when targets are presented in the right intact side, the RT pattern in exogenous orientation is abnormal: left neglect patients seem to show facilitation, instead of normal IOR, for repeated events occurring on the right, allegedly “normal” side [94]. A meta-analysis of results obtained in brain-damaged patients with the Posner paradigm revealed that (1) the disengage deficit is robust following peripheral cues but not following central cues, (2) the disengage deficit is large at shorter SOAs and decreases as SOA increases, and (3) the disengage deficit is larger in patients showing signs of unilateral neglect. The first two characteristics are typical of the operations of exogenous orienting; the third clearly links the disengage deficit to unilateral neglect [92]. Thus, the results of this meta-analysis give strong support to the hypothesis of a bias of exogenous orienting in left neglect.

Other component deficits of neglect might not necessarily be lateralized or directional problems. For example, it has been suggested that neglect results not only from an asymmetry in selective spatial attention but also from impairments in other, non-lateralized attentional components, such as arousal or vigilance [95]. Such non-lateralized deficits may be invoked to explain the fact that neglect patients are slower than normal individuals when responding to visual targets even in the ipsilesional, non-neglected space. The normal timing of attentional events also seems to be disrupted in neglect for centrally presented visual stimuli. When normal individuals have to identify two visual events appearing one shortly after another in the same spatial location, the second event goes undetected if presented in a time window of 100–450 ms after the first event ("attentional blink" [96]). Non-lateralized attentional impairments could account for the hemispheric asymmetry of unilateral neglect. Right-brain damage slows down RTs more than left-hemisphere lesions [97], which can be interpreted as an arousal deficit [98]. The preferential occurrence of a deficit of arousal after right, rather than left, brain damage might be one of the bases of the predominance in frequency and severity of contralesional neglect after right, as opposed to left, hemispheric lesions [99, 100]. One could speculate that a unilateral brain lesion generally delays the processing of information coming from the contralesional field. An additional, non-lateralized slowing of attentional operation, resulting from right-brain damage, might further hold back the processing of left stimuli, to the point of exceeding a deadline after which this information cannot affect behavior anymore [84].

Thus, an asymmetry of exogenous orienting, with rightward attentional shifts being easier than leftward shifts, compounded with non-lateralized deficits such as arousal problems, seems to accommodate the experimental evidence coming from most cases of left visual neglect [84].

Importantly, the primary regions damaged in neglect include the right ventral attention networks [101]. Moreover, damage to the long-range white matter pathways connecting parietal and frontal areas within the right hemisphere may constitute a crucial antecedent of neglect [19, 102, 103]. Thus, neglect would not result from the dysfunction of a single cortical region but from the disruption of large networks [19, 101]. Only a small number of studies utilized EEG and visually evoked potential measurements to study the neural basis of neglect, finding slowing down of activation [104] and abnormal components in late visual processing [105] reflecting perturbations in the bottom-up processing and feedback connections from higher visual areas [106]. Several fMRI studies were also conducted, showing that in these patients a lesioned VAN can induce an imbalance in DAN, with a relative hyperactivity of left-hemisphere networks [101]. However, many paramount questions concerning the mechanisms of neglect still remain open. For instance, the right-hemisphere dominance of spatial neglect is one of the most puzzling aspects of this syndrome [19, 101]. Another unresolved issue is the fact that many of the behavioral deficits characterizing neglect are traditionally associated with functions of the dorsal attention networks, but these may be anatomically spared in strokes that cause neglect [101].

It is therefore suggested that the pathological neural mechanisms at the base of neglect are very complex, resulting from disruption of the interaction between the two attention networks and an ensuing imbalance in their activation. Indeed, evidence shows that disconnection of the two attention networks is a major cause of neglect. As mentioned before, the SLF II, whose caudal cortical origin is in part shared with that of the SLF III in the IPL, connects the parietal component of the VAN to the prefrontal component of the DAN [56]. Thus, it is plausible that damage to the IPL [107], when accompanied by injury to the underlying white matter [108, 109], can produce severe and persisting signs of neglect because it can jointly disrupt the functioning of both the VAN (through SLF III disconnection) and the DAN (through SLF II damage). On the other hand, less extensive lesions, perhaps sparing a significant part of SLF II, might allow for intrahemispheric compensation mechanisms relying on the possibility of communication between VAN and DAN offered by SLF II. In this case, an initial imbalance between the dorsal frontoparietal networks, with the left-hemisphere DAN being relatively more active than its right-hemisphere counterpart, might subside after the acute phase, with consequent recovery from neglect signs [110]. Supporting these ideas, temporary inactivation of the SLF II fibers connecting the DAN and VAN in the human right hemisphere impairs the symmetrical distribution of visual attention [103]. Recently, a longitudinal study using MRI tractography found that the severity of neglect correlated with fractional anisotropy values (a measure of the directionality of the diffusion of water molecules inferring the structure of white matter fibers, such that the larger the fibers, the more directional the diffusion) in superior longitudinal fasciculus II/III for subacute patients and in its caudal portion for chronic patients [111]. The results confirm a key role of frontoparietal disconnection in the emergence and chronic persistence of neglect and demonstrate an implication of caudal interhemispheric disconnection in chronic neglect. Such disconnections may prevent frontoparietal networks in the left hemisphere from resolving the activity imbalance with their right-hemisphere counterparts, thus leading to persistent neglect.

How do these notions map on the hypotheses concerning the organization of the attention networks in the brain? A plausible model of intra- and interhemispheric interactions in neglect [112] stipulates that damage to right-hemisphere VAN causes a functional imbalance between the left and right DANs, with a hyperactivity of the left dorsal frontoparietal networks, which would provoke an attentional bias toward right-sided objects and neglect of left-sided items. Consistent with this hypothesis, suppressive TMS on left frontoparietal networks correlated with an improvement of patients' performance on cancelation tests [113]. However, evidence also suggests that the left, unimpaired hemisphere may be crucial for long-term recovery from neglect [111, 114]. Thus, the classical hemispheric rivalry hypothesis of neglect, according to which neglect symptoms result from a hyperactive left hemisphere [115], appears to be too simplistic to account for all the available data. Also, Singh-Curry and Husain [9] argued that the VAN is not only dedicated to salience detection in a stimulus-driven way [5] but is also responsible for maintaining attention on goals or task demands, which is a top-down process. In support of this proposal, functional MRI has suggested a role for the inferior frontal junction (parts of BA 9, 44, 6) in mediating interactions between bottom-up and top-down attention [47]. Furthermore, TPJ, the caudal node of the VAN, demonstrates increased BOLD response for behaviorally relevant distractors but not for nonrelevant but highly salient ones (but see [48, 116]). Thus, deficits in these nonspatial aspects of attention may lead to an exacerbation of the spatial bias in neglect patients [117].

4.4 Conclusion

- Attention is necessary for the production of coherent behavior, taking into account both internal goals and the dynamic external environment.
- There are several distinct attention processes, each subserved by different anatomical brain regions.
- Sustained attention in time depends on various cortical regions, like the right frontoparietal networks, the insula, and the anterior cingulate cortex, and on subcortical structures such as the thalamus and brain stem nuclei.
- Spatial selective attention relies on the integrated functioning of frontoparietal networks, which exhibit a right-hemispheric bias.
- Brain damage resulting in an impairment in the coordinated functioning of these frontoparietal networks may hamper the conscious perception of objects in space and lead to a significant disability for patients.
- Our knowledge of these systems is still too limited to develop a treatment for the whole range of attentional impairments, but it is expanding at fast pace, offering hope for the future.

References

1. Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron, 58*(3), 306–324.
2. Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology, 62*(1), 73–101.
3. Posner, M. I., & Boies, S. J. (1971). Components of attention. *Psychological Review, 78*(5), 391.
4. Parasuraman, R. (1998). The attentive brain: Issues and prospects. In R. Parasuraman (Ed.), *The attentive brain* (pp. 3–15). Cambridge, MA: The MIT Press.
5. Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201–215.
6. Bartolomeo, P. (2014). *Attention disorders after right brain damage: Living in halved worlds*. London: Springer.
7. Mackworth, N. H. (1956). Vigilance. *Nature, 178*(4547), 1375–1377.
8. Leclercq, M. (2002). Theoretical aspects of the main components and functions of attention. In M. Leclercq & P. Zimmermann (Eds.), *Applied neuropsychology of attention: Theory, diagnosis and rehabilitation* (pp. 3–55). New York: Psychology Press.
9. Singh-Curry, V., & Husain, M. (2009). The functional role of the inferior parietal lobe in the dorsal and ventral stream dichotomy. *Neuropsychologia, 47*(6), 1434–1448.
10. Allport, D. A. (1989). Visual attention. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 631–687). Cambridge, MA: MIT Press.
11. Chica, A. B., Bartolomeo, P., & Lupiáñez, J. (2013). Two cognitive and neural systems for endogenous and exogenous spatial attention. *Behavioural Brain Research, 237*, 107–123.
12. James, W. (1890). *The principles of psychology* (Vol. 1). New York: Henry Holt.
13. Yantis, S. (1995). Attentional capture in vision. In A. F. Kramer, G. H. Coles, & G. D. Logan (Eds.), *Converging operations in the study of visual selective attention* (pp. 45–76). Washington, DC: American Psychological Association.

14. LaBerge, D., Auclair, L., & Siéroff, E. (2000). Preparatory attention: Experiment and theory. *Consciousness and Cognition, 9*, 396–434.
15. Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology, 32*, 3–25.
16. Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review in Neurosciences, 18*, 193–222.
17. Sokolov, E. N. (1963). Higher nervous functions: The orienting reflex. *Annual Review in Physiology, 25*, 545–580.
18. Di Ferdinando, A., Parisi, D., & Bartolomeo, P. (2007). Modeling orienting behavior and its disorders with "ecological" neural networks. *Journal of Cognitive Neuroscience, 19*(6), 1033–1049.
19. Bartolomeo, P., Thiebaut de Schotten, M., & Chica, A. B. (2012). Brain networks of visuospatial attention and their disruption in visual neglect. *Frontiers in Human Neuroscience, 6*, 110.
20. Maylor, E. A., & Hockey, R. (1985). Inhibitory component of externally controlled covert orienting in visual space. *Journal of Experimental Psychology. Human Perception and Performance, 11*, 777–787.
21. Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma & D. Bouwhuis (Eds.), *Attention and performance X* (pp. 531–556). London: Lawrence Erlbaum.
22. Rafal, R. D., & Henik, A. (1994). The neurology of inhibition: Integrating controlled and automatic processes. In D. Dagenbach & T. H. Carr (Eds.), *Inhibitory processes in attention, memory and language* (pp. 1–51). San Diego: Academic Press.
23. Bartolomeo, P., & Lupiáñez, J. (Eds.). (2006). *Inhibitory after-effects in spatial processing: Experimental and theoretical issues on Inhibition of Return*. Hove: Psychology Press.
24. Posner, M. I., Rafal, R. D., Choate, L. S., & Vaughan, J. (1985). Inhibition of return: Neural basis and function. *Cognitive Neuropsychology, 2*, 211–228.
25. Berlucchi, G. (2006). Inhibition of return: A phenomenon in search of a mechanism and a better name. *Cognitive Neuropsychology, 23*(7), 1065–1074.
26. Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences, 4*(4), 138–147.
27. Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion patients. *Journal of Experimental Psychology: General, 123*(2), 161–177.
28. Macquistan, A. D. (1997). Object-based allocation of visual attention in response to exogenous, but not endogenous, spatial precues. *Psychonomic Bulletin and Review, 4*(4), 512–515.
29. Robertson, I. H., & Garavan, H. (2004). Vigilant attention. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed., pp. 563–578). Cambridge, MA: MIT Press.
30. Coull, J. T., Buchel, C., Friston, K. J., & Frith, C. D. (1999). Noradrenergically mediated plasticity in a human attentional neuronal networks. *NeuroImage, 10*(6), 705–715.
31. Fernandez-Duque, D., & Posner, M. I. (1997). Relating the mechanisms of orienting and alerting. *Neuropsychologia, 35*(4), 477–486.
32. Wilkins, A. J., Shallice, T., & McCarthy, R. (1987). Frontal lesions and sustained attention. *Neuropsychologia, 25*(2), 359–365.
33. Sturm, W., & Willmes, K. (2001). On the functional neuroanatomy of intrinsic and phasic alertness. *NeuroImage, 14*(1 Pt 2), S76–84.
34. Pardo, J. V., Fox, P. T., & Raichle, M. E. (1991). Localization of a human system for sustained attention by positron emission tomography. *Nature, 349*, 61–64.
35. Sturm, W., de Simone, A., Krause, B. J., Specht, K., Hesselmann, V., Radermacher, I., et al. (1999). Functional anatomy of intrinsic alertness: Evidence for a fronto-parietal-thalamic-brainstem networks in the right hemisphere. *Neuropsychologia, 37*(7), 797–805.
36. Sadaghiani, S., Scheeringa, R., Lehongre, K., Morillon, B., Giraud, A. L., & Kleinschmidt, A. (2010). Intrinsic connectivity networks, alpha oscillations, and tonic alertness: A simultaneous electroencephalography/functional magnetic resonance imaging study. *Journal of Neuroscience, 30*(30), 10243–10250.

37. Bartolomeo, P., Zieren, N., Vohn, R., Dubois, B., & Sturm, W. (2008). Neural correlates of primary and reflective consciousness of spatial orienting. *Neuropsychologia, 46*(1), 348–361.
38. Mottaghy, F. M., Willmes, K., Horwitz, B., Muller, H. W., Krause, B. J., & Sturm, W. (2006). Systems level modeling of a neuronal networks subserving intrinsic alertness. *NeuroImage, 29*(1), 225–233.
39. Sturm, W., Longoni, F., Fimm, B., Dietrich, T., Weis, S., Kemna, S., et al. (2004). networks for auditory intrinsic alertness: A PET study. *Neuropsychologia, 42*(5), 563–538.
40. Paus, T., Koski, L., Caramanos, Z., & Westbury, C. (1998). Regional differences in the effects of task difficulty and motor output on blood flow response in the human anterior cingulate cortex: A review of 107 PET activation studies. *NeuroReport, 9*(9), R37–R47.
41. Gaspar, P., Berger, B., Febvret, A., Vigny, A., & Henry, J. P. (1989). Catecholamine innervation of the human cerebral cortex as revealed by comparative immunohistochemistry of tyrosine hydroxylase and dopamine-beta-hydroxylase. *Journal of Comparative Neurology, 279*(2), 249–271.
42. Mesulam, M. M., Hersh, L. B., Mash, D. C., & Geula, C. (1992). Differential cholinergic innervation within functional subdivisions of the human cerebral cortex: A choline acetyltransferase study. *Journal of Comparative Neurology, 318*(3), 316–328.
43. Sarter, M., Givens, B., & Bruno, J. P. (2001). The cognitive neuroscience of sustained attention: Where top-down meets bottom-up. *Behavioral Brain Research, 35*(2), 146–160.
44. Moruzzi, G., & Magoun, H. W. (1949). Brainstem reticular formation and activation of the EEG. *Electroencephalography and Clinical Neurophysiology, 1*, 455–473.
45. Aston-Jones, G., & Cohen, J. D. (2005). An Integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience, 28*(1), 403–450.
46. Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science, 315*(5820), 1860–1862.
47. Asplund, C. L., Todd, J. J., Snyder, A. P., & Marois, R. (2010). A central role for the lateral prefrontal cortex in goal-directed and stimulus-driven attention. *Nature Neuroscience, 13*(4), 507–512.
48. Indovina, I., & Macaluso, E. (2007). Dissociation of stimulus relevance and saliency factors during shifts of visuospatial attention. *Cerebral Cortex, 17*(7), 1701–1711.
49. Patel, G. H., Yang, D., Jamerson, E. C., Snyder, L. H., Corbetta, M., & Ferrera, V. P. (2015). Functional evolution of new and expanded attention networks in humans. *Proceedings of the National Academy of Sciences of the United States of America, 112*(30), 9454–9459.
50. Chica, A. B., Bartolomeo, P., & Valero-Cabre, A. (2011). Dorsal and ventral parietal contributions to spatial orienting in the human brain. *Journal of Neuroscience, 31*(22), 8143–8149.
51. Bourgeois, A., Chica, A. B., Valero-Cabre, A., & Bartolomeo, P. (2013). Cortical control of Inhibition of Return: Exploring the causal contributions of the left parietal cortex. *Cortex, 49*(10), 2927–2934.
52. Bourgeois, A., Chica, A. B., Migliaccio, R., Thiebaut de Schotten, M., & Bartolomeo, P. (2012). Cortical control of inhibition of return: Evidence from patients with inferior parietal damage and visual neglect. *Neuropsychologia, 50*(5), 800–809.
53. Bourgeois, A., Chica, A. B., Valero-Cabré, A., & Bartolomeo, P. (2013). Cortical control of inhibition of return: Causal evidence for task-dependent modulations by dorsal and ventral parietal regions. *Cortex, 49*(8), 2229–2238.
54. Petrides, M., & Pandya, D. N. (1984). Projections to the frontal cortex from the posterior parietal region in the rhesus monkey. *Journal of Comparative Neurology, 228*(1), 105–116.
55. Schmahmann, J. D., & Pandya, D. N. (2006). *Fiber pathways of the brain*. New York: Oxford University Press.
56. Thiebaut de Schotten, M., Dell'Acqua, F., Forkel, S. J., Simmons, A., Vergani, F., Murphy, D. G. M., et al. (2011). A lateralized brain networks for visuospatial attention. *Nature Neuroscience, 14*(10), 1245–1246.

57. Bowers, D., & Heilman, K. M. (1980). Pseudoneglect: Effects of hemispace on a tactile line bisection task. *Neuropsychologia, 18*, 491–498.
58. Jewell, G., & McCourt, M. E. (2000). Pseudoneglect: A review and meta-analysis of performance factors in line bisection tasks. *Neuropsychologia, 38*(1), 93–110.
59. Toba, M. N., Cavanagh, P., & Bartolomeo, P. (2011). Attention biases the perceived midpoint of horizontal lines. *Neuropsychologia, 49*(2), 238–246.
60. Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences, 6*, 414–417.
61. Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences, 17*(1), 26–49.
62. Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*(1), 20–25.
63. Rizzolatti, G., & Matelli, M. (2003). Two different streams form the dorsal visual system: Anatomy and functions. *Experimental Brain Research, 153*(2), 146–157.
64. Kravitz, D. J., Saleem, K. S., Baker, C. I., & Mishkin, M. (2011). A new neural framework for visuospatial processing. *Nature Reviews Neuroscience, 12*(4), 217–230.
65. Catani, M., Jones, D. K., Donato, R., & Ffytche, D. H. (2003). Occipito-temporal connections in the human brain. *Brain, 126*(Pt 9), 2093–2107.
66. Chica, A. B., & Bartolomeo, P. (2012). Attentional routes to conscious perception. *Frontiers in Psychology, 3*(1), 1–12.
67. Carrasco, M., Ling, S., & Read, S. (2004). Attention alters appearance. *Nature Neuroscience, 7*(3), 308–313.
68. Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science, 229*(4715), 782–784.
69. Sundberg, K. A., Mitchell, J. F., Gawne, T. J., & Reynolds, J. H. (2012). Attention influences single unit and local field potential response latencies in visual cortical area v4. *Journal of Neuroscience, 32*(45), 16040–16050.
70. Anton-Erxleben, K., & Carrasco, M. (2013). Attentional enhancement of spatial resolution: Linking behavioural and neurophysiological evidence. *Nature Reviews Neuroscience, 14*(3), 188–200.
71. Treue, S. (2003). Visual attention: The where, what, how and why of saliency. *Current Opinion in Neurobiology, 13*(4), 428–432.
72. O'Connor, D. H., Fukui, M. M., Pinsk, M. A., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience, 5*(11), 1203–1209.
73. Puckett, A. M., & DeYoe, E. A. (2015). The attentional field revealed by single-voxel modeling of fMRI time courses. *The Journal of Neuroscience, 35*(12), 5030–5042.
74. Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*(4), 219–227.
75. Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.
76. Bisley, J. W., & Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annual Review of Neuroscience, 33*, 1.
77. Knight, R. T., Scabini, D., Woods, D. L., & Clayworth, C. C. (1989). Contributions of temporal-parietal junction to the human auditory P3. *Brain Research, 502*(1), 109–116.
78. Barcelo, F., Suwazono, S., & Knight, R. T. (2000). Prefrontal modulation of visual processing in humans. *Nature Neuroscience, 3*(4), 399–403.
79. Lhermitte, F., Turell, E., LeBrigand, D., & Chain, F. (1985). Unilateral visual neglect and wave P 300: A study of nine cases with unilateral lesions of the parietal lobes. *Archives of Neurology, 42*(6), 567–573.
80. Vallar, G. (1993). The anatomical basis of spatial hemineglect in humans. In J. Marshall & I. Robertson (Eds.), *Unilateral neglect: Clinical and experimental studies* (pp. 27–59). Hove: Psychology Press.

81. Karnath, H.-O., Ferber, S., & Himmelbach, M. (2001). Spatial awareness is a function of the temporal not the posterior parietal lobe. *Nature, 411*(6840), 950–963.
82. Gainotti, G., D'Erme, P., & Bartolomeo, P. (1991). Early orientation of attention toward the half space ipsilateral to the lesion in patients with unilateral brain damage. *Journal of Neurology, Neurosurgery, and Psychiatry, 54*, 1082–1089.
83. Denes, G., Semenza, C., Stoppa, E., & Lis, A. (1982). Unilateral spatial neglect and recovery from hemiplegia: A follow-up study. *Brain, 105*(3), 543–552.
84. Bartolomeo, P., & Chokron, S. (2002). Orienting of attention in left unilateral neglect. *Neuroscience and Biobehavioral Reviews, 26*(2), 217–234.
85. Bartolomeo, P., Chokron, S., & Gainotti, G. (2001). Laterally directed arm movements and right unilateral neglect after left hemisphere damage. *Neuropsychologia, 39*(10), 1013–1021.
86. Beis, J. M., Keller, C., Morin, N., Bartolomeo, P., Bernati, T., Chokron, S., et al. (2004). Right spatial neglect after left hemisphere stroke: Qualitative and quantitative study. *Neurology, 63*(9), 1600–1605.
87. Bartolomeo, P., D'Erme, P., & Gainotti, G. (1994). The relationship between visuospatial and representational neglect. *Neurology, 44*, 1710–1714.
88. Weintraub, S., Daffner, K. R., Ahern, G. L., Price, B. H., & Mesulam, M. M. (1996). Right sided hemispatial neglect and bilateral cerebral lesions. *Journal of Neurology, Neurosurgery, and Psychiatry, 60*(3), 342–344.
89. Andrade, K., Samri, D., Sarazin, M., Cruz De Souza, L., Cohen, L., Thiebaut de Schotten, M., et al. (2010). Visual neglect in posterior cortical atrophy. *BMC Neurology, 10*, 68.
90. Bartolomeo, P., Dalla Barba, G., Boissé, M. T., Bachoud-Lévi, A. C., Degos, J. D., & Boller, F. (1998). Right-side neglect in Alzheimer's disease. *Neurology, 51*(4), 1207–1209.
91. Silveri, M. C., Ciccarelli, N., & Cappa, A. (2011). Unilateral spatial neglect in degenerative brain pathology. *Neuropsychology, 25*(5), 554–566.
92. Losier, B. J., & Klein, R. M. (2001). A review of the evidence for a disengage deficit following parietal lobe damage. *Neuroscience and Biobehavioral Reviews, 25*(1), 1–13.
93. Posner, M. I., Walker, J. A., Friedrich, F. J., & Rafal, R. D. (1984). Effects of parietal injury on covert orienting of attention. *Journal of Neuroscience, 4*, 1863–1874.
94. Bartolomeo, P., Chokron, S., & Siéroff, E. (1999). Facilitation instead of inhibition for repeated right-sided events in left neglect. *NeuroReport, 10*(16), 3353–3357.
95. Robertson, I. H. (1993). The relationship between lateralised and non-lateralised attentional deficits in unilateral neglect. In I. H. Robertson & J. C. Marshall (Eds.), *Unilateral neglect: Clinical and experimental studies* (pp. 257–278). Hove: Lawrence Erlbaum Assoc.
96. Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology. Human Perception and Performance, 18*(3), 849.
97. Howes, D., & Boller, F. (1975). Simple reaction time: Evidence for focal impairment from lesions of the right hemisphere. *Brain: A Journal of Neurology, 98*(2), 317–332.
98. Posner, M. I., Inhoff, A. W., Friedrich, F. J., & Cohen, A. (1987). Isolating attentional systems: A cognitive-anatomical analysis. *Psychobiology, 15*(2), 107–121.
99. Heilman, K. M., Watson, R. T., & Valenstein, E. (1993). Neglect and related disorders. In K. M. Heilman & E. Valenstein (Eds.), *Clinical neuropsychology* (3rd ed., pp. 279–336). New York: Oxford University Press.
100. Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13*, 25–42.
101. Corbetta, M., & Shulman, G. L. (2011). Spatial neglect and attention networks. *Annual Review of Neuroscience, 34*, 569–599.
102. Bartolomeo, P., Thiebaut de Schotten, M., & Doricchi, F. (2007). Left unilateral neglect as a disconnection syndrome. *Cerebral Cortex, 17*(11), 2479–2490.
103. Thiebaut de Schotten, M., Urbanski, M., Duffau, H., Volle, E., Levy, R., Dubois, B., et al. (2005). Direct evidence for a parietal-frontal pathway subserving spatial awareness in humans. *Science, 309*(5744), 2226–2228.

104. Watson, R. T., Andriola, M., & Heilman, K. M. (1977). The electroencephalogram in neglect. *Journal of the Neurological Sciences, 34*(3), 343–348.
105. Watson, R. T., Miller, B. D., & Heilman, K. M. (1977). Evoked potential in neglect. *Archives of Neurology, 34*(4), 224–227.
106. Di Russo, F., Aprile, T., Spitoni, G., & Spinelli, D. (2008). Impaired visual processing of contralesional stimuli in neglect patients: A visual-evoked potential study. *Brain, 131*(Pt 3), 842–854.
107. Mort, D. J., Malhotra, P., Mannan, S. K., Rorden, C., Pambakian, A., Kennard, C., et al. (2003). The anatomy of visual neglect. *Brain, 126*(Pt 9), 1986–1997.
108. Doricchi, F., & Tomaiuolo, F. (2003). The anatomy of neglect without hemianopia: A key role for parietal-frontal disconnection? *NeuroReport, 14*(17), 2239–2243.
109. Verdon, V., Schwartz, S., Lovblad, K. O., Hauert, C. A., & Vuilleumier, P. (2010). Neuroanatomy of hemispatial neglect and its functional components: A study using voxel-based lesion-symptom mapping. *Brain, 133*(Pt 3), 880–894.
110. Corbetta, M., Kincade, M. J., Lewis, C., Snyder, A. Z., & Sapir, A. (2005). Neural basis and recovery of spatial attention deficits in spatial neglect. *Nature Neuroscience, 8*(11), 1603–1610.
111. Lunven, M., Thiebaut de Schotten, M., Bourlon, C., Duret, C., Migliaccio, R., Rode, G., et al. (2015). White matter lesional predictors of chronic visual neglect: a longitudinal study. *Brain, 138*, 746–760.
112. He, B. J., Snyder, A. Z., Vincent, J. L., Epstein, A., Shulman, G. L., & Corbetta, M. (2007). Breakdown of functional connectivity in frontoparietal networks underlies behavioral deficits in spatial neglect. *Neuron, 53*(6), 905–918.
113. Koch, G., Oliveri, M., Cheeran, B., Ruge, D., Lo Gerfo, E., Salerno, S., et al. (2008). Hyperexcitability of parietal-motor functional connections in the intact left-hemisphere of patients with neglect. *Brain, 131*(Pt 12), 3147–3155.
114. Pantano, P., Di Piero, V., Fieschi, C., Judica, A., Guariglia, C., & Pizzamiglio, L. (1992). Pattern of CBF in the rehabilitation of visual spatial neglect. *International Journal of Neurosciences, 66*, 153–161.
115. Kinsbourne, M. (1977). Hemi-neglect and hemisphere rivalry. In E. A. Weinstein & R. P. Friedland (Eds.), *Hemi-inattention and hemisphere specialization* (Vol. 18, pp. 41–49). New York: Raven Press.
116. Chica, A. B., Bourgeois, A., & Bartolomeo, P. (2014). On the role of the ventral attention system in spatial orienting. *Frontiers in Human Neuroscience, 8*, 235.
117. Husain, M., & Nachev, P. (2007). Space and the parietal cortex. *Trends in Cognitive Sciences, 11*(1), 30–66.

Part II
Modeling

Chapter 5
Attention and Signal Detection: A Practical Guide

Vincent P. Ferrera

> *A faint tap per se is not an interesting sound; it may well escape being discriminated from the general rumor of the world. But when it is a signal, as that of a lover on the window-pane, it will hardly go unperceived.*
>
> – William James [1]

5.1 Detection of Weak Signals

The ability to detect weak signals in the environment can have a profound impact on an organism's ability to survive and reproduce. This aspect of perception is therefore likely to have been optimized by natural selection. Part of this optimization may involve strategies to maximize performance by allocating scarce neural resources. The ability to allocate limited resources by selecting and prioritizing sensory information is often what is meant when people talk about selective attention [2]. The notion of a limited capacity filter has been invoked to explain why orienting attention to a particular location in space or a particular stimulus feature enhances detection and shortens response times. This view has given rise to imaginative metaphors such as the "spotlight" [3, 4] or "zoom lens" [5] of attention. An alternative, albeit less poetic, view considers attention from the standpoint of a decision-maker trying to make sense of noisy signals arising from multiple detectors. In this view, what is commonly referred to as "attention" may be a manifestation of the effect of uncertainty on the behavior of an ideal observer [6, 7]. While precise definitions are elusive, it is reasonable to say that attention includes a collection of computational strategies that enhance the detection and discrimination of weak signals and/or refine the behavioral response to such signals. These strategies might include increasing the signal-to-noise ratio of individual neurons, optimizing decision parameters, and identifying subsets of detectors (e.g., neurons) that are more reliable.

V.P. Ferrera (✉)
Department of Neuroscience, Columbia University, 1051 Riverside Drive, Unit 87, New York, NY 10032, USA
e-mail: vpf3@cumc.columbia.edu; vincent.ferrera@gmail.com

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_5

Stimulus detectability across the visual scene is one way to quantify perceptual salience. Salience maps are important for both human and machine vision systems as they indicate areas of heightened interest, attention, and action. Signal detection theory provides various strategies for computing salience maps. Salience maps computed using principles of signal detection can incorporate the effects of prior information (i.e., environmental cues or knowledge about target prevalence), observer bias, and/or economic value.

The conversion of sensory signals into percepts, decisions, and actions occurs over multiple stages of neural processing. At which level does attention act? Does attention affect the quality of incoming sensory information? Does it affect decision-making, response selection, or even later processes? This issue is part of the long-standing debate over "early" vs. "late" selection of signals [2, 8].

The idea behind early selection is that multiple stimuli compete for attention at an early stage of sensory processing. Attention biases this competition by enhancing the representation of behaviorally relevant stimuli [9]. Thus, when attention is directed toward a particular location or object, it improves the quality of sensory data acquired at the focus of attention. Improved quality means that the neural representation has higher fidelity, stronger signal, less noise. Better signal to noise should enhance the detectability of weak signals.

The "late" selection hypothesis holds that attention acts mainly on higher-order processes, leaving sensory representations largely intact. For example, attention can act at the level of response selection by adjusting decision criteria. An observer may have prior information that creates an expectation that the stimulus will occur at a particular time or place. This could be due to statistical regularities in the environment or to the presence of reliable cues, either natural or artificial. If a stimulus is expected to occur in a given place or at a given time, the observer may require less sensory evidence to report that it was present. Thus, they may lower their internal decision threshold or adopt a bias in favor of making a positive response. The expectation of a stimulus does not necessarily mean that the quality of the sensory evidence provided by that stimulus is better, but rather that the prior likelihood of the stimulus biases the observer to report that it is present.

To understand how attention might improve an observer's performance, it is useful to introduce the framework of signal detection theory (SDT [10, 11]). SDT provides a simple set of computations to select responses based on factors such as signal strength, stimulus probability, and the consequences of different responses. The underlying model for SDT is that signals in the environment cause changes in the internal state of an observer. Changes in internal state then guide categorical responses, such as "yes/no" or "seen/not seen." The SDT model affords a great deal of flexibility in mapping stimuli onto responses. Flexibility derives from the probabilistic relationship between external signals and internal states, and the criterion-dependent relationship between internal states and responses.

SDT clarifies the distinction between *stimulus detectability* and *response bias*. Detectability is a function of the sensory signal alone. It is the certainty with which an external event in the environment can be inferred from the internal state of the observer. Detectability depends only on the difference in the observer's internal state

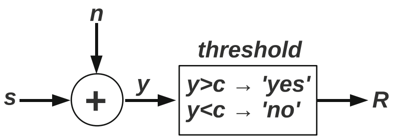


Fig. 5.1 Signal detection model. Noise, *n*, is added to external signal, *s*, to produce an internal state, *y*, which is compared to criterion, *c*, to determine response, *R*

Table 5.1 Confusion matrix for signal detection model

	Stimulus present	Stimulus absent
Respond "Yes"	*Hit*	*False alarm*
Respond "No"	*Miss*	*Correct rejection*

when the external stimulus is present as compared to when the stimulus is absent. It does not depend on the relationship between the observer's internal state and their response, which may be biased toward one alternative or the other, independently of the signal.

Formally, the premise of signal detection theory is that the internal state of an observer (y) is perturbed by an external signal (s) that is affected by noise (n). In the simplest case of additive noise, $y = s + n$ (Fig. 5.1). The internal state, y, is then compared to a threshold decision criterion, c, to generate a binary response of "yes/no" or "seen/not seen." The detectability of the stimulus is entirely determined by the characteristics of the signal and noise. However, the response can be biased depending on the level of the decision criterion. A change in criterion might cause the observer to report that the stimulus is present more or less often even though there is no real change in stimulus detectability.

In the most basic case, the stimulus takes one of two values (present or absent), and the response also has two possible values (yes, no). There are thus four possible outcomes (Table 5.1). Hits and correct rejections are both correct responses. Misses and false alarms are incorrect.

When the stimulus is present, it gives rise to an internal state drawn from a probability density function called the signal distribution (S). This distribution consists of signal + noise, as noise is always present. When the stimulus is absent, the internal state is drawn from the noise distribution (N). Internal state might correspond to the instantaneous firing rate of a neuron, or the number of action potentials fired in a specified time interval (spike count). Figure 5.2 (left panel) shows hypothetical examples of signal (blue bars) and noise (gray bars) distributions. The dashed vertical lines represent two different criterion values. For a given criterion value, the *hit rate* is the proportion of the signal distribution that is greater than the criterion. Similarly, the *false alarm rate* is the proportion of the noise distribution that is greater than the criterion. The miss rate is 1.0, hit rate, and the correct rejection rate is 1.0, false alarm rate. Observer performance is completely characterized by the rates of hits and false alarms.

To obtain a criterion-independent estimate of detectability, one can vary the criterion level through the entire range of states represented in the signal and noise

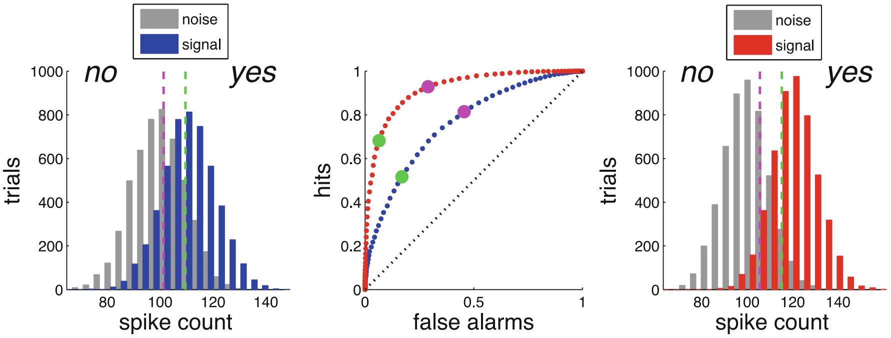


Fig. 5.2 (*Left*) Spike count histograms for stimulus present (*blue*) or absent (*gray*) trials. *Dashed vertical lines* represent different criterion values. (*Middle*) ROC curves. "Hits" are the number of hits divided by hits + misses. "False alarms" is the number of false alarms divided by false alarms + correct rejections. *Blue dots* are the ROC derived from the distributions in the *left panel. Red dots* are the ROC derived from the distributions in the *right panel. Green* and *magenta dots* are the points of the ROC curves corresponding to the criterion levels shown in the *left* and *right panels*. (*Right*) Histograms for a case with stronger signal

distributions. For each criterion level, the hit rate can be plotted against the false alarm rate (Fig. 5.2, middle). The resulting curve is called the "receiver operating characteristic" or ROC curve. The area under the ROC curve (AROC) is a measure of stimulus detectability across all criteria. Imagine drawing two random samples: one from the signal and one from the noise distribution. The area under the ROC curve is the probability that the sample drawn from the signal distribution is larger. In psychophysics, the ROC area is equal to the percentage of correct responses for an ideal observer in a two-interval forced choice experiment.

Based on how the ROC curve is constructed, it follows that changing the decision criterion, e.g., from the magenta to the green line in Fig. 5.2, only moves one along the ROC curve. The magenta and green dots superimposed on the blue curve in Fig. 5.2 (middle) are the hit and false alarm rates corresponding to the criterion levels in the left panel. Changing the criterion does not change the ROC curve itself. To do this, there must be a change in the amount of overlap between the signal and noise distributions. The right panel in Fig. 5.2 shows the distributions for a stronger signal with the same noise as in the right panel. The corresponding ROC curve is shown in red in the middle panel. The increased area under the red curve means that the signal can be more reliably detected.

It should be clear that stimulus detectability depends on the overlap of the signal and noise distributions, which in turn depends on two factors: the separation between the means of the signal and noise distributions, and the variances of those distributions. Attention can therefore improve detectability by increasing the former and/or reducing the latter. In SDT, these are the only two variables that affect the internal representation of signal quality. However, attention may also act by optimizing the decision criteria that determine the observer's response.

5.2 Effect of Stimulus Probability

The likelihood that a stimulus will occur during a given observation period is referred to as its prior probability. In an unbiased experiment, stimulus-present and stimulus-absent trials should occur in equal proportion, so that the prior probabilities of signal and noise are both 0.5 and the percent correct that can be achieved by guessing is 50 %. In most real-world environments, the signal and noise probabilities are not necessarily equal. Rare or novel signals may attract attention by an oddball effect. Frequent signals may result in sensory adaptation, thus weakening their internal representation. Response habituation can also play a role; if a stimulus is quite rare, then observers may fall into a habit of responding "no." This habitual response may cause observers to miss a rare stimulus if they are not vigilant. Some studies report that prevalence effects can result in miss rates of up to 50 % [12]. Other studies with medical images (chest x-rays) have reported that prevalence effects are negligible [13].

Observers can take advantage of variations in stimulus probability by adapting their decision criteria. These adjustments can be made without any explicit knowledge about stimulus probability itself. SDT naturally handles cases where stimulus probability is different from 0.5. In the unbiased case, the areas of the signal and noise probability distributions are both equal to 0.5. If there is a preponderance of signal trials, then the area of the signal distribution will be between 0.5 and 1.0, while the noise distribution will have area = 1.0 − signal area.

To understand how prior probability affects detectability, it is important to realize that the hit and false alarm rates are conditional probabilities. Specifically, hit rate is the probability that the internal state, y, is greater than the criterion, c, given that the stimulus is present: $p(y > c \mid S)$. Likewise, false alarm rate is conditioned on the absence of the stimulus, thus $p(y > c \mid N)$. Because hit rate is conditioned on the presence of the stimulus, changing the likelihood that the stimulus is present does not change the hit rate. For a given criterion level, the proportion of the signal distribution that is greater than the criterion is invariant to scaling of the distribution. The same goes for the false alarm rate. Hence, stimulus probability has no effect on the ROC curve and thus no effect on stimulus detectability.

Figure 5.3 illustrates this by showing cases where the signal has a low probability (left panel) or high probability (right panel). The ROC curves are the same for both conditions (middle). Another way to think about this is that the ability to detect a stimulus depends only on the strength and fidelity of its representation in the nervous system at the time the stimulus is present. It does not depend on the past history of the stimulus. This surprising feature of ROC curves is advantageous in areas like medical diagnosis because the probabilities of S and N are generally unknown (and difficult to measure). Thus, ROC curves are understood to provide a reliable metric for diagnostic efficacy independent of the relative prevalence of S.

It may seem counterintuitive that detectability is not affected by stimulus probability. Certainly, prior knowledge about the signal must confer some performance advantage, and it does. But the advantage derives from the fact that observers are

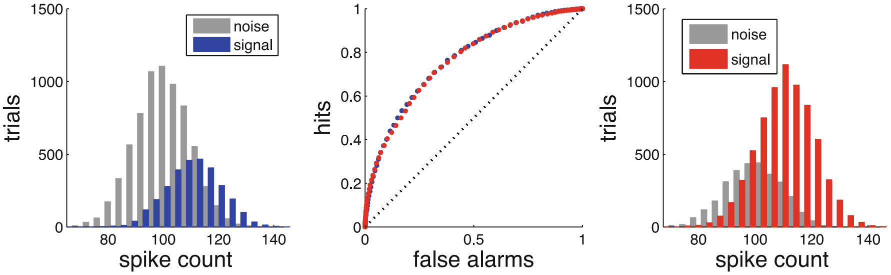


Fig. 5.3 (*Left*) Histogram of spike counts for stimulus-present (*signal*) and stimulus-absent (*noise*) trials with low stimulus probability. (*Middle*) ROC for (*blue dots*) and high (*red dots*) stimulus probability. (*Right*) Histograms for high stimulus probability

able to improve performance by altering their decision strategy. If the stimulus is more likely to appear than not, then there is an advantage to using a more liberal decision criterion (smaller value of c) for responding "yes." In this case, simply closing one's eyes and guessing that the stimulus is present or saying that it is always present will yield performance greater than 50 % correct. Signal detection theory can be used to determine the value of the decision criterion that will optimize performance (percent correct) for a given signal strength and probability. Analytically, the optimum criterion is the value of y that satisfies the following equation:

$$p\ (S) * f_s(y) = p\ (N) * f_n(y)$$

where $p(S)$ and $p(N)$ are the signal and noise prior probabilities (signal present or absent, respectively) and $f_s(y)$ and $f_n(y)$ are the unweighted signal and noise probability densities (see [14] for derivation).

Figure 5.4 illustrates simulations where the signal has low (top left) or high (top right) probability. Performance, in terms of percent correct detection, is plotted as a function of criterion level in the bottom left for the case of high (red) or low (blue) signal probability. The dashed vertical line indicates the criterion level that optimizes performance. The optimum criterion can be computed for any signal probability (bottom right). These simulations show that, while knowledge of stimulus probability does not affect detection, performance may nevertheless be improved by selecting the optimal decision criterion.

One of the chief complaints about SDT is that it seems to assume that the probability distributions are known with arbitrary precision. In practice, observers may not know the shapes of these distributions or the prior probabilities of signal and noise. However, there are simple iterative algorithms for adapting the decision criterion that produce near-optimal performance and are based only on quantities available to the observer, for example, their behavioral response and the feedback they receive (assuming feedback is given). First, note that Table 5.1 can be rearranged as follows (Table 5.2):

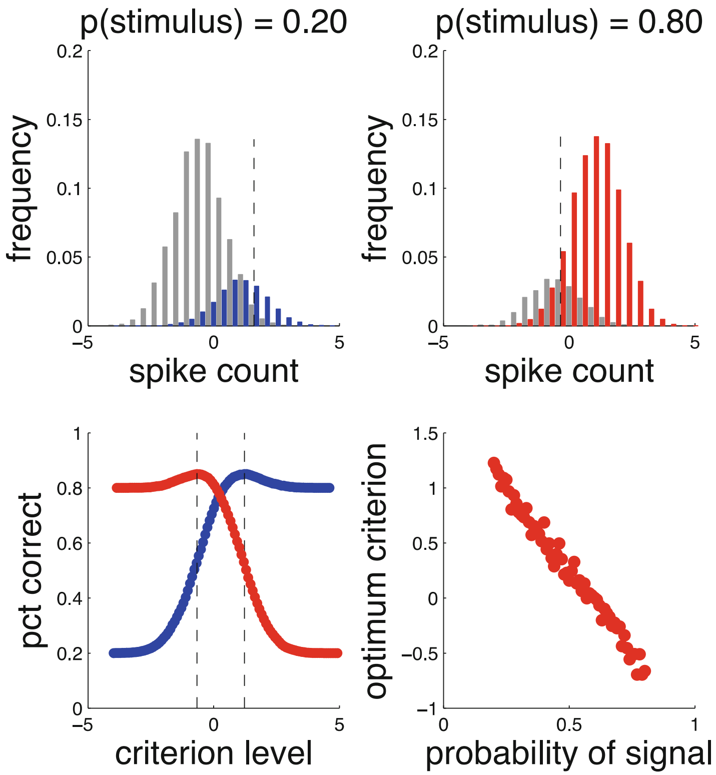


Fig. 5.4 (*Top row*) Spike count histogram for low (*blue*) and high (*red*) stimulus probabilities. (*Bottom row*) Performance as a function of criterion level and optimum criteria for different signal probabilities

Table 5.2 Confusion matrix for signal detection model reordered by response outcome

	Correct	Incorrect
Respond "Yes"	*Hit*	*False alarm*
Respond "No"	*Correct rejection*	*Miss*

Hence, an observer can deduce whether the stimulus was present (hit, miss) or absent (false alarm, correct rejection) based on the conjunction of their response (yes/no) and the outcome (correct/incorrect). The observer can use this information to estimate the prior likelihood of the stimulus. Knowing only their response and the outcome, the observer can optimize their decision criterion based on feedback.

An iterative algorithm for optimizing the decision criterion is the following: (1) after each "yes" response, the criterion level is incremented in proportion to the rate of signal-absent trials and, (2) after each "no," the criterion is decremented in proportion to the rate of signal-present trials. This can be quantified by the following updating rules:

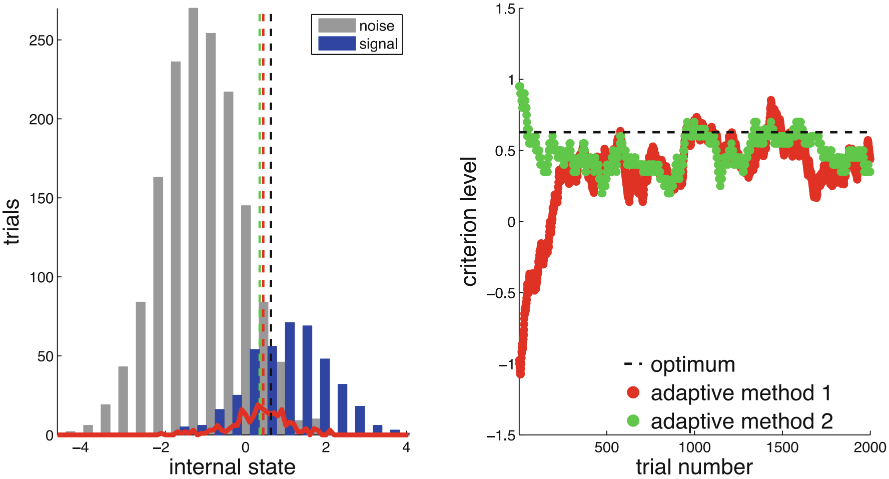


Fig. 5.5 Optimization of decision criterion. (*Left*) Signal and noise distributions with signal probability = 0.2. *Red* and *green vertical dashed lines* are final criterion values for two adaptive procedures. *Black dashed line* is optimal criterion. *Red line* is the distribution of error trials. (*Right*) Adaptive criteria as a function of trial number

1. If response = "yes," $c\,(n+1) = c(n) + k * p(N)$
2. If response = "no," $c\,(n+1) = c(n) - k * p(S)$

The criterion level on trial n is denoted by $c(n)$, k is the learning rate, $p(N)$ is the probability that the stimulus was absent, and $p(S)$ is the probability of stimulus present. These probabilities are not known in advance, but are continuously updated based on feedback. This method is stable and converges to the criterion value corresponding to an internal state that is equally likely for stimulus-present and stimulus-absent trials, i.e., $p(y \mid S) = p(y \mid N)$. An example of the algorithm's performance is shown in Fig. 5.5 (method 1).

If feedback ("correct" or "incorrect") is given after each trial, then the observer can deduce whether the outcome is a hit, miss, FA or CR, and thus whether or not the stimulus was present. Therefore, the information required to implement this procedure is available to the observer and does not require prior knowledge of stimulus probability.

There is another algorithm for adapting the decision criterion that uses only feedback on error trials. Specifically, after each false alarm, the criterion is incremented a small amount. After each miss, the criterion is decremented by the same quantity. This process converges on the criterion for which the miss and false alarm rates are equal, which is close to the optimum criterion for minimizing the error rate (Fig. 5.5, method 2). It is easy to show that this algorithm is stable: if the criterion value is too high, then misses outnumber false alarms and the criterion is decremented until the miss rate equals the false alarm rate; if the criterion is too low, false alarms outnumber misses resulting in a net increment. Furthermore, as long as the signal and noise distributions have positive area, there is always a criterion value for which misses = false alarms. This can be demonstrated by considering that, as the criterion goes from –infinity to infinity, the false alarms start at a finite, positive value and then decrease to zero. At the same time, misses increase from zero to a finite positive value. Therefore, the miss and false alarm curves must cross. This procedure is therefore guaranteed to converge and has the additional advantage that it does not rely on an estimate of stimulus probability. Procedures based only on correct responses can also be used, but tend to converge at criterion values that are far from the optimum.

5.3 Effect of Costs and Benefits for Various Outcomes

SDT has four classes of outcome: hits, misses, false alarms, and correct rejections. In real-life situations, each outcome has an associated cost or benefit. Misses and false alarms are both incorrect outcomes, but are not equally costly. If a person has a medical exam, the cost of a false negative (miss) might be that they will not receive treatment and their condition may worsen. The cost of a false positive is that they could receive a treatment that is unnecessary. One of these outcomes may be catastrophic (e.g., an infection that becomes life threatening), while the other is relatively benign (taking a superfluous course of antibiotics). Likewise, the benefit of a hit may be greater or less than a correct rejection.

If one can assign a numerical value to each outcome, then there is a formula for the expected value (EV) of each trial [14]:

$$EV = V_h * p(h) + V_m * p(m) + V_{\mathrm{fa}} * p\,(\mathrm{fa}) + V_{\mathrm{cr}} * p\,(\mathrm{cr})$$

where V is value, p is probability, and the subscripts denote the various outcomes. The values of misses and false alarms are typically negative as these outcomes represent costs. This formula also incorporates effects of prior stimulus probability as this affects the outcome probabilities.

Figure 5.6 shows an example of the effect of value-weighted outcomes on the optimal decision criterion. In the left panel are weighted outcomes for hits (green), correct rejections (blue), misses (black), and false alarms (red). The stimulus probability is 0.5. The dashed lines represent the balanced case where correct responses have a value of 1 and errors have a value of −1. The expected payoff and optimum criterion is shown in black in the right panel.

The heavy lines in the left panel represent a situation where hits have a value of 1.5, misses −1.5, false alarms –0.7, and correct rejections 0.7. The optimum criterion in this case (right panel, magenta) shifts to a smaller (more liberal) value. This results in more hits and fewer misses at the cost of more false alarms and fewer correct rejections, reflecting the relative value of these outcomes. The optimum criterion is the value of y that satisfies the following equation:

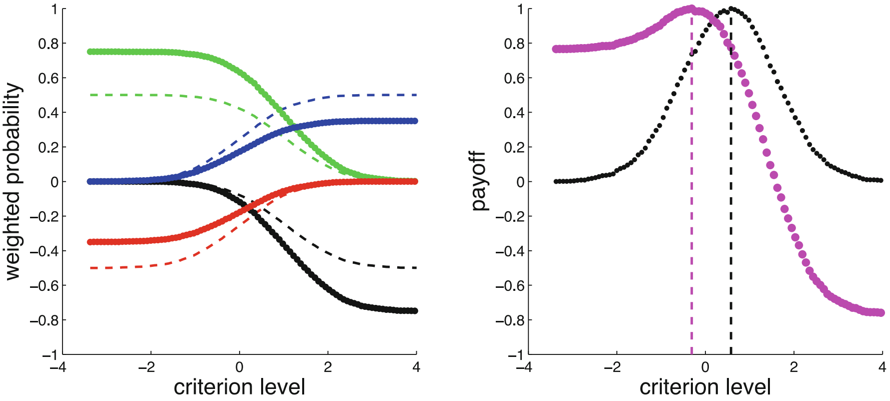


Fig. 5.6 Effect of payoffs on optimal decision criteria. (*Left*) All outcomes have same absolute value (*dashed lined*) or different values (*solid lines*). (*Right*) Total payoff vs. criterion for equal outcome value (*black*) and unequal value (*magenta*)

$$(V_h – V_m) * p(S) * f_s(y) = \left(V_{cr} – V_{fa}\right) * p(N) * f_n(y)$$

where $V_x(x = h, m, fa, cr)$ is the value of a hit, miss, false alarm, or correct rejection. Again, V_{m} and V_{fa} usually have negative values so that when all outcomes have equal weight (0.5), the terms $(V_h - V_m)$ and $(V_{cr} - V_{fa})$ can be replaced by 1.0.

The values assigned to different outcome classes may reflect economic value, such as subjective utility. They may also reflect emotional value (intensity) and valence (positive or negative). Stimuli that are associated with high outcome values may automatically attract attention, regardless of whether the outcome is positive or negative [15].

5.4 Effects of Pooling Over Multiple Detectors

The above considerations apply to the case of a single detector (e.g., a neuron) but can be readily extended to multiple detectors. In the simplest case, all detectors have the same inputs, sensitivity, and noise characteristics. All observations therefore have equal weight. One observation from each of two detectors is the same as two observations from one detector. However, even this simple case presents an opportunity to test different rules for pooling across detectors. Furthermore, we can examine how detectability improves with the number of detectors and observe the effects of correlations among detector responses. Studies in monkeys have found that the spike count correlation between nearby visual cortical neurons is roughly 0.1–0.2 [16, 17] and that these correlations are reduced by attention [18, 19].

To analyze the activity in multiple detectors, it is useful to first build an activity matrix. Each row in this matrix represents an individual detector (neuron), and each

Table 5.3 Activity matrix with each cell [$y(i,j)$] representing an average firing rate for neuron i on trial j

	Trial 1	Trial 2	Trial 3	...
Neuron 1	$y(1,1)$	$y(1,2)$	$y(1,3)$	...
Neuron 2	$y(2,1)$	$y(2,2)$	$y(2,3)$	...
Neuron 3	$y(3,1)$	$y(3,2)$	$y(3,3)$	...
...	...	...	...	...

column is a single observation period (trial). The values in each cell thus represent the activity of a single detector on a single trial (Table 5.3):

From this activity matrix, we can construct joint ROC functions in several ways. The simplest method is to treat each cell, $y(i,j)$, as an independent observation. The result is that the presence of multiple detectors increases the number of observations at a given time, but otherwise confers no improvement in detectability. In other words, if all neurons are equal, then adding neurons does not change the joint ROC curve. It is the same as simply gathering additional observations from a single neuron.

To gain any advantage from multiple neurons, the responses must be aggregated in some manner. One method is simply to average observations prior to constructing the ROC function. One can either average over trials for each neuron or over neurons within each trial. Both methods reduce the number of effective observations underlying the ROC, but have the advantage that the variability of those observations may be substantially reduced. Figure 5.7 shows an example with 12 neurons. In the left panel are the responses of a representative pair of neurons collected over many trials (blue dots = stimulus absent, red dots = stimulus present). The trial-to-trial responses are weakly correlated (Pearson's correlation coefficient = 0.2). The right panel shows the ROC function for each of the 12 neurons individually (blue lines) and for the ensemble (red dots) when activity in each trial is averaged across neurons before computing the ROC. The average ROC area for each neuron alone is 0.76, while the joint ROC has area = 0.86. In this example, all neurons have the same sensitivity, and their contributions are weighted equally. One could alternatively construct a weighted average such that the contribution of each neuron would be weighted by its reliability, for example, by dividing by the standard deviation or variance of the spike count distribution. This would result in a more Bayesian style of combining responses.

Figure 5.8 (left panel) shows how detectability increases with the number of neurons. When the response of each neuron is independent of the other neurons, detectability starts to saturate when there are about 32 neurons in the pool (black curve). The exact number of neurons at which saturation occurs is not fixed, but typically depends on the characteristics of the signal and noise, as well as the method of pooling [20].

The advantage of pooling responses across neurons is reduced when their activity is correlated. For example, if the degree of correlation for every pair of cells in the population is $r = 0.2$ (as depicted in Fig. 5.7, left panel), then the area under the joint ROC is represented by the red curve in Fig. 5.8 (left panel). Here, the optimum detectability reaches only 85 % of that obtained when there is no

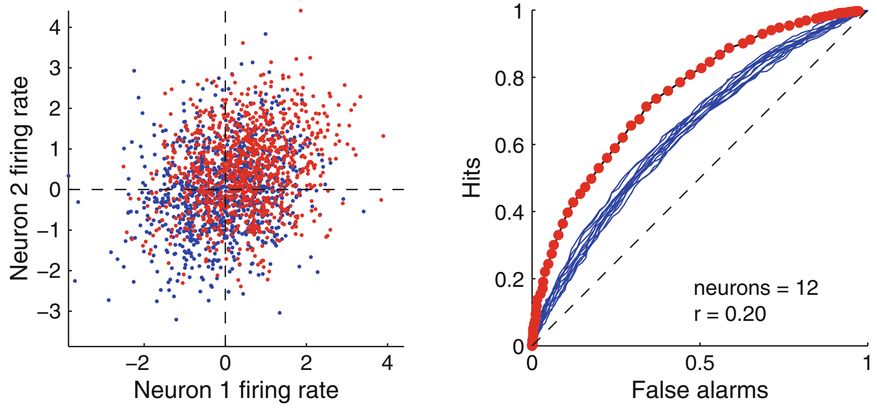


Fig. 5.7 (*Left*) Firing rates of two simulated neurons whose trial-to-trial responses are weakly correlated. *Blue dots* are for stimulus-absent, and *red dots* are for stimulus-present trials. (*Right*) ROC curves for 12 weakly correlated neurons. *Blue lines* are ROCs computed for each neuron individually. *Red dots* are the ROC curve when responses are pooled across all neurons

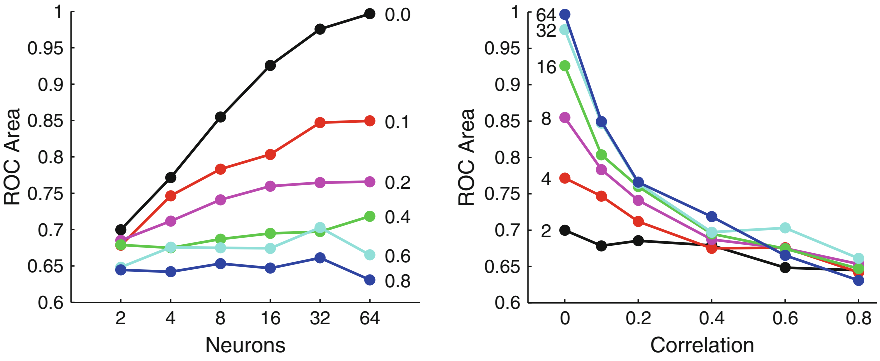


Fig. 5.8 Impact of neuronal pool size (*left panel*) and between-neuron response correlations (*right panel*) on detectability

correlation between neurons. When the correlation is 0.6 (green curve), there is almost no advantage of pooling. The same data are replotted as a function of correlation strength (Fig. 5.8, right panel). The curves for different pool sizes all come together between $r = 0.4$ and $r = 0.6$, indicating that the strength of correlation that eliminates the advantage of pooling in this case is about 0.5.

If positive correlations among neurons reduce the benefits of pooling, then it might be expected that negative correlations would have the opposite effect. Figure 5.9 shows an example where the pool size is 2 neurons and the trial-to-trial correlation in firing rate is −0.9. The negative correlation reduces the overlap of signal and noise distributions. The resulting joint ROC has an area of 0.95. This can be compared to the case where the correlation is 0 and the joint ROC area is

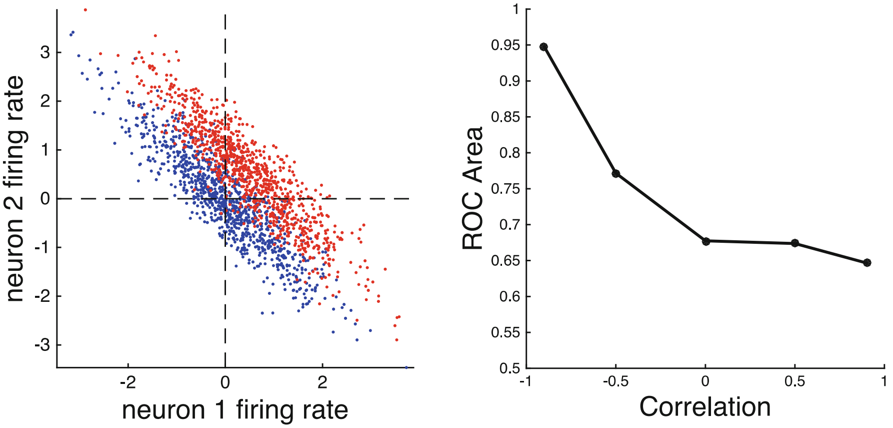


Fig. 5.9 Impact of negative correlations on detectability for a pool of 2 neurons. (*Right*) signal (*red*) and noise (*blue*) distributions. (*Left*) ROC area as a function of correlation coefficient

0.69. It would therefore seem that negative correlations are capable of producing huge improvements in detectability. However, there are some important caveats. Foremost among these is that for a pool size greater than 2, it is impossible for all the pairwise correlations to be negative. If neurons 1 and 2 are negatively correlated, and 2 and 3 are also negatively correlated, then 1 and 3 must be positively correlated. In other words, the correlation matrix must be positive semi-definite. As long as this constraint is satisfied, a pool of neurons with some negative correlations might provide significantly better detectability than a pool of independent, uncorrelated neurons.

For weaker stimuli, the advantages of increasing the size of the neuronal pool are greater; however, the effects of positively correlated activity are more devastating. For example, when there is no correlation between neurons, the improvement in detectability of a weak stimulus may not start to saturate until the pool size reaches over 1000 neurons. However, even a weak ($r = 0.1$) correlation can eliminate the advantage of pooling altogether.

Another method for pooling across neurons is to compute the probability of hits and false alarms for each neuron individually and then sum the probabilities across neurons. The probabilistically summed hit and false alarm rates can then be used to compute the joint ROC. The pooled probability of a hit and false alarm is as follows:

$$p\,(\text{hit}) = 1.0\text{–}\,[p\,(\text{miss}_1)\,\&p\,(\text{miss}_2)\,\&\ \ \ldots\ \ \&p\,(\text{miss}_\text{n})]$$

$$p\,(\text{fa}) = 1.0\text{–}\,[p\,(\text{cr}_1)\,\&p\,(\text{cr}_2)\,\&\ \ \ldots\ \ \&p\,(\text{cr}_\text{n})]$$

where the subscripts (1 ... n) index the individual neurons and n is the total number of neurons. The probability of a miss or correct rejection is calculated across all trials for each neuron. These calculations are done for each criterion level to construct the joint ROC curve. Note that because the firing rates may be correlated across neurons, the probabilities of misses or correct rejections are not independent. Thus, the right sides of these equations must be calculated using the formula for the joint probability of dependent events:

$$p\,(a\&b\&c\&d\&\ \ \ldots) = p(a) * p\,(b|a) * p\,(c|a\&b) * p\,(d|a\&b\&c)\ \ \ldots$$

These results suggest that attention can improve signal detection by reducing correlated activity among neurons or by selecting the responses of neurons whose activity is maximally uncorrelated (or negatively correlated). However, we have so far only dealt with the case of detectors with uniform sensitivity. In general, neurons may have different sensitivities (responsiveness) and baseline firing rates. Furthermore, the degree of correlated activity between pairs of neurons is likely to vary across the population, rather than being constant for all pairs as in the simulations presented here.

The above considerations apply to the case where all neurons in the pool respond to the signal. In this case, increasing the number of neurons in the pool increases stimulus detectability, though the marginal improvement may at times be small. However, it is unusual for all detectors to be sensitive to the stimulus. In general, a given stimulus will be represented by a small fraction of the relevant population of neurons, i.e., those whose front-end filtering properties (e.g., spatial and feature selectivity) are appropriately matched to the stimulus. We can call this the "signal pool." The rest of the neurons in the brain (the "noise pool") contribute nothing to detection of the stimulus. In fact, their activity is deleterious to performance as it represents background noise. One of the great problems of attention is how to select responses from the signal pool while ignoring or suppressing activity in the noise pool. The problem is compounded by the fact that individual neurons can switch from one pool to the other at any given time, depending on the stimuli present in the environment and the organism's behavioral goals.

5.5 Uncertainty and Cueing Effects

One of the most common behavioral paradigms for studying attention is to provide observers with prior information (a cue) about a target whose properties are uncertain [4, 21, 22]. For example, observers might be asked to detect a low contrast target presented at a location that is randomized from trial-to-trial, thus introducing spatial uncertainty. At some time before the target appears, a high contrast cue is presented at a location that is more or less predictive of the target location. Such cues often improve performance accuracy, but whether these improvements are due to enhanced stimulus detectability or reduction in decision uncertainty has been subject to much debate [23–27].

Figure 5.10 illustrates trial conditions from a task where the stimulus (a vertically oriented grating patch) can appear at one of two locations (essentially the same task as used by [21]). The subject's task is to report the presence of the target. On all trials, the cue is equally likely to occur at either location. On half the trials, there is no stimulus ("catch" trials). On the other trials, the stimulus is preceded by a cue that predicts where the stimulus is likely to appear (i.e., it may or may not be a "valid" cue). The predictiveness of the cue is referred to as its "validity." If the cue is 80 % valid, then the stimulus appears at the cued location on 80 % of the trials and at the uncued location on 20 % of the trials. If the cue is valid on 80 % of the signal-present trials and invalid on 20 %, then the signal probability is 0.4 at the cued location and 0.1 at the uncued location. The cue not only attracts attention but allows the observer to use a more liberal criterion for responding that the stimulus is present. This should result in a higher percent correct on cued vs. uncued trials. It may also lead to shorter reaction times.

Consider the responses of two detectors, one at the cued location and another at the uncued location. What level of performance can be achieved by combining responses from the two detectors with equal weight? Each detector experiences a signal probability of 0.25, because the stimulus is present on half the trials and its location is randomized. Figure 5.11 (top left) illustrates the theoretical percent correct (hits + correct rejections divided by total trials) for detecting the stimulus as a function of criterion level for both detectors when the cue validity is 50 % (i.e., the

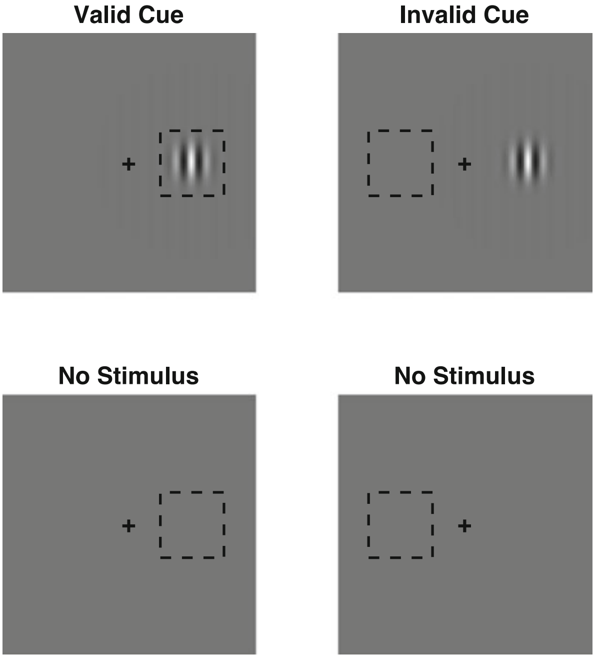


Fig. 5.10 Spatial cueing task with vertically oriented target

cue location is uncorrelated with stimulus location). Because the signal probability at each detector is 0.25, the optimal criterion (vertical dashed lines) is relatively conservative and is the same for both detectors.

The same calculations are shown for the case where cue validity is 80 % (Fig. 5.11, top right). Here, the optimal criterion is more liberal for the detector at the cue location (blue) because the signal probability (0.4) is higher at that location. The stimulus probability at the uncued location is only 0.1. This leads to the counterintuitive observation that the uncued detector can actually achieve better performance than the cued detector. This happens because an observer can use a very conservative criterion for the uncued detector. In fact, they can say "no" (i.e., reject the hypothesis that the stimulus is present at the detector location) on every trial and be correct 90 % of the time, regardless of the state of the detector.

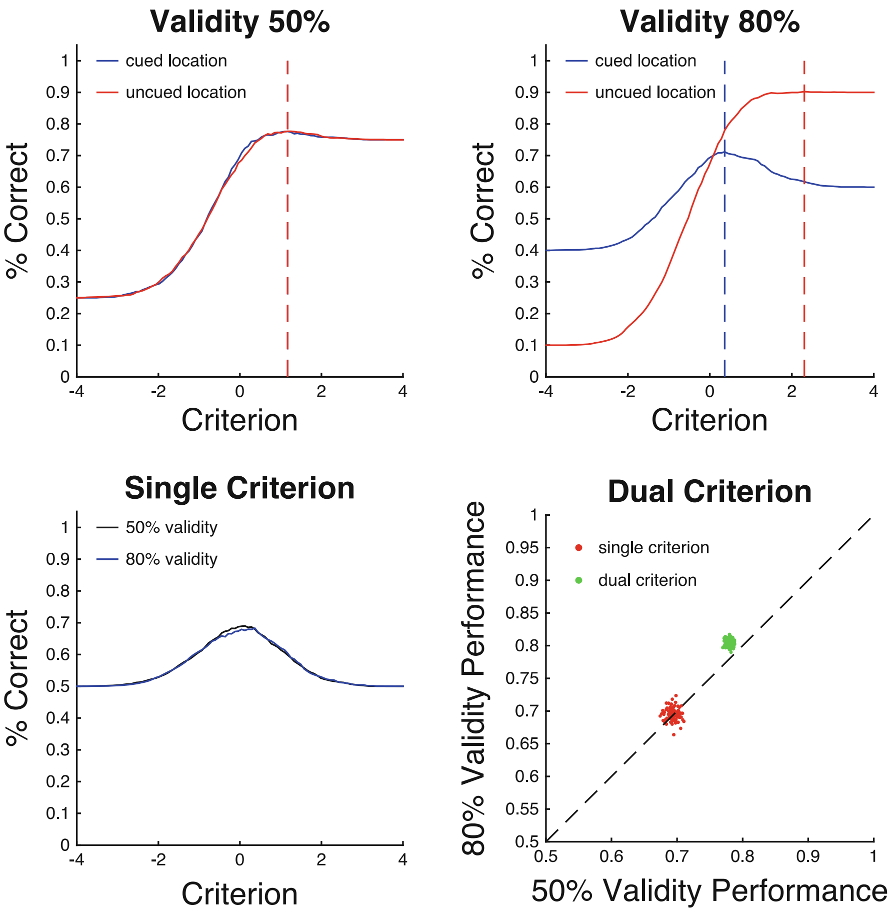


Fig. 5.11 Effects of cueing analyzed by SDT

It may appear that, in the case of 80 % cue validity, one should be able to achieve high performance simply by using the response of the better detector. Unfortunately, each detector only provides partial information (whether or not the stimulus is present at the detector location). Both detector responses must be combined to determine the overall response of the observer. Overall performance is calculated by adding up hits and correct rejections from all detectors and dividing by the total number of observations (number of trials x number of detectors). Surprisingly, if one is limited to using the same criterion for both detectors, then valid cues offer no advantage. The percent correct is the same for both 50 % and 80 % validity (Fig. 5.11, bottom left), and this is true regardless of criterion. In a sense, this is understandable as the cues provide information only about likely stimulus location, whereas the observer's job is to report stimulus presence.

One way that valid cues can yield an advantage is if the observer is allowed to choose the optimum criterion independently for each detector. Figure 5.11 (bottom right) shows performance for the case where the optimum criterion for each detector is used (green). This is compared to the case of a single criterion that optimizes performance for both detectors (red). The advantage of valid cues is small. This is partly due to the fact that the proportion of catch trials is large (50 %). Reducing the proportion of catch trials increases the performance advantage provided by valid cues. Whether or not subjects are capable of maintaining multiple decision criteria at the same time is an open question [28].

Other approaches to understanding cueing effects have been suggested. Cueing effects can be modeled using Bayesian statistics, which leads to similar conclusions about the advantages of valid cues [7]. In all of the above, the assumption is that valid cues affect the decision process but not signal quality. If valid cues enhance signal strength, then that advantage would add to the advantage one can achieve by adjusting decision criteria.

5.6 Signal Detection Over Time

Attention can increase the rate of information processing [29]. Hence, models are needed that account for both improvements in detectability and response time. However, the preceding discussion applies only to signals that are non-time-varying in the sense that they remain constant over the duration of a given observation period or "trial." The assumption is that, on each trial, the observer draws a single sample from the distribution of internal states of each detector and a decision is made based on those samples. These models can predict performance accuracy, but not the amount of time needed to respond at a given level of accuracy. Adding the dimension of time allows observers to draw multiple samples from each detector and to integrate the evidence provided by those samples, before reaching a decision.

In the 1940s, Wald [30], and others, developed the theory of sequential sampling as a way to calculate the incremental evidence provided by each sample and, thus, how many samples are needed for a given level of performance. If samples are drawn at a steady rate, this number corresponds to response time. Specifically, Wald developed the *sequential probability ratio test* (SPRT), which integrates the incremental information provided by each sample and also specifies a stopping rule, i.e., the amount of integrated evidence needed to achieve a given level of accuracy, defined in terms of percentage correct (hits and correct rejections) or incorrect (false alarms and misses). This test is derived from the Neyman-Pearson lemma which states that the likelihood ratio test maximizes the probability of detection for a given probability of false alarms [31].

The problem addressed by the SPRT is how to quantify the information in each sample so that it can be combined with other samples. The optimal way to do this is to start with the likelihood that a given sample, y, was drawn from the signal-present or signal-absent probability densities, i.e., $p(y|S)$ and $p(y|N)$. The next step is to compute the log of the likelihood ratio: $\underline{x} = \log[p(y|S)/p(y|N)]$. The quantity, x, represents the momentary evidence favoring hypothesis H1: signal present vs. H0: signal absent. The process is iterated by repeatedly drawing samples, calculating the log of the likelihood ratio, and adding that incremental evidence to the total evidence accumulated from previous samples:

$$x_{t+1} = x_t + \log\left[\frac{p\left(y_{t+1}|S\right)}{p\left(y_{t+1}|N\right)}\right].$$

The accumulation of evidence continues until x reaches a threshold value, or boundary. There are two boundaries: if x first reaches bound A, H1 is accepted (e.g., observer responds “yes”), and if x reaches bound B, H1 is rejected (response is “no”). The values of A and B are calculated to yield a predetermined level of performance accuracy. If alpha is the desired false alarm rate and beta the desired miss rate, then $A = \log[(1 - \text{beta})/\text{alpha}]$, and $B = \log[\text{beta}/(1 - \text{alpha})]$. The bounds can also be calculated in terms of the hit and correct rejection rates, as hit rate = 1 − beta and correction rejection rate = 1 − alpha.

The SPRT can be thought of as a one-dimensional diffusion-to-bound process [32, 33], wherein the decision variable, x, takes a random walk that starts at zero and ends at one of the two bounds. This can be written as $\mathrm{d}x/\mathrm{d}t = r + u(0,s)$, where r is the mean drift rate and u is the momentary noise represented by a random variable drawn from some distribution, typically a Gaussian with mean = 0 and standard deviation = s. The random element guarantees that, given enough time, x will hit one bound or the other even if the drift rate is zero. The diffusion parameters (r, s) as well as the bounds (A,B) can be fit to experimental data for accuracy and reaction time [34].

Figure 5.12 shows simulations based on the SPRT where the likelihood density functions are Gaussians. The outcomes can be classified as hits (blue) and correct rejections (red), as well as misses and false alarms (not shown). The proportions of correct and incorrect trials as well as the response time distributions for each class of outcome are fully determined by the log-likelihood ratio and the boundaries (Fig. 5.12, below).

The standard SDT notions of detectability and response bias are built into the SPRT. Detectability depends on the rate of evidence accumulation, drift rate, and the variance in drift rate or momentary noise. Response bias occurs when the bounds are

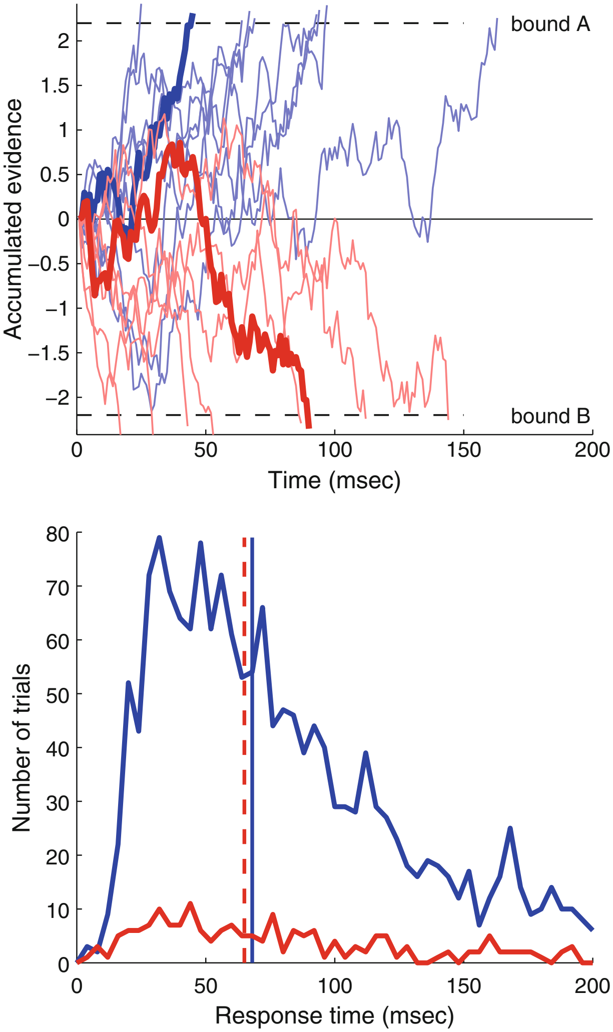


Fig. 5.12 Simulations of SPRT. (*Top*) *Blue lines* represent stimulus-present trials. *Red* are stimulus-absent trials. (*Bottom*) RT distributions for stimulus-present (*blue*) and stimulus-absent (*red*) trials

asymmetric so that the diffusion process starts from a position closer to one bound than the other. The SPRT effectively has two independent criteria, whereas the static SDT model has only one. In the SPRT, the outcome classes are more independent. For example, it is possible to maintain a constant hit rate while varying the false alarm rate. Thus, the trade-off between hits and false alarms that is characteristic of the static SDT model does not hold for the SPRT.

If attention increases signal quality by reducing signal to noise, the effect on the SPRT will be to increase the rate of evidence accumulation [35]. This is equivalent to improving stimulus detectability. Others have incorporated salience and economic value by modulating drift rate [36].

The SPRT also provides a solution to the problem of pooling responses across multiple detectors. The summation of log-likelihoods applies not only to the integration of multiple samples from a single detector, but also to the integration of individual samples from multiple detectors. Given a set of samples from multiple detectors, one can simply sum the log of the likelihood ratios to obtain an estimate of the evidence that a stimulus is present. This could be called the parallel probability ratio test (PPRT). This calculation can be performed at every moment in time. The conversion from raw samples to log-likelihoods takes into account the signal-to-noise of each detector and thus provides a common metric for integrating responses from detectors with different sensitivities, filtering properties, and noise characteristics.

Computing the SPRT in parallel across the visual scene using a 2D array of detectors results in a detectability salience map. This is illustrated in Fig. 5.13 with a 40 × 40 array of detectors. The signal can occur at one of two locations (lower left or upper right). Detectability (hit rate – false alarm rate) for each detector is plotted (black indicates detectors with high false alarms, white indicates detectors with high hit rate). The cueing paradigm described in Fig. 5.10 was implemented with two different cue validities. Cue validity is implemented by biasing the starting point of the decision processes at the cued and uncued locations [37]. When the cue validity is 0.5, the cue provides no information, the signal and noise distributions have equal area, and detectability is equal at the two locations. When the cue validity is 0.55, the signal appears at the cued location 55 % of the time and at the uncued location 45 % of the time. This enhances the detectability at the cued location and reduces detectability at the uncued location. Figure 5.13 plots stimulus detectability, but SPRT-computed salience can also be expressed in terms of response time. If detectability and response time are combined, it is possible to calculate the information processed by the observer in terms of bits/second

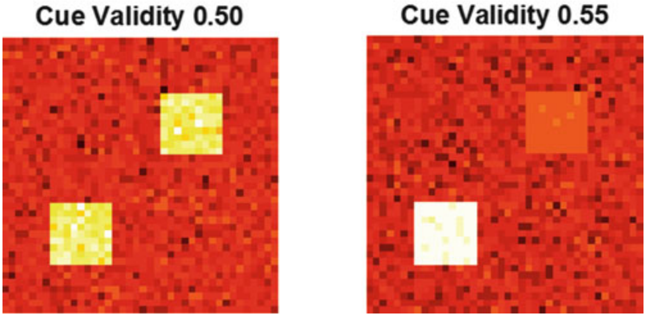


Fig. 5.13 Salience maps computed using SPRT. Both maps show the detectability of a signal that can occur at one of two locations. (*Left*) signal occurs at either location with equal probability. (*Right*) signal occurs at cued location (*lower left*) 55 % of the time and uncued location (*upper right*) 45 % of the time. Intensity indicates stimulus detectability

(e.g., [38]). Furthermore, the SPRT allows the observer to adjust the decision boundaries at the cued and uncued location, which should also affect the relative salience.

5.7 Conclusion

Signal detection theory provides a simple yet powerful framework for understanding how observers respond to weak signals in the environment. The theory makes a clear distinction between detection and response selection. Attention can improve signal detection by increasing the gain of sensory responses while reducing noise. For a fixed level of detectability, attention can further improve performance by optimizing decision criteria. When there are multiple detectors, attention can improve detectability by de-correlating responses and by selectively monitoring detectors that are more sensitive to the stimulus by virtue of their receptive field location, feature selectivity, or other properties.

Acknowledgements This work was supported by NIH grant MH059244. We thank Prof. Norma Graham and Dr. Greg Jensen for a critical reading of the manuscript.

References

1. James, W. (1890). *The principle of psychology*. New York: Henry Holt & Co.
2. Broadbent, D. (1958). *Perception and communication*. London: Pergamon Press.
3. Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological Review, 75*, 522–536.
4. Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology. General, 109*, 160–174.
5. Eriksen, C., & St. James, J. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics, 40*(4), 225–240.
6. Graham, N. V. S. (1989). *Visual pattern analyzers*. New York: Oxford University Press.
7. Eckstein, M. P., Peterson, M. F., Pham, B. T., & Droll, J. A. (2009). Statistical decision theory to related neurons to behavior in the study of covert visual attention. *Vision Research, 49*, 1097–1128.
8. Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review, 70*(1), 80–90.
9. Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.
10. Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
11. Mcmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd ed.). New York: Lawrence Erlbaum Associates, Psychology Press.
12. Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature, 435*, 439.

13. Gur, D., Rockette, H. E., Armfield, D. R., Blachar, A., Bogan, J. K., Brancatelli, G., Britton, C. A., Brown, M. L., Davis, P. L., Ferris, J. V., Fuhrman, C., Golla, S. K., Katyal, S., Lacomis, J. M., McCook, B. M., Thaete, F. L., & Warfel, T. E. (2003). Prevalence effect in a laboratory environment. *Radiology, 228*, 1–14.
14. Wickens, T. D. (2001). *Elementary signal detection theory*. New York: Oxford University Press.
15. Maunsell, J. H. (2004). Neuronal representations of cognitive state: Reward or attention? *Trends in Cognitive Sciences, 8*(6), 261–265.
16. Shadlen, M. N., Britten, K. H., Newsome, W. T., & Movshon, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience, 16*(4), 1486–1510.
17. Zohary, E., Shadlen, M. N., & Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature, 370*(6485), 140–143.
18. Cohen, M. R., & Maunsell, J. H. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience, 12*(12), 1594–1600.
19. Mitchell, J. F., Sundberg, K. A., & Reynolds, J. H. (2009). Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. *Neuron, 63*(6), 879–888.
20. Shadlen, M. N., Bitten, K. H., Newsome, W. T., & Movshon, J. A. (1996) A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience, 16*(4), 1486–1510.
21. Bashinski, H. S., & Bacharach, V. R. (1980). Enhancements of perceptual sensitivity as the result of selectively attending to spatial locations. *Perception & Psychophysics, 28*, 241–248.
22. Mertens, J. J. (1956). Influence of knowledge of target location upon the probability of observation of peripherally observable test flashes. *Journal of the Optical Society of America, 46*(12), 1069–1070.
23. Downing, C. J. (1988). Expectancy and visual-spatial attention: Effects on perceptual quality. *Journal of Experimental Psychology: Human Perception and Performance, 14*(2), 188–202.
24. Hawkins, H. L., Hillyard, S. A., Luck, S. J., Mouloua, M., Downing, C. J., & Woodward, D. P. (1990). Visual attention modulates signal detectability. *Journal of Experimental Psychology: Human Perception and Performance, 16*(4), 802–811.
25. Muller, H. J., & Findlay, J. M. (1987). Sensitivity and criterion effects in the spatial cueing of visual attention. *Perception & Psychophysics, 42*, 383–399.
26. Shaw, M. L. (1984). Division of attention among spatial locations: A fundamental difference between detection of letters and detection of luminance increments. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X* (pp. 109–121). Hillsdale: Erlbaum.
27. Müller, H. J., & Findlay, J. M. (1987). Sensitivity and criterion effects in the spatial cuing of visual attention. *Perception & Psychophysics, 42*(4), 383–399.
28. Gorea, A., & Sagi, D. (2005). Decision and attention. In L. Itti, G. Rees, & J. Tsotsos (Eds.), *Neurobiology of attention*. New York: Elsevier.
29. Carrasco, M., & McElree, B. (2001). Covert attention accelerates the rate of visual information processing. *Proceedings of the National Academy of Sciences, 98*(9), 5363–5367.
30. Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics, 16*(2), 117–186.
31. Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 231*(694–706), 289–337.
32. Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.
33. Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review, 116*(2), 283–317.
34. Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin and Review, 9*(3), 438–481.

35. Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *PNAS, 108*, 13852–13857.
36. Towal, R. B., Mormann, M., & Koch, C. (2013). Simultaneous modeling of visual saliency and value computation improves predictions of economic choice. *PNAS, 110*(40), E3858–E3867.
37. Sperling, G. (1984). A unified theory of attention and signal detection. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 103–181). London: Academic Press.
38. Teichert, T., Ferrera, V. P., & Grinband, J. (2014). Humans optimize decision-making by delaying decision onset. *PLoS One, 9*(3), e89638.

Chapter 6
Effects of Attention in Visual Cortex: Linking Single Neuron Physiology to Visual Detection and Discrimination

Vincent P. Ferrera

6.1 Introduction

Studies of neuronal activity in visual cortex have relied heavily on macaque monkeys as a model system. Macaques, like humans, are old world primates and range throughout Asia and North Africa. The macaque genus comprises 23 species, including *Macaca mulatta* (rhesus monkey), *Macaca fascicularis* (cynomolgus or "crab-eating" monkey), and *Macaca fuscata* (Japanese snow monkey). The most recent common ancestor of humans and macaques lived roughly 25 million years ago. Macaques are largely diurnal animals that have trichromatic color vision and a retina that is anatomically almost identical to humans. In particular, the macaque retina has a distinct fovea for high-acuity central vision.

Macaques explore their visual environment in much the same way as humans. They have forward-looking eyes whose monocular visual fields are largely overlapping, providing a large binocular field with excellent stereoscopic depth perception [1]. Their oculomotor behavior is similar to humans, particularly with regard to voluntary eye movements. Macaques have *vergence* eye movements that align the foveae of the two eyes on targets at a particular distance. They make rapid and frequent *saccades* to foveate objects of interest. They can track moving targets with *smooth pursuit*, a behavior that appears to be unique to primates (at least among mammals.)

V.P. Ferrera (✉)
Department of Neuroscience, Columbia University, 1051 Riverside Drive, Unit 87, New York, NY 10032, USA
e-mail: vpf3@cumc.columbia.edu; vincent.ferrera@gmail.com

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_6

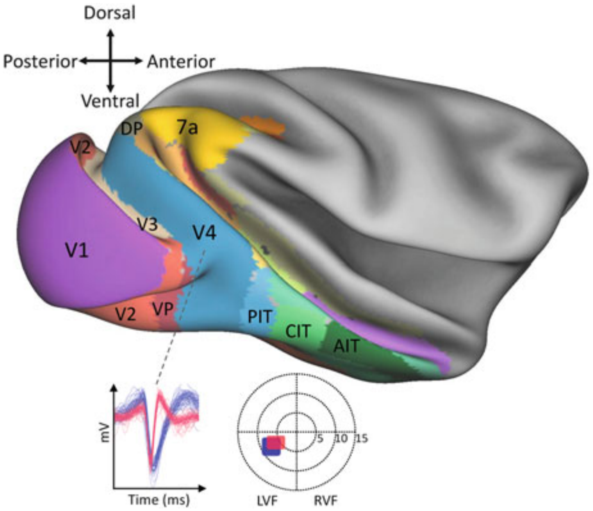


Fig. 6.1 Macaque cerebral cortex (lateral view, partially inflated) showing visual cortical areas *V1* (primary), *V2*, *V3*, *V4*, *DP* (dorsal posterior), *7a*, *VP* (ventral posterior), *PIT* (posterior inferotemporal), *CIT* (central inferotemporal), and *AIT* (anterior inferotemporal; after [2]). A microelectrode can be used to record neuronal activity at a precise cortical location (*dashed line*). Extracellular action potential waveforms for two simultaneously recorded neurons are shown in the *lower left* panel. The visual field locations and sizes of the receptive fields (*blue* and *red* squares) of the neurons are shown in the *lower right* panel. *LVF* left visual field, *RVF* right visual field

In the macaque monkey brain, there are 32 cortical areas that are involved in vision and visuomotor function [2]. For many of these areas, human homologues have been identified [3]. Macaques can be trained to perform simple tasks that involve visual detection, discrimination, and eye movements. The electrical activity of individual neurons can be recorded by fine metal microelectrodes inserted into the cerebral cortex while the animal is performing a visual task. Of the 32 visual areas in macaques, several have been studied extensively in behavioral paradigms that manipulate selective attention (Fig. 6.1). These studies have examined how attention affects receptive field properties as well as the sensitivity and reliability of neuronal responses. The current state of knowledge makes it possible to relate these neuronal response properties to psychophysical performance using simple computational models. The goal of the present chapter is to understand how attention alters the representation of information in visual cortex and thus affects an observer's ability to detect weak stimuli and to discriminate between similar stimuli.

Visual neurons are those that receive information directly or indirectly from the retina. The part of the environment that gives rise to light that falls onto the retina defines the visual field. Visual neurons typically do not respond to light that arises from anywhere in the visual field, but are sensitive to only a small region, called the *receptive field*. The receptive field for an individual neuron is the part of the retina within which changes in illumination cause changes in the electrical activity (typically, the firing rate) of the cell. If the eyes are not moving, the receptive field corresponds to a fixed region of visual space. Every cell in the visual system, from retina photoreceptors to cortical neurons, has a receptive field. The size of receptive fields generally increases along the visual hierarchy from retina to lateral geniculate to cortex and also with retina eccentricity (distance from the fovea). If a monkey is trained to fixate its gaze on a small target presented on a video display, then the borders of the receptive field can be easily mapped. This may be done by moving a spot or bar of light through the visual field and outlining the region where the stimulus causes a change in firing rate of the cell. Firing rate can be monitored qualitatively by amplifying the action potentials and playing them through an audio speaker. As long as the monkey is fixating, a particular stimulus other stimuli presented in the visual field will have a known spatial relationship with respect to the receptive field of a given neuron. Controlling the retina stimulus in this manner makes it possible to study the influence of *extraretinal* factors, such as attention, on the activity of visual neurons.

The receptive field of a neuron can be modeled mathematically as a spatial weighting function, which specifies the neuron's firing rate as a function of the retina position of a small spot of light. A visual neuron's sensitivity to light within the receptive field is not necessarily uniform, but may have subregions that are excited or inhibited by light. The spatiotemporal structure of the receptive field may confer selectivity for orientation and direction of motion. Different parts of the receptive field may also be sensitive to different wavelengths of light, giving rise to color selectivity. For current purposes, we will ignore the internal structure of visual receptive fields and simply model sensitivity within the spatial receptive field (RF) as a two-dimensional Gaussian:

$$\mathrm{RF}\,(\mathrm{x},\mathrm{y}) = \mathrm{A} + \mathrm{B} \times \exp\left[-\left(\left(\mathrm{x}-\mathrm{x}'\right)^2 + \mathrm{C} \times \left(\mathrm{y}-\mathrm{y}'\right)^2\right)/\mathrm{s}^2\right] \tag{6.1}$$

Here, (x′, y′) is the center of the receptive field, s is the spread or size of the RF (otherwise known as the *space constant*), B is the overall gain or sensitivity, C determines the aspect ratio (length/width), and A is a constant that accounts for the baseline firing of the cell in the absence of a stimulus. Many studies of the effects of attention on the activity of visual neurons have examined changes in spatial parameters that correspond to shrinking or expanding of the receptive field. These are modeled as changes in the space constant, s. Other studies have documented shifts of the RF center (x′, y′) and changes in overall sensitivity (B) and background firing (A).

To understand how attention-related changes in receptive field properties affect stimulus detectability and discriminability, it is necessary to consider the statistics of neuronal responses, i.e., the variability in neuronal firing when the same stimulus is presented repeatedly under the same conditions.

To a first approximation, cortical neurons fire at purely random times. Their firing can be modeled as a Poisson process where the probability of an action potential at any given time is determined by a rate parameter, r, and is independent of the time of occurrence of any other action potentials. The interspike intervals (times between two successive action potentials) follow a Poisson distribution. The number of action potentials in a fixed time window (spike count) is also Poisson distributed. Spike count variability can be quantified by the Fano factor [4], which is the variance in spike count divided by its mean. For a Poisson process, the Fano factor is always around 1.0 as the variance scales in direct proportion to the mean spike count.

Poisson firing statistics represent an ideal case that is never achieved in reality. In particular, a purely Poisson neuron could have infinitely small interspike intervals, which are biophysically impossible. Real neurons have refractory periods – a short window of time following a spike during which the cell is unable to fire another spike (absolute refractory period) or has an elevated threshold for firing (relative refractory period). Refractory periods are easy to incorporate into simulations that generate pseudo-Poisson spike trains using random number generators [5]. Refractory periods cause neuronal firing to become more regular (lower variance in interspike intervals). Any finite refractory period therefore reduces the Fano factor below 1.0. A number of studies have documented sub-Poisson variability in macaque visual cortex and in higher, attention-related cortical areas [6–10].

Figure 6.2 shows simulated Poisson-like spike trains generated by an algorithm that incorporates an absolute refractory period. In the top-left panel are spike trains where the refractory period is equal to 0 and below that the spike count histogram and Fano factor versus mean spike count. The right column shows spike trains with the same average rate, but a longer refractory period, making both the interspike intervals and spike counts much more regular.

It has been found empirically that attention can reduce neuronal variability [11], but the reduction is small and not always statistically significant [12]. It seems intuitive that reduced variability should improve the ability to detect and discriminate stimuli. One of the goals of the models presented below is to test whether this is indeed the case.

A simple model of the response of an individual visual neuron can be obtained by using Eq. 6.1 to provide the input to a Poisson spike generating process. This is illustrated in Fig. 6.3 which shows the mean rate according to a one-dimensional reduction of Eq. 6.1 (Fig. 6.3, top) and the Poisson spike counts (Fig. 6.3, bottom) generated when a stimulus is present ($A = 5$, $B = 10$) or absent ($A = 5$, $B = 0$). Detectability can be computed for each stimulus position as the overlap (area under ROC) of the stimulus-present and stimulus-absent spike count distributions. Note that this is not a complete neuronal model as it does not include contrast nonlinearities, adaptation, or other factors that affect firing. Real visual neurons tend to have sigmoidal contrast response functions, and their contrast sensitivity may be modulated by attention [13, 14]. However, the current model is adequate for testing effects of changes in sensitivity or variability for briefly presented stimuli of fixed contrast.

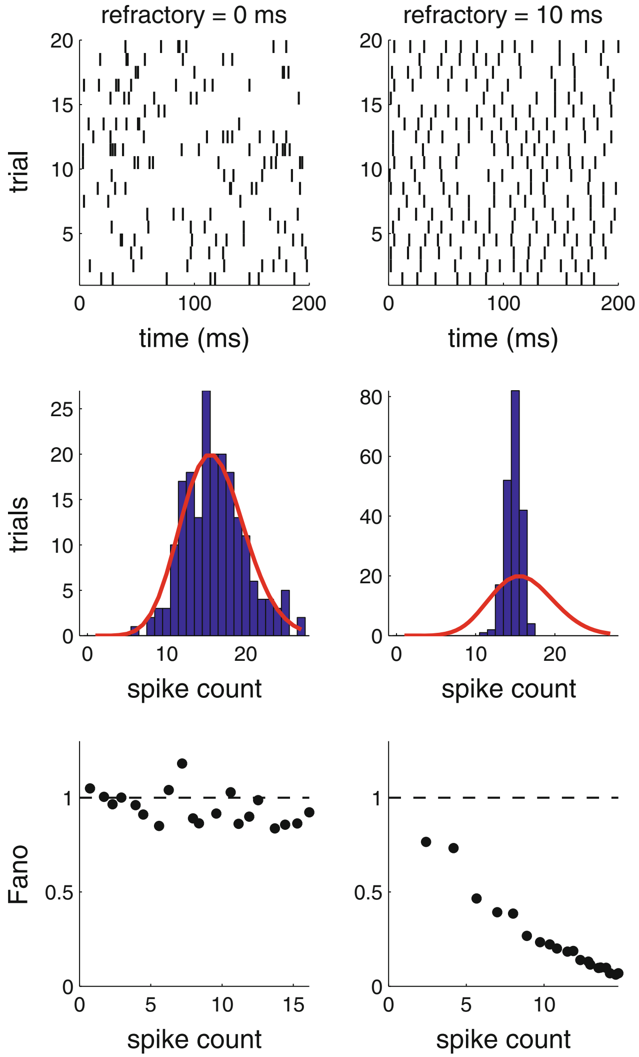


Fig. 6.2 Neuronal firing statistics. (*Left*) Pure Poisson process. (*Right*) Poisson with refractory period. *Top row* shows 20 spike trains for each model. *Middle row* shows spike count distributions for several hundred trials. *Bottom row* shows Fano factor versus spike count

6.2 Effects of Attention on Neuronal Responses

Moran and Desimone [15] published one of the first studies of the effect of attention on neurons in macaque visual cortex. They trained monkeys to fixate a small spot presented in the center of a video display. Eye movements were monitored so that visual stimuli could be presented at known positions on the retina. While

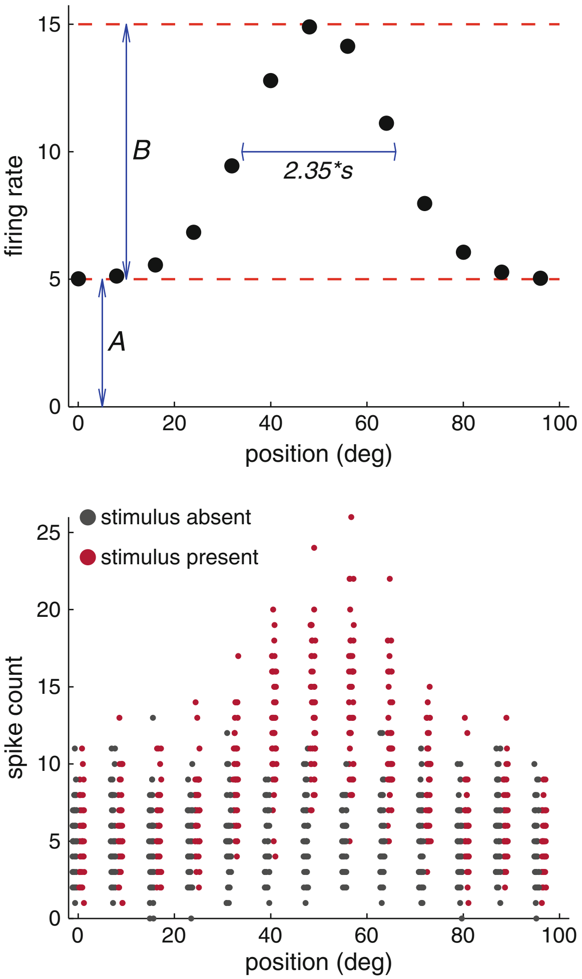


Fig. 6.3 Minimal model of a single neuron. (*Top*) Mean firing rate as a function of position. (*Bottom*) Distributions of spike counts for stimulus-absent and stimulus-present conditions as a function of stimulus position in receptive field

the monkeys fixated, two stimuli were presented, and the monkeys were rewarded for responding to one stimulus. The monkeys are presumed to have attended to rewarded stimulus and to have ignored the other.

Moran and Desimone recorded from neurons in visual area V4 and in the inferior temporal (IT) cortex. Neural responses were quantified as changes in firing rate (action potentials per second), while visual stimuli were presented to the animal. The receptive fields of the neurons were in the peripheral visual field and were large enough that two stimuli could be presented inside the receptive field and the monkey could still discriminate them. If both stimuli were in the receptive field of the neuron, the cell responded well to the attended stimulus, but weakly to the unattended stimulus. The experimenters could therefore compare the response to the same stimulus when it was attended or unattended. Generally, the response to the

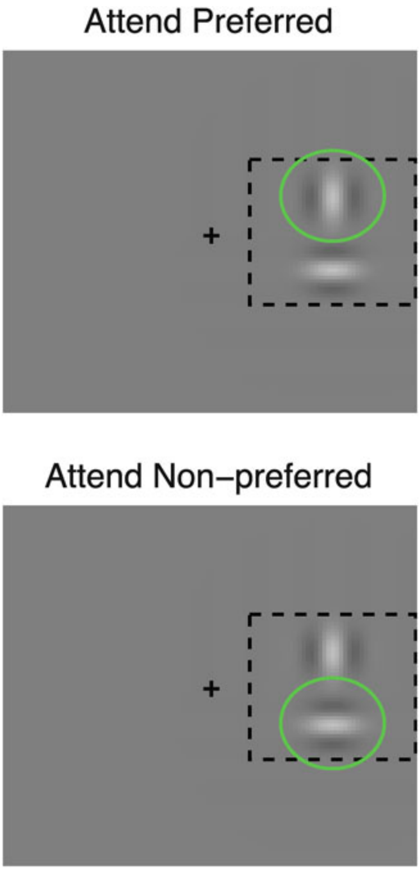


Fig. 6.4 Attention task used by Reynolds et al. [16]. Monkeys were trained to fixate their gaze at the center of the display (+) while the activity of a visual neuron was recorded. The receptive field of the neuron is indicated by the *dashed box*. Two stimuli were presented inside the receptive field, and the monkey was rewarded for responding to one or the other. The attended stimulus is indicated by the *green circle* (this cue was not presented to the animal)

stimulus was greater when it was attended. If one stimulus was inside the receptive field and the other was outside, the effect of attention was reduced as compared to when both were inside the receptive field.

A later study by Reynolds et al. [16] expanded on this result. Reynolds' study used the strategy of placing two oriented bar stimuli in the receptive field of a V4 neuron (Fig. 6.4). Neurons in V4 tend to be selective for stimulus orientation. The orientation of one of the bars was matched to the cell's preferred orientation and evoked a strong response. The other stimulus was at a non-preferred orientation. In the absence of attention, the neuronal response when both stimuli were presented together was the average of the response to either stimulus alone. When the monkey was rewarded for attending to one stimulus or the other, the cell behaved as if there was only one stimulus in the receptive field; if the monkey attended the preferred stimulus, the response was greater than the average; if he attended the non-preferred stimulus, the response was less than the average. Thus, attention caused the cells to shift from a response-averaging mode to a winner-take-all mode. These results are consistent with a shrinking of the receptive field around the attended stimulus

[17, 18]. Since Moran and Desimone's [15] paper, a large number of studies have demonstrated changes in the receptive field spatial weighting function that are correlated with attention.

Attention can change the overall gain of visual responses. This was demonstrated for neurons in visual area V4 by McAdams and Maunsell [12, 19]. This study examined orientation tuning. Orientation selectivity is reasonably described by Gaussian-shaped tuning function. McAdams and Maunsell placed an oriented grating pattern in the receptive field of a V4 neuron (Fig. 6.5). They then recorded responses to stimuli of various orientations and compared the orientation-tuned responses when attention was directed toward the stimulus inside the RF or to a similar stimulus well outside the receptive field. They modeled V4 responses using

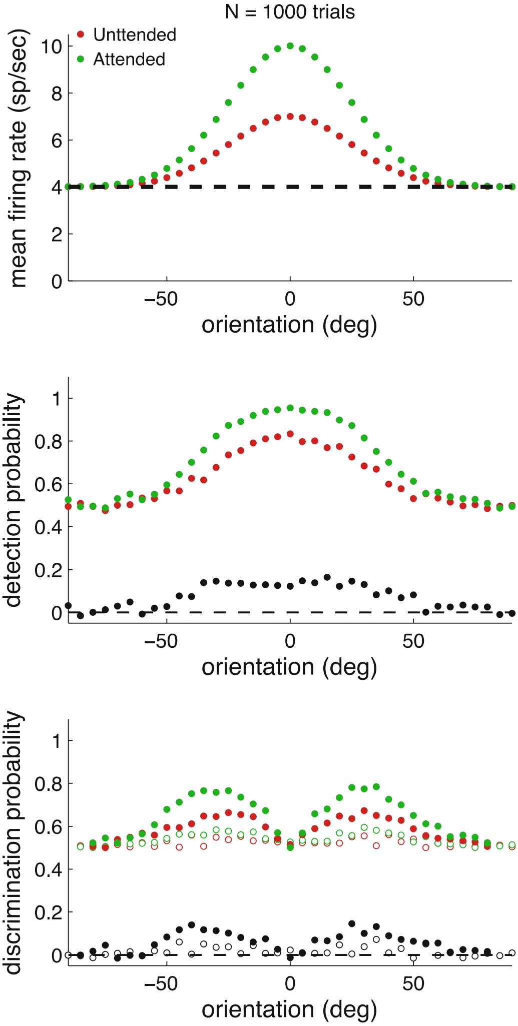


Fig. 6.5 Effects of attention on response of a single neuron. (*Top*) Mean rates when attention is directed outside the neuron's receptive field (*red*) and inside the RF (*green*). (*Middle*) Effect of attention on detectability as a function of stimulus position. *Black dots* are difference attended (*green*) – unattended (*red*). (*Bottom*) Effect of attention on stimulus discriminability

an equation similar to Eq. 6.1 and concluded that attention mainly affects the overall gain (B). Importantly, they also measured activity in the absence of a stimulus and found that attention did not affect the baseline (undriven) firing rate (A).

These few studies provide enough information to simulate the effects of attention in the model introduced previously. In this model, the receptive field equation (Eq. 6.1) is used to determine the mean firing rate for a small spot of light presented at any position in the receptive field. This mean rate is then fed into a function that generates a pseudo-Poisson spike train for a fixed time interval (1.0 s). This spike train can be purely Poisson (refractory period = 0), or can have a finite refractory period, resulting in sub-Poisson variability. For each stimulus, a large number of spike trains are generated and the total spike count for each train is used as the measure of neuronal response. One can then use principles of signal detection theory to determine the ability of one or more simulated neurons to detect or discriminate visual stimuli, given the trial-to-trial variability in the neuronal responses. Figure 6.5 (top) shows the mean response of a single visual neuron (reduced to one dimension) as a function of stimulus location. The effect of attention is modeled as an overall gain factor, G, applied to the stimulus-driven response, so that

$$\mathrm{RF}\,(\mathrm{x}) = \mathrm{G} \times \left\{\mathrm{A} + \mathrm{B} \times \exp\left[-\left(\left(\mathrm{x} - \mathrm{x}'\right)^2\right)/\mathrm{s}^2\right]\right\} \tag{6.2}$$

The responses in Fig. 6.5 show the cases where G = 1.0 (red, attention outside receptive field) and G = 2.0 (green, attention inside RF). In the absence of a stimulus, the response is simply RF(x) = G*A, where A is the baseline firing rate.

Detectability and discriminability are computed by applying signal detection theory to the spike count distributions for each stimulus. Detectability is defined as the area under the ROC curve computed with stimulus-present and stimulus-absent trials. The effect of attention on detectability is shown in Fig. 6.5 (middle). Even though attention increases the driven firing rate by twofold, the maximum change in detection probability is only 0.1. It should be noted that an attentional gain of 2.0 is unusual. Typically, attention enhances neuronal responses by increasing mean firing rate from 20 % to 40 %. For many cells, attention actually reduces responses.

In the simulation shown in Fig. 6.5, the baseline firing rate in the unattended condition was 4 spikes/sec, and the maximum firing rate was 10 spikes/sec. This value for maximum firing rate is on the low end of the range for cortical neurons. Values of 30 spikes/sec or greater are more typical for responses to optimal stimuli. Hence, the low ratio of max firing rate to baseline can be thought of as representing the response to suboptimal or weak stimuli. Detection probability in the unattended condition starts to saturate at 1.0 (perfect performance) when the maximum firing rate is about 3 times the baseline rate. Attention cannot improve performance when detection rates in the unattended condition are already optimal. Thus, attention should have the greatest effect on detectability for weak or suboptimal stimuli or cells that simply have low signal to noise even for optimal stimuli.

Stimulus discriminability is defined as the ROC area computed for pairs of similar stimuli. In these simulations, the difference between neighboring orientations was 5°. For the parameters used in the simulations of Fig. 6.5, attention had little effect on discriminability for neighboring orientations (open symbols). However, for cruder discriminations (stimuli separated by 20°.), discrimination performance was better overall (Fig. 6.5, bottom, filled symbols) and was more strongly enhanced by attention. Further simulations showed that as the maximum firing rate is increased, attention had a greater effect on improving discrimination performance. For example, at a maximum firing rate of 50 spikes/sec (keeping all other parameters the same), the best discrimination for neighboring orientations improved from 74 % correct to 82 %. Hence, while attention improves detection performance mainly for neurons with low signal to noise, it improves discrimination for neurons with high signal strength. This suggests that different tasks might reveal attention effects on different subpopulation of neurons.

The effects of attention on psychophysical performance predicted by the model are fairly modest. Using realistic parameters, attention improves detection and discrimination rates by a maximum of about 10 %. Such small changes in performance are far below what is typically reported in the literature. For example, [20] found that attention could produce up to fourfold improvements in contrast sensitivity. We will consider two factors that could bridge this gap. The first is the effect of attention on undriven (stimulus-absent) activity. The second is the effect of attention on spike count variability.

So far, we have assumed that attention affects firing rates proportionately for both stimulus-present (driven activity) and stimulus-absent (undriven or baseline activity) conditions. This point is disputed. Some studies report that attention affects background firing rates [21], while others [12, 19] reported that attention did not affect undriven activity. When undriven activity is held constant in the model, so that attention enhances activity only in the presence of a stimulus, the affects of attention on detection are greatly increased. Repeating the simulations of Fig. 6.5 with a constant baseline, the improvement in detection probability goes from 0.1 to 0.3. This is a large enough improvement to account for actual psychophysical performance. Thus, the issue of whether attention affects baseline activity is critical for understanding improvements in detection performance. However, in the model, baseline firing rate plays no role in discrimination performance.

Now we can address the issue of attention-related changes in spike count variability. As noted above, some studies have reported that attention can reduce trial-to-trial variability in firing activity [11]. Here, we reduce spike count variability by introducing a refractory period. There is a caveat to this approach: for any two spike trains with the same underlying rate, the one with the longer refractory period will have a lower spike count. Thus, it is important to equalize spike count when assessing the effects of regularity. Figure 6.6 (top) shows Fano factor as a function of mean spike count for spike trains with no refractory period (red) and with a refractory period of 10 ms (green). Note that refractoriness causes Fano factor to decrease with mean spike count, being reduced by about half for the highest firing rate.

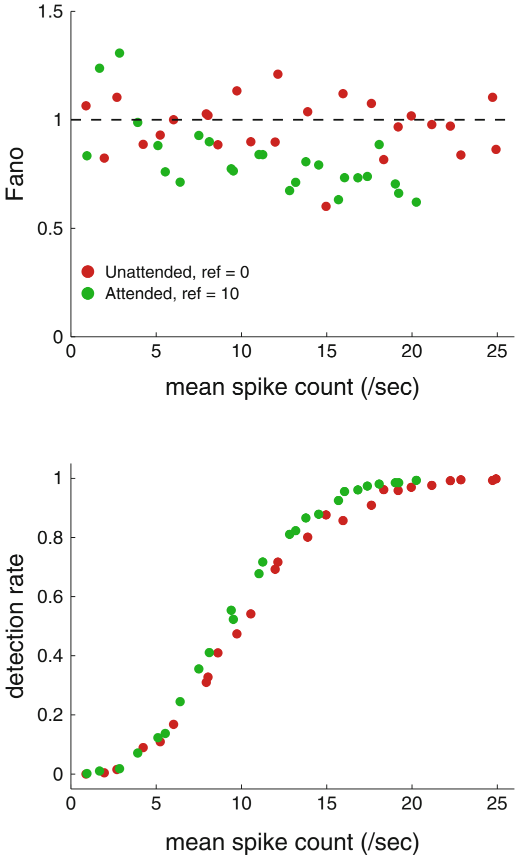


Fig. 6.6 Effects of attention modeled as changes in trial-to-trial spike count variability. (*Top*) Fano factor as a function of mean spike count for refractory period = 0 (*red*) or 10 (*green*). (*Bottom*) Detection rate as a function of mean spike count. Same convention as *top panel*

The effect of spiking regularity on detection rate is shown in Fig. 6.6 (bottom). Again, the red dots are for spike trains with zero refractory period; the green are for a refractory period of 10 ms. The same refractory period was used for both stimulus-present and stimulus-absent conditions, although there is some evidence that stimulus onset itself is accompanied by a reduction in spike count variability [22]. What is evident from Fig. 6.6 is that a reduction in variability improves detection rates, but only by about 5–10 %. The improvement is greatest when the signal to noise is relatively weak, such that the maximum firing rate is about twice the baseline firing. When the maximum firing rate increases beyond this, detection rates saturate and spike count regularity has no effect. The simulations were carried out with a baseline of 10 spikes/sec. Changing the baseline firing rate shifts the curves left and right, but the same principles apply.

While spike count regularity alone results in some enhancement of detectability, it has a smaller effect on discrimination performance. Figure 6.7 shows simulations of a neuron whose receptive field is modeled as a one-dimensional Gaussian function of position, with preferred position at 50°. The left panel shows Fano

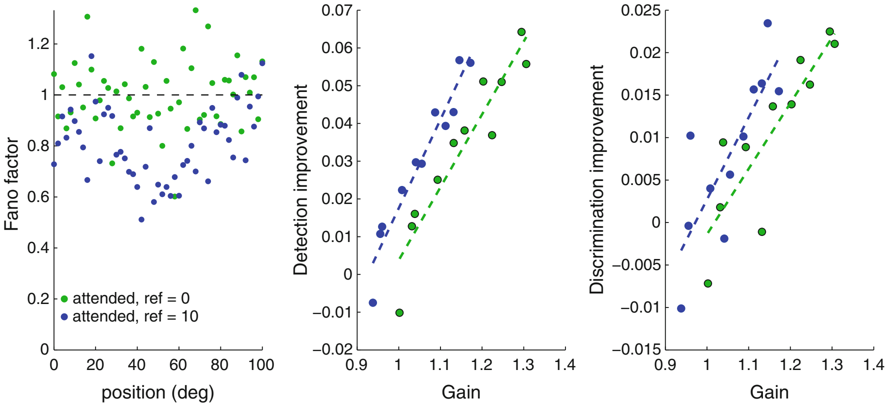


Fig. 6.7 Effects of spike count variability on discrimination performance. (*Left*) Fano factor as a function of stimulus position for unattended (*green*, refractory period = 0) and attended (*green*, refractory period = 1). (*Middle*) Effects of variability on detection. Gain is attended response/unattended response. Detection improvement is the difference in detection rate for each position (attended – unattended) averaged over all positions. *Dashed lines* are best fit linear regressions. (*Right*) Effects of variability on detection. Same conventions as *middle* panel

factors as a function of stimulus position in the receptive field for no refractory period (green) or a refractory period of 10 ms (blue). There is a substantial, stimulus position-dependent decrease in Fano factor. However, the decrease in variability is accompanied by a proportional decrease in mean spike count due to the refractoriness of the cell. In other words, the overall response of the cell is scaled, including the mean spike count and variance. The decrease in variability leads to an improvement in detectability of a few percent. Detection rate is calculated by computing the ROC for stimulus-present versus stimulus-absent conditions and assuming that attention does not affect baseline firing (either average rate or variability) in the absence of the stimulus. Discriminability is based on the spike count distributions for neighboring stimuli. Since refractoriness scales both distributions proportionately, firing regularity only has a small effect on ROC area. When the attention-related improvement in discrimination performance is plotted as a function of actual gain (Fig. 6.7, right), the improvement in performance is quite small (green, refractory period = 0; blue, refractory period = 10 ms).

To summarize, attention can affect the gain of neuronal responses as well as their reliability. Large changes in response gain lead to only modest improvements in detection and discrimination rates. If baseline activity is unaffected by the gain change, then much larger increases in detection rates are achievable, but there is no effect on discrimination. Improving reliability by incorporating a refractory period into the spike train generator has a small effect on detection and an even smaller effect on discrimination. One caveat is that refractoriness always reduces both the variance and mean of the spike counts. Other methods that reduce variability without changing mean rate were not explored.

6.3 Effects of Attention Across Multiple Neurons

When considering the effects of attention across multiple neurons, there is a general expectation that such effects will be stronger and/or more reliable. This expectation may be frustrated for several reasons. Having more neurons can improve signal processing, but it also means that there will be more noise due to random firing from neurons that are not sensitive to the stimulus. Indeed the problem of selective attention is not only one of selecting the most relevant stimulus, but, perhaps more importantly, selecting the most relevant neurons.

To model the effects of attention across multiple neurons, consider an array of neurons that are identical except for the location of their receptive field centers. Instead of the scalar attentional gain factor in the single neuron model described above, attention is modeled as a gain field [G(x), [18]] that ranges over the entire visual field:

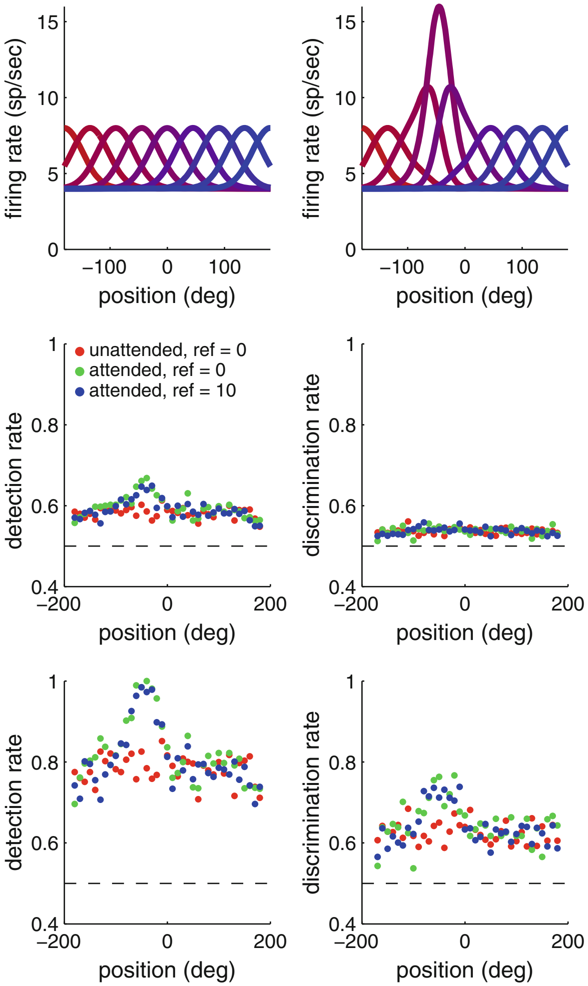
$$\mathrm{G}(\mathrm{x}) = 1.0 + \mathrm{a} \times \exp\left(-\left(\mathrm{x} - \mathrm{x}'\right)^2/\mathrm{s}^2\right) \tag{6.3}$$

where a is the attentional enhancement, x is visual field location, x′ is the focus of attention, and s is the spread of attention. Figure 6.8 shows the effect of attentional gain enhancement on an array of neurons that have identical tuning width and sensitivity, but different receptive field centers. Tuning curves for the unattended case are shown in the left. On the right are tuning curves with a maximum attentional gain of 2.0. Attention does not enhance the activity of cells whose preferred locations are remote from the focus of attention. It should be noted that the attentional gain field not only enhances the response of cells at the focus of attention, but it also distorts the tuning functions. Such shifting of receptive fields has been documented in visual areas V4 [23] and MT [18].

In the case of a single neuron, it was found that if attention enhanced both undriven and stimulus-driven activity, there was little improvement in detection or discrimination performance. Here, we test what happens if attention does not affect baseline firing in either the stimulus-absent or stimulus-present conditions. When we simulate this condition, it turns out that attention has little effect on detection (Fig. 6.8, middle left) and no effect on discrimination (Fig. 6.8, middle right). This result holds over a wide range of signal strengths (maximum firing rate re: baseline). To obtain even a small increase in detectability requires an attention gain of about 4x or greater. Discriminability does not improve for any gain level. There was no effect of refractoriness on detection or discrimination rates.

The finding that attentional gain has little effect on detection may seem counter-intuitive. However, it makes perfect sense. There are nine neurons in the simulation, and, as can be seen in Fig. 6.8, attention only affects 3 of them. For any given stimulus, most of the cells do not respond at all. Yet, all of the cells must be included when computing detectability, even if they are unmodulated by attention, or not even driven by the stimulus. The reason for this is that the stimulus has an equal probability of occurring at any location and this location is not known in advance.

Fig. 6.8 (*Top*) Tuning functions for an array of model neurons in the unattended condition (*left*) and with attention focused at −50°. (*Middle*) detection and discrimination rates. (*Bottom*) Detection and discrimination rates when responses are pooled across neurons



Thus, at any given time, most of the cells are simply contributing noise. This not only dilutes the effect of attention; it can negate the effect altogether.

However, we have yet to consider the issue of pooling activity across neurons. By this, we mean how signals from different neurons are combined when computing the joint ROC. None of the multi-neuron simulations discussed above included any pooling; each response was considered as an independent observation and was weighted equally in the ROC analysis. One way to pool responses is to compute the mean spike count across all neurons in the model on each trial. Thus, the data are reduced from 3 dimensions (neuron × stimulus × trial) to only 2 (stimulus × trial). This averaging is done before the ROC area is computed. The effect of this kind of pooling is that the neurons that are sensitive to the stimulus tend to pull up the average response of the ensemble. On the other hand, when there is no stimulus, averaging across neurons has little effect because they all have the same baseline activity.

We thus consider a model in which attention has no effect on baseline activity, but activity on every trial is pooled by averaging across all neurons. The results are shown in the bottom row of Fig. 6.8. For this model, attention enhances both detection (Fig. 6.8 bottom left) and discrimination (Fig. 6.8, bottom right). As in previous simulations, reducing trial-to-trial variability has no effect (compare green dots, refractory period = 0, to blue, refractory period = 10 ms).

Averaging over all neurons is an extreme form of pooling that is not physiologically or anatomically plausible. It requires that all of the sensory neurons converge onto a single decision neuron. However, one can imagine a pooling function that computes a weighted average of responses over a limited spatial extent so that only cells with similar receptive field locations are combined. This agrees well with how the visual cortex is wired and the fact that receptive fields get larger as one traverses the cortical hierarchy from primary visual cortex (V1) to V2, V3, V4, and IT.

To appreciate how attention affects the representation of information in visual cortex, we can use some of the aforementioned ideas to construct "neural" images of simple stimuli. Figure 6.9 shows simulations of a 2D array of model neurons. The input image consists of two vertically oriented Gabor patterns embedded in random noise. The green circle (Fig. 6.9, left) indicates the focus of attention, but was not present in the image used for the simulations. Each model neuron comprised a Gaussian spatial weighting function that represented the neuron's receptive field. Each receptive field was approximately 1/20th the size of the image in linear dimension. There were approximately 200 × 200 neurons whose RF centers were distributed to cover the entire image. The response of each neuron was computed by calculating the inner product of the weighting function and the part of the image within the receptive field. This number was used as the rate parameter for a Poisson spike generation function. Each pixel in Fig. 6.9 (middle and right) represents the resulting spike count for a single neuron. The middle panel of Fig. 6.9 illustrates a condition where attention increased the gain of the response at the attended location. The right panel shows a condition where the gain was constant across the image,

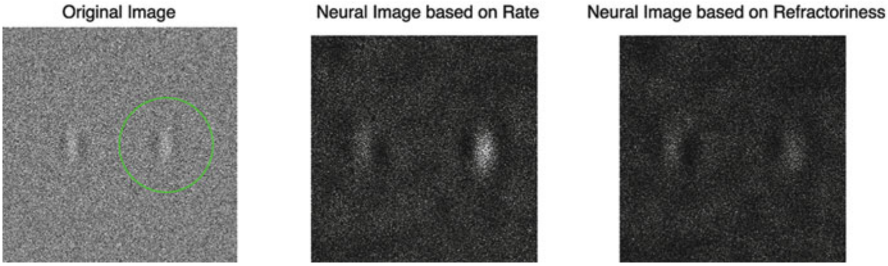

Original Image
Neural Image based on Rate
Neural Image based on Refractoriness

Fig. 6.9 Neural images created by computing the responses of a 2D array of model neurons. (*Left*) Original stimulus. The *green circle* indicates the focus of attention and was not present in the image used for model simulations. (*Middle*) A 2D array of model neurons. Attention increases the gain of the response in the attended region. Pixel intensity represents firing rate. (*Right*) A 2D array with a constant response gain across location, but increased refractoriness at the attended location

but the refractoriness of the cells was increased at the attended location. The result of increasing refractoriness is that there is less variability across cells that have the same input. The simulations suggest that increasing the gain has a pronounced effect on salience, whereas reducing variability through refractoriness has little effect. These neural images can be converted to detectability maps by running multiple trials with and without the stimulus and computing ROC functions for each neuron.

The simulations in this chapter have explored attentional gain control and reliability and how these affect detection and discrimination performance. Some features of the model that turned out to be important are (1) that attention enhances stimulus-driven responses but not baseline activity and (2) that responses are pooled over multiple neurons. Pooling of responses across neurons reduces variability and can have a pronounced effect on performance. One feature that was of only modest importance was trial-to-trial spike count variability; when variability is reduced by refractoriness there is little effect on detection or discrimination performance. Relatively, few empirical studies have investigated effects of attention on neural detection and discrimination thresholds [12, 13, 24, 25]. Fewer still have related changes in neural responses to behavioral thresholds [26]. This is an area that warrants further investigation and can profit from approaches that combine computational modeling and neurophysiological experimentation.

Acknowledgements This work was supported by NIH grant MH059244. Dr. Gaurav Patel assisted with the preparation of Fig. 6.1.

Literature Cited and Further Reading

1. DeAngelis, G. C., Cumming, B. G., & Newsome, W. T. (1998). Cortical area MT and the perception of stereoscopic depth. *Nature, 394*(6694), 677–680.
2. Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*(1), 1–47.
3. Van Essen, D. C., Lewis, J. W., Drury, H. A., Hadjikhani, N., Tootell, R. B., Bakircioglu, M., et al. (2001). Mapping visual cortex in monkeys and humans using surface-based atlases. *Vision Research, 41*(10–11), 1359–1378.
4. Fano, U. (1947). Ionization yield of radiations. II. The fluctuations of the number of ions. *Physical Review, 72*(1), 26.
5. Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge: MIT Press.
6. Bair, W., & Koch, C. (1996). Temporal precision of spike trains in extrastriate cortex of the behaving monkey. *Neural Computation, 8*(6), 1185–1202.
7. Huang, X., & Lisberger, S. G. (2013). Circuit mechanism revealed by spike-timing correlations in macaque area MT. *Journal of Neurophysiology, 109*(3), 851–866.
8. Maimon, G., & Assad, J. A. (2009). Beyond poisson: Increased spike-time regularity across primate parietal cortex. *Neuron, 62*(3), 426–440.
9. Osborne, L. C., Bialek, W., & Lisberger, S. G. (2004). Time course of information about motion direction in visual area MT of macaque monkeys. *The Journal of Neuroscience, 24*(13), 3210–3222.
10. Ferrera, V. P. (2015). Smooth pursuit preparation modulates neuronal responses in visual areas MT and MST. *Journal of Neurophysiology, 114*(1), 638–649.

11. Mitchell, J. F., Sundberg, K. A., & Reynolds, J. H. (2007). Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron, 55*(1), 131–141.
12. McAdams, C. J., & Maunsell, J. H. R. (1999). Effects of attention on the reliability of individual neurons in monkey visual cortex. *Neuron, 23*(4), 765–773.
13. Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron, 26*, 703–714.
14. Williford, T., & Maunsell, J. H. (2006). Effects of spatial attention on contrast response functions in macaque area V4. *Journal of Neurophysiology, 96*(1), 40–54.
15. Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science, 229*(4715), 782–784.
16. Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience, 19*(5), 1736–1753.
17. Roberts, M., Delicato, L. S., Herrero, J., Gieselmann, M. A., & Thiele, A. (2007). Attention alters spatial integration in macaque V1 in an eccentricity-dependent manner. *Nature Neuroscience, 10*(11), 1483–1491.
18. Womelsdorf, T., Anton-Erxleben, K., & Treue, S. (2008). Receptive field shift and shrinkage in macaque middle temporal area through attentional gain modulation. *Journal of Neuroscience, 28*(36), 8934–8944.
19. McAdams, C. J., & Maunsell, J. H. R. (1999). Effects of attention on orientation-tuning functions of single neurons in Macaque cortical area V4. *Journal of Neuroscience, 19*(1), 431–441.
20. Zenger, B., Braun, J., & Koch, C. (2000). Attentional effects on contrast detection in the presence of surround masks. *Vision Research, 40*, 3717–3724.
21. Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology, 77,* 24–42.
22. Churchland, M. M., Yu, B. M., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., Newsome, W. T., Clark, A. M., Hosseini, P., Scott, B. B., Bradley, D. C., Smith, M. A., Kohn, A., Movshon, J. A., Armstrong, K. M., Moore, T., Chang, S. W., Snyder, L. H., Lisberger, S. G., Priebe, N. J., Finn, I. M., Ferster, D., Ryu, S. I., Santhanam, G., Sahani, M., Shenoy, K. V. (2010). Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nature Neuroscience, 13*(3), 369–378.
23. Connor, C. E., Preddie, D. C., Gallant, J. L., & Van Essen, D. C. (1997). Spatial attention effects in macaque area V4. *Journal of Neuroscience, 17*(9), 3201–3214.
24. Boudreau, C. E., Williford, T. H., & Maunsell, J. H. (2006). Effects of task difficulty and target likelihood in area V4 of macaque monkeys. *Journal of Neurophysiology, 96*(5), 2377–2387.
25. Moore, T., & Chang, M. H. (2009). Presaccadic discrimination of receptive field stimuli by area V4 neurons. *Vision Research, 49*(10), 1227–1232.
26. Cook, E. P., & Maunsell, J. H. R. (2002). Attentional modulation of behavioral performance and neuronal responses in middle temporal and ventral intraparietal areas of Macaque Monkey. *Journal of Neuroscience, 22*(5), 1994–2004.
27. Bushnell, C., Goldberg, M. E., & Robinson, D. L. (1981). Behavioral enhancement of visual responses in monkey cerebral cortex. I. Modulation in posterior parietal cortex related to selective visual attention. *Journal of Neurophysiology, 46*, 755–772.
28. Carrasco, M., Ling, S., & Read, S. (2004). Attention alters appearance. *Nature Neuroscience, 7*, 308–313.
29. Connor, C. E., Gallant, J. L., Preddie, D. C., & Van Essen, D. C. (1996). Responses in area V4 depend on the spatial relationship between stimulus and attention. *Journal of Neurophysiology, 75*, 1306–1309.
30. Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate cortex. *Philosophical Transactions of the Royal Society of London, Series B, 353*, 1245–1255.
31. Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.

32. Fries, P., Reynolds, J. H., Rorle, A. E., & Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective attention. *Science, 291*, 1560–1563.
33. Martinez-Trujillo, J. C., & Treue, S. (2002). Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron, 35*, 365–370.
34. Martinez-Trujillo, J. C., & Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology, 14*, 744–751.
35. McAdams, C. J., & Maunsell, J. H. R. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology, 83*, 1751–1755.
36. Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of Neurophysiology, 70*(3), 909–919.
37. Motter, B. C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience, 14*(4), 2178–2189.
38. Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience, 27*, 611–647.
39. Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature, 395*, 376–381.
40. Seidemann, E., & Newsome, W. T. (1999). Effect of spatial attention on the responses of area MT. *Journal of Neurophysiology, 81*, 1783–1794.
41. Skottun, B. C., De Valois, R. L., Grosof, D. H., Movshon, J. A., Albrecht, D. G., & Bonds, A. B. (1991). Classifying simple and complex cells on the basis of response modulation. *Vision Research, 31*(7–8), 1079–1086.
42. Spitzer, H., Desimone, R., & Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance. *Science, 240*(4850), 338–340.
43. Treue, S., & Maunsell, J. H. R. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature, 382*(6591), 539–541.
44. Treue, S., & Martinez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature, 399*(6736), 575–579.
45. Treue, S., & Martinez Trujillo, J. C. (1999). Reshaping neuronal representations of visual scenes through attention. *Current Psychology of Cognition, 18*(5–6), 951–972.
46. Treue, S., & Maunsell, J. H. R. (1999). Effects of attention on the processing of motion in macaque visual cortical areas MT and MST. *Journal of Neuroscience, 19*(17), 7603–7616.
47. Treue, S. (2001). Neural correlates of attention in primate visual cortex. *TINS, 24*(5), 295–300.

Chapter 7
Modeling Attention in Engineering

Matei Mancas

7.1 Attention in Computer Science: Idea and Approaches

There are two main approaches to attention modeling in computer science. The first one is based on the notion of "saliency," while the second one is based on the idea of "visibility." The number of papers and the amount of work is dramatically different between these two approaches, and the models based on saliency are by far more spread than the visibility models in computer science.

The notion of "saliency" implies a competition between "bottom-up" or exogenous and "top-down" or endogenous information. The idea of bottom-up saliency maps is that the sight or gaze of people will direct to areas which, in some way, stand out from the background based on novel or rare features. This bottom-up saliency can be modulated by top-down information based on memory, emotions or, goals. The eye movements can be computed from the saliency map by using winner-take-all [10] or more dynamical algorithms [18, 25].

The second approach to attention modeling is based on the notion of "visibility" which assumes that people look to locations that will lead to successful task performance. Those models are dynamic and intend to maximize the information acquired by the eye (the visibility) of eccentric regions compared to the current eye fixation to solve a given task (which can also be simply free viewing). In this case, top-down information is naturally included in the notion of task along with the dynamic bottom-up information maximization. The eye movements are in this approach directly an output from the model and do not have to be inferred from

M. Mancas (✉)
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: matei.mancas@umons.ac.be

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_7

a “saliency map” which is considered as a surface giving the posterior probability (following each fixation) that the target is at each scene location [7].

7.2 Visibility Models

Compared to other Bayesian frameworks, like the one of [30], visibility models have one main difference. The saliency map is dynamic even for static images, as it will change depending on the eye fixations and not only the signal features: of course, given the resolution drop-off from the fixation point to the periphery, it is clear that some features are well identified in some eye fixation, while less or even not visible during other eye fixations. At the contrary of saliency models, visibility models make explicit the resolution variability of the retina: in that way, an attention map is “recomputed” at each new fixation, as the feature visibility changes at each of these fixations.

Najemnik and Geisler [28] found that an ideal observer based on a Bayesian framework can predict eye search patterns including the number of saccades needed to find a target, the amount of time needed, as well as the saccade spatial distribution.

Other authors like [19] proposed a visibility model capable to predict the eye fixations during the task of reading. Reninger used similar approaches for the task of shape recognition. Tatler [34] introduces a tendency of the eye gaze to stay in the middle of the scene to maximize the visibility over the image (which reminds the centered preference for natural images or centered Gaussian bias illustrated in Fig. 7.9).

The visibility models are much more used in the case of strong tasks, and few of them are applied to free viewing which is considered as a week task [7].

7.3 Saliency Approaches: Bottom-Up Methods

While visibility models are more used in cognitive sciences and with strong tasks, in computer science, bottom-up approaches use features extracted only once from the signal independently from the eye fixations, such as luminance, color, orientation, texture, object relative position, or even simply neighborhoods or patches from the signal. Once those features are extracted, all the existing methods are essentially based on the same principle: looking for contrasted, rare, surprising, novel, worthy to learn, less compressible, maximizing the information areas. All those definitions are actually synonyms, and they all amount to searching for some unusual features in a given context which can be spatial or temporal. In the following, we provide examples of contexts used for different kinds of signals.

7.3.1 Still Images

The literature is very active concerning still image saliency models. While some years ago only some labs in the world were working on this topic, nowadays hundreds of different models are available. Those models have various implementations and technical approaches even if initially they all derive from the same idea.

It is thus very hard to find a simple taxonomy which classifies all the methods. Some attempts of taxonomies proposed an opposition between "biologically driven" and "mathematically based" methods with a third class including "top-down information." This approach implies that only some methods can handle top-down information while all bottom-up methods could use top-down information more or less naturally. Another difficult point is to judge the biological plausibility which can be obvious for some methods but much less for the others. Another criterion is the computational time or the algorithm complexity, but it is very difficult to make this comparison as all the existing models do not provide cues about their complexity. Finally, a classification of methods based on center-surround contrast compared to information theory-based methods does not take into account different approaches as the spectral residual one, for example. Other taxonomies will also be introduced in the next chapters as, for example, the dependence on image features. Here, we show a taxonomy of the saliency methods which is based on the context that those methods take into account to exhibit signal novelty. In this framework, there are three classes of methods.

The first one is pixel's surroundings: here a pixel, a group of pixels, or a patch is compared with its surroundings at one or several scales.

A second class of methods will use as a context the entire image and compare pixels or patches of pixels with other pixels or patches from other locations in the image but not necessarily in the surroundings of the initial patch. Some models even use more than one image as a context: an entire dataset can be used here.

Finally, the third class will take into account a context which is based on a model of what the normality should be.

In the following sections, these three classes of models are illustrated.

7.3.1.1 Context: Pixel's Surroundings

This approach is initially based on a biological motivation. Its origins come from the work of [17] on attention modeling. The main idea is to compute visual features at several scales in parallel, to apply center-surround inhibition, combination into conspicuity maps (one per feature), and finally to fuse them into a single saliency map. There are a lot of models derived from this approach which mainly use local center-surround contrast as a local measure of novelty. A good example of this family of approaches is the Itti's model (Fig. 7.1) [10] which is the first implementation of the Koch and Ullman model. It is composed of three main steps. First, three types of static visual features are selected (colors, intensity, and orientations) at several scales. The second step is the center-surround inhibition

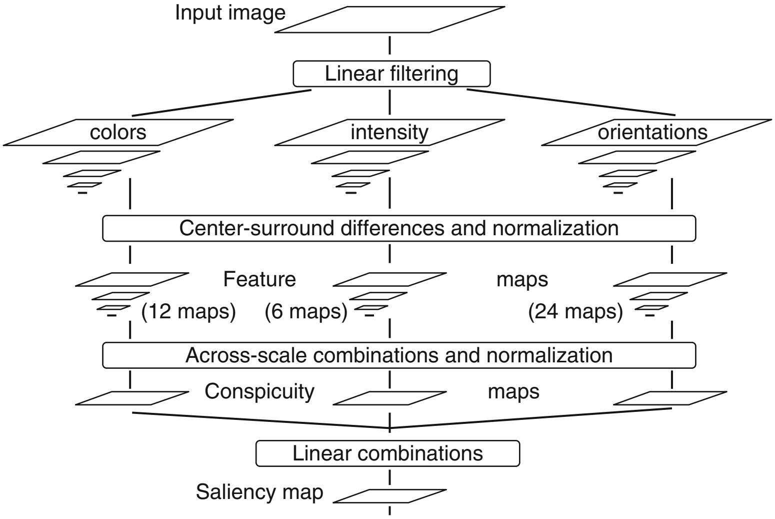


Fig. 7.1 Model of [10]. Three stages: center-surround differences, conspicuity maps, inter-feature fusion into saliency map (Adapted from [10])

which will provide high response in case of high contrast, while it will have low response in case of low contrast. This step results in a set of feature maps for each scale. The third step consists in an across-scale combination, followed by normalization to form "conspicuity" maps which are single multiscale contrast maps for each feature. Finally, a linear combination is made to achieve inter-feature fusion. Itti proposed several combination strategies: a simple and efficient one is to provide higher weights to conspicuity maps which have global peaks much bigger than their mean. This is an interesting step which integrates global information in addition to the local multiscale contrast information.

This implementation proved to be the first successful approach of attention computation by providing better predictions of the human gaze than chance or simple descriptors like entropy. Following this success, most of the computational models of bottom-up attention use the comparison of a central patch to its surroundings as a novelty indicator.

7.3.1.2 Context: The Whole Image or a Dataset of Images

In this approach, the context which is used to provide a degree of novelty or rarity to image patches is not necessarily the surroundings of the patch but can be other patches in its neighborhood or even anywhere in the image or an image database. The idea can be divided in two steps. First, local features are computed in parallel from a given image. The second step measures the likeness of a pixel or a neighborhood of pixels to other pixels or neighborhoods within the image.

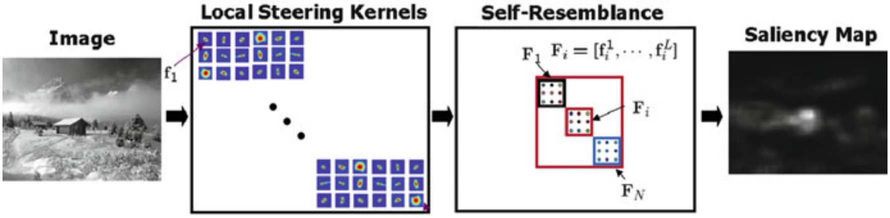


Fig. 7.2 Model of [32]. Patches at different locations in the image are compared (Adapted from [32])

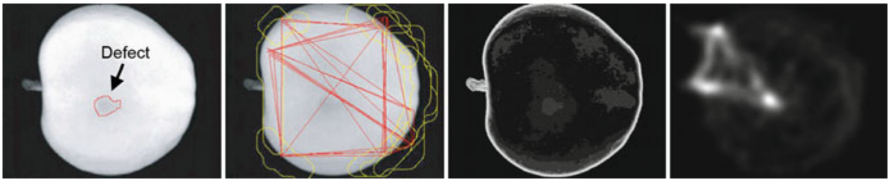


Fig. 7.3 Difference between locally contrasted and globally rare features. *Left image*: an apple with a defect in *red*. *Second image*: [10]. *Third image*: [24]. *Right image*: mouse tracking (ground truth)

This kind of visual saliency is called "self-resemblance." A good example is shown in Fig. 7.2. The model has two steps. First, it proposes to use local regression kernels as features. Second, it proposes to use a nonparametric kernel density estimation for such features, which results in a saliency map consisting of local "self-resemblance" measure, indicating likelihood of saliency [32].

Mancas [21] and Riche et al. [31] focus on the entire image. These models are designed to detect saliency in the areas which are globally rare and locally contrasted. After a feature extraction step, both local contrast and global rarity of pixels are taken into account to compute a saliency map. An example of the difference between locally contrasted and globally rare features is given in Fig. 7.3. On the left, there is the initial image of an apple with a defect in red; the second image shows the fixations predicted by [10] where the locally contrasted apple edges are well detected while its less contrasted but rare defect is not. The third image shows [24] which detected the apple edges, but also the defect. Finally, the rightmost is the mouse-tracking result for more than 30 users. Boiman and Irani [4] look for similar patches and relative positions of these patches in an image database which provide more cues about what should be normal. The use of a database might be viewed as an introduction of top-down information.

7.3.1.3 Context: A Model of Normality

This approach is probably less biologically motivated than most of the other implementations. The context which is used here is a model of what the image

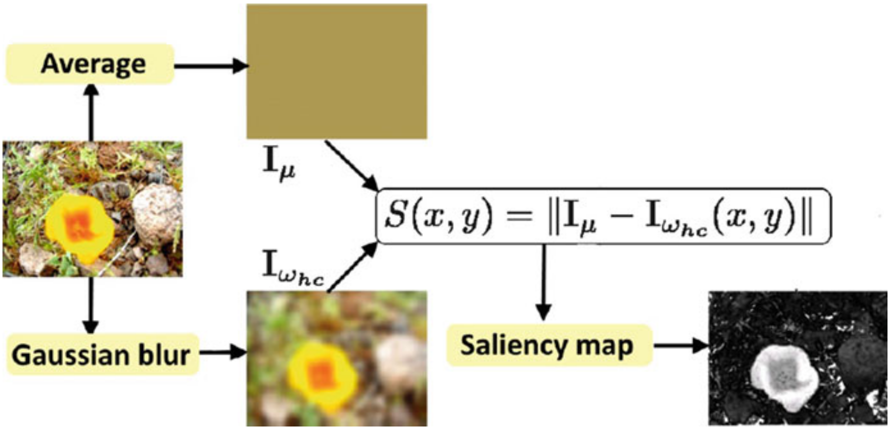

Average
$\mathbf{I}_{\mu}$
$\mathbf{I}_{\omega_{hc}}$
$S(x,y) = \|\mathbf{I}_{\mu} - \mathbf{I}_{\omega_{hc}}(x,y)\|$
Gaussian blur
Saliency map

Fig. 7.4 Achanta et al. [1] use a model of the mean image (Adapted from [1])

should be: if things are not like they should be, this can be surprising, thus attracting people's attention. Achanta et al. [1] proposed a very simple attention model (Fig. 7.4): first, the color space is converted from RGB to Lab; second, the Euclidean distance is computed between a Gaussian filtered version of the input image and the average Lab vector of the input image. The mean image used is a kind of model of the image statistics: pixels which are far from those statistics are more salient. This model is mainly useful in salient object detection.

Another approach to "normality" can be found in [8], where the authors proposed a spectral model that is independent of any features. As it is known that natural images have a $\frac{1}{f}$ decreasing Fourier log-spectrum, the difference between the log-spectrum of the image and its smoothed log-spectrum (spectral residual) is reconstructed into a saliency map. Indeed, a smoothed version of the log-spectrum is closer to an $\frac{1}{f}$ decreasing log-spectrum as small variations are removed. This approach is almost as simple as [1] but much more efficient in predicting eye fixations.

More details on still image saliency modeling can be found in the Chaps. 8 and 9.

7.3.2 *Videos*

Part of the static models have been extended to video. As shown in Fig. 7.5, it is the case of [32] where the time dimension is introduced by replacing square spatial patches by 3D spatiotemporal cube patches where the third dimension is the time. Also, Itti's model was generalized with the addition of motion features and flickering to the initial spatial set of features containing luminance, color, and orientations.

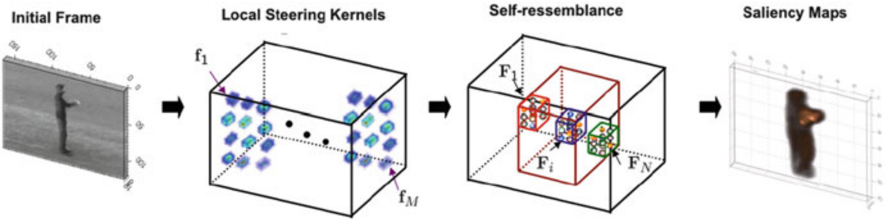


Fig. 7.5 Seo and Milanfar [32] generalized to video by introducing the spatiotemporal cubes (Adapted from [32])

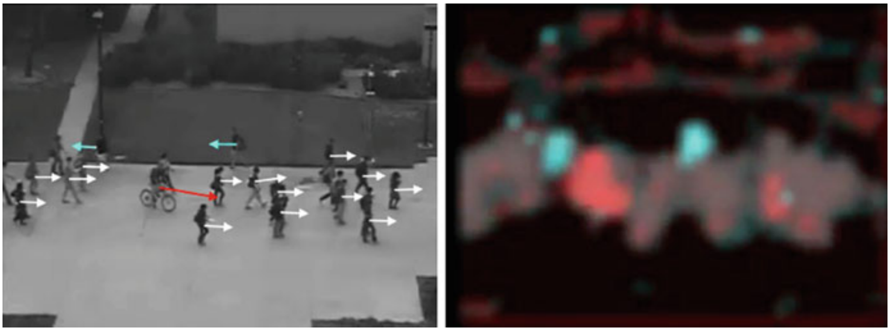


Fig. 7.6 Detection of salient motion compared to the rest of motion. *Red* motion is salient because of unexpected speed. *Cyan* motion is salient because of unexpected direction [26]

Those models mainly show that important motion is well detected. Other models like [26] have developed a bottom-up saliency map to detect abnormal motion. The proposed method is based on a multiscale approach using features extracted from optical flow and global rarity quantification to compute bottom-up saliency maps. The model exhibits promising results from a few moving objects to dense crowds with increasing performance (Fig. 7.6). The idea here is to show that motion is most of the time salient, but within motion, some motion areas are more interesting than others.

More details on video saliency modeling can be found in Chap. 10.

7.3.3 Extension to 3D

3D saliency modeling is an emerging area of research which was boosted by two main evolutions.

First is the arrival of affordable RGB-D cameras which provide both classical RGB images and a depth map describing pixel distance from the camera. In terms of computational attention, this depth information is very important. For example, in all models released up to now, movement perpendicular to the plane of the camera could not be taken into account, while now it is directly available in the depth map. Those cameras (e.g., MS Kinect) provide a whole set of new features to the community through the depth map but also through the available point cloud and its 3D geometric features (surface normals, curvature, compactness, convexity, etc.).

The second event is the arrival of 3D printers which democratized the 3D models used to print objects. 3D models are more easily available, and libraries like PCL [2] can handle 3D point clouds, convert formats, and compute features from those point clouds.

Most of the 3D saliency models are extensions of still image models. Some use the 3D meshes based on Itti's approach, others just add the depth as an additional feature, while recent models are based on the use of point clouds. More details can be found in the Chap. 17.

As 3D saliency models are mainly extensions of 2D models, depending on the extended model, the different features can be taken into account locally and/or globally on the 3D objects.

7.3.4 Audio Signals

There are very few auditory attention models compared to visual attention models. One of the main issues is that it is not easy to find easy ground truth in the audio domain (contrary to eye tracking for visual attention). Also, the audio modality taken alone is much less informative on the scene than the visual modality. However, we can classify existing models into different categories.

The first one represents the local context for audio signals. As shown in Fig. 7.7, Kayser et al. [14] compute auditory saliency maps based on Itti's visual model (1998). First, the sound wave is converted to a time-frequency representation

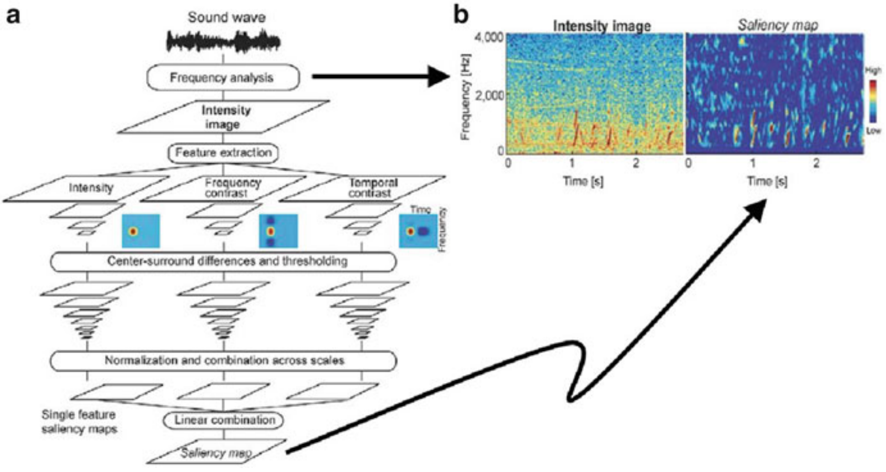


Fig. 7.7 Kayser et al. [14] audio saliency model inspired from Itti (Adapted from [14])

("intensity image"). Then three auditory features are extracted on different scales and in parallel (intensity, frequency contrast, and temporal contrast). For each feature, the maps obtained at different scales are compared using a center-surround mechanism and normalized. The center-surround maps are fused across scales achieving saliency maps for individual features. Finally, a linear combination builds the saliency map which is then reduced to one dimension to be able to fit on the one-dimensional audio signal.

Another approach to compute auditory saliency map is based on following the well-established approach of Bayesian Surprise in computer vision [9]. An auditory surprise is introduced to detect acoustically salient events. First, a short-time Fourier transform (STFT) is used to calculate the spectrogram. The surprise is computed in the Bayesian framework. This surprise approach represents the "normality" context for audio signals.

In the case of audio signal, there is no real "global" context as the time dimension has no real boundaries as the spatial dimensions have. A global context will be a long period of time in the past.

Mancas et al. [22] directly use as features the amplitude of the STFT and quantify their rarity compared to a long audio history. The model detects sudden and unexpected changes of audio textures and focuses the attention of a surveillance operator to sound segments of interest in audio streams that are monitored.

7.3.5 Mixing Video and Audio Signals

The superior colliculus (SC) is the brain structure which directly communicates with the eye motor command in charge of eye orientation. Its originality is to integrate information coming from different sensory areas but mainly visual and auditory. The information within the SC (Fig. 7.8) has a retinotopic representation. Visual information is displayed on the superficial layers and the auditory information on the deeper layers [15]. Once in the same coordinate system, multisensory information will be fused in order to take a decision on the eye movement. The main task of the SC is thus to direct the eyes onto the "important" areas of the surrounding space in terms of both vision and sounds and mix those two modalities.

Some attempts in mixing visual (still, video, and 3D) to audio signals saliency showed that the result is much more complex than expected. The final result is NOT the simple addition of visual and audio saliency taken together and it also depends on the scene (natural, social, action, etc.) [5, 16].

Basically, the visual modality seems to take the lead of attention unless the audio event is congruent spatially AND temporally with an image object/action. In this case, the audio has a great impact on the global attention. Given the retinotopic representation in the superior colliculus, a correspondence between the audio and visual location in the same time range is necessary for the fusion to be effective. This task should also be easier in the future as arrays of microphones which also

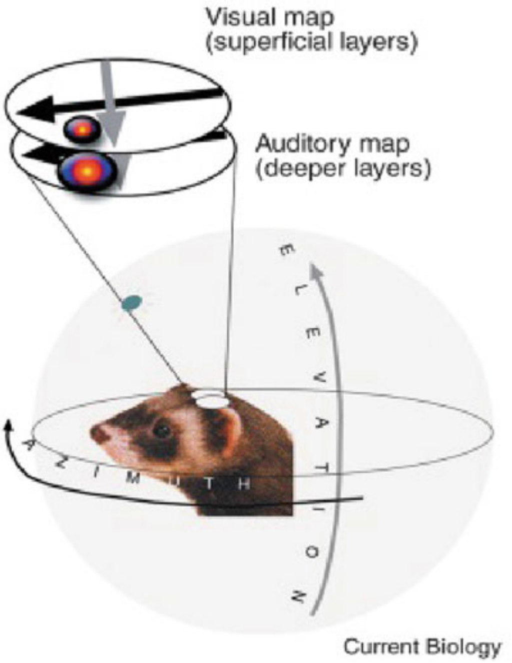


Fig. 7.8 Data fusion within the superior colliculus (Adapted from [15])

provide the direction of a sound are available together with the RGB and depth map on low-cost sensors as the MS Kinect.

More details on mixing audio and visual saliency can be found in the Chap. 16.

7.4 Saliency Models: Including Top-Down Information

Top-down is endogenous information and comes from the inner world (information from memory, their related emotional level, and also the task-related information). The separation between bottom-up and top-down information is far from being clear. Depending on the viewpoint and the definitions, some notions can be considered as bottom-up or top-down.

One can say that top-down is not involved if the memory/learning is not involved. In this case, all the hard-wired features which might be low level (luminance, color, orientation, motion direction), mid-level (object basic properties as the size, centered Gaussian as a default context), or high level (face detection, people detection) which involve specific brain areas but do not need memory and learning are bottom-up. An interesting point is that if bottom-up attention might be considered as common to a given species attention embodiment (e.g., humans) as it is hard-wired, it is not fully the case. Indeed, as the cognitive capabilities may vary, bottom-up information might also vary within the population. A color-blind person, for example, will have a different perception even without using any learning or memory, so a different bottom-up filter on the acquired data.

Top-down involves learning and memory and will deal with specific contexts (e.g., websites, adds, etc.), object recognition (face recognition, people recognition, specific animal or object), or a given task coming from inner needs (looking for the keys, etc.). Top-down information is a specialization of attention which implies important differences in attention focus between members of a given species (e.g., humans) depending on personal life experiences, mood, etc.

It is thus interesting that face detection can be considered as bottom-up (face feature detection does not necessary need memory and might be located in a specific brain area, the fusiform gyrus [27]) while face recognition is clearly top-down as it directly uses memory to remember a specific person.

In practice, three main families of top-down information can be added to bottom-up attention models.

The first one mainly deals with learned normality in a given context which can come from the experience from the current signal if it is time varying, or from previous experience (tests, databases) for still images.

The second approach is about task modeling which can either use object recognition-related techniques or which can model the usual location of those objects of interest.

The third one uses learning to extract both bottom-up and top-down information from eye-tracking results on a dataset of images.

7.4.1 Top-Down as Learned Normality

Concerning still images, the "normal" gaze behavior can be learned from the "mean observer." Eye-tracking techniques can be used on several users, and the average of their gaze on a set of natural images can be computed. This was achieved by several authors as it can be seen on Fig. 7.9. Bruce and Judd et al. [13] used eye trackers, while [20] used mouse-tracking techniques to compute this mean observer. In all cases, it seems clear that, for natural images, the eye gaze is attracted by the center of the images. This information is not top-down as it is generic enough not to be learned.

This centered distribution seems logical as natural images are taken using cameras and the photographer will naturally tend to locate the objects of interest

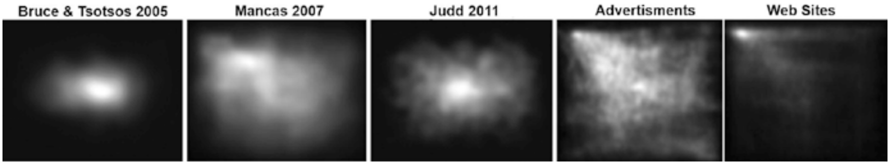


Fig. 7.9 Three models of the mean observer for natural images on the *left*. *The two right images*: model of the mean observer on a set of advertising and website images

in the center of the picture. Another point is that the objects in the center of the visual field are the ones one might interact with; they are then more important than the others.

This observation for natural images is very different from more specific images which use a priori knowledge and which are top-down. In [6], the author shows that the centered distribution mainly follows an horizontal axis for landscapes while it follows both horizontal and vertical directions for images of interiors. Mancas [21] showed using mouse tracking that gaze density is very different on a set of advertisements and on a set of websites as displayed on Fig. 7.9 on the two right images. This is partly due to a priori knowledge that people have about those images. For example, when viewing a website, the upper part has high chances to contain the logo and title, while the left part should contain the menu. During images or video viewing, the default template is the one of natural images with a high weight on the center of the image. If supplemental knowledge is known about the image, the top-down information will modify the mean behavior toward the optimized gaze density. Those top-down maps can highly influence the bottom-up saliency map, but this influence is variable. In [21], it appears that top-down information seems more important in the case of websites than advertisements and natural images. Other kinds of models can be learned from videos, especially if the camera is still. It is possible to accumulate motion patterns for each extracted feature which provides a model of normality. As an example, after a given period of observation, one can say: here moving objects are generally fast (first feature: speed) and going from left to right (second feature: direction). If an object, at the same location, is slow and/or going from right to left, this is surprising given what was previously learned from the scene; thus attention will be directed to this object. This kind of considerations can be found in [23]. It is possible to go further and to have different cyclic models in time. In a metro station, for example, normal people behavior when a train arrives in the station is different from the one during the waiting period in terms of people direction, speed, density, etc. In the literature (mainly in video surveillance), the variations in time of the normality models are learned through HMMs (hidden Markov models) [11].

For 3D signals, another information is the proximity of objects. For natural images, centered objects also attract our attention because they might be the ones we will interact with as they are in the center of the visual filed. In the same way, a close object is more likely to attract attention as it is more likely to be the first that we will have to interact with. In real world, the default context is a mix between a centered Gaussian and proximity value: centered close objects are the most important while far objects on the sides the less.

7.4.2 Top-Down as a Task

While the previous section dealt with attention attracted by events which lead to situations which are not consistent with the knowledge acquired about the scene, here we focus on a second main top-down cue which is a visual task ("Find the keys!”). This task will also have a huge influence on the way the image is attended, and it will imply object recognition (“recognize the keys”) and object usual location (“they could be on the floor, but never on the ceiling”).

7.4.2.1 Object Recognition

Object recognition can be achieved through classical methods or using points of interest (like SIFT, SURF, etc., [3]) which are somehow related to saliency. Some authors integrated the notion of object recognition into the architecture of their model like [29]. They extract the same features as for the bottom-up model, from the object, and learn them. This learning step will provide weight modification for the fusion of the conspicuity maps which will lead to the detection of the areas which contain the same feature combination as the learned object. More about object recongition and slaiency can be found in Chap. 19.

7.4.2.2 Object Location

Another approach is in providing with a higher weight the areas from the image which have a higher probability to contain the searched object. Several authors as [30] developed methods to learn objects’ location. Vectors of features are extracted from the images and their dimension is reduced by using PCA (principal component analysis). Those vectors are then compared to the ones from a database of images containing the given object. Figure 7.10 shows the potential people location that has been extracted from the image. This information, combined with bottom-up saliency, leads to the selection of a person sitting down on the left part of the image.

7.4.3 Task, Context, and Learning

Recently, learning the salient features becomes more and more popular: the idea here is not to find the rare regions, but to find an optimal description of those rare

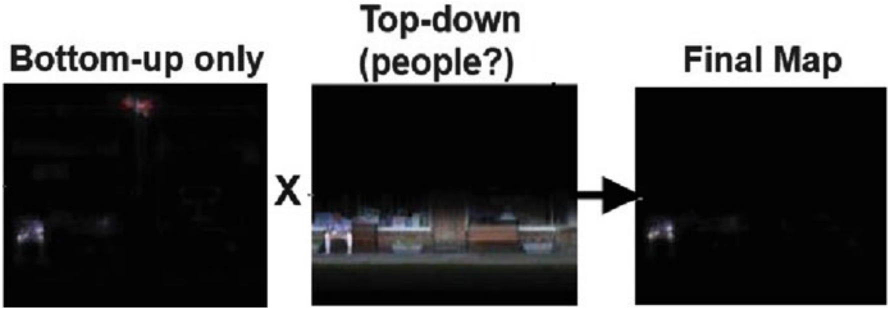


Fig. 7.10 Bottom-up saliency model inhibited by top-down information to select only salient people (Adapted from [30])

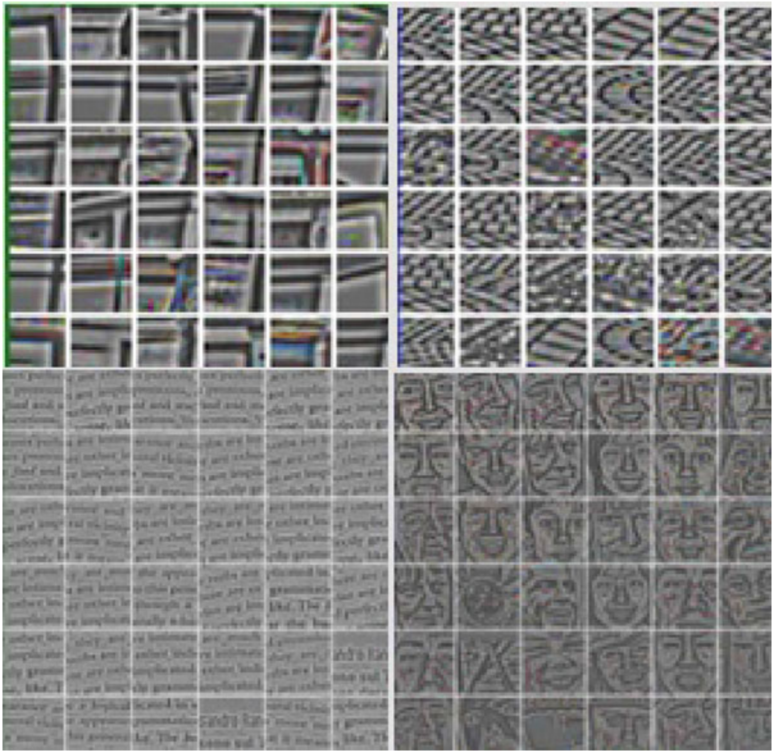


Fig. 7.11 Deep learning of salient features: at the second layer (mid-level features, *top row*) and at the third layer (high-level features, *bottom row*) (Adapted from [33])

regions which are already known from eye tracking or mouse tracking ground truth. The learning is based on deep neural networks, sparse coding and pooling based on large images datasets where the regions of interest are known. The most attended regions based on eye-tracking results are used to train classifiers which will extract the main features of these areas.

The use of deep neural networks greatly improved those techniques which are now able to extract meaningful middle- and high-level features which can describe the best salient regions [33]. Figure 7.11 shows examples of interesting feature extraction in the context of the training set which was here the MIT dataset [12]. This dataset contains general purpose images and free viewing; thus, specific top-down information is not included. The top row of the figure shows the features after the second layer. One can see mid-level features like corners or textures which naturally pop out from learning. More interestingly higher-level features such as text-like texture, faces, circular objects, and man-made structures are learned in the third layer. Those features might be considered top-down even if generic face detection, for example, can also be considered as bottom-up. These features are then mixed with weights which are again learned from the ground truth into saliency maps.

An interesting thing with this kind of approach is that it can be tailored to datasets where specific contexts (like outdoor pictures) or specific tasks (looking for wild animals) are taken into account. In that case, the initial feature learning phase could exhibit features which are more related to this context and task and which integrate both bottom-up and top-down information. However, a drawback of these methods is that they are too much taylored to the training dataset. Analyzing advertising images using a model trained to natural images can provide bad results. The extensive use of deep learning can lead to a loss of genericity of the saliency model. The future in this direction is probably in a mix of deep learning and more classical pipelines.

7.5 Modeling Attention in Computer Science

In computer science, there are two families of models: some are based on feature visibility and others on the concept of saliency maps, the latter approach being the most prolific.

For saliency-based bottom-up attention, the idea is the same for all the models: find areas in the image which are the most surprising in a given context. Three main types of contexts can be found: a local one mainly focusing on contrast, a global one quantifying the feature rarity, and a normality based which uses normal forms in image or Fourier space.

Saliency models can be also applied to video, audio, and even 3D signals. When mixing audio and visual signals, the influence of the audio seems to be taken into account only if it is congruent with a visual event.

Finally, a set of top-down features which can influence the saliency-based models are reviewed. While some of them are in fact bottom-up (centered Gaussian, face detection, etc.), others are real top-down features (context related, object and face recognition, object location).

In the next chapters, the saliency-based models will be described for still images, for videos, but also for 3D and multimedia models. A strong validation of still and video models is also done to see how effective the models are.

References

1. Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami. http://www.cvpr2009.org/
2. Aldoma, A., Marton, Z.-C., Tombari, F., Wohlkinger, W., Potthast, C., Zeisl, B., Rusu, R. B., Gedikli, S., & Vincze, M. (2012). Point cloud library. *IEEE Robotics & Automation Magazine*. ISSN 1070-9932/12.
3. Bay, H., Ess, A., Tuytelaars, T., & Gool, L. V. (2008). Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU), 110*(3), 346–359.

4. Boiman, O., & Irani, M. (2007). Detecting irregularities in images and in video. *International Journal of Computer Vision, 74*(1), 17–31.
5. Coutrot, A., & Guyader, N. (2014) . How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of Vision, 14*(8), 5.
6. Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision, 8*(2), 6.
7. Geisler, W. S., & Cormack, L. (2011). Chapter 24: Models of overt attention. In *The Oxford handbook of eye movements*. Oxford/New York: Oxford University Press.
8. Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR '07*, Minneapolis (pp. 1–8).
9. Itti, L., & Baldi, P. F. (2006). Modeling what attracts human gaze over dynamic natural scenes. In L. Harris & M. Jenkin (Eds.), *Computational vision in neural and machine systems*. Cambridge, MA: Cambridge University Press.
10. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.
11. Jouneau, E., & Carincotte, C. (2011). Particle-based tracking model for automatic anomaly detection. In *IEEE International Conference on Image Processing (ICIP)*, Brussels.
12. Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, Kyoto (pp. 2106–2113). IEEE.
13. Judd, T., Ehinger, K., Durand & Torralba, A. (2009). Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, Kyoto (pp. 2376–2383).
14. Kayser, C., Petkov, C., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology, 15*, 1943–1947.
15. King, A. J. (2004). The superior colliculus. *Current Biology, 14*(9), R335–R338.
16. King, A. J. (2009). Visual influences on auditory spatial learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1515), 331–339.
17. Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Hum Neurobiol, 4*(4), 219–227.
18. Le Meur, O., & Liu, Z. (2015). Saccadic model of eye movements for free-viewing condition. *Vision Research, 116*, 152–164.
19. Legge, H., Klitz, M., & Tjan (2002). Mr.chips 2002: New insights from an idealobserver model of reading. *Vision Research, 42*(18), 2219–2234.
20. Mancas, M. (2007). *Computational attention towards attentive computers*. Louvain la Neuve: Presses Universitaires de Louvain.
21. Mancas, M. (2009). Relative influence of bottom-up and top-down attention. In *Attention in cognitive systems* (Lecture notes in computer science, Vol. 5395). Berlin/Heidelberg: Springer.
22. Mancas, M., Couvreur, L., Gosselin, B., & Macq, B. et al. (2007) . Computational attention for event detection. In *Proceedings of the Fifth International Conference Computer Vision Systems*, Rio de Janeiro.
23. Mancas, M., & Gosselin, B. (2010). Dense crowd analysis through bottom-up and top-down attention. In *Proceedings of the Brain Inspired Cognitive Systems (BICS 2019).*
24. Mancas, M., Gosselin, B., & Macq, B. (2007). Perceptual image representation. *Journal on Image and Video Processing, 2007*, 3–3. http://dx.doi.org/10.1155/2007/98181.
25. Mancas, M., Pirri, F., & Pizzoli, M. (2011). From saliency to eye gaze: Embodied visual selection for a pan-tilt-based robotic head. In *Proceedings of the 7th International Symposium on Visual Computing (ISVC)*, Las Vegas.
26. Mancas, M., Riche, N., & J. Leroy, B. G. (2011). Abnormal motion selection in crowds using bottom-up saliency. In *IEEE ICIP*.
27. McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience, 9*(5), 605–610.

28. Najemnik, J., & Geisler, W. (2005). Optimal eye movement strategies in visual search. *Nature*, 387–391.
29. Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45*(2), 205–231.
30. Oliva, A., Torralba, A., Castelhano, M., & Henderson, J. (2003). Top-down control of visual attention in object detection. In *Proceedings of the International Conference on Image Processing, 2003 (ICIP 2003)* (Vol. 1, pp. I – 253–6 vol.1).
31. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013). Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication, 28*(6), 642–658.
32. Seo, H. J., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision, 9*(12). http://www.journalofvision.org/content/9/12/15.abstract.
33. Shen, C., & Zhao, Q. (2014). Learning to predict eye fixations for semantic contents using multi-layer sparse network. *Neurocomputing, 138*, 61–68.
34. Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*.

Chapter 8
Bottom-Up Visual Attention for Still Images: A Global View

Fred Stentiford

8.1 Introduction

Studies in neuroscience [1] are suggesting that human visual attention is enhanced through a process of competing interactions among neurons representing all of the stimuli present in the visual field. The competition results in the selection of a few areas of attention and the suppression of irrelevant material. It means that people and animals are able to spot outstanding patterns in a scene perhaps no part of which they have seen before and attention is drawn in general to the anomalous object in the scene. This makes visual attention a vital element in the survival of all creatures that have evolved since the Cambrian explosion [2] when vision first appeared on the Earth.

The ensemble of mechanisms grouped under the term attention swings into action before we are even conscious of anything strange. Indeed there is a *pre-attentive* period often less than 100 ms during which low-level processes rapidly identify image regions that deserve attention, and it has been the subject of considerable research. Treisman [3] describes experiments that reveal pre-attentive behaviour in human vision. She points out a ‘masking effect’ that depends upon the presence elsewhere of other elements sharing the local distinctive property. A locally salient feature can be suppressed by more distant structures in the image. Single distinctive features such as colour or orientation promote immediate saliency, but if these properties are cojoined, the search for a target is more difficult. Treisman describes several examples of images that exhibit the pop-out effect, some of which behave asymmetrically. For example, the time taken to find a circle crossed by an intersecting line is independent of the number of identical circles in the display,

F. Stentiford (✉)
University College London, Torrington Place, London, WC1E 7JE, UK
e-mail: f.stentiford@ucl.ac.uk

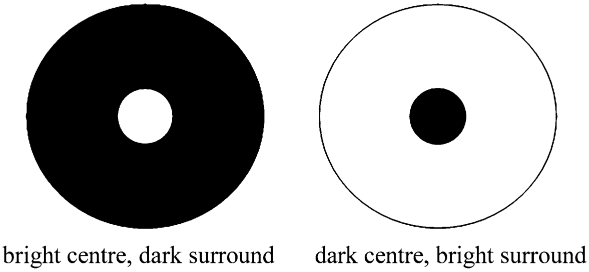
M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_8

whereas the time taken to find a circle among circles with lines increases linearly with the number of distractors.

It is known that a process of *centre-surround suppression* takes place in receptive fields in primate vision V1 [4] that detects local luminance contrast and enhances the edges that surround objects in the visual field. This mechanism is incorporated in most models of attention and is usually represented by inner and outer circular image regions that respond when contrast levels are significantly different (Fig. 8.1).

Computer models of visual attention aim to imitate aspects of the behaviour of the human visual system. The models identify image regions that attract our attentioneither directly by our gaze or covertly in our peripheral vision. Points in these regions are assigned saliency scores according to particular measures and the results displayed as *saliency maps* (Fig. 8.2). The appearance of saliency

Fig. 8.1 High contrast centre-surround image regions



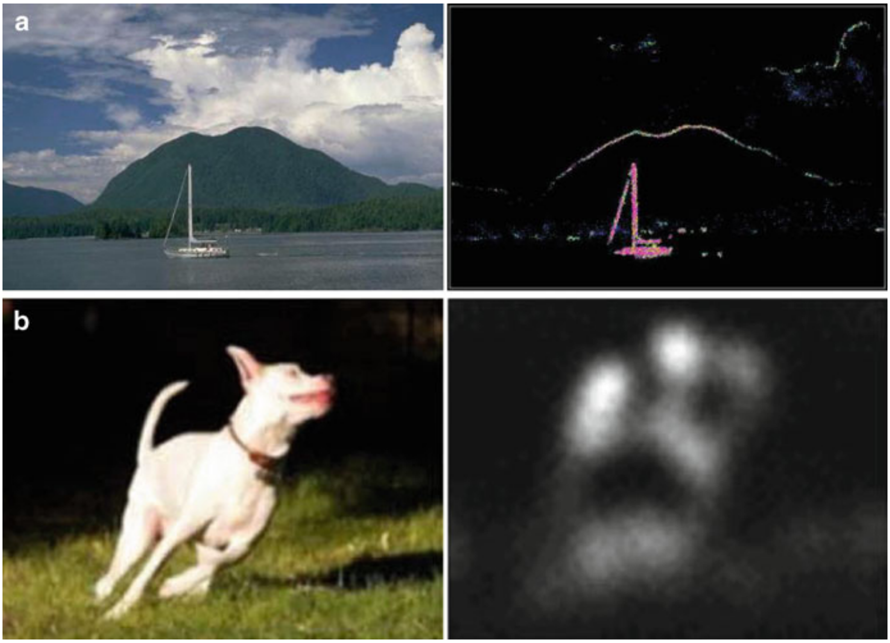


Fig. 8.2 Image and corresponding saliency maps (**a**) [5], (**b**) [6]

maps depends not only the balance between global and local measurements in the algorithms but also on the content of the images which may or may not suit the style of analysis. Many models are strictly bottom-up, that is, they rely totally on the information contained within the images in question. Others incorporate top-down methods that allow the statistics of related images to influence the parameters that determine local saliency values. In the extreme, top-down attention becomes recognition when attention is solely directed at a particular class of object and the features characterising those objects are used as a template in the calculation of the saliency measure. Saliency maps can predict eye fixations, by assigning a probability to points in an image where most people will look. However, human gaze behaviour is normally task driven, and therefore saliency maps are agnostic about the sequencing of eye movements and more complex modelling mechanisms are necessary [7].

The potential benefits that can arise from a model of attention are manifold and include applications to visual inspection in manufacturing processes, medical diagnosis, spotting security breaches, removing redundancy in data, various targeting applications and many others. The next section briefly outlines some common models of attention, highlighting some advantages and weaknesses.

8.2 Categorisation Schemes for Attention Models

Several authors have attempted to classify and categorise attention models against a set of criteria that separate the methods they employ into groups that highlight the selected features. Motion is certainly a powerful attentive factor that outweighs and obscures other features that characterise attention. It has the advantage that it can be detected easily by measuring local changes rather than any more complex processing. Top-down approaches may draw particular attention to specific sequences in time.

Object-based models dependent on Gestalt factors such as closure and symmetry can be contrasted with models that rely purely on spatial measurements. This viewpoint is related to a top-bottom perspective in which objects themselves become top-down targets for attention. Task definitions can shift human attention in a dramatic fashion, in the extreme not to even see the object of attention [8]. In a similar fashion, a task is a top-down influence which when modelled targets attention towards an object search or an interactive role.

Models that make use of concepts from physiology and neuroscience fall into a biologically inspired category. These models use mathematical frameworks that reflect current theories of the human visual system. Successful modelling of human behaviour could provide better understanding of the actual mechanisms involved. However, the various mathematical implementations of cognitive models may be categorised in other ways.

Both decision-theoretic models and Bayesian approaches carry out statistical analyses to detect regions of interest. The methods rely upon the use of features that distinguish foreground from background and can also incorporate top-down information. These approaches can be described as cognitive as aspects are reflected in biological cell systems.

It is plausible to assume that a salient object represents a concentration of information relative to the surrounding background. In this way, information theoretic models measure the rarity of features present in regions to determine saliency. Again statistics are used to spot unusual structure, but it is worth noting that the probabilities of rare features are derived from very few samples and therefore are statistically less reliable than those derived from features that are common.

Another attempt at characterising saliency transforms images into the frequency domain, the idea being that saliency is easier to detect in the new space. It is likely that relevant salient features dependent on spatial frequency distributions will be emphasised, but other potentially more salient features in the original image will be suppressed.

Some models employ graphical representations, but this aspect may be the only factor that is common to the different approaches because quite diverse functionality is assigned to nodes and edges. However, it can allow solutions to become large and complex, but does maintain a direct link with structure present in the image.

The categorisation used in this chapter identifies approaches that extract structure as part of the saliency measure as opposed to relying more on other features that are selected to characterise saliency. Of course the dividing line is not always clear as with any categorisation rule, but the use of spatial relationships and low-level features provides a useful way of distinguishing the various methods. It is apparent that the categorisation of models of attention is not clear cut because most methods fall into several categories and this may therefore confuse rather than inform. However, some categories will be appropriate and helpful in the context of specific applications.

8.3 Computational Models

Investigations into the physical operation of the visual cortex are very difficult not just because of the complexity but also the lack of suitable tools that can monitor in real time the potential interactions of multitudes of individual neurons. Current multi-electrode techniques can record simultaneously spikes from a few hundreds of neurons [9], and this number is sure to increase over the next few years. Nevertheless, we will not be able to predict and model the operation of neurons unless we know in detail how they operate in normal circumstances both individually and in concert. Functional magnetic resonance imaging (fMRI) scans are certainly providing scientists with valuable insights into brain function, but they are very blunt instruments when it comes to comprehending the precise firing sequences of neurons. This is presenting a barrier to our understanding of human vision which can be met in part by making use of computational models that reflect the outward behaviour of the visual system. It enables theories to be tested against behavioural data and perhaps provide justification for the design of new behavioural experiments. A wide ranging survey of the state of the art in visual attention modelling can be found in Borji et al. [10].

Bottom-up approaches that rely on preselected features characterising saliency are now contrasted with methods that place more importance upon structure detection. Balancing the assumptions associated with the selection of feature measurements against any unforeseen restrictions they place on potential future applications is a challenge to be met by this research.

8.3.1 A Priori Feature-Based Methods

Saliency is frequently modelled by combinations of values of image parameters such as intensity, colour, orientation, size and others. Particular local structures such as edges, curvature, corners, shape and location are also considered relevant measures of saliency.

Itti et al. [11] define a system which models visual search in primates (Fig. 8.3). Features based upon linear filters and centre-surround structures encoding intensity, orientation and colour are used to construct a saliency map that reflects areas of high attention (Fig. 8.4). Supervised learning is suggested as a strategy to bias the relative weights of the features in order to tune the system towards specific target detection

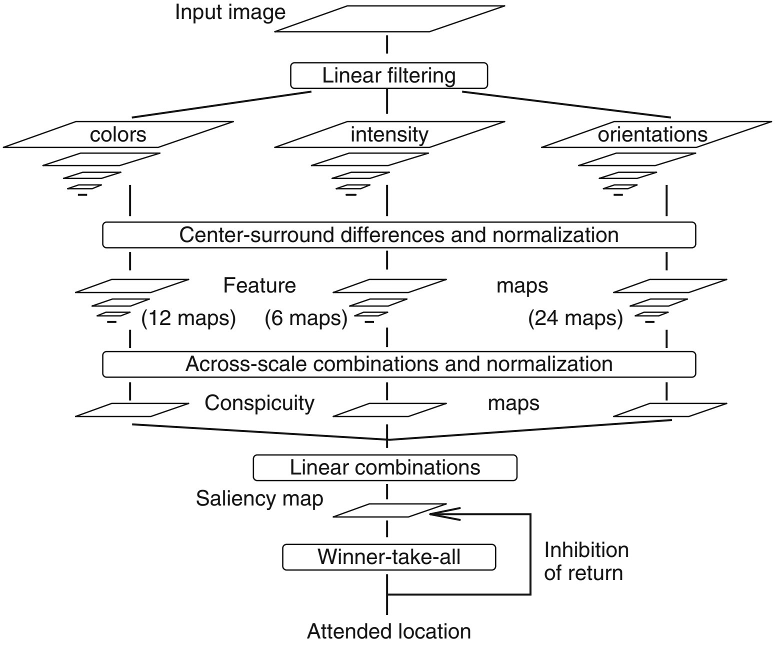


Fig. 8.3 Model architecture from Itti [11]

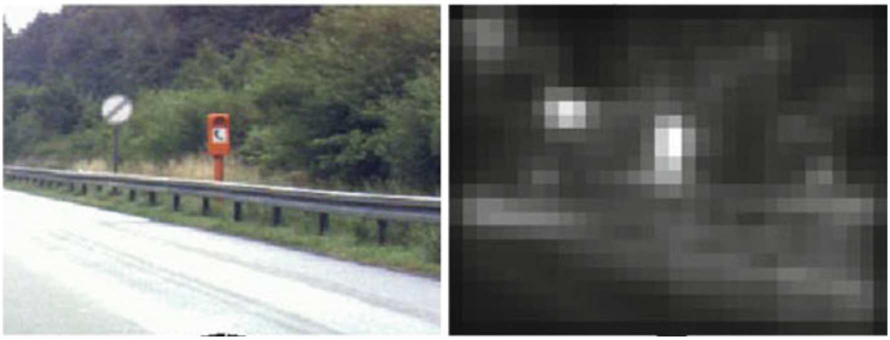


Fig. 8.4 Image and corresponding saliency map from Itti et al. [11]

tasks. However, the method of combining information from each of these filters is difficult and may not function well on certain categories of image. Itti's work has provided a basis for performance comparisons reported in many papers on visual attention. Han et al. [12] extended the Itti model by using a Markov random field. The computational visual attention mechanisms are integrated with region growing techniques. Gao et al. [13] use the feature decomposition of Itti et al., and saliency is determined from the discrimination obtained from the mutual information between centre and surround.

Tsotsos [14] presents a pyramidal processing attention model as a means of reducing the complexity arising from a large number of selected features. A winner-takes-all strategy is imposed on the processing layers in the pyramid so that the more salient objects are identified by location and features at the top of the pyramid. Provision is made to offset a boundary effect in the pyramidal structure that lays emphasis on central items even if they are less significant than peripheral items. Tsotsos highlights features such as size, luminance, edge contrast and orientation as possible features for defining saliency in static images, but there is little guidance on how these might be selected or combined.

Osberger et al. [15] identify perceptually important regions by first segmenting images into homogeneous regions and then scoring each area using five intuitively selected measures. These measures are grey-level contrast, size, shape, central image location and image border location. The approach is heavily dependent upon the success of the segmentation, and in spite of this, it is not clear that the method is able to identify important features in faces such as the eyes. Luo et al. [16] also devise a set of intuitive saliency features and weights and use them to segment images to depict regions of interest. Some higher-level priors are used such as skin colour, and selected images are used to normalise feature measurements.

The study by Le Meur et al. [17] lays emphasis on the considerable bias of observers towards looking at the central parts of images where perhaps the photographer usually places the subject. Le Meur et al. also take account of visual masking in their model as it is known that the differential sensitivity of the human visual system is dependent on the absolute values of parameters such as spatial frequency. Gopalakrishnan et al. [18] apply features based on colour and orientation to characterise salient regions, whereas Valenti et al. [19] employ features based on the edges of colour regions and their curvature.

Cheng et al. [20] use a distance metric in Lab colour space to measure contrast between regions and estimate saliency. The number of colours is minimised to reduce computation. The approach is used to segment salient objects. Achanta [21] again uses the Lab colour space but blurs the images with a Gaussian kernel and uses the difference with the original image to identify salient regions that are also easy to segment.

J. Zhang et al. [22] construct a set of Boolean feature maps using the Lab colour space. Connected regions touching the image border are ignored. This has the effect of introducing a centre bias and obtains good results on the MIT benchmark data [23].

L. Zhang et al. [24] gather statistics from a set of natural scenes to train sets of features used to estimate saliency. Saliency is indicated if features in a region are comparatively rare in the background. In a similar fashion, Bruce et al. [25] use 3600 natural images to prepare a set of basis functions and identify saliency using the likelihood of content within a region on the basis of the surround.

Ritchie et al. [26] derive six feature maps from the image, three maps from colour measurements and three from orientation measurements. Each feature map is subjected to Gaussian decomposition into four scales and a rarity map produced. The maps are then merged, and the method yields saliency maps that compare favourably with other approaches on the MIT data. The work highlights the effects that high-level recognition can have on fixation maps.

It might be argued that methods employing the intuitively inferred properties of images could be to some extent reflecting top-down information into the images being analysed. In this same sense, an attention mechanism that is driven by a specific task is also making use of top-down information and previously acquired experience. Several approaches fall into this category such as Liu et al. [27] who use an intuitively selected set of features including multiscale contrast, centre-surround measures and colour distribution to train a classifier to identify salient objects.

Oliva et al. [28] construct contextual features that guide attention towards specific targets such as people. However, detecting such irregularities as salient necessitates top-down knowledge of what characterises the images of people. Kavak et al. [29] utilises a learning-based saliency model and also employs both low-level features and high-level object-based features. A centre bias is introduced to improve performance.

Vig et al. [30] employ a training set of salient and non-salient gaze-labelled regions to construct an ensemble of convolutional network models. The method obtained good results on the MIT benchmark data. This work was developed by Kummerer et al. [31] who trained a high-dimensional feature space that had been previously optimised for object recognition and achieved an improved performance. Torralba [32] computes local salience using features derived from RGB, six orientations and four scales at each pixel. This is then modulated with contextual features trained on specific attentive objects such as people, paintings and mugs.

The operation is therefore significantly governed by features derived from top-down guidance. Judd [33] trains a 33-feature support vector machine to obtain a performance that approaches that of a human. He notes the importance of including object detectors among the features such as for faces and text, as these figure strongly in the fixation data. He also incorporates a centre prior, but it should be noted that during the collection of eye tracking data, all users were asked to fixate centrally before viewing images, and this requirement will have a powerful influence on the position of many of the subsequent fixations.

Garcia-Diaz [34] uses a hierarchically whitened feature space, where the norm of the vector displays the variability and serves as a saliency metric to measure how far a pixel feature vector deviates from the rest of the data. Some centre bias was removed from data by randomising the start point for eye tracking.

Many feature-based methods incorporate some form of top-down input either through supervised learning or a targeted choice of features. Whether features such as red or colour are higher or lower in the top-down scale is open to debate depending on the relevance in particular applications. However, the main attraction of this approach is the freedom to select features that are known to characterise saliency in general. The difficulty with preselecting features in this manner is that they cannot always anticipate the properties of a yet-to-be-seen attentive object.

8.3.2 Structural-Based Methods

All attention models make use of low-level measurements, but many seek commonality of structure within this data to isolate regions worthy of attention rather than only rely upon weighted combinations of feature measurements.

Erdem et al. [35] compare regions within a certain distance of each other using covariance matrices of image features to determine saliency. It is noted that this approach fails to identify saliency arising from the absence of local structure located elsewhere in the image. In a similar fashion, Seo et al. [36] measure local patch similarity using neighbouring feature matrices, but as with Erdem, the saliency of objects not possessing features present globally is not detected.

Fang et al. [37] divide the image into patches and identify saliency where a patch differs from those found elsewhere in the image while attaching a greater weight to patches that are closest. The similarity of patches is based on measurements of colour, intensity and orientation. In related work, Chen et al. [38] identify salient regions by detecting local groups of similar pixels but which form only a small percentage of the image.

Boiman et al. [39] consider higher-level structure and search for patch ensembles common to a database and the candidate image. Regions that cannot be composed from ensembles in the database are considered irregular. Patch configurations are compared according to their descriptors q^i and their relative positions with respect to an origin point C (Fig. 8.5).

Goferman [40] computes the saliency of a pixel by comparing the surrounding patch with others in the image. A patch is salient if it differs in colour as well as

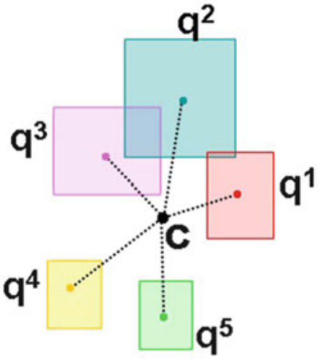


Fig. 8.5 Ensembles of patches [39]

being physically close to others in the image. The saliency is averaged over four scales and increased in value where pixels are close to others with high saliency to indicate ‘context’. In addition the saliency map is given a centre bias. The approach is particularly sensitive to edges. Borji [41] also employs local and global patch comparisons, but obtains improved performance by combining maps obtained from RGB and Lab colour channels. Each patch is represented by a set of basis functions optimised over a training set. Results are evaluated using the shuffled AUC formula which only uses the locations of fixation points and therefore ignores the effects of any centre bias in the data. Borji emphasises the potential benefits of top-down recognition on performance.

Harel et al. [42] propose a graphical model in which nodes correspond to image locations and the edges represent feature-based measures of dissimilarity between those nodes. Regions possessing high concentrations of dissimilarity are associated with saliency.

Stentiford [43] compares local groups of isolated pixels with others in the same relative position elsewhere in the image to detect unusual structure and hence a measure of saliency. The pixels are selected randomly, and the process does not involve the preselection of features or subsequent training.

Hou et al. [6] rely on frequency domain processing in which the difference between the original log spectrum and a prior averaged spectrum is transformed back into the spatial domain as the saliency map. Hou et al. [44] continued the spectrally based approach by first computing the sign of the discrete cosine transform of the image and constructing a saliency map by taking the inverse transform and smoothing the result.

Kadir et al. [45] measure the entropy of the local distribution of image intensity across a number of scales. High entropy indicates high local complexity and hence high saliency. This may not be the case where the salient structures possess lower entropy.

Lindenberg [46] provides a framework for detecting salient blob-like objects without relying on a priori information. He stresses that not all significant image structures are blobs. His research makes the assumption that structures that are significant in scale space will also be perceptually significant. Although this may

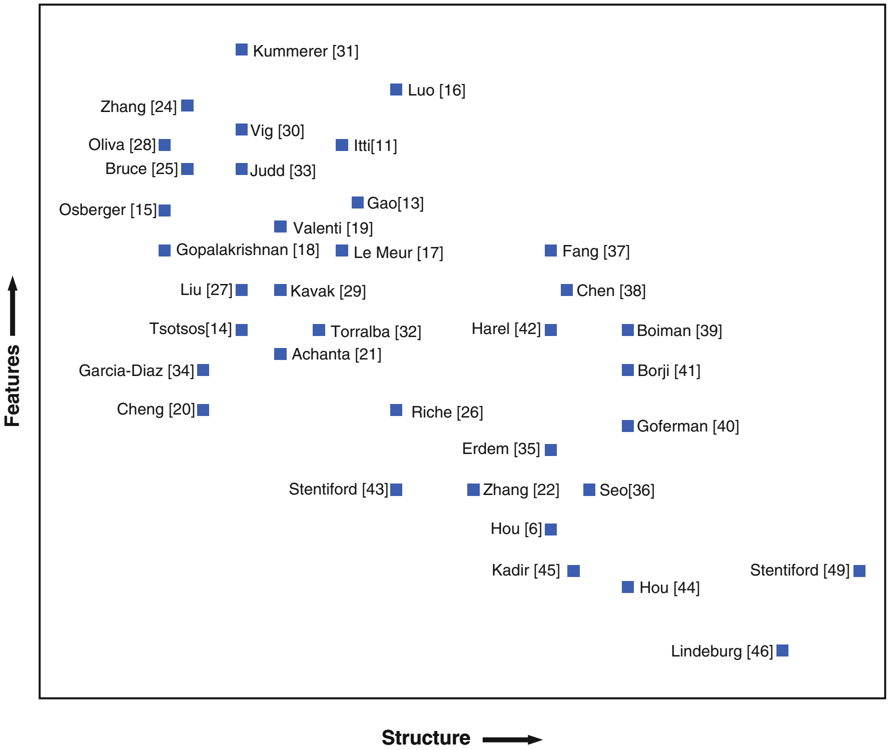


Fig. 8.6 Relative importance of features and structure in referenced papers

be true for some blob configurations, it does not apply to all, as, for example, no provision is made for the attention-suppressing effect of surrounding similar configurations as demonstrated by Treisman [3].

In summary, Fig. 8.6 displays how the various approaches to the analysis of saliency described in this chapter make relative use of features and structure. No method can avoid making use of some low-level measurements, and equally all approaches take some account of spatial aspects. However, there is a spread between these two extremes which may be interpreted as showing that authors have yet to agree on future research directions in this field.

8.4 A Closer Look at Models

8.4.1 Feature Based

The feature-based and structurally based models both have their advantages and disadvantages, and it is therefore worthwhile examining example models in more detail in order to highlight differences and any outstanding issues.

The Itti model [11] uses a biologically plausible architecture and is related to Treisman's feature integration theory [3]. Low-level feature measurements of colour channels, intensities and orientations are extracted from images over a range of scales. The scales (0–8) start with the original image, and dimensions are decreased by a power of 2, ending with a ratio of 1:256. Analogously to visual receptive fields, each measurement is fed through a centre-surround mechanism in which a combination is used of a finer scale in the centre (c) and a course scale in the surround (s). Six feature maps for each measurement are produced using scale combinations $(c,s) = \{(2,5),(2,6),(3,6),(3,7),(4,7),4,8)\}$.

An intensity image I is produced from the red (r), green (g) and blue (b) measurements where $I = (r + g + b)/3$ and Gaussian pyramids $I(\sigma)$ formed for each of the scales ($\sigma = 0, \ \ldots \ , 8$). Six feature maps are produced from the intensity values. The *rgb* channels are first normalised by intensity, and four other colour channels are defined as follows: $R = r - (g + b)/2$, $G = g - (r + b)/2$, $B = b - (r + g)/2$ and $Y = (r + g)/2 - |r - g|/2 - b$, where negative values are set to zero. Corresponding Gaussian pyramids $R(\sigma)$, $G(\sigma)$, $B(\sigma)$, $Y(\sigma)$ are then created. Centre-surround features are calculated by summing point differences between centre and surround interpolating to the finer scale where appropriate.

The colour opponency of red-green and blue-yellow in human vision is modelled using the same centre-surround mechanism where $R(\sigma){:}G(\sigma)$ and $B(\sigma){:}Y(\sigma)$ differences are used as values in the centre and surround regions. This yields six more feature maps from each of the two colour opponents. The orientation sensitivity of receptive fields is represented by orientation Gabor pyramids, each sensitive to the preferred angles 0, π/4, π/2, 3π/4. This produces a further 24 feature maps that contrast orientations between centre and surround.

The fusion of the 42 feature maps is difficult because saliency indicated in a few of the feature maps can be suppressed when they are all combined (Fig. 8.7). Equally, chance reinforcements between maps can produce spurious indications of

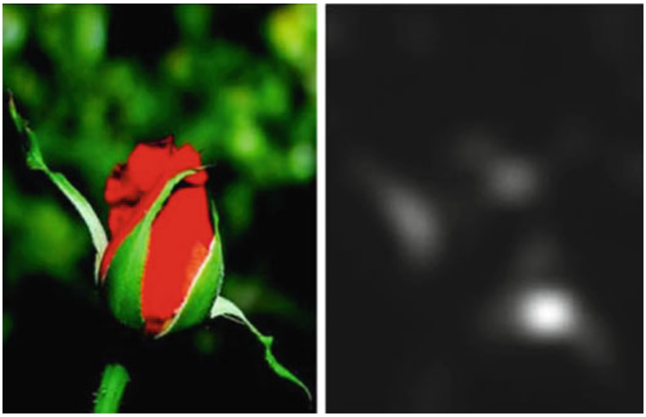


Fig. 8.7 Image and Itti saliency map taken from [37]

saliency. To overcome some of these problems, feature map values are normalised to a fixed range, and maps possessing a few strong maxima values are weighted against those whose local maxima do not differ significantly. The feature maps for each of intensity, colour and orientation are then summed across scales and combined to form three intermediate maps which are then averaged to form the final saliency map.

The principal limitation of a feature-based model is the prior selection of features that may not detect salient aspects such as in this case corners and other structural items. The complexity of features can be increased, and this is exemplified in the next section.

8.4.2 Structure Based

It is apparent that salience arises not just from colour, brightness and local orientation but also from structural features especially if they have not been seen before. As an example, the work of Boiman et al. [39] is aimed at detecting unusual irregularities that cannot be characterised beforehand and which therefore represent salience. Ensembles of patches (Fig. 8.5) taken from an observed image are compared with ensembles in a database to determine whether the observed structure has occurred before or whether it is strange or irregular.

Regions (typically 50 × 50 pixels) of an image are broken into 7 × 7 pixel patches taken from Gaussian pyramids at multiple scales. A spatial gradient is assigned to each pixel, producing a normalised 49-vector d^i for each patch i. The similarity of vectors (d^i, d^j) is given by

$$P\left(d^i, d^j\right) = \alpha \ \ \exp\left(-\left(d^i - d^j\right)^T S_D^{-1}\left(d^i - d^j\right)\right)$$

where α is a constant and S_D is a constant covariance matrix. Boiman defines the similarity of the positions l_y^i, l_x^j of patch pixels relative to origins c_y, c_x associated with ensembles y, x, respectively, as

$$P\left(l_y^i, c_y, l_x^j, c_x\right) = \beta \ \ \exp\left(-\left(\left(l_y^i - c_y\right) - \left(l_x^j - c_x\right)\right)^T S_L^{-1}\left(\left(l_y^i - c_y\right) - \left(l_x^j - c_x\right)\right)\right)$$

where β is a constant and S_L is a constant covariance matrix.

A correspondence md_y^i is set up between similar patches in the observed image y and those in the database x:

$$md_y^i = \max_{d_x^j} P\left(d_y^i, d_x^j\right) P\left(d_x^j, l_x^j\right)$$

where $P\left(d_x^j, l_x^j\right) = \left\{ \frac{1 \quad d_x^j, l_x^j \in x}{0 \quad \text{otherwise}} \right.$

High values of md_y^i identify the locations of patches in the database that are similar to the ith patch in y.

Patches that are not only similar but also possess the same relative position with respect to the ensemble origin are identified by mc_y^i

$$mc_y^i = \max_{l_x^j} P\left(l_y^i, c_y, l_x^j, c_x\right) md_y^i$$

This yields a set of candidate origins c_x^i in the database corresponding to each of the patches in y, and hence the best matching ensemble in x may be determined from

$$M = \max_{c_x = c_x^i} mc_y^i$$

A number of heuristics are introduced to reduce the size of an impossibly large search space. Patch ensembles are grown starting from a single patch, and additional patches are added if they match a database structure. Secondly, computation is reduced by considering coarse scales during the initial stages, but if matches are not found, different processing rules have to be followed using a finer scale. The method is applied to the problem of detecting suspicious behaviour, and top-down information from a previously collected database is used to determine saliency. The results are encouraging, but the examples reported only illustrate the algorithms. The work has been developed further and has been used to measure self-similarity [47] and therefore could be used to model bottom-up attention.

Although the approach is intuitively appealing, it is not clear how a database should be set up best to reflect a particular problem. A large number of examples of patch ensembles are proposed, but these may not capture all the ‘normal’ situations which may later be misidentified as irregular. In addition the specific features used to describe patches may not be able to represent certain structures in a way that enables them to be matched. Finally, without prior knowledge, the actual selection of patches to a large extent is random and therefore must introduce irrelevant information which has to be processed and could degrade performance if it is not subsequently ignored.

8.4.3 Background Identification

Saliency is difficult to characterise in general because surprise [48] cannot be predicted! Selecting features such as colour, brightness and orientation to measure saliency cannot offer a guarantee of success because the chosen features may not be appropriate for the salient region in question. Boiman and others recognise that other factors can affect visual attention that include structural relationships.

Human visual attention in a still image is governed by the relationship between the background and the salient object. Methods therefore are needed that separate foreground from background that do not employ *a priori* information or make any assumptions regarding properties of saliency. Background regions can be identified by recognising the self-similarity that they exhibit. This means that salient regions can be identified by the *absence* of self-similarity and avoids the need to preselect features to characterise saliency itself which by its very nature is unpredictable.

Stentiford [49] takes this approach in which pairs of pixels $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in region 1 of an image and pairs of pixels in region 2 $(\boldsymbol{x}_m, \boldsymbol{x}_n)$ of the image match if brightness, local gradient orientation and relative orientation lie within certain thresholds where

$$|u^{x_i} - u^{x_m}| < \delta, \quad |u^{x_j} - u^{x_n}| < \delta \text{ and } u^{x_k} \text{ is the brightness at pixel } \boldsymbol{x}_k \tag{8.1}$$

$$|\theta_i - \theta_m| < \varepsilon_1, \ \left| \ \theta_j - \theta_m \right| < \varepsilon_1 \tag{8.2}$$

and θ_k is the local gradient orientation of pixel $\boldsymbol{x}_k$

$$\text{and } \frac{\left(\boldsymbol{x}_j - \boldsymbol{x}_i\right) \bullet (\boldsymbol{x}_n - \boldsymbol{x}_m)}{\left|\boldsymbol{x}_j - \boldsymbol{x}_i\right| * |\boldsymbol{x}_n - \boldsymbol{x}_m|} \geq \lambda \tag{8.3}$$

The inner product in Eq. (8.3) constrains the difference in slopes between the pairs of points in each region to be less than a certain angle $\varepsilon_2 \geq \left|\varphi_{ij} - \varphi_{nm}\right|$ where $\lambda = \cos \ \varepsilon_2$. These conditions for match are scale invariant and partly orientation invariant $o(\varepsilon_2)$. The matching of the pairs of pixels $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ and $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ has greater reliability if the pair $\boldsymbol{x}_k$ and $\boldsymbol{x}_i$ also match as this shows that the properties of all three points match according to (8.1) and (8.2) and are in the same relative angular position according to (8.3) in both regions.

The three pixels are represented by nodes in a fully connected graph or *clique* (Fig. 8.8) with edges representing their angular relationship. The matching of the relative orientation of points in (8.3) reflects the structure present in both locations and is at the same time scale invariant. Greater reliability is obtained through the

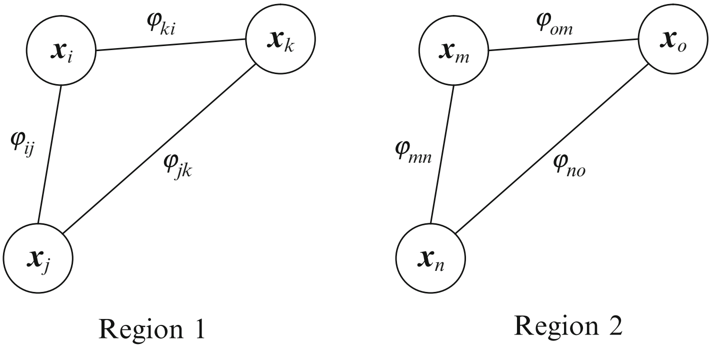


Fig. 8.8 Matching cliques of size 3

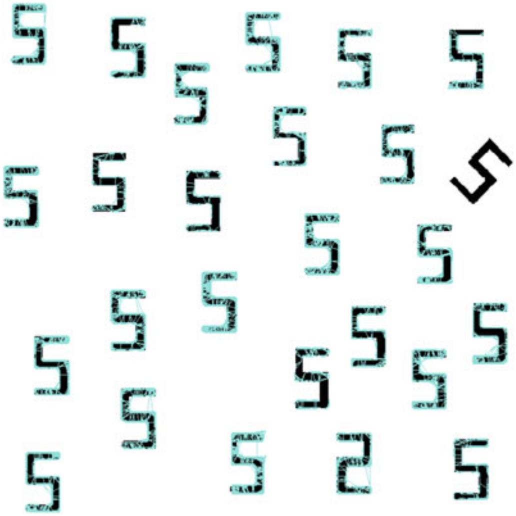


Fig. 8.9 Matching cliques fitting background 5s

associative power of larger *maximal cliques* that may be traded off against the precision of the thresholds δ, ε_1, ε_2 and thereby obtain more flexible matching. A maximal clique is obtained when it is not possible to add more matching nodes.

The similarity measure is used to analyse black and white images exhibiting pop-out effects. Gradient orientations θ_k are quantised into just four values (0°, 90°, 180°, 270°), and matching points possess the same value. The threshold for ε_2 is set at 20°. Intensities in (8.1) are not used with black and white images. In practice the check on the ε_2 relative orientation threshold is only applied to the four closest nodes because as distance increases, virtually all nodes satisfy the condition if the first four do. The images are compared with themselves, and an additional restriction is placed on matching pixel separations:

$$\left|\boldsymbol{x}_i - \boldsymbol{x}_j\right| < R \quad \text{and} \quad \left|\boldsymbol{x}_m - \boldsymbol{x}_n\right| < R \tag{8.4}$$

where R is chosen to limit the size of the regions being compared. Figure 8.9 shows maximal cliques matching pairs of identical background shapes but not the tilted shape. For clarity each clique is represented by coloured lines joining only the three nearest points in each clique. The ‘2’ does not pop out and is matched because the top and bottom sections of the ‘2’ match the bottom and top, respectively, of the background ‘5’s.

The identification of background in this approach is strongly structural and does not make use of any training stages save that of analysing the image itself. More generally, attention is also dependent on prior knowledge in the sense that a familiar pattern such as a face will pop out regardless of other structure in the background. It is worth noting that the approach to measuring saliency by identifying maximal matching cliques within a single image has been applied to the task of measuring the similarity of *different* images. In this case, it was sufficient to match gradient orientation and relative orientation to obtain face recognition [50].

8.5 Summary and Conclusions

This chapter has not attempted to produce a comprehensive survey of all research on computational attention because it is a very large and varied field. Indeed categorising the various approaches is difficult with many associated drawbacks in each case. Nevertheless, certain key methods for identifying saliency in images have been categorised above according to the extent that features and structure have been employed. There is a concern that preselected features thought to characterise saliency cannot always measure surprise and therefore any move away from this approach may be of interest in the future. A purely structural method has been used to illustrate how identifying background is sufficient to reveal saliency without characterising it beforehand, although this still may not take account of top-down influences.

This chapter has exposed several issues:

- Feature-based methods work well but not on images that are not reflected in the specific features used. This also applies to structural approaches as well.
- Top-down approaches are relevant if salient objects, such as text or faces, are present.
- Centre bias in eye tracking data is an important factor when assessing the performance of models of human attention on still images.

Understanding the nature of attention in human vision is fundamental to the future of computer vision whether it is based on features, structure or higher-level recognition. A framework that reflects visual behaviour both in recognition and attention is an exciting target for research in this area and could yield new questions for human vision itself.

References

1. Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London, Series B, 353*, 1245–1255.
2. Parker, A. (2003). *In the blink of an eye: How vision sparked the big bang of evolution*. Pemeus, Publishing, New York.
3. Treisman, A. (1988). Preattentive processing in vision. In Z. Pylyshyn (Ed.), *Computational processes in human vision: An interdisciplinary perspective*. Norwood: Ablex Publishing Corporation.
4. Nothdurft, H.-C., Gallant, J. L., & Van Essen, D. C. (1999). Response modulation by texture surround in primate area V1: Correlates of "popout" under anesthesia. *Visual Neuroscience, 16*, 15–34.
5. Stentiford, F. W. M. (2001). An estimator for visual attention through competitive novelty with application to image compression. In Picture Coding Symposium, Seoul, 25–27 Apr 2001.
6. Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis.
7. Butko, N. J., & Movellan, J. R. (2010). Infomax control of eye movements. *IEEE Transactions on Autonomous Mental Development, 2*(2), 1–17.

8. Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*(9), 1059–1074.
9. Stevenson, I. H., Rebesco, J. M., Hatsopoulos, N. G., Haga, Z., Miller, L. E., & Kording, K. P. (2009). Bayesian inference of functional connectivity and network structure from spikes. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 17*, 203–213.
10. Borji, A., & Itti, L. (2012). State-of-the-art in visual attention modelling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*, 185–207.
11. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.
12. Han, J., Ngan, K. N., Li, M. J., & Zhang, H.-J. (2006). Unsupervised extraction of visual attention objects in color images. *IEEE Transactions on Circuits and Systems for Video Technology, 16*(1), 141–145.
13. Gao, D., & Vasconcelos, N. (2007). Bottom-up saliency is a discriminant process. In Proceedings of the International Conference on Computer Vision, Rio de Janeiro.
14. Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence, 78*, 507–545.
15. Osberger, W., & Maeder, A. J. (1998). Automatic identification of perceptually important regions in an image. In Proceedings of the 14th IEEE International Conference on Pattern Recognition, 16–20 Aug 1998, Brisbane.
16. Luo, J., & Singhal, A. (2000). On measuring low-level saliency in photographic images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island.
17. Meur, O. L., Callet, P. L., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(5), 802–817.
18. Gopalakrishnan, V., Hu, Y., & Rajan, D. (2009). Salient region detection by modelling distributions of color and orientation. *IEEE Transactions on Multimedia, 11*, 892–905.
19. Valenti, R., Sebe, N., & Gevers, T. (2009). Isocentric color saliency in images. In Proceedings of the IEEE International Conference on Image Processing, Cairo.
20. Cheng, M.-M., Mitra, N. J., Huang, X., Torr, P. H. S., & Hu, S.-M. (2015). Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(3), 569–582.
21. Achanta, R., Hamami, S., Astrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1597–1604, Miami Beach.
22. Zhang, J., & Sclaroff, S. (2013). Saliency detection: A Boolean map approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 153–160, Portland.
23. MIT saliency benchmark, http://saliency.mit.edu/#anchor_submit
24. Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7), 32, 1–20.
25. Bruce, N. D. B., & Tsotsos, J. K. (2009). Saliency, attention and visual search: An information theoretic approach. *Journal of Vision, 9*(3), 5, 1–24.
26. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013). RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication, 28*, 642–658.
27. Liu, T., Sun, J., Zheng, N., Tang, X., & Shum, H. Y. (2007). Learning to detect a salient object. In Proceedings of the International Conference on Computer Vision and Pattern Recognition, Minneapolis.
28. Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. In Proceedings of the International Conference on Image Processing, Barcelona.
29. Kavak, Y., Erdem, E., & Erdem, A. (2013). Visual saliency estimation by integrating features using multiple kernel learning. In Proceedings of the 6th International Symposium on Attention in Cognitive Systems, Beijing.

30. Vig, E., Dorr, M., & Cox, D. (2014). Large-scale optimization of hierarchical features for saliency prediction in natural images. In Proceedings of the International Conference on Computer Vision and Pattern Recognition, Columbus.
31. Kummerer, M., Theis, L., & Bethge, M. (2014). Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet. arkXiv preprint arkXiv:1411.1045
32. Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real world scenes: The role of global features on object search. *Psychological Review, 113*(4), 766–786.
33. Judd, T., Ehinger, K., Durand, F., & Torraolba, A. (2009). Learning to predict where humans look. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2106–2113, Kyoto.
34. Garcia-Diaz, A., Leboran, V., & Fdez-Vidal, X. R. (2012). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision, 12*(6), 17, 1–22.
35. Erdem, E., & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision, 13*(4), 11, 1–20.
36. Seo, H. J., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision, 9*(12), 15, 1–27.
37. Fang, Y., Lin, W., Lee, B.-S., Lau, C.-T., Chen, Z., & Lin, C.-W. (2012). Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum. *IEEE Transactions on Multimedia, 14*, 187–198.
38. Chen, C., Tang, H., Lyu, Z., Lyang, H., Shang, J., & Serem, M. (2014). Saliency modeling via outlier detection. *Journal of Electronic Imaging, 23*(5).
39. Boiman, O., & Irani, M. (2005). Detecting irregularities in images and in video. In Proceedings of the International Conference on Computer Vision, Beijing.
40. Goferman, S., Zelnik-Manor, L., & Tal, A. (2012). Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(10), 1915–1926.
41. Borji, A., & Itti, L. (2012). Exploiting local and global patch rarities for saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 478–485, Providence.
42. Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In Proceedings of the Neural Information Processing Systems, Vancouver.
43. Stentiford, F. W. M. (2001). An estimator for visual attention through competitive novelty with application to image compression. In Proceedings of the Picture Coding Symposium, Seoul, pp. 101–104.
44. Hou, X., Harel, J., & Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(1), 194–201.
45. Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision, 45*, 83–105.
46. Lindeberg, T. (1993). Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision, 11*, 3. http://www.scholarpedia.org/article/Computational_models_of_visual_attention.
47. Schectman, E., & Irani, M. (2007). Matching local self-similarities across images and videos. In Proceedings of the IEEE Conference on CVPR, Minneapolis.
48. Itti, L., & Baldi, P. (2009). Bayesian surprise attracts attention. *Vision Research, 49*, 1295–1306.
49. Stentiford, F. W. M. (2013). Saliency identified by absence of background structure. In SPIE Human Vision & Electronic Imaging XVIII, San Francisco.
50. Stentiford, F. W. M. (2014). Face recognition by detection of matching cliques of points. In Image Processing Machine Vision Applications VII Conference, IS&T/SPIE Electronic Imaging 2014, San Francisco.

Chapter 9
Bottom-Up Saliency Models for Still Images: A Practical Review

Nicolas Riche and Matei Mancas

There is an increasing interest to utilize human visual attention abilities on computational systems. This is especially the case for computer vision which needs to select the most relevant parts within a large amount of data. Therefore, modeling visual attention, particularly the bottom-up part, has been a very active research area over the past 20 years. Many different models of visual bottom-up attention are now available online. They take as input natural images and output a saliency map which gives the probability of each pixels to grab our attention. In this chapter, a state of the art of saliency-based models has been done. Those models will be used in the next chapters for explaining the validation of saliency models in computer science.

9.1 Background

The taxonomy proposed in this section is very simple and based on the historical development of saliency models. It is the most efficient one to present the study and to validate the saliency models which will be detailed in the next chapters. We distinguish two big classes of models, corresponding to different types of outputs.

Chronologically, the first algorithm type is mostly inspired from the psychological and neurobiological theories. It uses eye tracking data (fixation map) as ground truth. This is why we call models corresponding to this type **eye tracking (ET)-based models** in the present chapter. The purpose of this class of models is to

N. Riche (✉) • M. Mancas
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: nicolas.riche@umons.ac.be; matei.mancas@umons.ac.be

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_9

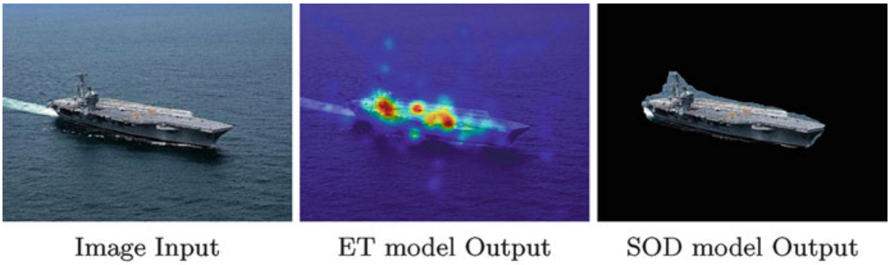


Fig. 9.1 The proposed taxonomy. Two saliency model types: ET models which predict the gaze distribution and SOD models which detect salient object

correlate the saliency map with the gaze distribution and predict the eye fixations as shown in the second column of Fig. 9.1.

In a second time, due to the requirement of computer vision applications like seam carving [1], object detection, and segmentation [2], a second class of models appeared. They are called **salient object detection (SOD)-based models** because their purpose is to detect salient objects as displayed in the third column of Fig. 9.1. They use manual annotations (binary masks which highlight the salient objects) as ground truth.

9.1.1 Eye Tracking (ET)-Based Models

The models based on eye tracking are all different but a similar structure can be found. This structure consists in the following three main steps:

- Feature extraction.
- Attentive process for saliency computation.
- Fusion to build a single saliency map.

The first step almost always represents low-level (local orientations, texture, colors, curvatures, intensity) and mid-level (horizon, faces, objects) feature extraction from still images. These extractions from the input image can be performed with single (original image resolution) or multiple (blurs and subsamples) scales to build feature maps. Then, an attentive process for saliency computation is applied on each feature map. This attentive process is often a technique from image processing which attempts to model preattentive theories. It can be local (patch) or global (entire image) but also applied with single or multiple scales. Some of the most popular operations are center-surround algorithm [3], rarity mechanism (self-information) [4], entropy [5], spectral transformation [6], and graph-based model [7]. Finally, the last step consists of merging all the obtained maps into a single saliency map. To do this, normalization and linear/nonlinear combination are computed to represent the saliency of each image pixel.

It is important to notice that this structure only uses the stimuli (RGB or grayscale input images) to compute the saliency map. More recently, some authors have also quantified biases of viewers which are looking at static and dynamic natural scenes. It was found that, for images, people tend to focus in the center of the image [8]. This is why some models add a centered 2D Gaussian bias to model the gaze pattern. As explained in [8], five causes can explain that fixations have a high probability to be in the center of an image: photographer bias, motor bias, viewing strategy, orbital reserve, and screen center.

A complete overview of eye tracking-based models is available in [9] where authors present nearly 65 models. In this section, only the ones used to explain the validation framework proposed in the next chapters are presented. A constraint is that these models must be available online.

We will focus on those models using descriptive sheets. A descriptive sheet has six basic elements: the name of the model, the year, the authors, the publications, a general figure, and a description. The purpose is to summarize and provide readers with keys to a better understanding of all elements used during the validation framework.

Moreover, in order to compare eye tracking-based models, four characteristics have been chosen and added into the descriptive sheets, following the color convention introduced by the colored keywords describing each characteristic below, for reader's convenience.

- The first characteristic divides models based on their **approaches**. Indeed, some models have a **global** approach which is applied to the entire image while others compute a saliency map with a **local** approach which is applied to a picture patch.
- The second one classifies models which use as **post processing** the **center** bias of gaze of people in free viewing. To model this bias, some models apply a 2D centered Gaussian bias to highlight the center of the saliency map.
- Third, based on [9], A. Borji et al. present a categorization of saliency models comparing their attentive **mechanism** to obtain saliency map. The eight proposed categories of methods are **cognitive, graphical, spectral, information theory, pattern classification, Bayesian, decision theory, and other** models.
- Finally, the last characteristic shows how the **stimuli** are used. Some models take into account all the channels in the **color** images while others just need the **grayscale**.

The 19 eye tracking-based models which are represented by their acronyms in Fig. 9.2 will be describe in the following of this section and use in the validation framework.

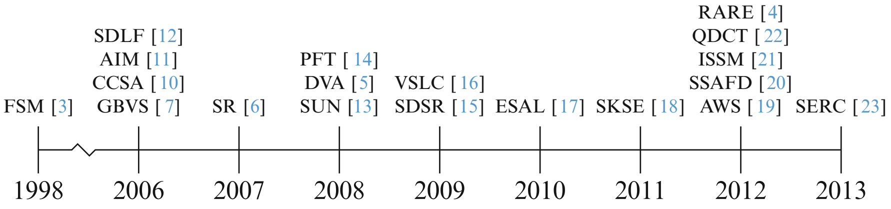


Fig. 9.2 Chronological overview of eye tracking-based models

The proposed timeline of these models shows that most algorithms have been released over the past 10 years. For most models, the ones with default parameters given by authors have been kept while for few other models, some specifications have to be changed. In these cases, the choices are described in details in the descriptive sheet.

FSM: Feature-Based Saliency Model (1998)

Characteristics: **local** | **/** | **cognitive** | **color**
Authors: L. Itti, C. Koch and E. Niebur [3].

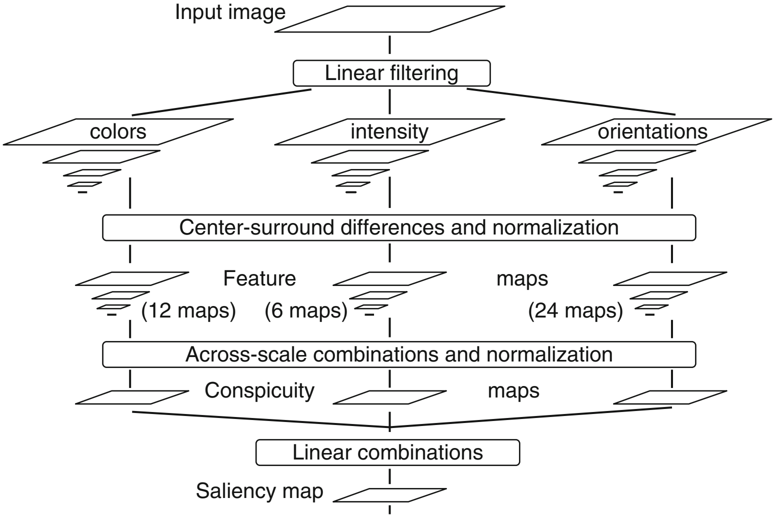


Fig. 9.3 Overview of the FSM model. From top to bottom: input image, feature extraction, center-surround differences, linear combinations, and saliency map (Adapted from [3])

Description: This model, which has been the basis of later models, was the first implementation of the Koch and Ullman attention model [24] and consists in three steps. First, an input image is subsampled into a Gaussian pyramid, and each pyramid level is decomposed in three types of static features (colors, intensity, and orientations). In the second step, center-surround feature maps are constructed from the static features. The center-surround filters provide high response in case of high contrast and low response in case of low contrast. In each channel, maps are summed across scale and normalized to form conspicuity maps which are single contrast maps for each channel. Finally, a linear combination is computed to build the saliency map (Fig. 9.3).

GBVS: Graph-Based Visual Saliency (2006)

Characteristics: **local** | **center** | **graphical** | **color**
Authors: J. Harel, C. Koch and P. Penora [7].

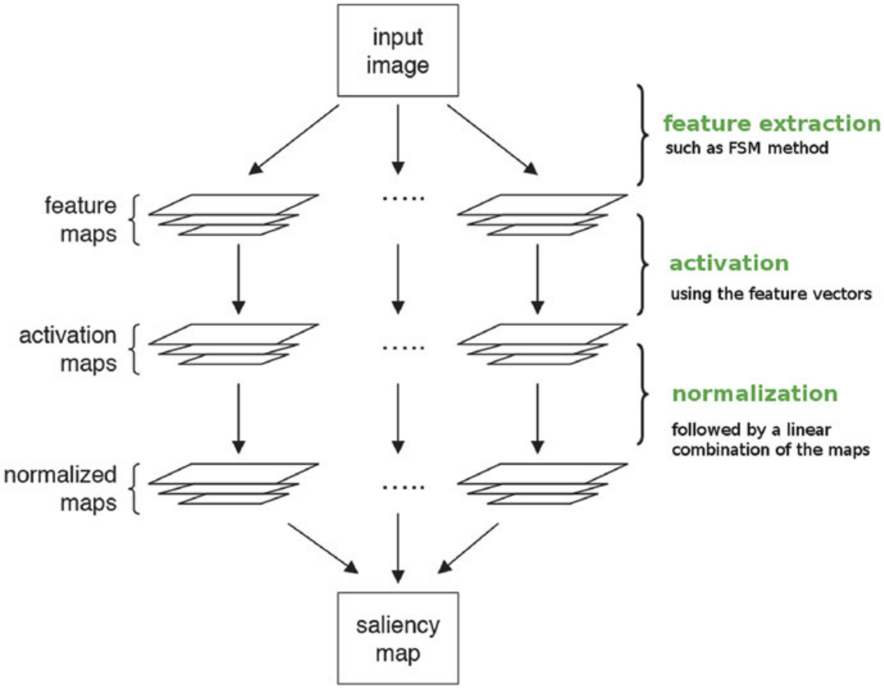


Fig. 9.4 Schematic representation of the GBVS model: input image, feature extraction, activation, normalization followed by a linear combination, and saliency map (Inspired by [7])

Description: This model introduced a graph-based method to compute visual saliency. First, the same feature maps than in the FSM model are extracted. It leads to three multiscale feature maps: colors, intensity, and orientations. Then, a fully connected graph is built over all grid locations of each feature map, and a weight is assigned between nodes. This weight depends on the spatial distance and the value of the feature map between nodes. Finally, each graph is treated as Markov chains to build an activation map where nodes which are highly dissimilar to surrounding nodes will be assigned high values. All activation maps are merged into the final saliency map. Again here, only locally contrasted features are integrated over the image; the model is thus mainly based on local context (Fig. 9.4).

CCSA: Coherent Computational Saliency Approach (2006)

Characteristics: **local** | / | **cognitive** | color
Authors: O. Le Meur, P. Le Callet, D. Barba and D. Thoreau [10].

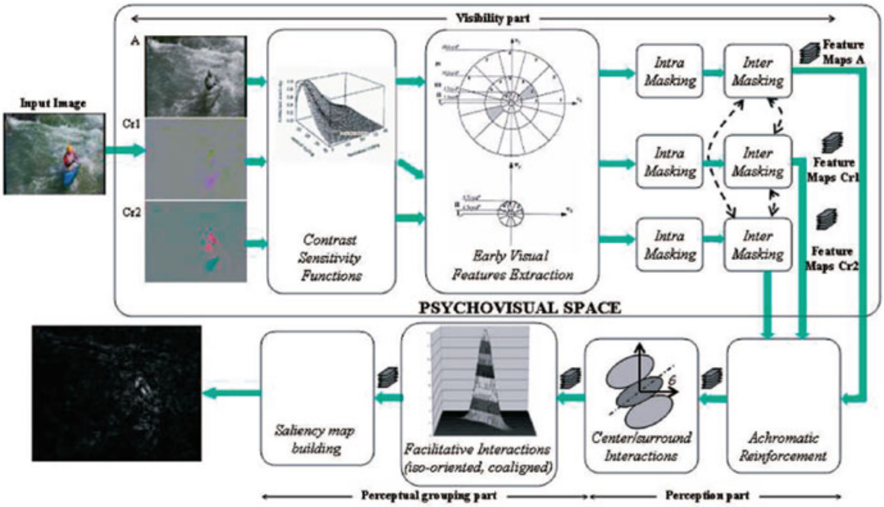


Fig. 9.5 Overview of the CCSA model. First row, from left to right: input image and the visibility part into the psychovisual space. Second row, from right to left: the perception part with center-surround interactions, the perceptual grouping part, and saliency map (Adapted from [10])

Description: This cognitive model is directly based on the current understanding of the human visual system (HVS) behavior. Three aspects of the vision are processed: visibility, perception, and perceptual grouping. The visibility part simulates the limited sensitivity of our human visual system. Visual data is normalized and grouped into a psychovisual space. The perception is used to suppress the redundant visual information by simulating the behavior of cortical cells. Two mechanisms are involved in this part: achromatic reinforcement by chromatic context and center-surround suppressive interaction. Perceptual grouping refers to the human visual ability to group and bind visual features and build a saliency map (Fig. 9.5).

AIM: Attention Based on Information Maximization (2006)

Characteristics: **local** | / | **information** | color
Authors: N. Bruce and J. Tsotsos [11].

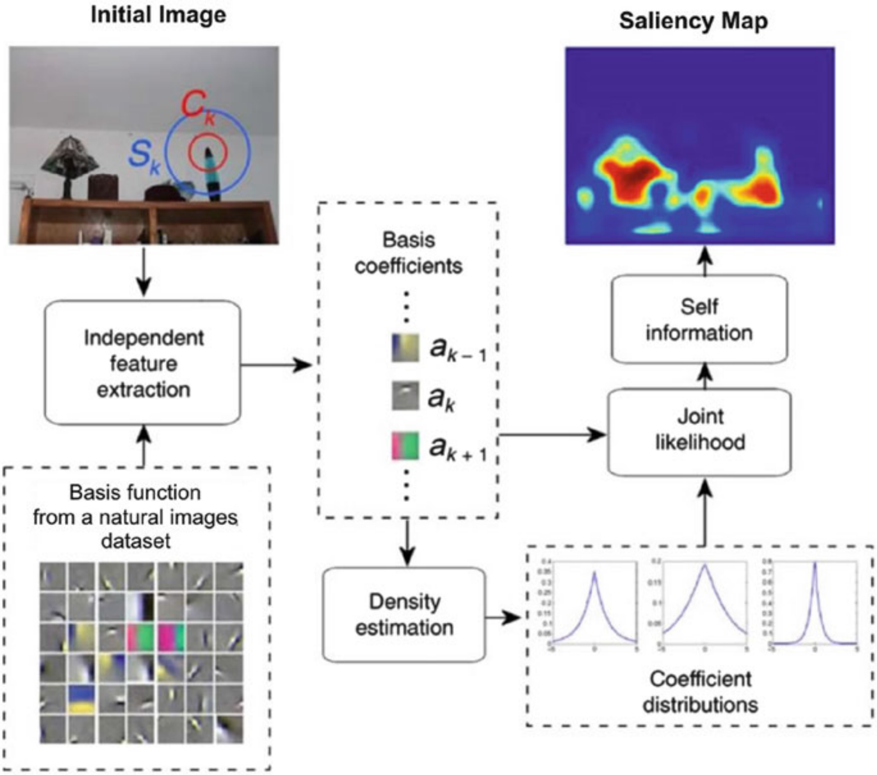


Fig. 9.6 Schematic representation of AIM model. *Left*: independent feature extraction from input image and basis functions. *Middle*: basis coefficients and density estimation. *Right*: coefficient distributions, joint likelihood, self-information, and saliency map (Adapted from [25])

Description: This model detects visual Attention based on Information Maximization (AIM). Shannon's self-information measure is used to compute the saliency. First, a patch C_k (red circle in Fig. 9.6, top left) and its neighborhood S_k (blue circle) are projected on a large sample of 7 × 7 RGB patches drawn from natural images (basis functions). The basis coefficients are obtained by performing an independent component analysis (ICA), and their probability density functions are estimated to compute the joint likelihood. The saliency value is inversely proportional to the joint likelihood. The saliency of a local image region is thus computed as the information conveyed by that region relative to its surroundings.

SDLF: Saliency Detection by Using Local Features (2006)

Characteristics: **local** | / | **bayesian** | **grayscale**
Authors: A. Torralba, A. Oliva, M. Castelhano and J. Henderson [12].

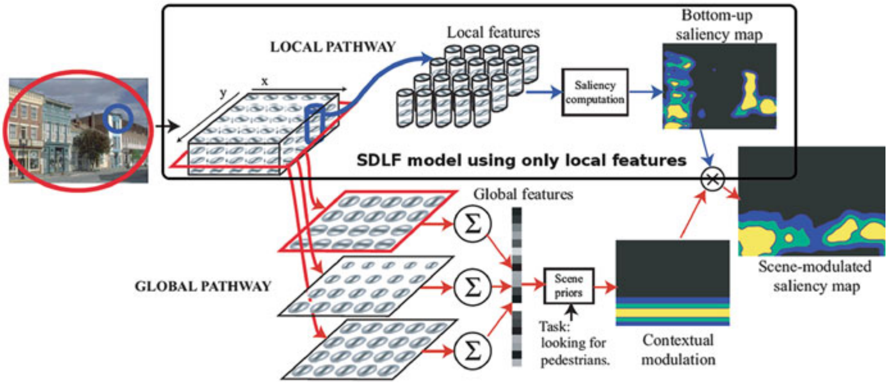


Fig. 9.7 Overview of the Torralba's model. From left to right: input image, two parallel pathways—one computes local features to build saliency map (top middle), while the other computes global features to build scene priors (bottom middle) and scene-modulated saliency map (Adapted from [12])

Description: This method proposes to analyze the image in two parallel pathways. One pathway computes local features (saliency). The second pathway is a global approach. It takes into account the contextual modulation and can be seen as the modeling of the top-down part of visual attention by computing scene priors. This model uses a Bayesian framework that integrates both image saliency and scene priors.

The SDLF algorithm considered here is the purely bottom-up saliency map without the task scene priors obtained by the global pathway. The local pathway represents each spatial location independently and provides a measure of how unlikely it is to find a set of local measurements within the image. To do this, a steerable pyramid which is a linear multiscale and multi-orientation image decomposition is employed. This local representation is used to compute image saliency (Fig. 9.7).

SR: Spectral Residual (2007)

Characteristics: **global** | / | **spectral** | **grayscale**
Authors: X. Hou and L. Zhang [6].

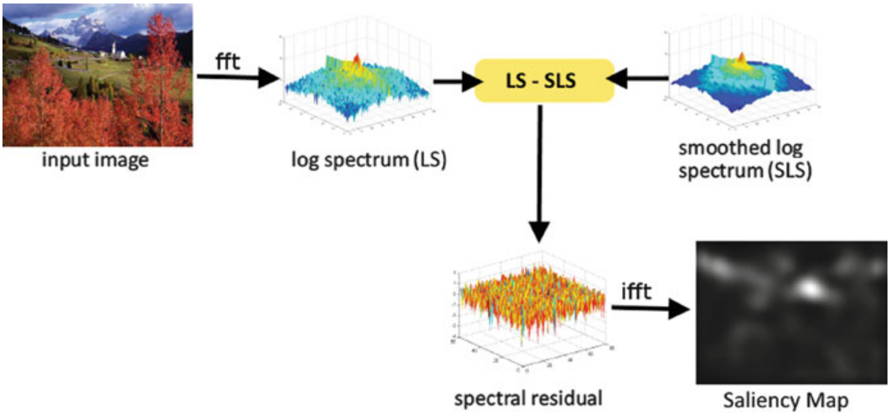


Fig. 9.8 Schematic representation of SR model. First row: input image, log-spectrum (LS) and smoothed log-spectrum (SLS). Second row: spectral residual and saliency map (Adapted from [6])

Description: The SR model is independent from any feature. In this method, the first step is to compute the image Fourier spectrum (the amplitude and phase maps). Then, the log-spectrum of the amplitude map is calculated. A filtering amplitude map is also computed by multiplying the log-spectrum map with a local average filter. The spectral residual map is obtained by subtracting these last two maps. The saliency map is obtained through Fourier transform inversion. It should be noted that the phase spectrum is preserved during the process. The idea is that if the image log-spectrum is far from the $1/f$ of natural images (image filtered spectrum), there is something abnormal which deserves attention (Fig. 9.8).

SUN: Saliency Using Natural Image Statistics (2008)

Characteristics: **local** | / | **bayesian** | **color**
Authors: L. Zhang, M. Tong, T. Marks, H. Shan and G. Cottrell [13].

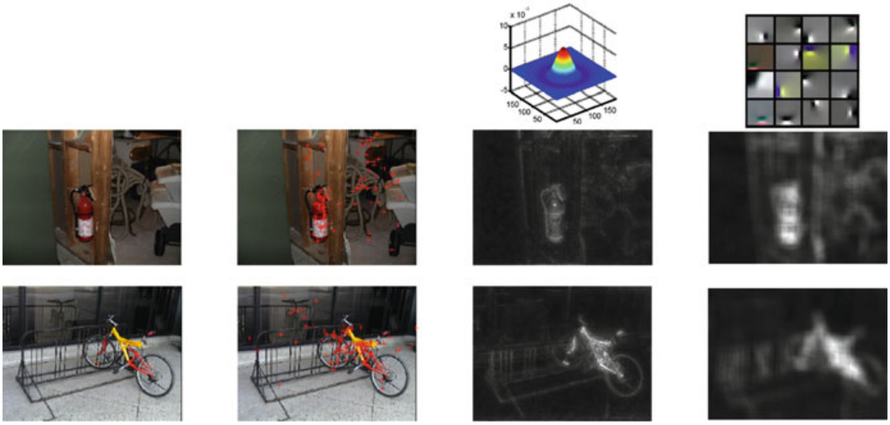


Fig. 9.9 Overview of the SUN model. From left to right: input images, images with the eye fixations, and the two methods; third column features are computed with differences of Gaussians (DoG) and fourth column with independent component analysis (ICA) (Adapted from [13])

Description: This saliency model using natural statistics (SUN) proposes a Bayesian framework from which bottom-up saliency emerges naturally as the self-information of visual features. In this method, notions similar to SDLF (Bayes formula) and AIM (local self-information) models are found. The Bayesian framework is composed of three terms: self-information, log-likelihood, and location prior. The first term (bottom-up) is independent of the target while the two others (top-down) depend on target.

The saliency map is reduced here to the self-information (bottom-up). Two methods have been implemented. First, the features are calculated as outputs of linear filters, such as DoG filters. Second, the features are calculated as the outputs to filters learned from natural images using ICA. SUN with ICA (Method 2) used here outperforms SUN with DoG filters (Method 1). These output maps are computed on a set of 138 images of natural scenes. An estimation of the probability distribution is obtained over the observed values of each of the features. The self-information measure is applied on statistics from this database of natural images (among which the current image is not present). Those images act like typical "normal" images and difference from the statistics of those images might attract attention (Fig. 9.9).

DVA: Dynamic Visual Attention (2008)

Characteristics: **local** | / | **information** | **color**
Authors: X. Hou and L. Zhang [5].

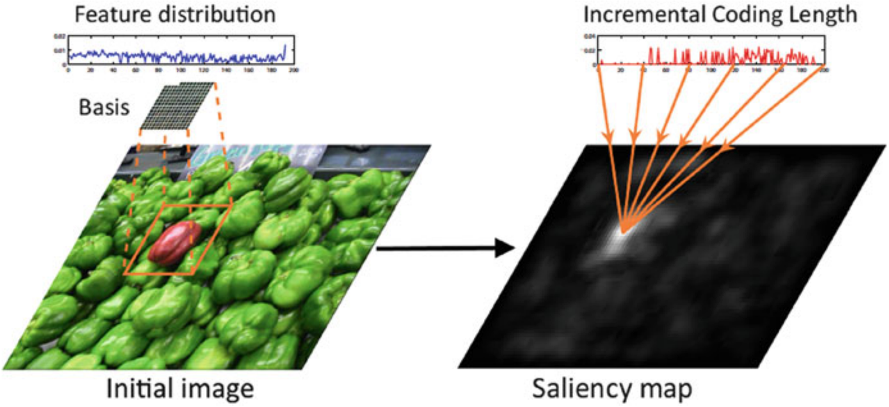


Fig. 9.10 Schematic representation of DVA model: the initial image with a feature distribution (*left*) and the corresponding saliency map with the incremental coding length of this feature distribution (*right*) (Adapted from [5])

Description: This model, based on the rarity of features, introduced the incremental coding length (ICL) approach to measure the perspective entropy gain of each feature. Motivated by the sparse coding strategy discovered in primary visual cortex, an image patch is first represented as a linear combination of sparse coding basis functions. The activity ratio of a feature is its average response to image patches. The activity of the feature ensemble is considered as a probability function. Then, each feature is evaluated with respect to its incremental coding length (ICL). The ICL of one feature is defined as the entropy gain of the ensemble during the activity increment of this feature. In accordance with the general principle of predictive coding, they redistribute energy to features according to their ICL contribution: frequently activated features receive less energy than rarer features. Finally, the saliency of a region is obtained by summing up the activity of all features in that region (Fig. 9.10).

PFT: Phase Fourier Transform (2008)

Characteristics: **global** | / | **spectral** | **grayscale**
Authors: C. Guo, Q. Ma and L. Zhang [14].

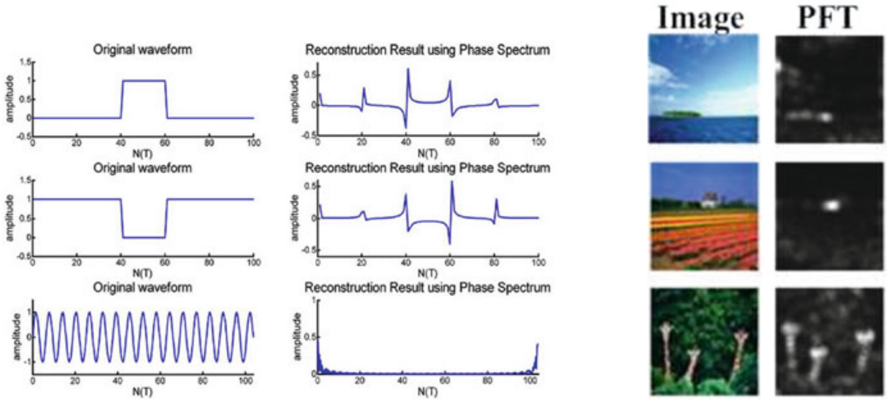


Fig. 9.11 Example of saliency maps from PFT algorithm on three input images (*right*) and reconstruction obtained by the phase spectrum alone on three one-dimensional waveforms (*left*). When there are many varying sinusoidal components (pulse), the reconstruction contains the largest spikes (Adapted from [14])

Description: This method is based on the SR model which uses the spectral residual of the amplitude spectrum to obtain the saliency map. PFT proposes to use the phase spectrum instead of the amplitude. The key idea is that the amplitude spectrum specifies how much of each sinusoidal component is present in an image while the phase information specifies where each of the sinusoidal components resides within it. The location with less periodicity or less homogeneity indicates where the interesting areas are and helps in obtaining the saliency map (Fig. 9.11).

SDSR: Saliency Detection by Self-Resemblance (2009)

Characteristics: **local** | / | **information** | **grayscale**
Authors: H. J. Seo and P. Milanfar [15].

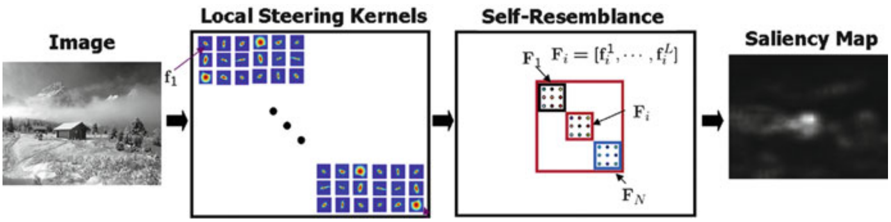


Fig. 9.12 Overview of SDSR saliency detection system. Local regression kernels capture the underlying local structure of a grayscale image (*left*), and a self-resemblance measure is obtained by using a nonparametric kernel density estimation and indicates the likelihood of saliency. A saliency map is built on this measure (*right*) (Adapted from [15])

Description: This bottom-up model proposes a Saliency Detection by Self-Resemblance (SDSR). The implementation by Seo and Milanfar consists in two parts. First, they describe local image structure at each pixel by local regression kernels as features (matrix of local descriptors). The underlying hypothesis is that eye fixations are driven by local feature contrast and these local descriptors are robust with image distortions and noisy images. In a second step, they quantify the likeness of each pixel to its surroundings and use matrix cosine similarity, which results in a saliency map consisting of local self-resemblance measures. Even if patches of the image are compared on a wider space than only surround, they are not compared on the entire image (Fig. 9.12).

VSLC: Visual Saliency Based on Lossy Coding (2009)

Characteristics: **local** | / | **information** | **color**
Authors: Y. Li, Y. Junchi and Z. Yue [16].

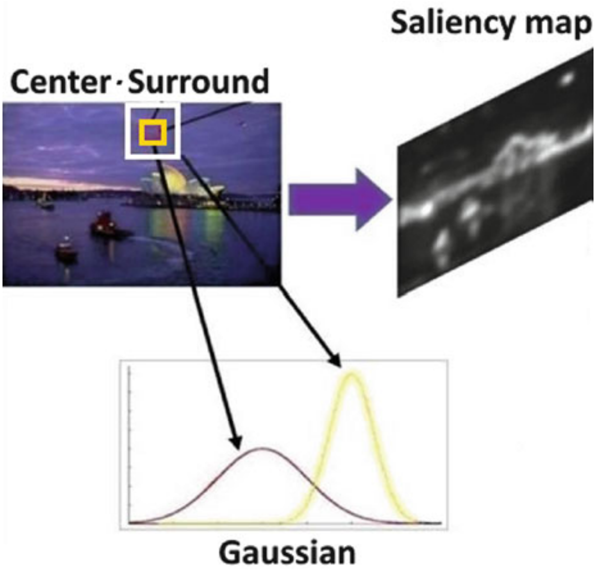


Fig. 9.13 Schematic representation of the VSLC algorithm: the saliency map (*right*) is computed by a local center-surround mechanism (*left*) which approximates the conditional entropy with the lossy coding length of multivariate Gaussian data (*below*) (Adapted from [26])

Description: This method computes visual saliency based on lossy coding (VSLC). This definition of visual saliency is strictly local. The saliency is measured as the minimum conditional entropy, which represents the uncertainty of the center-surround local region, when the surrounding area is given and the perceptional distortion is considered. The conditional entropy is approximated by the lossy coding length of multivariate Gaussian data. The final saliency map is accumulated by pixels (Fig. 9.13).

ESAL: Extended Saliency (2010)

Characteristics: **global** | **/** | **graphical** | **color**
Authors: T. Avraham and M. Lindenbaum [17].

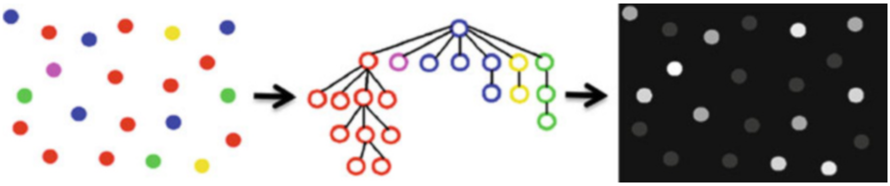


Fig. 9.14 ESAL algorithm on a color synthetic image. From left to right: the synthetic image, the tree where each node is colored according to the corresponding candidates, the computed saliency map based on self-similarities (Adapted from [17])

Description: The ESAL model proposes a static saliency model based on self-similarities (Fig. 9.14). It is built on three observations:

1. The number of target candidates (salient patches) is usually small. So the model is region based. The image is divided into segments, which are the candidates for attention. The initial probability for each candidate gives preference to small number of expected targets.
2. There is a correlation between visual similarity and target-nontarget labels. So two visually similar candidates are likely to both be objects of interest or not. The visual similarity between candidates is measured from their feature space distance. Each is represented as a vector of features (texture and color). A short distance between the two vectors indicates that the corresponding candidates are visually similar and infers the correlations between the corresponding labels.
3. Natural scenes are often composed of clustered structural units. The data is clustered into a mixture of multivariate Gaussians. The saliency of each candidate is deduced by marginalization.

The algorithm is essentially a method for estimating the probability that a candidate is a target.

SKSE: Sparse Sampling-Kernel Density Saliency Estimation (2011)

Characteristics: **local** | **center** | **bayesian** | **color**
Authors: R. H. Tavakoli, E. Rahtu and J. Heikkilä [18].

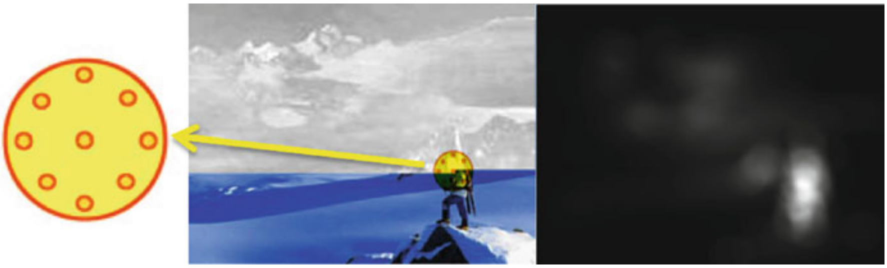


Fig. 9.15 An example of saliency map obtained using the SKSE method (*right*) for the input image (*middle*). The procedure of applying a window is also illustrated (*middle*) with a pixel and its selected surrounding samples in a window (*left*) (Adapted from [18])

Description: This SKSE method measures saliency with a simple center-surround mechanism for still images. The proposed algorithm is based on estimating saliency by local contrast. The distributions of features are estimated using sparse sampling and kernel density estimation. A general Bayesian framework defines saliency map and implicitly includes center bias. This method is fast in comparison to other similar approaches and is able to run in real time (Fig. 9.15).

AWS: Adaptive Whitening Saliency (2012)

Characteristics: **global** | / | **other** | **color**
Authors: A. Garcia-Diaz, V. Leborán, X. Fdez-Vidal and X. Pardo [19].

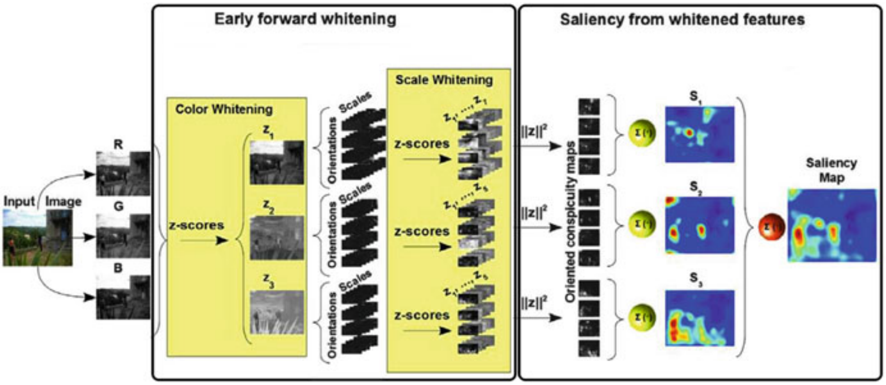


Fig. 9.16 Schematic representation of AWS mechanism: an early forward whitening applied on RGB input image (*left*) and saliency map computed from whitened features (*right*) (Adapted from [27])

Description: This model of bottom-up saliency is based on the variability in local energy as a measure of saliency. First, the chromatic components are approximated with a chromatic decomposition and whitening from RGB images. A bank of log-Gabor filters is then applied on chromatic components. Each representation is decorrelated by using whitening and a distance is computed to build oriented conspicuity maps. The final saliency map is obtained by summing these maps. The decorrelation is a global operation which considers the whole image (Fig. 9.16).

SSAFD: Saliency on Scale-Space Analysis in Frequency Domain (2012)

Characteristics: **global** | / | **spectral** | **color**
Authors: J. Li, M. Levine, X. An, X. Xu and H. He [20].

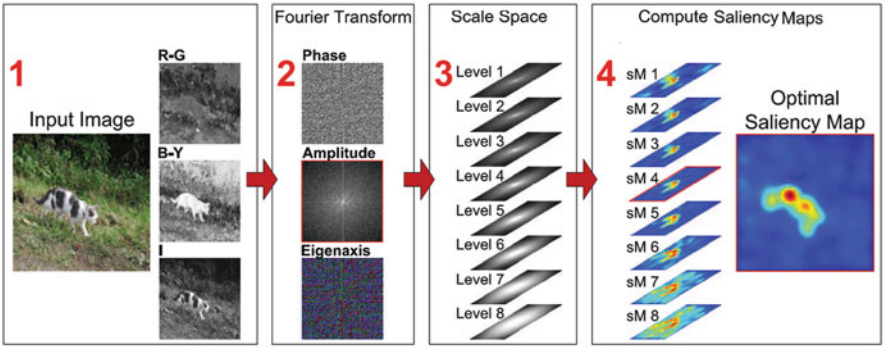


Fig. 9.17 Overview of the SSAFD method: the feature matrices are computed to form a hypercomplex matrix (1). A Fourier transform is performed (2) and a spectrum scale space is obtained by smoothing the amplitude (3). Finally, the saliency map is built by selecting the best saliency scale-space maps (4) (Adapted from [20])

Description: This mechanism considers saliency detection as a frequency domain analysis problem. First, feature maps are extracted from a color image which is converted into I, red-green, blue-yellow feature maps to form a hypercomplex matrix. Second, a Fourier transform is applied on this matrix and outputs the amplitude, the phase, and the eigenaxis spectrum. Third, spectrum scale space is obtained by smoothing the amplitude spectrum with Gaussian kernels at different levels. Finally, the saliency map is obtained by selecting the best saliency maps produced by the spectrum scale space (Fig. 9.17).

ISSM: Image Signature Saliency Model (2012)

Characteristics: **global** | **/** | **spectral** | **grayscale**
Authors: X. Hou, J. Harel and C. Koch [21].

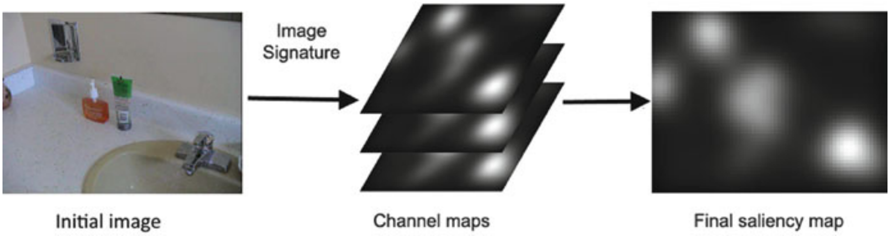


Fig. 9.18 Schematic representation of the ISSM method: an image signature is computed on each channel of the input image. The final saliency map is obtained by summing the results of the three channels (Adapted from [21])

Description: The ISSM method introduces a simple image descriptor referred to as the image signature. Given an input image, first, three color channels are extracted. Both RGB or CIE LAB color spaces can be used.

CIE LAB is chosen here as it was designed to closely mimic how human vision is believed to perceive color. Then image signature is computed on each channel to suppress background and detect the foreground of an image. To do that, a discrete cosine transform (DCT) is applied to each channel. Then, to approximately isolate the foreground, the sign of each DCT component, equivalent to phase for a Fourier decomposition, is stored and inversely transformed back into the spatial domain. The amplitude information across the entire frequency is discarded. A 2D Gaussian is then applied to blur the results, and the final saliency map is obtained by summing the results of the three channels (Fig. 9.18).

QDCT: Quaternion DCT Image Signature Saliency (2012)

Characteristics: **global** | / | **spectral** | **color**
Authors: B. Schauerte and R. Stiefelhagen [22].**spectral** |

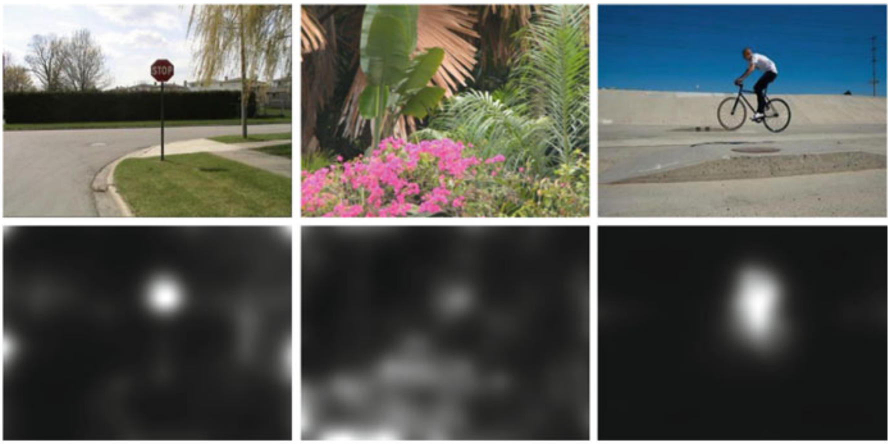


Fig. 9.19 Validation of the QDCT saliency model. First row: original images. Second row: saliency map obtained, thanks to the QDCT method (Adapted from [28])

Description: This QDCT model extends the previous proposed work on DCT-based image signatures (ISSM model) which defined the saliency using the inverse DCT of the signs in the cosine spectrum. In the QDCT model, the scalar definition of image signatures is transferred to quaternion images. Quaternions are used to represent and process color images (in CIE LAB color space) in a holistic framework and, subsequently, the quaternion DCT (QDCT) and signum function are applied to calculate the visual saliency. The signum function for quaternions can be considered as the quaternion direction (Fig. 9.19).

RARE: Multiscale Rarity-Based Saliency Algorithm (2012)

Characteristics: **global** | / | **information** | **color**
Authors: N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin and T. Dutoit [4].

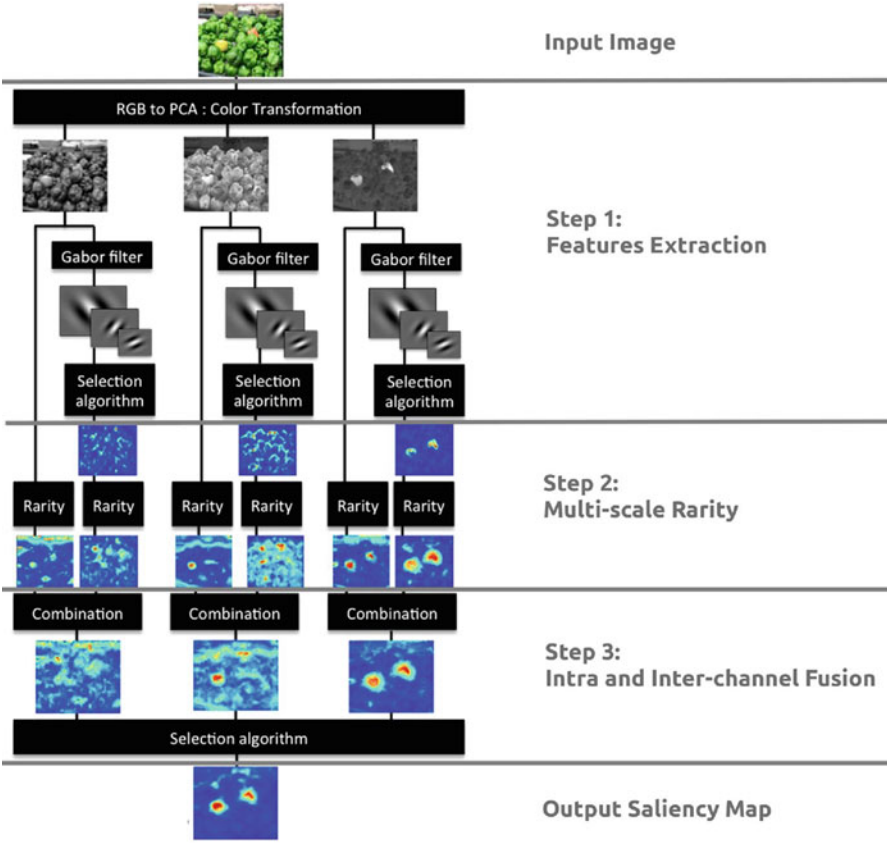


Fig. 9.20 Diagram of our proposed model. First, from the input image, color and orientation features are extracted in parallel or sequentially. Then, for each feature, a multiscale rarity mechanism is applied. Finally, two fusions (intra- and inter-channel) are made from the rarity maps to provide the final saliency map (Adapted from [4])

Description: The RARE mechanism has three main steps. First, the authors extract low-level color and medium-level orientation features. Afterwards, a multiscale rarity mechanism is applied. This rarity mechanism is the key of RARE. Indeed, a feature is not necessary salient alone, but only in a specific context. The mechanism of multiscale rarity allows to detect both locally contrasted and globally rare regions in the image. Finally, they fuse rarity maps into a single final saliency map (Fig. 9.20).

SERC: Saliency Estimation Using Region Covariances (2013)

Characteristics: **local** | **center** | **other** | **color**
Authors: E. Erdem and A. Erdem [23].

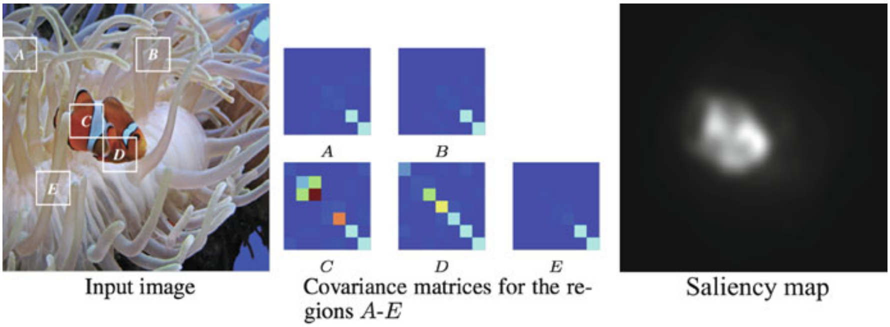


Fig. 9.21 SERC model proposes to use covariance matrices (*middle*) of input image patches (*left*) as meta-features for saliency estimation (*right*) (Adapted from [23])

Description: The SERC method investigates a better way than the commonly used linear combinations to merge maps which produce the master saliency map. The authors propose to use covariance matrices of simple image features as meta-features for saliency estimation. As low dimensional representations of image patches, region covariances provide nonlinear integration of different features by modeling their correlations. The input image is first decomposed into non-overlapping regions, and then the saliency of each region is measured by examining its surrounding regions. The salient regions are those that are highly dissimilar to their neighboring regions in terms of their covariance (second-order statistics) representations based on color, orientation, and spatial features. Moreover, to improve the detection, first-order statistics (mean) can be also used to capture saliency of an image region with respect to its surroundings (Fig. 9.21).

9.1.2 Salient Object Detection (SOD)-Based Models

Recently, salient object detection(SOD)-based models have attracted a lot of interest due to the explosion of computer vision applications like seam carving [1], object detection, or image segmentation [2]. The purpose of the models is to separate the salient object from the image background [29].

As explained in [30, 31], there is a strong relationship between where people look in scenes (fixation maps) and what they choose as the most salient object when they are explicitly asked (binary masks). Therefore, the most salient object is the one that attracts the highest fraction of fixations.

This finding justifies to use the notion of visual attention modeling to locate salient object or region in a scene. However, as seen in next chapters, the databases chosen for evaluation must have complex images and not just one single object with a clean background in order not to make solely foreground/background separation.

However, the approaches presented in Sect. 9.1.1 work well in finding fixation locations but they have not been able to accurately detect where salient objects should be. Therefore, a second wave of models have emerged by following the works in [32–34].

As explained in [35], most of these algorithms have two main steps: detect the most salient objects and segment the accurate boundary of these objects. A complete overview of salient object detection-based models is available in [35]. In their paper, the authors present a taxonomy divided into three categories:

1. SOD with intrinsic cues (36 models).
2. SOD with extrinsic cues (20 models).
3. Other algorithms (9 models).

In this section, only the ones used for our study and validation are exposed and detailed using descriptive sheets. Therefore, a constraint is that these models must be available online.

In order to compare salient object detection-based models, four characteristics have been chosen and added into the descriptive sheets, following the color convention introduced by the colored keywords describing each characteristic below, for reader's convenience.

- The first characteristic divides models based on their **approaches**. As seen in Sect. 9.1.1, some models have a **global** approach which is applied to the entire image while others compute a saliency map with a **local** approach which is applied to a picture area. Some models can also use **both** approaches.
- The second one classifies models which use as **prior** the **superpixel** segmentation. Some models practice a superpixel segmentation to improve the accurate boundary of the detected salient object.
- Third, models are classified in two categories depending on the **input** types used to compute their saliency map: all the **pixels** or **patches** which summarize the information.
- Finally, the last characteristic shows how the **stimuli** are used. Some models take into account all the channels in the **color** images while others use information from the **JPEG bit stream**.

The nine salient object detection-based models which are represented by their acronyms in Fig. 9.22 will be describe in the following of this section and use in the studies introduced in next chapters.

The proposed timeline of these models shows that most algorithms have been released over the past 5 years as a second wave in the modeling of visual attention. Indeed, these models are more recent than eye tracking-based saliency algorithms and their goal are to detect salient objects.

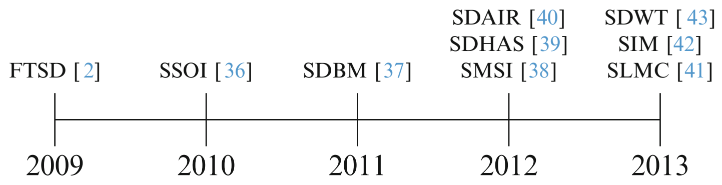


Fig. 9.22 Chronological overview of salient object detection-based models used in the next chapters

FTSD: Frequency-Tuned Saliency Detection (2009)

Characteristics: **global** | **/** | **pixels** | **color**
Authors: R. Achanta, S. Hemami, F. Estrada and S. Susstrunk [2].

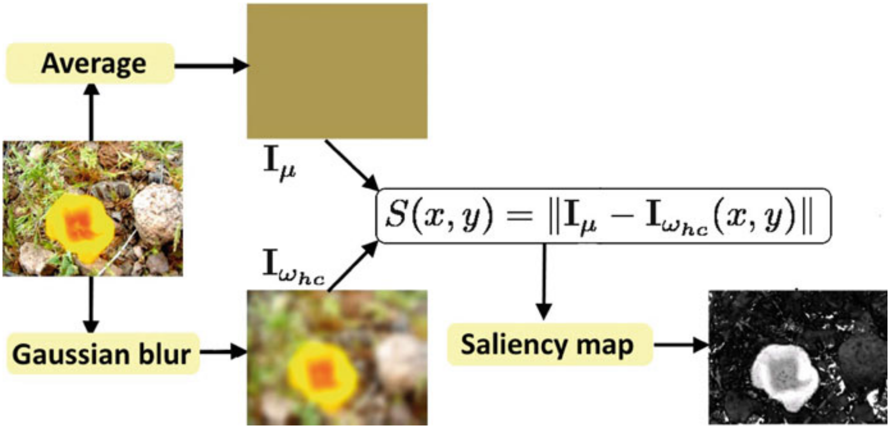

Average
$\mathbf{I}_\mu$
$\mathbf{I}_{\omega_{hc}}$
$S(x,y) = \|\mathbf{I}_\mu - \mathbf{I}_{\omega_{hc}}(x,y)\|$
Gaussian blur
Saliency map

Fig. 9.23 Schematic representation of the FSTD method: RGB input image (*left*), CIE LAB color space of the Gaussian filtered input image (*bottom middle*), the average CIE LAB of the input image (*top middle*), and saliency map (*right*) (Adapted from [2])

Description: This algorithm is a very simple model based on local color and luminance feature contrast. First, the input RGB image is transformed to CIE LAB color space. Second, the CIE LAB image is blurred with a Gaussian kernel to eliminate noise and texture details from the original CIE LAB image. Finally, the saliency map is computed by using euclidean distance between the Gaussian-filtered and the original image. The Gaussian-filtered image eliminates small objects and provides an idea about how the image appears to the eyes at a first glance. Objects which are very different from this normal image will attract attention (Fig. 9.23).

SSOI: Segmenting Salient Objects from Images (2010)

Characteristics: **local** | **/** | **pixels** | **color**
Authors: E. Rahtu, J. Kannala, M. Salo and J. Heikkilä [36].

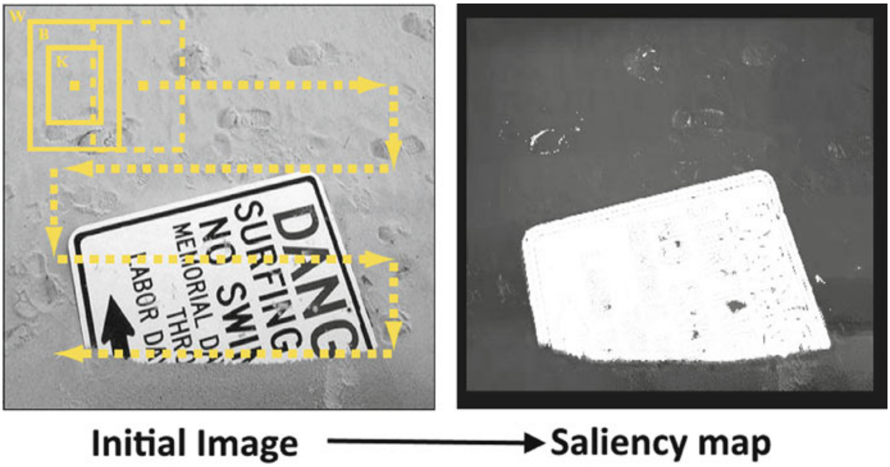


Fig. 9.24 Illustration of SSOI saliency map computation: the *yellow* sliding window applied on the input image (*left*) and the saliency map based on local feature contrast (*right*) (Adapted from [36])

Description: This model introduces a new salient object segmentation method based on Bayesian inference. A sliding window is applied on the image. For each window, a contrast is computed between the distribution of some features (such as illumination or color spaces) in an inner window and the distribution in the collar of the window. The proposed saliency measure is formulated using a statistical framework with these local feature contrasts. At the end, the framework combined them with a conditional random field (CRF) model which is a probabilistic model for labeling and segmenting data to provide the single saliency map (Fig. 9.24).

SDBM: Saliency Detection Based on Bayesian Model (2011)

Characteristics: **global** | **superpixels** | **pixels** | **color**
Authors: Y. Xie and H. Lu [37].

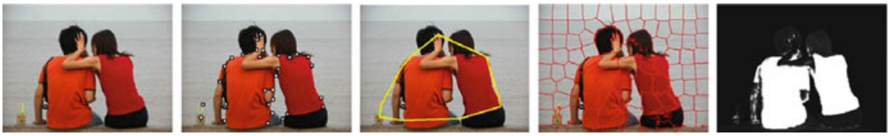


Fig. 9.25 Illustration of the SDBM method. From left to right: the input image, the detected salient points, the convex hull based on salient points, the superpixels, and the SDBM saliency map (Adapted from [37])

Description: This method proposes a new computational saliency detection model which is implemented with a coarse to fine strategy under the Bayesian framework. First, the authors extract salient points from the RGB input image to detect the corner of the salient object. Second, a convex hull is used to enclose the salient points after eliminating the points near the boundary and gives a coarse location of the salient region. Based on this rough salient region, they formulate the saliency computation as a Bayesian inference problem for estimating the posterior probability at each pixel of the image and obtain the final saliency map. The prior saliency distribution is based on superpixels and the obtained rough region (Fig. 9.25).

SMSI: Saliency Map Based on Sampling an Image (2012)

Characteristics: **local** | **/** | **pixels** | **color**
Authors: T. N. Vikram, M. Tscherepanow and B. Wrede [38].

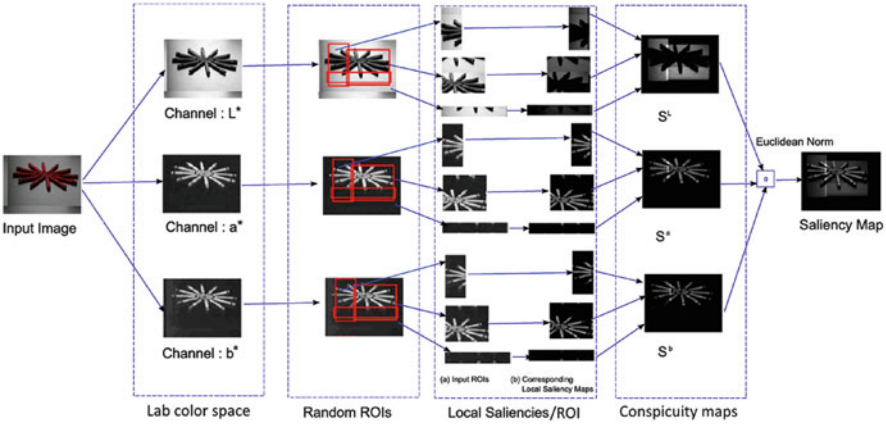


Fig. 9.26 An illustration of SMSI saliency model. From left to right: RGB input image, CIE LAB color space feature maps, local saliencies on random windows, conspicuity maps, and saliency map (Adapted from [38])

Description: This algorithm proposes to compute local saliencies over random rectangular regions of interest. To do that, an image I is first subjected to a Gaussian filter in order to remove noise and abrupt onsets. Second, it is converted into the CIE LAB space and decomposed into channels. Third, N random sub-windows are generated over each of the channels. Fourth, a center-surround map is defined for each channel as the sum of the absolute differences of the pixel intensity values to the mean intensity value of the random sub-windows. The final saliency map is computed as the euclidean norm of center-surround values over different channels (Fig. 9.26).

SDHAS: Saliency Detection on HSV and Amplitude Spectrum (2012)

Characteristics: **local** | / | **patches** | **color**
Authors: Y. Fang, W. Lin, B Lee, C. Lau, Z. Chen and C. Lin [39].

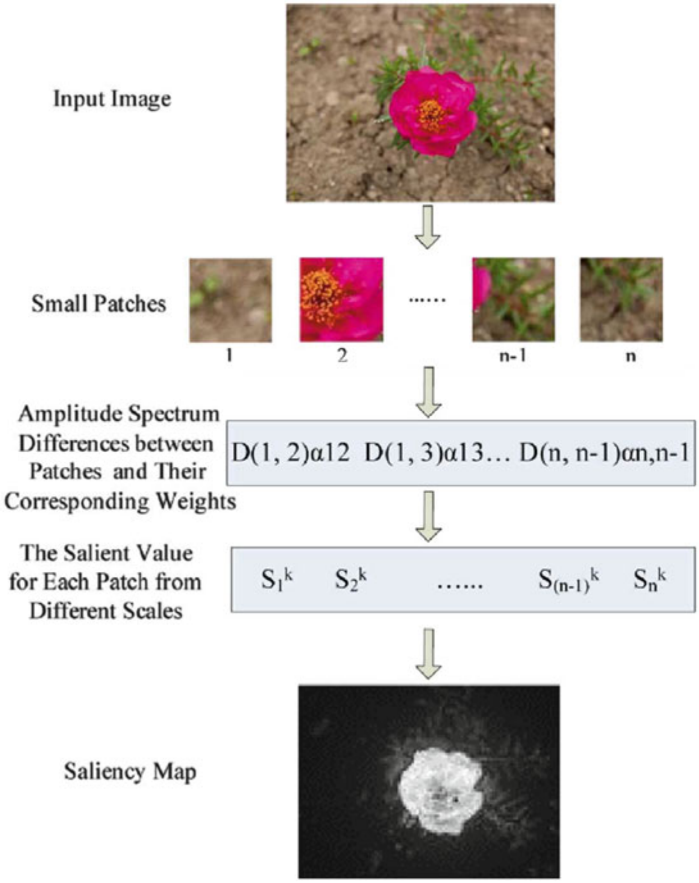

Input Image
Small Patches
1
2
......
n-1
n
Amplitude Spectrum Differences between Patches and Their Corresponding Weights
D(1, 2)α12 D(1, 3)α13... D(n, n-1)αn,n-1
The Salient Value for Each Patch from Different Scales
S_1^k S_2^k ...... $S_{(n-1)}^k$ S_n^k
Saliency Map

Fig. 9.27 Schematic representation of the SDHAS algorithm. Top to bottom: input image, patches from the input image, amplitude spectrum differences from patches and their corresponding neighbors, salient values for each patch, and saliency map (Adapted from [39])

Description: This algorithm is a new saliency detection model based on the human visual sensitivity and the amplitude spectrum of quaternion Fourier transform (QFT). First, the input image is divided into small patches. The model then computes the amplitude spectrum of QFT to represent the color, intensity, and orientation distributions of each image patch. The saliency value of each patch is obtained by computing the quaternion Fourier transform amplitude spectrum differences between a patch and its neighbor patches. The weights for these differences are determined by the human visual sensitivity and the final saliency map is influenced by the image patch size and the scale (Fig. 9.27).

SDAIR: Saliency Detection for Adaptive Image Retargeting (2012)

Characteristics: **global** | / | **patches** | **jpeg bit stream**
Authors: Y. Fang, Z. Chen, W. Lin, C. Lin [40].

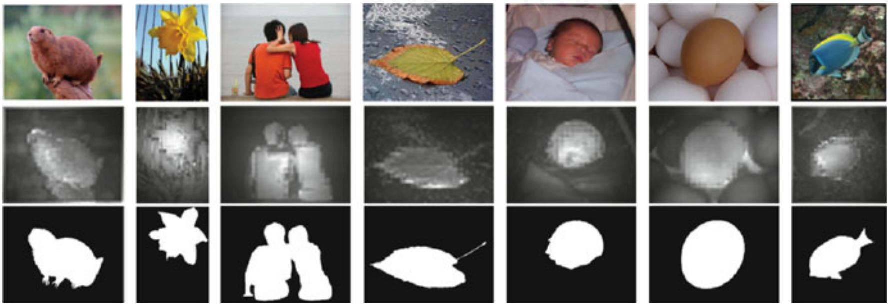


Fig. 9.28 Validation of the SDAIR saliency detection algorithm. First row: original images. Second row: the SDAIR saliency model. Last row: the ground truths (Adapted from [40])

Description: This mechanism proposes a novel saliency detection model in the compressed domain. The authors extract the saliency information for the image from the JPEG bit stream. The intensity, color, and texture features of the image are derived and extracted directly from the discrete cosine transform (DCT) coefficients in the JPEG bit stream. Then the Hausdorff distance is used to calculate the difference between two vectors of texture feature from two DCT blocks. The saliency map is obtained by integrating feature maps using a coherent normalization-based fusion method. Based on this model, an adaptive image retargeting algorithm can be designed (Fig. 9.28).

SLMC: Saliency via Low- and Mid-Level Cues (2013)

Characteristics: **global** | **superpixels** | **pixels** | **color**
Authors: Y. Xie, H. Lu and M. Yang [41].

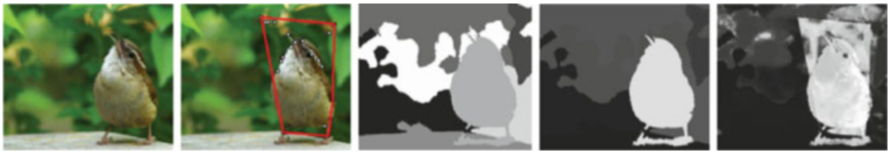


Fig. 9.29 The SLMC model proposes a Bayesian framework by exploiting low- and mid-level cues. Left to right: the original image, the Harris points detection and a convex hull applied on these points, the clustering results, the prior probability map, and finally the saliency map (Adapted from [41])

Description: This algorithm proposes to detect salient objects within a Bayesian framework by exploiting low- and mid-level cues. First, a coarse saliency region is obtained using a convex hull on Harris points. The likelihood probability is then computed based on the center-surround principle between the inner region and the outer one. For estimating the posterior probability at each pixel of the image, the prior distribution is then computed by mid-level cues like superpixels which are used to analyze the saliency information. A Laplacian sparse subspace clustering (LSSC) method groups superpixels. Finally, the Bayesian visual saliency map is computed based on the results of the superpixel clustering and the coarse saliency region (Fig. 9.29).

SIM: Saliency for Image Manipulation (2013)

Characteristics: **local** | **/** | **pixels** | **color**
Authors: R. Margolin, L. Zelnik-Manor and A. Tal [42].

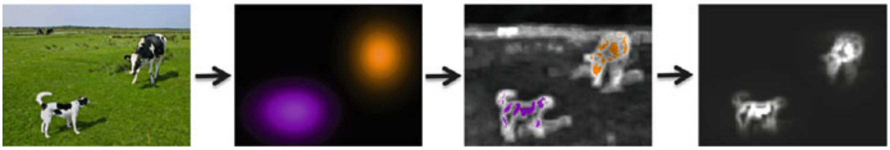


Fig. 9.30 Illustration of the SIM algorithm: from left to right: the input image, the multiple dominant object detection map, the distinctness map and the saliency map (Adapted from [42])

Description: The SIM algorithm proposes an approach for saliency detection based on four principles: pixel distinctness, pixel reciprocity, object association, and multilayer saliency. First, the authors compute the pixel distinctness where a pixel is considered distinct if its surrounding patch does not appear elsewhere in the image. Second, assuming that distinctive pixels are salient, a pixel reciprocity effect is computed. The distinctness map is updated with the reciprocity effect in order to assume that pixels in the neighborhood of distinctive pixels are more likely to be salient as well. Third, multiple dominant objects are detected and a method predicts their locations. Finally, the single saliency map combines patch distinctness with the object probability map. Due to the observation that a single saliency map is insufficient, a multilayer saliency map is built by varying degrees of abstraction. The final saliency map discards to small objects and noisy background (Fig. 9.30).

SDWT: Saliency Detection Based on Wavelet Transform (2013)

Characteristics: **both** | **/** | **pixels** | **color**
Authors: N. Imamoglu, W. Lin and Y. Fang [43].

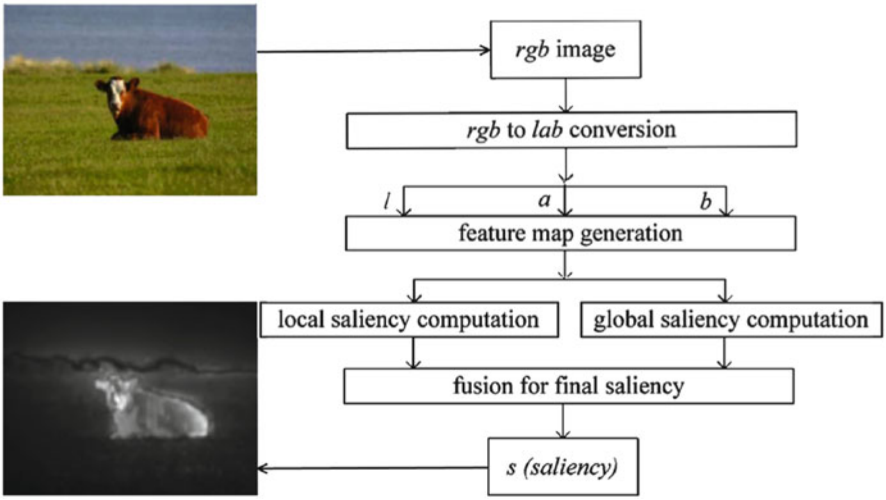

rgb image
rgb to *lab* conversion
l
a
b
feature map generation
local saliency computation
global saliency computation
fusion for final saliency
s (saliency)

Fig. 9.31 Schematic representation of the SDWT model. Top to bottom: RGB input image, feature map generation, local and global saliency computation, fusion, and saliency map (Adapted from [43])

Description: This mechanism first converts RGB to CIE LAB color space. Then, a 2D Gaussian filter is applied to remove noise. Third, a wavelet transform with increasing frequency bandwidths is employed to create the multiscale feature maps which can represent different features from edge to texture. After obtaining the feature maps, the method calculates the global distribution of local features to obtain both a global saliency map and a local saliency map by fusing the feature maps at each level without normalization operation. The final saliency map is a linear combination of these two maps (Fig. 9.31).

9.2 Conclusion: A Taxonomy of the Algorithms

Saliency models have been presented in this chapter with a simple taxonomy based on the historical development of methods: nineteen eye tracking-based algorithms as well as nine salient object detection-based models. This taxonomy has been constructed to present the studies and the validations of the saliency models which will be detailed in the next chapters but is not sufficient to classify models according to their structure.

This is why, in order to compare the saliency models inside each category, some characteristics have been added into the descriptive sheets. However, as explained in [44], the diversity of models makes taxonomy and comparison in the field of visual attention particularly difficult. The purpose of this section is to provide readers with a global view of each model characteristic.

9.2.1 Comparison of Eye Tracking-Based Models

Table 9.1 summarizes the four characteristics which have been chosen to compare nineteen eye tracking-based models. It shows which of the four characteristics each model owns.

In order to provide an idea of pros and cons of each characteristic, some observations have to be performed. The first characteristic divides models based on their global or local approach. The local approach has the advantage to properly detect high contrast while the global one highlights features which are different but not necessarily highly contrasted. The second characteristic classifies models according to their use (or not) of the center bias of gaze. This technique is particularly efficient when there are no particular salient regions or objects (e.g., landscapes) into the still images. Third, we use the categorization of Borji et al. [9] for saliency models which compare the attentive mechanism to obtain saliency map. This is the most popular taxonomy and some correlations with other features can be performed. Indeed, most of spectral techniques use the global approach, while cognitive, Bayesian, and information categories use the local one.

Finally, the last characteristics show if the still images are exploited with information from color or grayscale channels. Most of psychophysical theories show the importance of color during the visual attentive process. However, some techniques such as spectral transformation or orientation extraction cannot make use of all channels and only exploit the grayscale information.

To complete this analysis, the classical multidimensional scaling (MDS) [45] technique has been applied. MDS is a technique allowing to reduce the number of dimensions (N) necessary to convey or display the information contained in a distance matrix. In this chapter, it is a way to visualize in 2D ($N = 2$) the similarity level between the models. A distance matrix first needs to be calculated from the four characteristics. Table 9.2 shows an example of how we calculate a distance

Table 9.1 Comparison of nineteen eye tracking-based saliency models on four characteristics

	Approach	Post processing	Mechanism [9]	Stimuli
FSM	local		cognitive	color
GBVS	local	center	graphical	color
CCSA	local		cognitive	color
AIM	local		information	color
SDLF	local		bayesian	gray
SR	global		spectral	gray
SUN	local		bayesian	color
DVA	local		information	color
PFT	global		spectral	gray
SDSR	local		information	gray
VSLC	local		information	color
ESAL	global		graphical	color
SKSE	local	center	bayesian	color
AWS	global		other	color
SSAFD	global		spectral	color
ISSM	global		spectral	gray
QDCT	global		spectral	color
SERC	local	center	other	color
RARE	global		information	color

Table 9.2 Example of weight assignments for the calculation of a distance between two saliency models (AIM and SR) based on the four characteristics

	Lo	Go	PP	Co	Gr	In	Ba	Sp	Ot	Co	Gr
AIM	1	0	0	0	0	1	0	0	0	1	0
SR	0	1	0	0	0	0	0	1	0	0	1
D	1	1	0	0	0	1	0	1	0	1	1

($D = 6$) between two saliency models (AIM and SR). For each model, a weight of 1 is assigned to each characteristic the model owns and 0 otherwise. The distance is the sum of each weighted characteristic.

A distance matrix can be built by calculating each pairwise distance, and the MDS algorithm assigns two coordinates for each model so that the between-model distances are preserved as well as possible.

We can see from Fig. 9.32 a 2D MDS representation based on still image characteristics. The coordinates of this representation are components that represent a combination of characteristics. The first coordinate substantially corresponds to the first feature. Indeed, on one side (left), saliency models with local approach appear to have distances in the same range relatively to other models. On the other side (right), saliency models with global approach also seem to have distances in the same range. The second coordinate substantially corresponds to the last characteristic. Indeed, on one side (top), saliency models with color stimuli as input

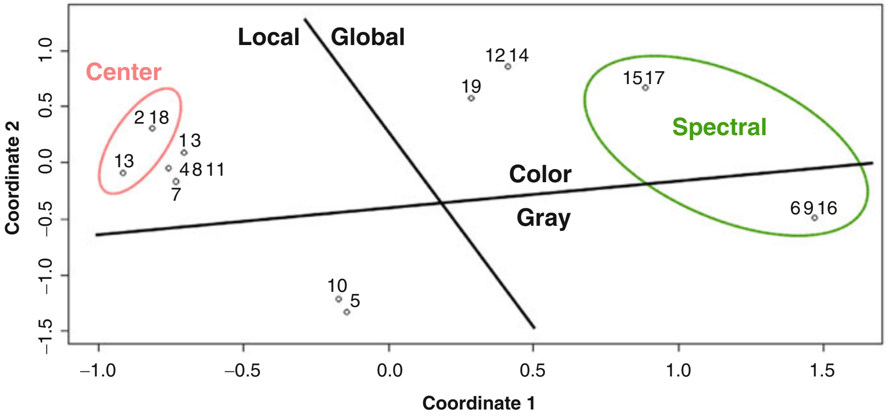


Fig. 9.32 MDS based on characteristics for nineteen ET models: 1.FSM / 2.GBVS / 3.CCSA / 4.AIM / 5.SDLF / 6.SR / 7.SUN / 8.DVA / 9.PFT / 10.SDSR / 11.VSLC / 12.ESAL / 13.SKSE / 14.AWS / 15.SSAFD / 16.ISSM / 17.QDCT / 18.SERC / 19.RARE. The first coordinate substantially corresponds to the local/global class while the second substantially represents color/grayscale as input. Two clusters can also be observed: center bias and spectral approach

are very close, while on the other side (bottom), saliency models with grayscale stimuli as input appear to have distances in the same range. We can also observe two clusters: one represents models with 2D centered Gaussian bias (models: 2, 13, and 18), while the other contains models with spectral mechanism (models: 6, 9, 15, 16, and 17).

9.2.2 *Comparison of Salient Object Detection-Based Models*

Table 9.3 summarizes the four characteristics which have been chosen to compare the nine salient object detection-based models. It shows which of the four characteristics each model owns.

As in Sect. 9.2.1, in order to give an idea of pros and cons of each chosen characteristic, some considerations have to be conducted. As in Sect. 9.2.1, the first characteristic compares the local approach which detects clearly contrast in images against the global approach which highlights features which are different but not necessarily highly contrasted. Some models use both complementary approaches. The second characteristic classifies models which take advantage or not of the superpixel segmentation which extracts perceptually homogeneous regions. A drawback of this method is the parameter adjustments which can often provide over- or under-segmentation of the scene. The third and fourth characteristics show how the stimuli are exploited. Indeed, respectively, we investigate if the saliency models use all the pixels of an image or patches to summarize the informations and if the still images are exploited with color information or directly with the JPEG bit stream information. The algorithms which use all the color pixel information must provide a more accurate contour of salient objects.

To complete this analysis, the same classical multidimensional scaling (MDS) technique as proposed above has been realized. The distances between models to compute this MDS are calculated from the four characteristics of Table 9.3. The purpose is to have a better visualization of the level of similarity between SOD models.

We can observe from Fig. 9.33 the 2D MDS representation based on still image characteristics. The coordinates of this representation are components that represent a combination of characteristics. The first coordinate substantially corresponds to the first characteristic. Indeed, on one side (left), saliency models with local approach appear to have distances in the same range, while on the other side (right), saliency models with global approach seem very close. We can also see the superpixel SP clusters (models: 3 and 7).

Table 9.3 Comparison of nine salient object detection-based models on four characteristics

	Approach	**Prior**	**Input**	**Stimuli**
FTSD	global		pixels	color
SSOI	local		pixels	color
SDBM	global	superpixels	pixels	color
SMSI	local		pixels	color
SDHAS	local		patches	color
SDAIR	global		patches	jpeg bit stream
SLMC	global	superpixels	pixels	color
SIM	local		pixels	color
SDWT	both		pixels	color

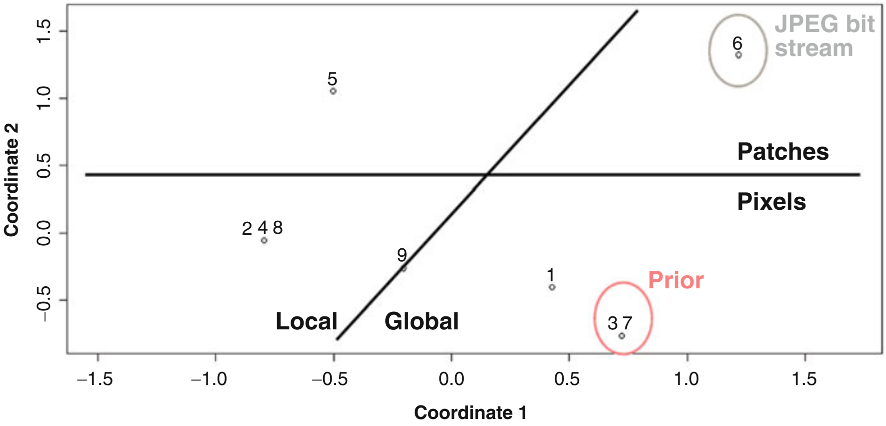


Fig. 9.33 Multidimensional scaling of nine salient object detection-based models based on characteristics in 2D: 1. FTSD / 2. SSOI / 3. SDBM / 4. SMSI / 5. SDHAS / 6. SDAIR / 7. SLMC / 8. SIM / 9. SDWT. The first coordinate substantially corresponds to the local/global class while the second substantially represents patch/pixels as input. The SP cluster can also be observed

9.3 Summary

- Nineteen models for eye tracking have been presented using descriptive sheets and will be use in the validation framework in the next chapters.
- Nine models for object segmentation have been introduced. They will be used in the studies in the next chapters.
- In order to compare the models, different characteristics have been chosen and classified them into some classes.
- A list of static state-of-the-art saliency models which are available online can be found from the Computational Attention Group of TCTS lab at http://tcts.fpms.ac.be/attention.

References

1. Avidan, S., & Shamir, A. (2007). Seam carving for content-aware image resizing. *ACM Transactions on graphics (TOG), 26*(3), 10. ACM.
2. Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1597–1604). IEEE, Miami.
3. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 20*(11), 1254–1259.
4. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013). Rare2012: A multiscale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication, 28*(6), 642–658.
5. Hou, X., & Zhang, L. (2008). Dynamic visual attention: Searching for coding length increments. In *Proceedings of Neural Information Processing Systems (NIPS)* (Vol. 5, p. 7), Vancouver.
6. Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis.
7. Harel, C. K. J., & Perona, P. (2006). Graph-based visual saliency. In *Proceedings of Neural Information Processing Systems (NIPS)*, Vancouver.
8. Tseng, P.-H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*(7), 4.
9. Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 35*(1), 185–207.
10. Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 28*(5), 802–817.
11. Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. *Proceedings of Neural Information Processing Systems (NIPS), 18*, 155–162.
12. Antonio Torralba, M. C., Oliva, A., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review, 113*(4), 766–786.
13. Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7), 32.
14. Guo, C., Ma, Q., & Zhang, L. (2008). Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8). IEEE, Anchorage.

15. Seo, H. J., & Milanfar, P. (2009). Nonparametric bottom-up saliency detection by self-resemblance. In *Conference on Computer Vision and Pattern Recognition (CVPR), 1st International Workshop on Visual Scene Understanding(ViSU)*. June 2009, Miami.
16. Yin Li, J. Y., & Zhou, Y. (2009). Visual saliency based on conditional entropy. In *The Asian Conference on Computer Vision (ACCV)*.
17. Avraham, T., & Lindenbaum, M. (2010). Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 32*(4), 693–708.
18. Tavakoli, H. R., Rahtu, E., & Heikkilä, J. (2011). Fast and efficient saliency detection using sparse sampling and kernel density estimation. In *Image analysis* (pp. 666–675). Berlin/Heidelberg: Springer.
19. Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R., & Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision, 12*(6), 17.
20. Li, J., Levine, M. D., An, X., Xu, X., & He, H. (2013). Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 35*(4), 996–1010.
21. Hou, X., Harel, J., & Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 34*(1), 194–201.
22. Schauerte, B., & Stiefelhagen, R. (2012). Predicting human gaze using quaternion dct image signature saliency and face detection. In *Proceedings of the 12th IEEE Workshop on the Applications of Computer Vision (WACV)/IEEE Winter Vision Meetings*, Breckenridge, Jan 2012 (pp. 9–11).
23. Erdem, E., & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision, 13*(4), 11.
24. Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of intelligence* (pp. 115–141). Dordrecht: Springer.
25. Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*(3), 5.
26. Li, Y., Zhou, Y., Xu, L., Yang, X., & Yang, J. (2009). Incremental sparse saliency detection. In *2009 16th IEEE International Conference on Image Processing (ICIP)* (pp. 3093–3096). IEEE, Cairo.
27. Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing, 30*(1), 51–64.
28. Schauerte, B., & Stiefelhagen, R. (2012). Quaternion-based spectral saliency detection for eye fixation prediction. In *Computer Vision–ECCV 2012* (pp. 116–129). Berlin/Heidelberg: Springer
29. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., & Shum, H.-Y. (2011). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 33*(2), 353–367.
30. Borji, A. (2015). What is a salient object? A dataset and a baseline model for salient object detection. *IEEE Transactions on Image Processing, 24*(2), 742–756.
31. Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The secrets of salient object segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 280–287). IEEE, Columbus.
32. Ma, Y.-F., & Zhang, H.-J. (2003). Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the Eleventh ACM International Conference on Multimedia* (pp. 374–381). ACM, Berkeley.
33. Liu, F., & Gleicher, M. (2006). Region enhanced scale-invariant saliency detection. In *International Conference on Multimedia and Expo (ICME)* (pp. 1477–1480). IEEE, Toronto.

34. Hu, Y., Rajan, D., & Chia, L.-T. (2005). Robust subspace analysis for detecting visual attention regions in images. In *Proceedings of the 13th Annual ACM International Conference on Multimedia* (pp. 716–724). ACM, Singapore.
35. Borji, A., Cheng, M.-M., Jiang, H., & Li, J. (2014). Salient object detection: A survey. *arXiv preprint arXiv:1411.5878*.
36. Rahtu, E., Kannala, J., Salo, M., & Heikkilä, J. (2010). Segmenting salient objects from images and videos. In *The European Conference on Computer Vision (ECCV)* (pp. 366–379). Springer, Heraklion.
37. Xie, Y., & Lu, H. (2011). Visual saliency detection based on bayesian model. In *International Conference on Image Processing (ICIP)* (pp. 645–648). IEEE, Brussels.
38. Vikram, T. N., Tscherepanow, M., & Wrede, B. (2012). A saliency map based on sampling an image into random rectangular regions of interest. *Pattern Recognition, 45*(9), 3114–3124.
39. Fang, Y., Lin, W., Lee, B.-S., Lau, C.-T., Chen, Z., & Lin, C.-W. (2012). Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum. *IEEE Transactions on Multimedia (MM), 14*(1), 187–198.
40. Fang, Y., Chen, Z., Lin, W., Lin, C.-W. (2012). Saliency detection in the compressed domain for adaptive image retargeting. *IEEE Transactions on Image Processing (TIP), 21*(9), 3888–3901.
41. Xie, Y., Lu, H., & Yang, M.-H. (2013). Bayesian saliency via low and mid level cues. *IEEE Transactions on Image Processing (TIP), 22*(5), 1689–1698.
42. Margolin, R., Zelnik-Manor, L., & Tal, A. (2013). Saliency for image manipulation. *The Visual Computer, 29*(5), 381–392.
43. Imamoglu, N., Lin, W., & Fang, Y. (2013). A saliency detection model using low-level features based on wavelet transform. *IEEE Transactions on Multimedia (MM), 15*(1), 96–105.
44. Bylinskii, Z., DeGennaro, E. M., Rajalingham, R., Ruda, H., Zhang, J., & Tsotsos, J. K. (2015). Towards the quantitative evaluation of visual attention models. *Vision Research, 116*, 258–268.
45. Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. New York: Springer Science & Business Media.

Chapter 10
Bottom-Up Saliency Models for Videos: A Practical Review

Nicolas Riche and Matei Mancas

10.1 Background

Research on visual saliency initially focused on still images rather than on video content. However, in the recent years, an increasing demand of video saliency appeared for some applications like gaming, editing, video retargeting, smart TV, robot navigation, surveillance, etc. Therefore, remarkable progress has been made first in the understanding on eye tracking data with dynamical stimuli and, in a second time, in the modeling process.

There are fundamental differences between videos and still images. For example, each video frame is only observed during a fraction of a second, while a still image can be viewed much longer. Some videos can feature varying camera motion such as tilting, panning, zooming, etc. For this reason, videos are probably viewed differently by human observers than still images, and some comprehensive comparative studies have emerged. In [1], for example, the authors study the influence of tasks on gaze behavior in static and dynamic scenes. In [2], the gaze on static and dynamic scene is compared; it also shows that the center bias decreases with dynamic stimuli.

In terms of modeling, static models have first been extended to video. This is the case for GBVS, SDSR, NMPT, or SSOI where authors added dynamic features to their models. Though these existing models are major contributions, video saliency estimation methods should then differ substantially from image saliency methods.

N. Riche (✉) • M. Mancas
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: nicolas.riche@umons.ac.be; matei.mancas@umons.ac.be

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_10

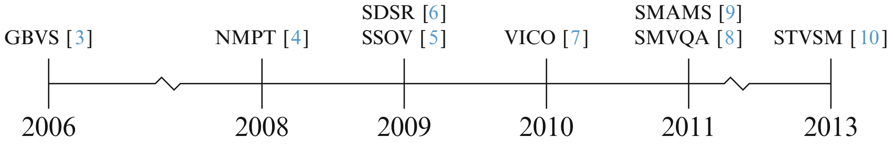


Fig. 10.1 Chronological overview of salient models for videos

Indeed, camera motions has a great impact on saliency estimation, and models need to be specifically designed to manage the temporal aspect. This is the case for STVSM or SMQVA.

In this section, the video attention models which will be used in the next chapters for saliency validation are described and discussed. In order to compare salient models for videos, four characteristics have been chosen and added into the descriptive sheets, following the color convention introduced by the colored keywords describing each characteristic below, for reader's convenience.

- The first characteristic such as for still images divides models based on their **approaches**. Some models have a **global** approach which is applied to the entire image, while others compute a saliency map with a **local** approach which is applied to a picture area.
- The second one classifies models which use or not **prior** information. As an improvement, some models practice some top-down factors (**TD**), a 2D centered Gaussian bias, a face recognition algorithm, or a segmentation at the end of the process.
- Third, the kind of **features** used to compute the saliency map classified the models. Indeed, some only use **static** features (colors, texture, etc.), while others compute **dynamical** features (motion, flicker, etc.). Some models can use **both** features.
- Finally, the last characteristic is similar to the last one for still images and shows if the **stimuli** are exploited either with all their channels (**color** images) or with just the **grayscale** information.

The eight saliency models for videos which are represented by their acronyms in Fig. 10.1 will be described in the following of this section and used in the validation framework.

GBVS: Graph-Based Visual Saliency (2006)

Characteristics: **local** | **HL** | **static** | **color**
Authors: J. Harel, C. Koch, and P. Penora [3].

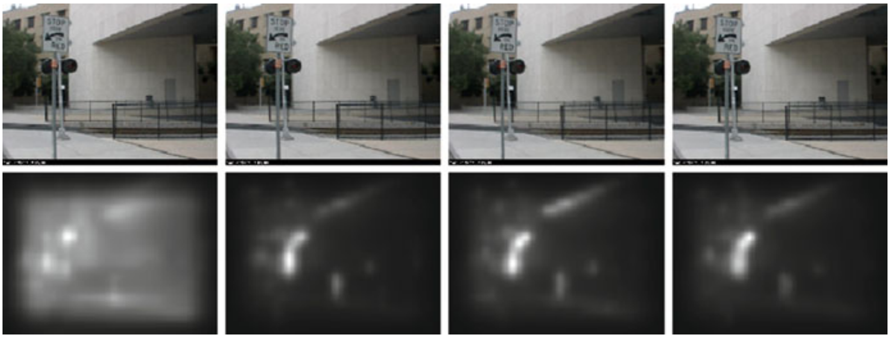


Fig. 10.2 Illustration of the GBVS method. On the first row, four frames of a video sequence. On the second row, the corresponding saliency maps (Inspired by [3])

Description: This model uses an approach similar to the model having the same name [3] for static scenes to create feature maps at multiple spatial scales and propose a Graph-Based Visual Saliency model (GBVS). There are again three main steps (Fig. 10.2), but during the feature extraction step, motion, and flicker channels can be added to compute the saliency maps of some video sequences. The algorithm then builds a fully connected graph over all grid locations of each feature map (intensity, orientation, color such as RGB or Derrington, Krauskopf, and Lennie (DKL) color space, motion, and flicker). Weights are assigned between nodes that are inversely proportional to the similarity of feature values and their spatial distance. A centered Gaussian is used to take advantage of the center bias and to improve the results.

NMPT: Nick's Machine Perception Toolbox (2008)

Characteristics: **local** | **/** | **static** | **color**
Authors: N. Butko, L. Zhang, G. Cottrell, and J. Movellan [4].

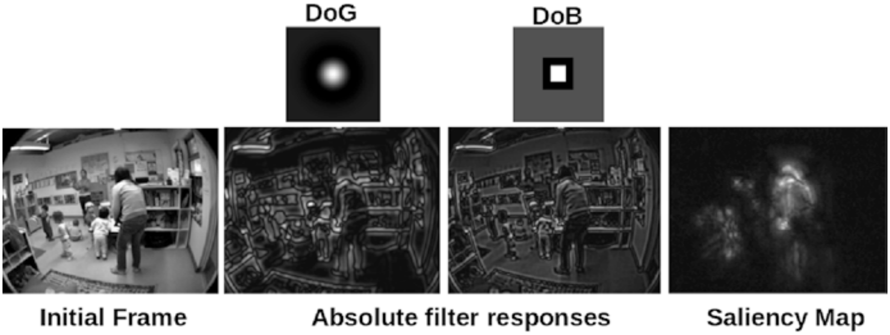


Fig. 10.3 The NMPT model computes saliency map using spatiotemporal filters on grayscale frame (left). The filters and their outputs are shown for the difference of Gaussian filter (second and third columns) and difference of Boxes approximation (fourth and fifth columns) (Adapted from [4])

Description: This algorithm proposes a fast approximation to dynamic scenes of the visual saliency model for still images proposed in [11] and called SUN (see Fig. 10.3). It introduces spatiotemporal filters and fits a generalized Gaussian distribution to the estimated distribution for each filter response. Spatiotemporal filters can be tuned with different settings to use only spatial, use only temporal, and use color contrast to be efficient and similar to the human visual system (HVS). The probability distributions of these spatiotemporal features were learned from a set of videos from natural environments. This model calculates its features and estimates the bottom-up saliency for each point.

SSOV: Segmenting Salient Objects for Videos (2009)

Characteristics: **local** | **/** | **static** | **color**
Authors: E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä [5].

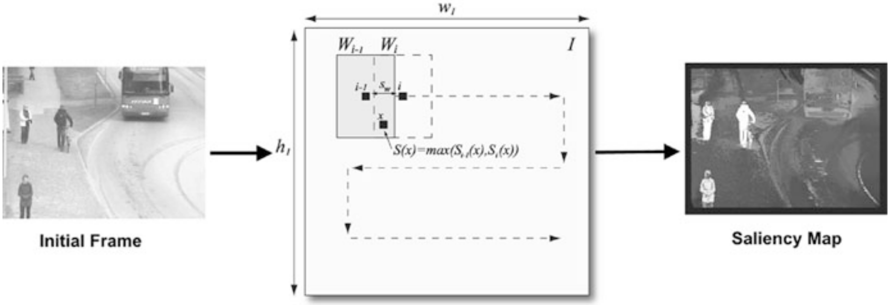


Fig. 10.4 Illustration of the SSOV method. From left to right: initial frame, an example of the sliding window applied to compute the saliency values, and saliency map (Adapted from [12])

Description: In order to adapt SSOI [5] from static scenes to video sequences, the CIE LAB perceptual color information of each frame is combined with the magnitude of the optical flow as input features at several scales (see Fig. 10.4). The optical flow was computed using an available online implementation [13]. The proposed saliency measure is formulated using a statistical framework and local feature contrast in motion, illumination, and color information. The final salient segments were computed using the energy function in the conditional random field (CRF) segmentation model for videos. The model is multiscale and does not require training, but the weight between the color space and motion intensity components has to be defined manually.

SDSR: Saliency Detection by Self-Resemblance (2009)

Characteristics: **local** | **/** | **static** | **grayscale**
Authors: H. J. Seo and P. Milanfar [6].

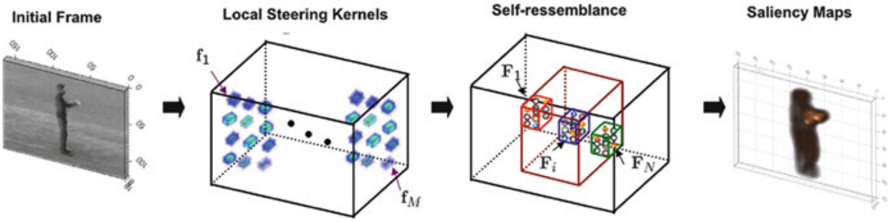


Fig. 10.5 Illustration of the SDSR method. From left to right: the grayscale video, space-time local steering kernels to compute feature maps from a space-time neighborhood, the self-resemblance algorithm, and the final space-time saliency map (Adapted from [6])

Description: The SDSR model is an approach similar to the model having the same name [14]. It uses local regression kernels as features (see Fig. 10.5). Kernel density estimation that estimates the distribution of the features in a patch is then applied. In statistics, the kernel density estimation is a nonparametric way to estimate the probability density function of a random variable. The time dimension is added to the static model to obtain a 3D local steering kernel to manage the case of video sequences. This model has the advantage to be robust to noise and other systemic perturbation.

VICO: VIsual COmpetitive Attention Model (2010)

Characteristics: **local** | **/** | **static** | **color**
Authors: M. Da Silva, V. Courboulay, A. Prigent, and P. Estraillier [7].

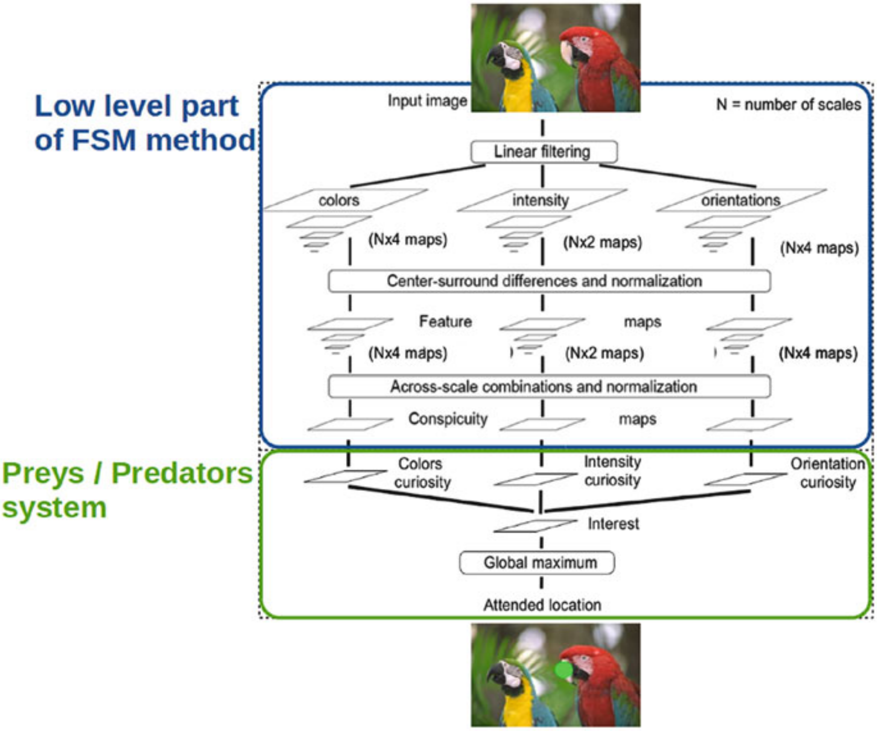


Fig. 10.6 Illustration of the VICO model. From top to bottom: input image, low level of the FSM method, preys-predators system, and attention location (Adapted from [7])

Description: This approach proposes a new version of the FSM model [15] for static scenes (see Fig. 10.6). The second part of FSM classical fusion is replaced by using preys-predators systems to merge conspicuity maps. The results reveal that preys-predators systems can help modeling visual attention and allow fast map generation while improving saliency map accuracy. VICO simulated the scan path of an observer across the frames of a video. Therefore, to obtain a density map at each frame, the model needs to be run multiple times (corresponding to the number of viewers by database) on the same video.

SMVQA: Salient Motion for Video Quality Assessment (2011)

Characteristics: **global** | **/** | **dynamic** | **grayscale**
Authors: D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, and D. Kukolj [8].

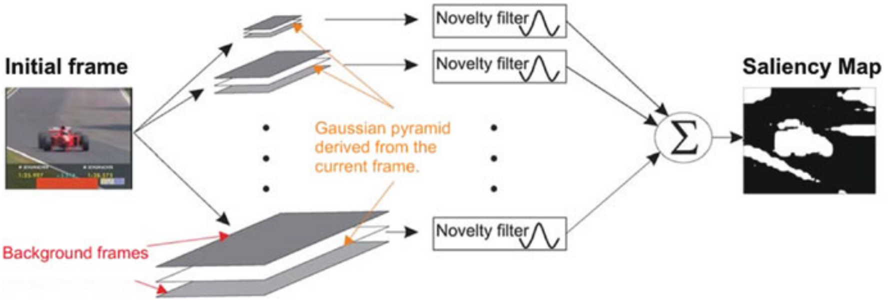


Fig. 10.7 Illustration of the SMVQA model. From left to right: initial frame, Gaussian pyramids derived from the current frame, novelty filters, sum, and saliency map (Adapted from [8])

Description: The SMVQA motion-based salient model has three main steps (Fig. 10.7): first, it uses a multiscale Gaussian pyramid derived from the current frame and two background frames as described in [16]. Novelty temporal filters are then performed on each pyramid level to indicate the extent to which the current frame differs from the background frames. Finally, the single saliency map is obtained by summing the score of the pixels from the filter outputs at different scales, and a modified Z-score test is used to detect the outliers in the frame. By efficiently managing the temporal information, this model detects cross-scale motion consistency, outlier, and temporal coherence on each frame and handles also videos with camera motion.

SMAMS: Saliency Models for Abnormal Motion Selection (2011)

Characteristics: **global** | **/** | **dynamic** | **grayscale**
Authors: M. Mancas, N. Riche, J. Leroy, and B. Gosselin [9].

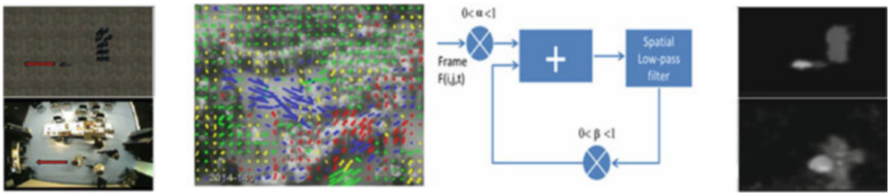


Fig. 10.8 Illustration of the SMAMS model. From left to right: synthetic and real video frames, optical flow applied on a frame, schematization of the 3D low-pass filtering, and the saliency maps for the corresponding input video frames (Adapted from [9])

Description: This algorithm proposes a model that detects abnormal motion. The SMAMS architecture has four main steps: first, motion features are extracted with an optical flow and output velocity and direction feature maps. Those two features are then spatiotemporally averaged with a 3D low-pass filter. The spatiotemporal averages separate each feature map into five bins at two different scales. Third, a self-information algorithm is computed for each map to highlight rare motion as salient. Indeed, the motion which is the most different in terms of speed and direction will have a higher saliency value as it is considered as abnormal. Finally, a fusion mechanism merges channels to give a single saliency map. As illustrated in the Fig. 10.8, some movements can be more salient than others. The model is effective for complex videos or dense crowds. Nevertheless, the model does not include any static cues as colors, for example.

STVSM: Spatiotemporal Visual Saliency Model (2013)

Characteristics: **local** | **HL** | **both** | **grayscale**
Authors: S. Marat, A. Rahman, D. Pellerin, N. Guyader, and D. Houzet [10].

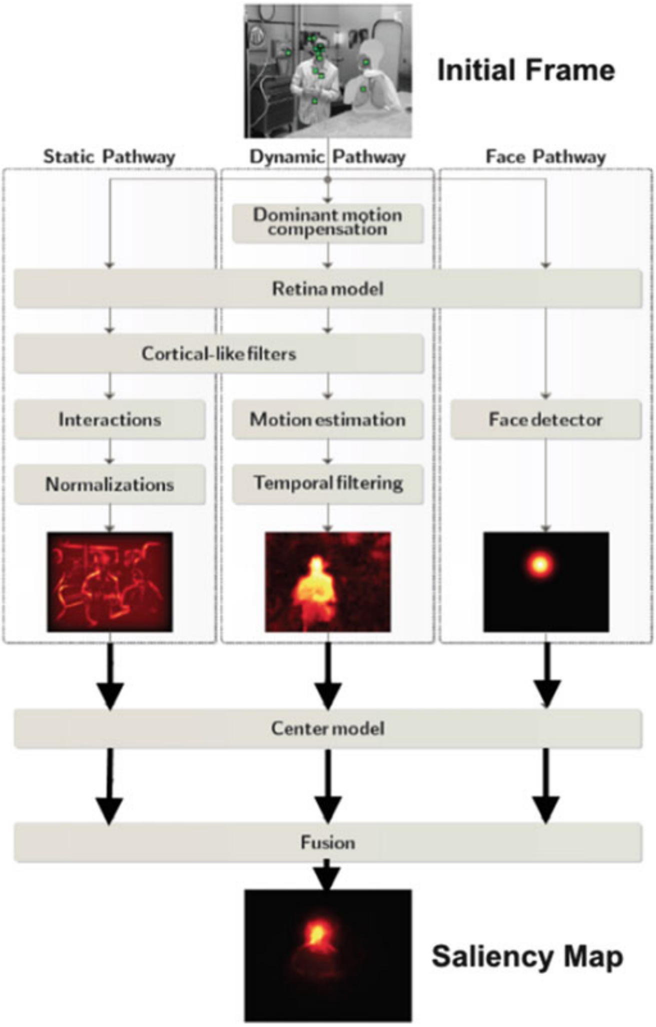


Fig. 10.9 Illustration of the STVSM model. Three pathways are computed from the grayscale input frame. From left to right: the static, the dynamic, and the face ones. A 2D centered Gaussian is then applied on each one before merging them to build the final saliency map (Adapted from [10])

Description: The STVSM model [10] is inspired by the biology of the visual system and breaks down each frame of a video into three maps: a *static* saliency map emphasizes regions that differ from their context in terms of luminance, orientation, and spatial frequency. A *dynamic* saliency map emphasizes moving regions with values proportional to motion amplitude. A *face* saliency map emphasizes areas where a face is detected with a value. Finally, a 2D centered Gaussian is applied on each map and fuses all of them into a single saliency map (see Fig. 10.9).

ST-RARE: Spatiotemporal Multiscale Rarity Mechanism (2013)

Characteristics: **global** | / | **both** | **color**
Authors: M. Decombas, N. Riche, F. Dufaux, B. Pesquet-Popescu, M. Mancas, B. Gosselin, and T. Dutoit *et al.* [17].

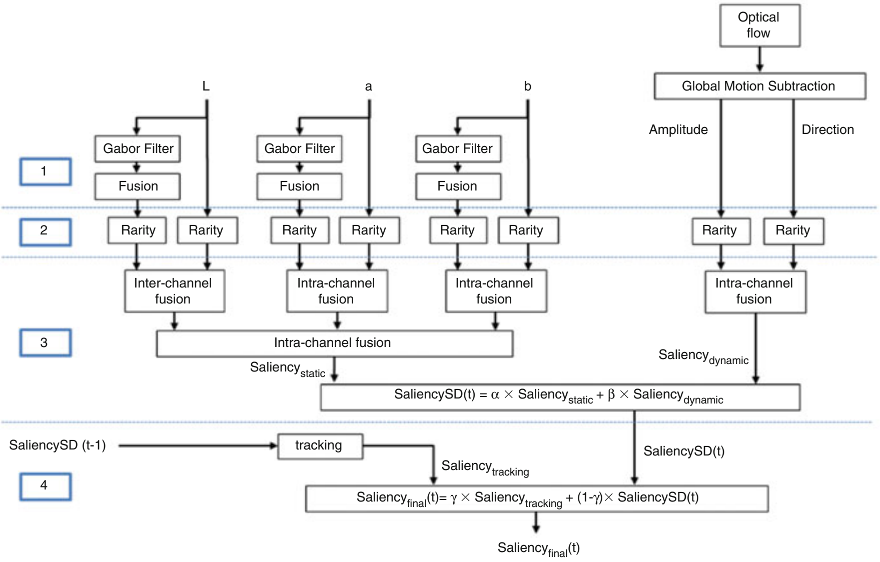


Fig. 10.10 Overview of the ST-RARE saliency model. From top to bottom: (1) feature extraction, (2) multiscale rarity mechanism, (3) fusion steps, and (4) tracking and temporal filtering (the static features are on the left, while the dynamic features are on the right) (Adapted from [17])

Description: The ST-RARE model combines spatial and temporal information to provide the map saliency (see Fig. 10.10). First, six spatial feature maps, three low level (which are the colors from the first path) and three medium level (the orientation and texture information coming from the Gabor filters), and two temporal feature maps: Motion amplitude and direction are extracted from video frame. Then, a multiscale is used on each feature map, and a fusion algorithm provides the saliency map. The last step is the temporal tracking framework in order to improve temporal coherence and robustness.

STRAP: Spatiotemporal Multiscale Rarity Algorithm with Priors (2013)

Characteristics: **global** | **HL** | **both** | **color**
Authors: M. Decombas, N. Riche, F. Dufaux, B. Pesquet-Popescu, M. Mancas, B. Gosselin, and T. Dutoit [17].
Description: STRAP is a new saliency model based on a spatiotemporal rarity mechanism and integrating prior information. It builds upon ST-RARE and includes several novel contributions: (1) a temporal motion compensation over a sliding window. In this way, neighboring frames can be jointly analyzed to increase temporal robustness; (2) color and frequency-based low-level priors are used together with the rarity mechanism, and the fusion algorithm is optimized to this new architecture; (3) high-level priors like a centered Gaussian or face detection are then combined with the saliency results; and (4) a spatiotemporal segmentation is finally used to improve the accuracy of the results and better detect the objects of interest. The method pipeline can be seen on Fig. 10.11.

10.2 Conclusion: A Taxonomy of the Algorithms

Saliency models for videos which will be used in the validations in the next chapters have been presented with descriptive sheets to provide readers with a global view of each model. However, as seen for still images in the previous chapter, this is not sufficient to classify dynamic models according to their structure. This is why some characteristics have been added into the descriptive sheets. Table 10.1 summarizes these four characteristics which have been chosen to compare the ten state-of-the-art saliency models for videos. It shows which of the four characteristics each model owns.

In order to provide an idea of pros and cons of each characteristic, some observations have to be conducted. The first characteristic for still images compares the local approach which detects clearly contrast in images against the global approach which highlights features which are different but not necessarily highly contrasted. The second characteristic classifies models which use or not top-down information. Saliency models can add some modules at the end of the process considered as top-down factors such as a 2D centered Gaussian, a face detector, or a segmentation algorithm. The purpose is to better detect the salient areas and therefore to improve the scores. However, if these modules are inappropriately used,

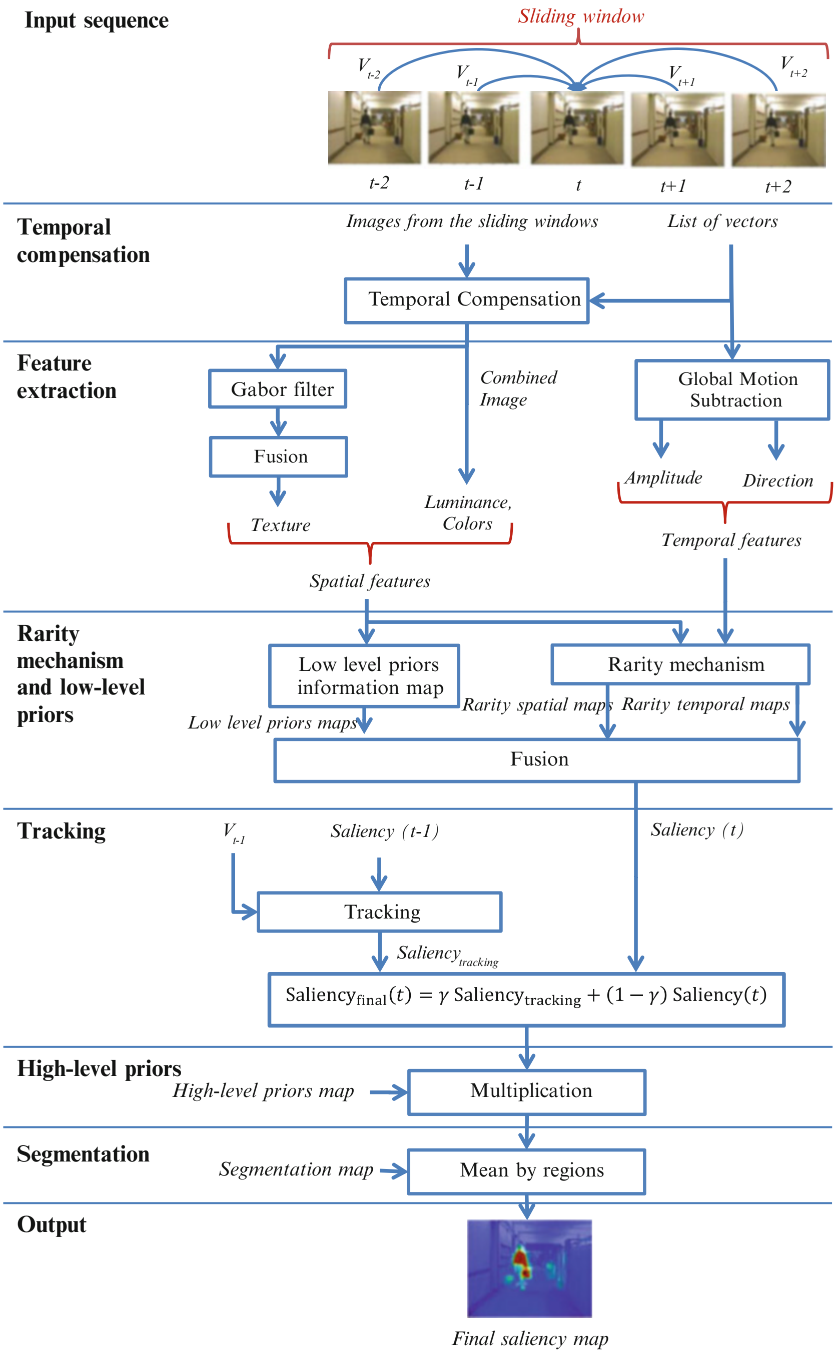


Fig. 10.11 Overview of the STRAP saliency model. From top to bottom: (1) temporal compensation, (2) feature extraction, (3) multiscale rarity mechanism and priors, (4) tracking, (5) high-level priors, and (6) segmentation (Adapted from [17])

Table 10.1 Comparison of eight saliency models for videos on seven characteristics

	Approach	Prior	Feature	Stimuli
GBVS	local	TD	static	color
NMPT	local		static	color
SSOV	local		both	color
SDSR	local		static	gray
VICO	local		static	color
SMVQA	global		dynamic	gray
SMAMS	global		dynamic	gray
STVSM	local	TD	both	gray
ST-RARE	global		both	color
STRAP	global	TD	both	color

they will do just the opposite. It is important to correctly weigh the 2D centered Gaussian, to adjust the parameters of the segmentation algorithm, or to choose a face detector with a lower false-positive rate.

The third characteristic shows which kinds of features are extracted to compute the saliency maps. Some models only use static features, while others compute only dynamical features. Finally, some models can combine both kinds of features. This last class of models is able to predict salient areas when there is or no motion in the videos, while models with only static features cannot detect motion, and models with only dynamic features cannot detect salient areas when there is no motion in the videos. Finally, the last characteristic shows how the stimuli are exploited: with color or grayscale informations. Although, most of the psychophysical theories show the importance of color during the visual attentive process and the color information is used in many saliency models, its contribution for saliency modeling in videos was less clear. However, some studies such as [18] show the importance of color information which helps to better predict fixation distribution in videos than models which only exploit the grayscale information.

To complete this comparison, the classical multidimensional scaling (MDS) technique similar to the one exposed in Chap. 9 has been chosen. The distances of this MDS are computed from the characteristics of Table 10.1. The purpose is to have a better visualization of the level of similarity between saliency models for videos.

We can see from Fig. 10.12 a 2D MDS model representation based on video characteristics. The coordinates of this representation are components that represent a combination of characteristics. The first coordinate substantially corresponds to the first characteristic. Indeed, on one side (right), saliency models with local approach appear to have distances in the same range relative to other models. On the other side (left), saliency models with global approach also seem to have distances in the same range. The second coordinate substantially corresponds to the last characteristic. Indeed, on one side (bottom), saliency models with color stimuli as input are very close, while on the other side (top), saliency models with grayscale stimuli as input appear to have distances in the same range. These observations divide the presented models into four categories (from C1 to C4 on Fig. 10.12).

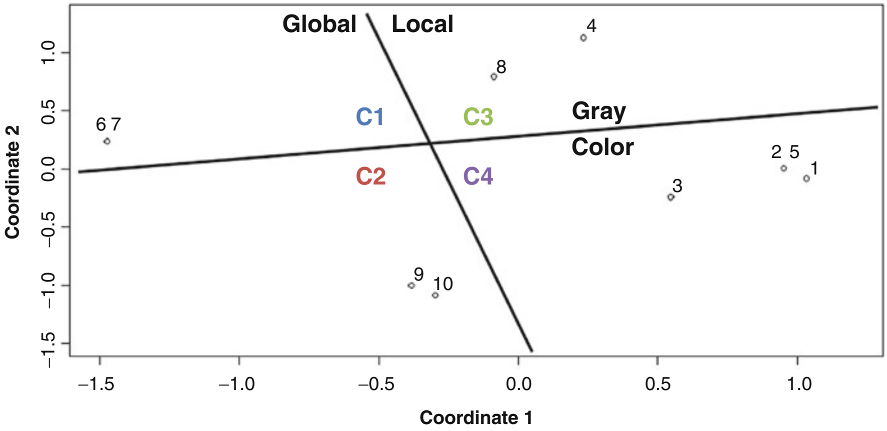


Fig. 10.12 Multidimensional scaling of ten saliency models for videos based on characteristics in 2D: (1) GBVS, (2) NMPT, (3) SSOV, (4) SDSR, (5) VICO, (6) SMVQA, (7) SMAMS, (8) STVSM, (9) ST-RARE, and (10) STRAP. The first coordinate substantially corresponds to the local/global class, while the second substantially represents color/grayscale as input

10.3 Summary

- Ten models for videos are described using descriptive sheets and will be used in the validation framework in the next chapters.
- Some models like GBVS and SSOV are extensions from 2D, while others are temporal models.
- In order to compare the models, different characteristics have been chosen and classified them into several classes.
- A list of dynamic state-of-the-art saliency models which are available online can be found from the Computational Attention Group of TCTS Lab at http://tcts.fpms.ac.be/attention.

References

1. Smith, T. J., & Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision, 13*(8), 16.
2. Nguyen, T. V., Xu, M., Gao, G., Kankanhalli, M., Tian, Q., & Yan, S. (2013). Static saliency vs. dynamic saliency: A comparative study. In *Proceedings of the 21st ACM International Conference on Multimedia* (pp. 987–996). ACM.
3. Harel, C. K. J., & Perona, P. (2006). Graph-based visual saliency. *Proceedings of Neural Information Processing Systems (NIPS).*
4. Butko, N. J., Zhang, L., Cottrell, G. W., & Movellan, J. R. (2008). Visual saliency model for robot cameras. In *International Conference on Robotics and Automation (ICRA)* (pp. 2398–2403). IEEE.

5. Rahtu, E., Kannala, J., Salo, M., Heikkilä, J. (2010). Segmenting salient objects from images and videos. In *The European Conference on Computer Vision (ECCV)* (pp. 366–379). Springer.
6. Seo, H. J., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Jounal of Vision, 9(12)*(15), 1–27.
7. Da Silva, M. P., Courboulay, V., Prigent, A., & Estraillier, P. (2010). Evaluation of preys/predators systems for visual attention simulation. In *The International Conference on Computer Vision Theory and Applications (VISAPP)* (Vol. 2, pp. 275–282). INSTICC.
8. Culibrk, D., Mirkovic, M., Zlokolica, V., Pokric, M., Crnojevic, V., & Kukolj, D. (2011). Salient motion features for video quality assessment. *IEEE Transactions on Image Processing (TIP), 20*(4), 948–958.
9. Mancas, M., Riche, N., Leroy, J., & Gosselin, B. (2011). Abnormal motion selection in crowds using bottom-up saliency. In *International Conference on Image Processing (ICIP)* (pp. 229–232). IEEE.
10. Marat, S., Rahman, A., Pellerin, D., Guyader, N., & Houzet, D. (2013). Improving visual saliency by adding "face feature map" and "center bias". *Cognitive Computation, 5*(1), 63–75.
11. Zhang, L., Tong, M. H., Marks, T. K., Shan, H., Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7), 32.
12. Rahtu, E., & Heikkilä, J. (2009). A simple and efficient saliency detector for background subtraction. In *2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)* (pp. 1137–1144). IEEE.
13. Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., & Bischof, H. (2009). Anisotropic huber-l1 optical flow. *Proceedings of the British Machine Vision Conference (BMVC), 1*(2), 3.
14. Seo, H. J., & Milanfar, P. (2009). Nonparametric bottom-up saliency detection by self-resemblance. In *Conference on Computer Vision and Pattern Recognition (CVPR), 1st International Workshop on Visual Scene Understanding(ViSU)*, June 2009.
15. Itti, L., Koch, C., Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 20*(11), 1254–1259.
16. Crnojević, V., Antić et al., B. (2009). Multiscale background modelling and segmentation. In *International Conference on Digital Signal Processing (DSP)* (pp. 1–6). IEEE.
17. Decombas, M., Riche, N., Dufaux, F., Pesquet-Popescu, B., Mancas, M., Gosselin, B., & Dutoit, T. (2013). Spatio-temporal saliency based on rare model. In *International Conference on Image Processing (ICIP)* (pp. 3451–3455). IEEE.
18. Hamel, S., Guyader, N., Pellerin, D., & Houzet, D. (2014). Color information in a model of saliency. In *Proceedings of the European Signal Processing Conference (EUSIPCO)* (pp. 226–230). IEEE.

Chapter 11
Databases for Saliency Model Evaluation

Nicolas Riche

The comparison between saliency algorithms needs two prerequisites: a dataset of stimuli with a ground-truth on which the algorithms can be compared and a metric which measures in an objective way how close an algorithm and the ground-truth are.

This chapter focuses on the stimuli datasets and the ground-truths. In computer vision, the databases for the modeling of visual attention contain two kinds of ground-truth: eye movement recording and salient region labeling. The stimuli are still images or videos.

11.1 Introduction

In this section, an overview of the databases which are available online is exposed under three categories: 1) still images along with eye tracking data, 2) still images along with salient object detection and 3) videos. It is important to note that all databases were collected with different experimental settings. Some studies [1] investigate on how the type of stimulus (e.g., fractal, website, advertising, and natural images) affects saliency models. On the other hand, there

N. Riche (✉)
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: nicolas.riche@umons.ac.be

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_11

are papers [2] which investigate how experiments influence gaze during scene observation with different viewing tasks (e.g., free viewing where observers are looking at the images or videos without any task or task-based viewing where observers are asked to perform a task while looking to the images or videos like finding people for example). In the proposed saliency assessment framework, we focus on free-viewing databases on natural images because we want to test saliency models which modelize the bottom-up part of visual attention.

The first databases used to validate saliency models had eye-tracking data as ground truth. A complete overview of these datasets is available in [3]. In this paper, 15 databases on images are given. Hereinafter, a first section will present the most popular datasets and a second section the ones we will use for the validation in the next chapters.

Main datasets: In 2006, N. Bruce proposed a database called Toronto [4]. It contains 120 natural scene images with free-viewing eye movement recordings from 20 users. Each image has been seen during 4 s. Data consists of a variety of indoor and outdoor scenes, some with very salient items, others with no particular regions of interest. In 2006, O. Lemeur created a new database [5] with 27 color images. Each image was seen by 40 observers for 15 s. In 2007, M. Cerf proposed a specific database with a lot of faces in images, called FiFA [6]. The purpose was to demonstrate that faces attract visual attention. The database contains 200 images viewed by 8 observers during 2 s. The probability to find a fixation on faces within the first two fixations is over 80 %. In 2009, U. Engelke built a visual attention database [7] for image quality. Forty-two images (14 images with 3 levels of quality) have been viewed by 15 observers during 12 s. The purpose was to validate that salient image regions contribute to objective image quality metrics. In the same year, T. Judd collected a large database of eye-tracking data [8]. One thousand and three random images from Flickr creative commons and LabelMe [9] have been seen by 15 subjects during 3 s. There are approximately 77 % landscape images and 23 % portrait images. In 2010, S. Ramanathan built a database [10] with 758 images. Each image has been viewed by 25 subjects during 5 s. Face, portrait, nude, action, affect-variant group, and other concepts were the diverse themes covered in the dataset. In 2011, T. Judd proposed the MIT low-resolution saliency database [11]. The purpose was to study how image resolution affects consistency in eye fixations across observers. To do that, 168 images (21 images with 8 resolutions) have been seen by 8 subjects during 3 s. The main observation was that the center bias increases as image resolution is reduced. In 2011, J. Li collected eye tracking of 235 color images viewed by 21 users [12]. This dataset is divided into six categories about the size of the salient objects.

Datasets used for validation: We now focus on the databases used in the validation framework described in the next chapters. A short overview based on descriptive sheet template is provided for each dataset.

TORONTO Dataset (2006)

Authors: N. Bruce and J. Tsotsos [4]

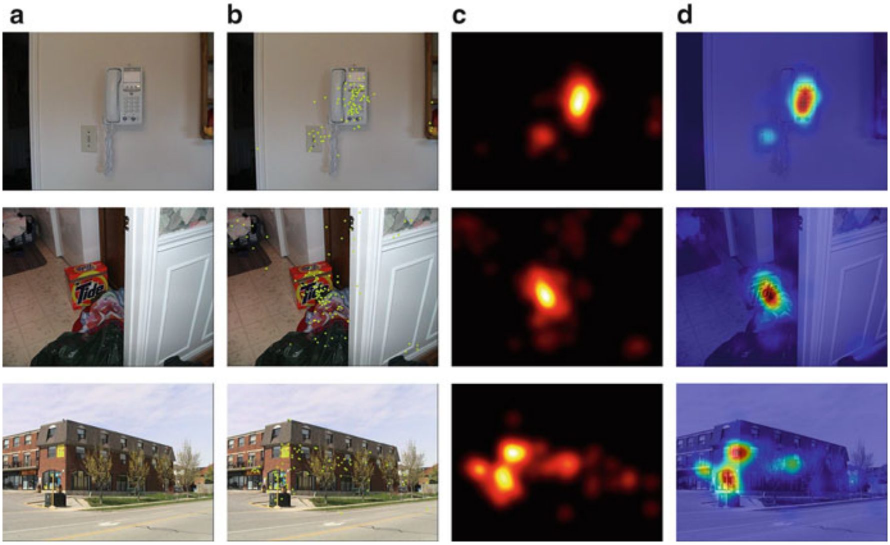


Fig. 11.1 Illustration of the TORONTO dataset with rows 1–2 (indoor scenes) and row 3 (outdoor scene) with no particular regions of interest (**a**) Original images (**b**) Fixation maps (**c**) Density maps (**d**) Heat maps

Description: This database contains the free viewing of 120 different color images from eye-tracking experiments. Images were presented in random order for 4 s each with a white screen between each pair of images. Data consists of a variety of indoor and outdoor scenes, some with very salient items, others with no particular regions of interest. It was collected from 20 subjects (Fig. 11.1).

MIT Dataset (2009)

Authors: T. Judd, K. Ehinger, F. Durand, and A. Torralba [8]

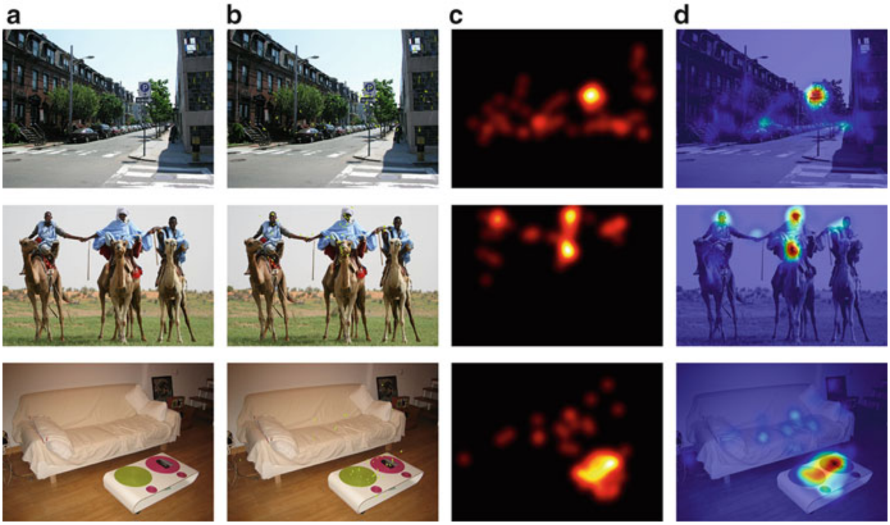


Fig. 11.2 Illustration of the MIT dataset with row 1 (street scenes), row 2 (image with animals and persons), and row3 (indoor scenes) (**a**) Original images (**b**) Fixation maps (**c**) Density maps (**d**) Heat maps

Description: This database contains 1003 random images from Flickr creative commons and LabelMe [9]. Data was collected from 15 subjects who freely viewed these images during 3 s separated by 1 s of viewing a gray screen. There are several categories such as text, faces, and indoor or outdoor scenes for approximately 77 % landscape images and 23 % portrait images (Fig. 11.2).

KOOTSTRA Dataset (2009)

Authors: G. Kootstra and L. Schomaker [13]
Description: This database has 99 natural images viewed by 31 observers during 5 s. The data consists of five different categories: 19 images of natural symmetrical objects, 12 images of animals in a natural environment, 12 images of street scenes, 16 images of buildings, and 40 images of natural environments. All these images were taken from the McGill database [14] (Fig. 11.3).

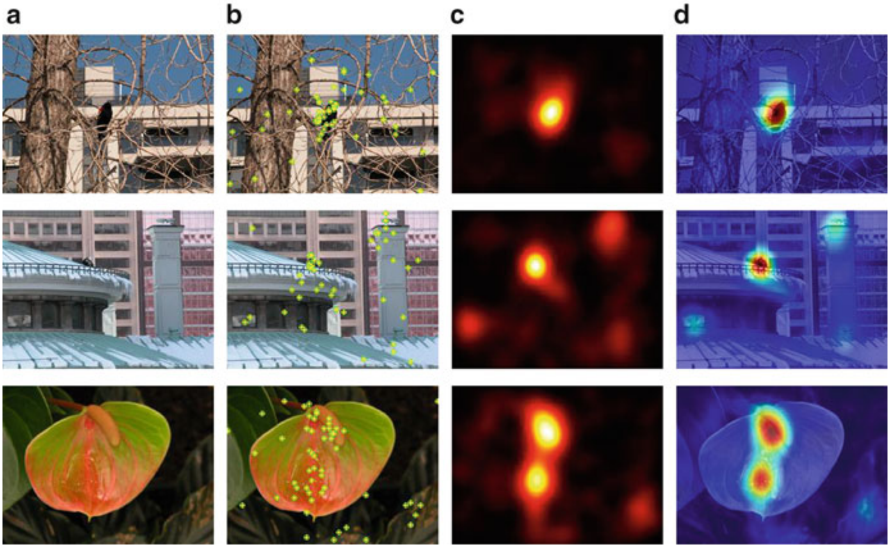


Fig. 11.3 Illustration of the KOOTSTRA dataset with row1 (image with animals), row2 (building scenes), and row3 (symmetrical images such as flowers) (**a**) Original images (**b**) Fixation maps (**c**) Density maps (**d**) Heat maps

IMGSAL Dataset (2011)

Authors: J. Li, L. Martin, A. Xiangjing, and H. Hangen [12]
Description: This database contains 235 images with both large and small salient objects. The images are divided into six categories: 50 images with large salient objects, 80 with intermediate salient objects, 60 with small salient objects, 15 with cluttered backgrounds, 15 with repeating distractors, and 15 images with both large and small salient objects (Fig. 11.4).

11.1.1 Salient Object Detection (SOD)-Based Datasets with Still Images

The first ground truth used to validate saliency models is eye-tracking data. Since 2007, a second kind of ground truth has appeared. With the development of various saliency-based applications, salient object detection has emerged. Therefore, a second ground truth used for model validation is the ability of the models to detect salient objects in natural scenes. The salient objects in images can be annotated with two types of masks: bounding boxes (rectangles around the objects) or pixel-wise (accurate contours of the objects) binary masks.

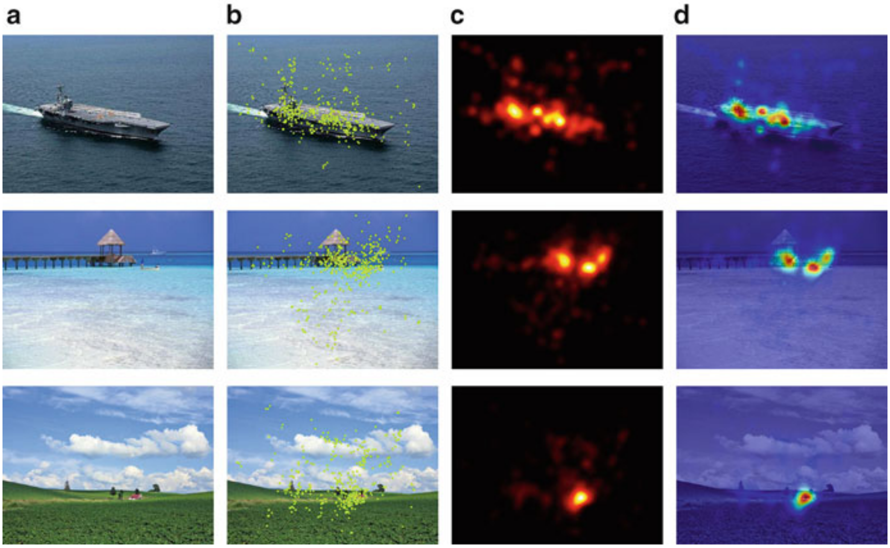


Fig. 11.4 Illustration of the IMGSAL dataset with three sizes of salient objects in natural scenes: large (row1), medium (row2), and small (row3) objects. (**a**) Original images (**b**) Fixation maps (**c**) Density maps (**d**) Heat maps

As discussed in [15] and [16], the first attempts to build these kind of databases faced two main drawbacks: the number of objects and the background. Indeed, most of these images contain only one object with a simple background. While one single object is more linked to the foreground/background notion, the concept of salient object exists when multiple objects are present in a scene. Therefore, new versions of these databases have been built with more than one object and complex backgrounds. They often use the same images as the ones that already have eye tracking data. The most recent and complete databases have thus two ground truth: the eye-tracking data and the binary masks around salient objects.

An overview of the most widely used datasets is available in [16]. In this paper, 20 databases on images have been listed. Hereinafter, a first section will present the most used datasets for SOD and a second section the ones that we will use for the study in the next chapters.

Main datasets: In 2007, the first dataset with a large number of images manually annotated with bounding boxes has been released and called MSRA [17]. It contains two parts: the salient objects have been shown by bounding boxes from 3 subjects on 20,000 images. Among them, 5000 images have been selected and annotated by 9 subjects. In 2009, R. Achanta selected 1000 images among the 5000 ones and proposed a database [18] which contains binary pixel-wise object masks for each image. This is one of the most popular datasets, but images have only one salient object and clean background.

More recently, some eye-tracking databases have been selected to be annotated and partially fixed the issue of having only one salient object and clean background. In 2011, J. Li proposed a new database by providing pixel-wise objects from 19 subjects in addition to eye movement recording on the 235 images of ImgSal database [12]. This is one of the first databases with both binary masks and human fixations. The drawback is the limited number of images in the database (235) compared to MSRA (5000), for example.

In 2013, A. Borji created masks for two other well-known eye-tracking databases: Toronto-A [15, 19] and MIT-A [15, 19]. While Toronto-A proposes 120 annotated images, MIT-A provides 900 images with masks. These masks were created by two participants which manually outlined objects, and the most salient one was selected by the peak of the human fixation map. In the same year, with the same motivation to solve the issues of [18], Q. Yan extended his Complex Scene Saliency Dataset (CSSD – 200 images) to a larger dataset (ECSSD) [20] with 1000 images. These images are collected from BSD300 [21], VOC dataset [22], and the Internet to represent more general situations that natural images fall into. Five users produce the ground-truth pixel-wise masks. Finally, C. Yang proposed the most complete database in 2013 called DUT-OMRON [23]. This database is the only one which has the eye fixations, salient objects bounding box, and the pixel-wise salient objects segmentation ground truth. 5000 images have been seen by five subjects during 2 s. Each image has been annotated by five participants which can draw several rectangles to enclose most salient objects in the image and the authors provide pixel-wise ground truth for all images. Although this database is a major contribution, five users are not yet enough to provide an accurate eye-tracking distribution. Moreover, a lot of images have only one single object.

In 2014, Y. Li proposed a new database called PASCAL-S [24] with both ground truths. This dataset is built on [25]. It contains 850 natural images viewed by 8 subjects during 2 s. To build the masks, the authors first manually perform a full segmentation and then ask 12 subjects to label the salient objects by clicking on them. The final saliency value of each segment is the total number of clicks it receives, divided by the number of observers. In this kind of ground truth, grayscale masks can be built based on these ratios. In the same year, J. Xu built a very complete database called OSIE [26]. Seven hundred images have been seen by 15 viewers during 3 s and 5551 objects have been segmented with precise contours. Moreover, this database proposed 12 semantic attributes (emotion, touch, smell, etc.).

Datasets used in this study: We now focus on the database used in the studies proposed in the next chapter. A short overview based on a descriptive sheet template is provided for each dataset.

IMGSAL Dataset (2011)

Authors: J. Li, L. Martin, A. Xiangjing, and H. Hangen [12]

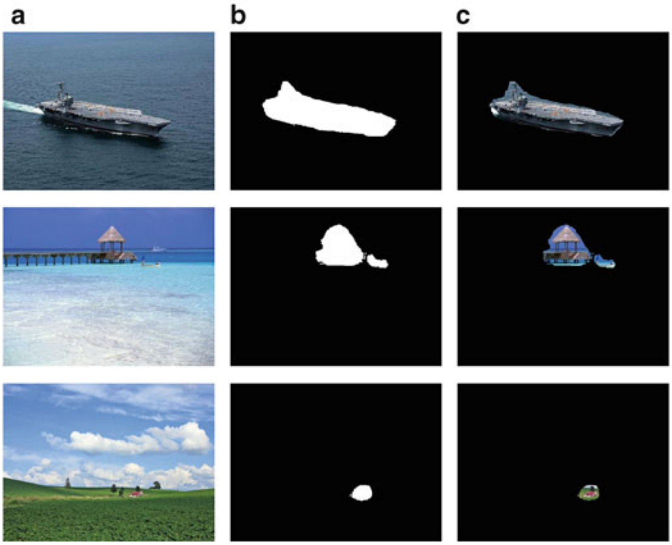


Fig. 11.5 Illustration of the IMGSAL dataset on the three sizes of objects in natural scenes: large (row1), medium (row2), and small (row3) objects (**a**) Original images (**b**) Mask maps (**c**) Object maps

Description: The database, called *ImgSal* in the previous paragraph, was labeled by 19 naive subjects. This labeling process is inspired by LabelMe [9] and Hou's method [27]. Images were presented one by one in a random manner and each observer was sat in front of the screen at a distance of six times the image width. After having viewed the image, the subject labeled the image immediately (Fig. 11.5).

11.1.2 ET and SOD-Based Datasets with Videos

Saliency video datasets are less well developed. Only two databases exist for evaluating salient object detection models in video [16], and there are few databases containing both video and eye-tracking data [3]. In Winkler's paper, ten databases on videos with eye tracking are presented [3]. More video datasets are required in the literature for the coming years, especially with both ground truths. Hereinafter, a first section will present the most used datasets and a second section the ones we will use for the validation in the next chapters.

Main datasets: The most widely used is CRCNS-ORIG Itti's video database released in 2004 [28–31]. This database contains 50 video clips along with eye-tracking data for eight viewers. The video contains complex stimuli like TV programs, outdoor videos, or video games.

In 2009, J. Li built PKU-RSD (Regional Saliency Dataset) as explained in [32]. A total number of 431 short videos have been annotated by 23 subjects with bounding boxes. The videos contain various scenes like surveillance, news, or cartoons. In 2010, M. Dorr provided a large database with eye movement on natural and Hollywood movies but also static images [33]. So, this is a dataset with both images and videos.

In 2011, Y. Wu proposed the second database with salient region ground truth [34]. It contains 32 video segments collected from the Internet. All the frames have been annotated with object-bounding boxes. In the same year, P. Mital proposed the DIEM database [35] which contains 85 videos along with corresponding eye data collected from 250 participants but who did not necessarily viewed all the videos.

In 2012, S. Mathe complemented two large-scale video datasets: Hollywood-2 [36] and UCF Sports [37] with human eye movements to build a new database called Actions in the Eye [38]. This is the first video eye tracking with significant size: 92 h of video and each frame viewed by 16 subjects (12 active for task recognition and 4 passive for free viewing). In the same year, an eye-tracking database, called Standard Video [39] with 12 standard videos used in image compression and quality estimation viewed by 15 people, was proposed.

Datasets used for validation: We now focus on the database used in the validation framework. A short overview based on a descriptive sheet template is provided for each dataset.

SVS Dataset (2012)

Authors: H. Hadizadeh, M. Enriquez, and I. Bajic [39]

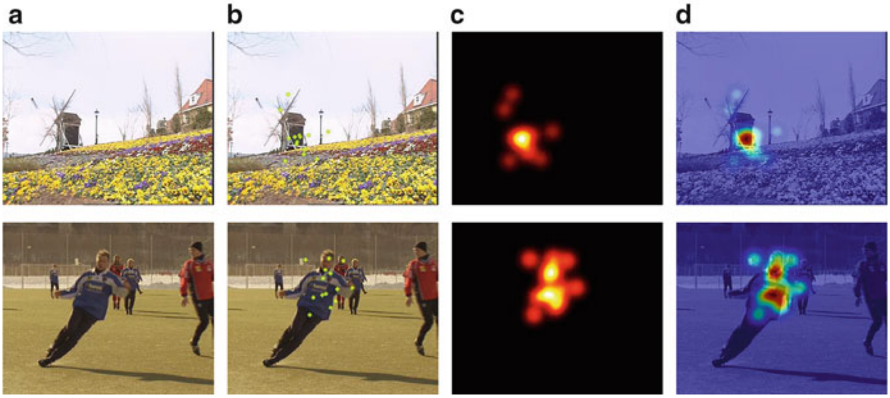


Fig. 11.6 Illustration of the SVS dataset: landscape (row1) and people (row2). (**a**) Original images (**b**) Fixation maps (**c**) Density maps (**d**) Heat maps

Description: The database is composed of 12 video sequences, 6 of them including people. It encompasses varying characteristics, including moving camera, cluttered background, and complex motion. Sequences are stored in YUV format, with 4:2:0 chrominance sampling. The uncompressed format avoids artifacts introduced by compression that may adversely disturb saliency models. The purpose is to evaluate where someone looks at when discovering the content for the first time (Fig. 11.6).

Extended SVS Dataset (2015)

As shown in Sect. 11.1.2, only two databases exist for salient object detection and only on compressed videos. To remedy this situation, based on the uncompressed videos of [39], the database has been extended here in terms of ground truth. Indeed, one limitation of the human fixation ground-truth data is that the eye-tracking data sometimes highlights the border of a salient object which can lead to the assignment of high saliency scores to not only the object but also to the surrounding background. To evaluate the precision of the salient object detection, ground-truth data adapted to an object-based approach is required.

For this purpose, ten manually segmented binary masks are added to the two already existent to complement the database. Binary masks are estimated for all the frames, and the whole salient object is segmented even though parts of it might be more salient in terms of gaze detection. Figure 11.7 illustrates the database with

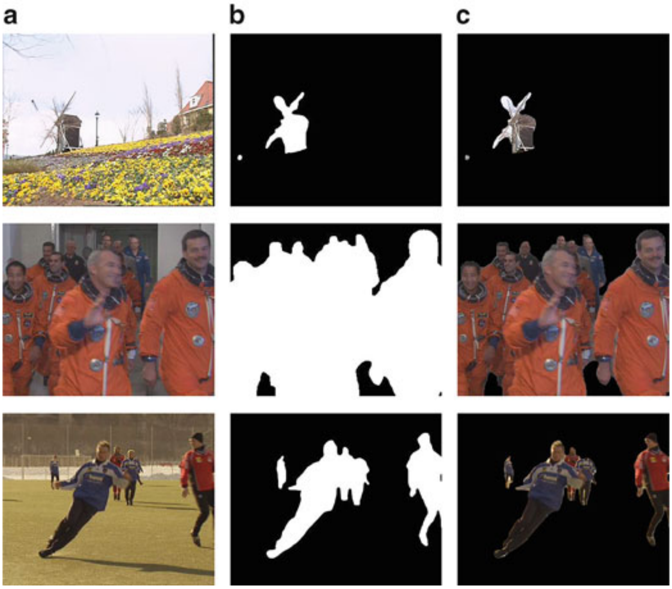


Fig. 11.7 Extended SVS dataset [39] with interesting object labeling. (**a**) Original images (**b**) Mask maps (**c**) Object maps

the original sequences, the binary masks that we provide, and the salient object. The SVS-extended database is, at author's best knowledge, the only video dataset providing both eye-tracking and binary masks along with uncompressed video sequences.

ASCMN Dataset (2012)

All databases presented in Sect. 11.1.2 contain some videos with different kinds of motion in the scene (the sequences are extracted mostly from TV programs, Hollywood movies, standard video databases, or video games).

However, none of them has been designed specifically to contain anomalous motion which would attract attention in the presence of other motion and enable testing of dynamic-saliency models. The proposed ASCMN database attempts to fill this gap. It contains videos obtained from other databases including the Itti's CRCNS database, Vasconcelo's database [40], and a standard complex-background video surveillance database [41]. These have been extended with Internet crowd movies and proprietary videos from a crowd database. The database is divided into five classes of movies as described in Table 11.1.

In addition to videos for which eye-tracking data has previously been published in existing databases, the classes cover a new type of videos, lacking in the other databases – videos that contain motion abnormalities and crowd motion. Also, eye-tracking data for complex-background surveillance videos included in ASCMN has not previously been published. Sample frames for different classes of videos are shown in Fig. 11.8.

The ASCMN database, therefore, provides data which covers a wider spectrum of video types, than the existing databases, and accumulates previously published videos suitable for dynamic saliency model evaluation. ASCMN database contains 24 videos, together with eye-tracking data from 13 viewers, acquired using a commercial eye-tracking system [42].

Table 11.1 The five classes of videos contained into the ASCMN database

Video classes	Description	Videos Nb.
ABNORMAL	Some moving blobs have different speed or direction compared to the main stream: Fig. 11.8 line 1	2, 4, 16, 18, 20
SURVEILLANCE	Classical surveillance camera with no special motion event: Fig. 11.8 line 2	1, 3, 5, 9
CROWD	Motion of more or less dense crowds: Fig. 11.8 line 3	8, 10, 12, 14, 21
MOVING	Videos taken with a moving camera: Fig. 11.8 line 4	6, 19, 22, 24
NOISE	No motion during several seconds followed by sudden important motion: Fig. 11.8 line 5	7, 11, 13, 15, 17, 23

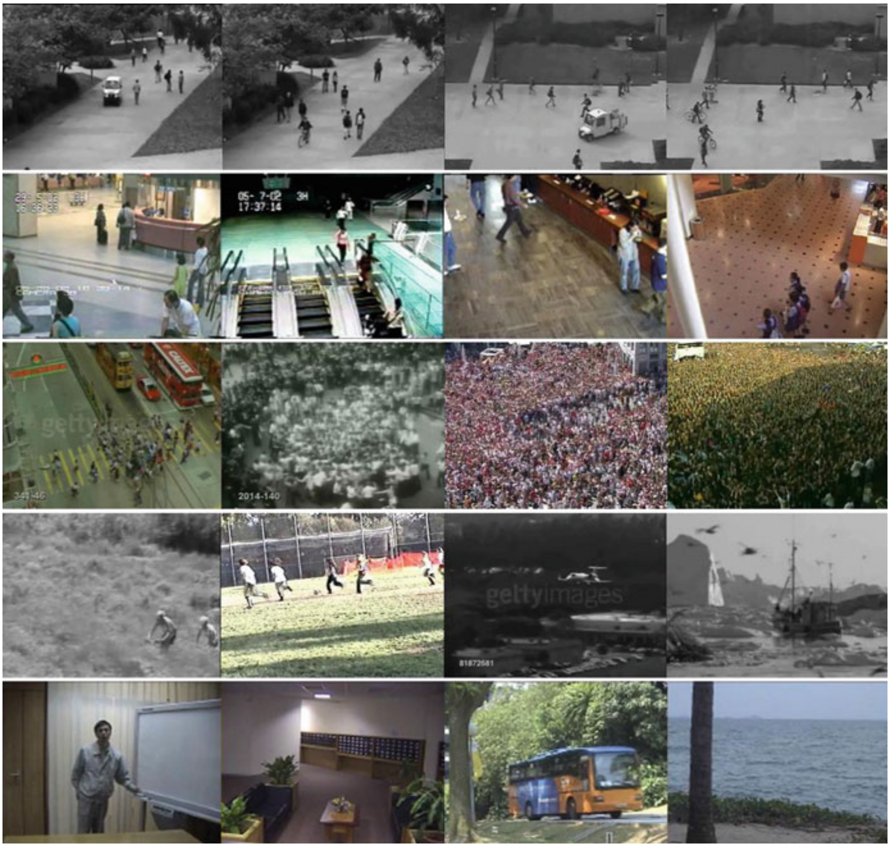


Fig. 11.8 The five classes of videos from the ASCMN database. First line represents ABNORMAL motion with bikes and cars which are faster than people, for example. The second line shows SURVEILLANCE classical motion with nothing really salient in terms of movement. The third line shows CROWD motion with increasing density from left to right. Line four shows MOVING camera videos. Line five displays videos with long periods of NOISE (frames 2, 4) and a sudden appearance of a salient object (frames 1, 3)

This system allows small head movements and is thus less intrusive than other eye-tracking systems, making the viewer feel more comfortable. The viewers are PhD students and researchers ranging from 23 to 35 years old, both males and females. The eye gaze positions are recorded and superimposed on the initial video for all the viewers, as shown in the second column of Fig. 11.9. This data is low-pass filtered to obtain a "heat map" which can also be superimposed on corresponding video frame (Fig. 11.9, right column). This post-processing step is

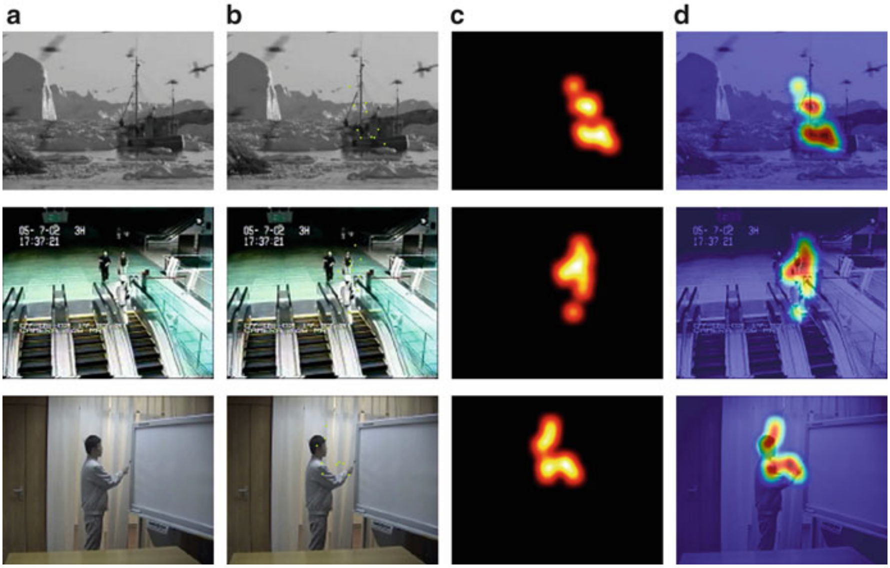


Fig. 11.9 First column: original images. Second column: aggregated eye-tracking results (each yellow cross is the position of the eye gaze of one viewer). The third column shows density maps. Right column contains smoothed gaze location producing "heat maps" superimposed to the corresponding frame. (**a**) Original images (**b**) Fixation maps (**c**) Density maps (**d**) Heat maps

useful in estimating the mean gaze density, eliminating the outliers and providing higher weight to the focus points common to several users.

11.2 Comparisons

In order to compare and have a global view of the six datasets which are described in this chapter, Table 11.2 shows three features: the stimuli (images or videos), the number of stimuli proposed in the database, and a special property (extra notes).

Various properties of natural images such as indoor, outdoor, or symmetrical scenes are proposed for a total of 1457 images. A large variety of videos are also available from uncompressed to compressed movies, with moving or static cameras, etc. A total of 36 videos (13,354 frames) are proposed.

In terms of ground truth, all the databases have eye-tracking data. Some of them have an additional ground truth (binary masks) to complement the validation. Two features have be chosen to compare the databases as shown in Table 11.3: the number of observers and the durations (in seconds) when the informations were available.

Table 11.2 Stimuli-based comparison for the six presented datasets

Name	Stimuli	Nb stimuli	Extra notes
Toronto	Images	120	Indoor–outdoor scenes
MIT	Images	1003	Natural images
Kootstra	Images	99	Symmetrical images
ImgSal	Images	235	Size of salient objects
ASCMN	Videos	24	Five kinds of motion
SVS	Videos	12	Uncompressed videos

Table 11.3 Ground-truth-based comparison for the six presented datasets

Name	Observers	Durations (s)	Additional ground truth
Toronto	20	4	
MIT	15	3	
Kootstra	31	5	
ImgSal	21		Binary masks
ASCMN	13	0.067	
SVS	15	0.04	Binary masks

The number of observers varies from 13 to 31 subjects and the durations depend of the kind of stimuli. There are fundamental differences between videos and still images such as the duration. Indeed, each video frame is only observed a fraction of a second (depending on the frame rate of the video), while a still image can be viewed during a longer period of time (from 3 to 5 s). Finally, two databases (one for still images and one for videos) have additional binary masks to detect salient objects.

11.3 Conclusions

There are many databases in the literature which provide stimuli and ground-truth data and which are freely available online. These datasets can be classified in many ways, including the number of images, number of videos, number of participants, type of ground truth, or type of stimuli (natural images, portraits, websites, advertisements, movies, news, cartoons, etc.).

It shows the importance of choosing appropriate characteristics and experimental settings for a validation framework. The authors need to clarify why they choose these databases and ground truth. The proposed framework validation intends to assess bottom-up saliency models. So, only free-viewing databases have been chosen. Moreover, the assessment will be done on color natural images and various videos.

11.4 Summary

- In order to validate saliency models (see next chapters), a database with stimuli and ground truth is needed.
- The first type of stimuli and ground truth which was used is still images and eye-tracking data.
- Another ground truth is based on salient objects segmentation. Manual segmentations (bounding-box or pixel-wise segmentations) can be used. While some datasets exhibit only the big centered object segmentation, others are more complex and close to the reality.
- Some datasets provide both eye-tracking and object segmentation ground truth.
- In addition to still images, video databases are available mainly with eye-tracking data. Very few video datasets provide both eye-tracking and object segmentation ground truth.
- A list of databases which are available online for visual attention can be found from the Computational Attention Group of TCTS lab at http://tcts.fpms.ac.be/attention.

References

1. Ye, B., Sugano, Y., & Sato, Y. (2014). Influence of stimulus and viewing task types on a learning-based visual saliency model. In *ETRA*, Safety Harbor (pp. 271–274).
2. Smith, T. J., & Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision, 13*(8), 16.
3. Winkler, S., & Ramanathan, S. (2013). Overview of eye tracking datasets. In *QoMEX*, Klagenfurt am Wörthersee (pp. 212–217).
4. Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In *Advances in Neural Information Processing Systems* (Vol. 18, pp. 155–162). Vancouver, Canada.
5. Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(5), 802–817.
6. Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. In *Advances in Neural Information Processing Systems* (pp. 241–248). Vancouver, Canada.
7. Engelke, U., Maeder, A., & Zepernick, H. (2009). Visual attention modelling for subjective image quality databases. In *IEEE International Workshop on Multimedia Signal Processing (MMSP'09)*, Rio de Janeiro (pp. 1–6). IEEE.
8. Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *IEEE 12th International Conference on Computer Vision 2009*, Kyoto (pp. 2106–2113). IEEE.
9. Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision, 77*(1–3), 157–173.
10. Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., & Chua, T.-S. (2010). An eye fixation database for saliency detection in images. In *Computer Vision–ECCV 2010*, Heraklion (pp. 30–43). Springer.

11. Judd, T., Durand, F., & Torralba, A. (2011). Fixations on low-resolution images. *Journal of Vision, 11*(4), 14.
12. Li, J., Levine, M., An, X., & He, H. (2011). Saliency detection based on frequency and spatial domain analyses. In *Proceedings of the British Machine Vision Conference* (pp. 86.1–86.11). BMVA Press. http://dx.doi.org/10.5244/C.25.86.
13. Kootstra, G., & Schomaker, L. R. (2009). Prediction of human eye fixations using symmetry. In *The 31st Annual Conference of the Cognitive Science Society (CogSci09)*, Amsterdam (pp. 56–61). Cognitive Science Society.
14. Olmos, A., & Kingdom, F. A. A. (2004). McGill calibrated colour image database (pp. 05–08). http://tabby.vision.mcgill.ca. Last accessed 2011.
15. Borji, A. (2015). What is a salient object? A dataset and a baseline model for salient object detection. *IEEE Transactions on Image Processing, 24*(2), 742–756.
16. Borji, A., Cheng, M.-M., Jiang, H., & Li, J. (2014). Salient object detection: A survey. arXiv preprint arXiv:1411.5878.
17. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., & Shum, H.-Y. (2011). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(2), 353–367.
18. Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami (pp. 1597–1604). IEEE.
19. Borji, A., Sihite, D. N., & Itti, L. (2013). What stands out in a scene? A study of human explicit saliency judgment. *Vision Research, 91*, 62–77.
20. Yan, Q., Xu, L., Shi, J., & Jia, J. (2013). Hierarchical saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, Portland (pp. 1155–1162). IEEE.
21. Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV 2001)*, Vancouver (Vol. 2, pp. 416–423). IEEE.
22. Everingham, M., Van Gool, L., Williams, C., Winn, J., & Zisserman, A. (2012). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision, 111*(1), 98–136.
23. Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M.-H. (2013). Saliency detection via graph-based manifold ranking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, Portland (pp. 3166–3173). IEEE.
24. Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The secrets of salient object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, Columbus (pp. 280–287). IEEE.
25. Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision, 88*(2), 303–338.
26. Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision, 14*(1), 28.
27. Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, Minneapolis.
28. Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing, 13*(10), 1304–1318.
29. Itti, L. (2006). Quantitative modelling of perceptual salience at human eye position. *Visual Cognition, 14*(4–8), 959–984.
30. Carmi, R., & Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research, 46*(26), 4333–4345.
31. Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *Journal of Vision, 6*(9), 4.

32. Li, J., Tian, Y., Huang, T., & Gao, W. (2009). A dataset and evaluation methodology for visual saliency in video. In *IEEE International Conference on Multimedia and Expo (ICME 2009)*, New York (pp. 442–445). IEEE.
33. Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision, 10*(10), 28.
34. Wu, Y., Zheng, N., Yuan, Z., Jiang, H., & Liu, T. (2011). Detection of salient objects with focused attention based on spatial and temporal coherence. *Chinese Science Bulletin, 56*(10), 1055–1062.
35. Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation, 3*(1), 5–24.
36. Marszalek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami (pp. 2929–2936). IEEE.
37. Rodriguez, M. D., Ahmed, J., & Shah, M. (2008). Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Anchorage.
38. Mathe, S., & Sminchisescu, C. (2012). Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *Computer Vision–ECCV 2012*, Florence (pp. 842–856). Springer.
39. Hadizadeh, H., Enriquez, M. J., & Bajic, I. V. (2012). Eye-tracking database for a set of standard video sequences. *IEEE Transactions on Image Processing, 21*(2), 898–903.
40. Mahadevan, V., & Vasconcelos, N. (2010). Spatiotemporal saliency in dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(1), 171–177.
41. Li, L., Huang, W., Gu, I.-H., & Tian, Q. (2004). Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing, 13*(11), 1459–1472.
42. Machines, S. (2015). Facelab commercial eye tracking system. http://www.seeingmachines.com/. Accessed 04 May 2015 (online).

Chapter 12
Metrics for Saliency Model Validation

Nicolas Riche

Different scores have been used in the literature to validate saliency models. While reviews of databases [1] or saliency models [2, 3] exist, reviews of similarity metrics are harder to come by. In this chapter, we will explain the standard measures used to evaluate the salient object detection and eye tracking models. The metrics presented here will be used in our study and validation in the next chapters. While some metrics focus on eye scanpath [4], here we will deal with approaches involving 2D maps. As it was described in the previous chapter, there are two ground truths to validate saliency maps. The first one is based on salient object segmentation (using bounding boxes or pixel-wise segmentation) and the other one is based on eye-tracking data. In Sect. 12.1 we present similarity metrics for salient object detection ground truth while in Sect. 12.2 we focus on metrics for eye-tracking data. We finally conclude on existing metrics for saliency maps evaluation.

12.1 Literature Review of Metrics for Object Detection

In this section, all metrics that have been used to assess salient object detection models are presented. Indeed, there are several ways to measure the agreement between salient object detection models and binary masks (bounding boxes or pixel-wise masks). Sometimes, metrics do not agree with each other.

However, contrary to the eye tracking-based category, all the salient object detection benchmarks use very close gold standard location-based metrics. Moreover, in

N. Riche (✉)
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: nicolas.riche@umons.ac.be

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_12

Table 12.1 Definitions of four concepts to compute precision-recall and FPR/TPR

	Reference results	
Predicted results	**TP:** Correct result	**FP:** Unexpected result
	FN: Missing result	**TN:** Correct absence of result

85 % of the publications on salient object detection model, the authors use one gold standard metric (F-score from precision-recall curve) to compare their models to other state-of-the-art models.

12.1.1 Location-Based Metrics: Focus on Location of Salient Regions and Binary Masks

For all location-based metrics, we retrieve the concept and terminology from a confusion matrix: true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs) that compare the predicted results (saliency map) with the reference results (binary mask). Therefore, the saliency maps need to be converted to a binary map. To do that, several thresholds are defined. Table 12.1 illustrates their definition.

From this concept, two metrics can be computed. A new one, the F-score, calculates a score from precision-recall, and the area under the receiver operating characteristic (AUC), like eye tracking-based metrics, computes a score from the false- and true-positive rate. All these notions will be described in detail below.

A third metric called MAE exists in the literature as described in [3]. The purpose is to consider the true negative (TN) when a pixel is correctly marked as non-salient.

Finally, recently, we find some variations of F-score which propose a weighted calculation of precision and recall. The objective is to provide a more reliable evaluation. In [5], the authors start by identifying three causes of inaccurate evaluation: interpolation flaw, dependency flaw and equal-importance flaw. By amending these three assumptions, they propose a new reliable measure available for images.

F: F-Score from Precision-Recall (2009)

Authors: R. Achanta, S. Hemami, F. Estrada and S. Susstrunk [6].

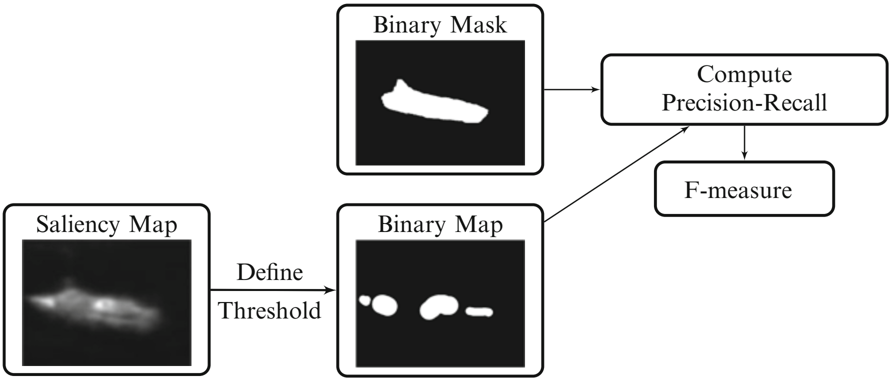


Fig. 12.1 F-score between saliency map and binary mask

Description: Many authors like [7–10] also used F-score metric to compare saliency maps and binary masks (Fig. 12.1). Precision is the number of relevant points compared with the total number of points found (Eq. 12.1 (left)). Recall is the number of relevant points compared with the total number of important points in the reference (Eq. 12.1 (right)):

$$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{Recall} = \frac{tp}{tp + fn} \tag{12.1}$$

A usual way to combine precision and recall is to use the F-score defined as in Eq. 12.2 where as suggested by several salient object detection benchmarks [6], β^2 is set to 0.3 to give more importance to the precision value:

$$F - \text{score} = \frac{(1 + \beta^2) * \text{Precision} * \text{Recall}}{\beta^2 * \text{Precision} + \text{Recall}} \tag{12.2}$$

AUC: Area Under the ROC Curve (2011)

Authors: J. Li, L. Martin, A. Xiangjing and H. Hangen [11].

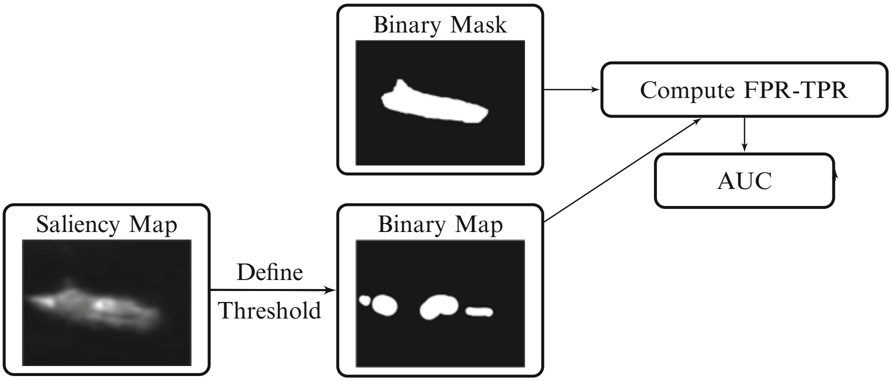


Fig. 12.2 AUC between saliency map and binary mask

Description: Many authors like [12, 13] also used AUC metric to compare saliency maps and binary masks (Fig. 12.2). The true-positive rate, also called sensitivity, measures, as the recall, the proportion of true positive under all the positive reference results (Eq. 12.3 (left)). The false-positive rate measures the proportion of false positive under all the negative reference results (Eq. 12.3 (right)):

$$TPR = \frac{tp}{tp + fn} \qquad FPR = \frac{fp}{fp + tn} \tag{12.3}$$

A usual way to combine them is to plot the true-positive rate (TPR) vs. the false-positive rate (FPR) to form an ROC curve. Then, the area under the ROC can be computed.

12.2 Literature Review of Metrics for Eye Tracking

In this section, all the similarity metrics that have been used to assess eye tracking saliency models are presented. Contrary to salient object detection validation, a gold standard metric doesn't exist, and a lot of metrics have been proposed to validate eye tracking saliency models.

Therefore, we propose here a taxonomy to classify them. The classification is related to the nature of the similarity metric and can be divided into three categories: value-based metrics which focus on saliency map values at eye gaze positions, distribution-based metrics which focus on saliency and gaze statistical distributions and location-based metrics which focus on location of salient regions at gaze positions.

All these metrics will be described in detail in this section and will be used in the next chapter to study their similarity. They take two distributions as input: the prediction (noted SM for saliency map) and the ground truth (noted FM for fixation map).

It is important to note that a discrete fixation map is used for location-based and value-based metrics while a continuous one is used for distribution-based metrics. The continuous fixation map is deduced by convolving the fixation map with a 2D Gaussian function. The parameters of this function depend on the database.

12.2.1 Value-Based Metrics: Focus on Saliency Map Values at Eye Gaze Positions

This first category of metrics compares values or amplitudes of the saliency maps with the corresponding eye fixation maps.

Three similarity metrics are proposed and described in the following subsections.

NSS: Normalized Scanpath Saliency (2005)

Authors: R. Peters, A. Iyer, L. Itti and C. Koch [14].

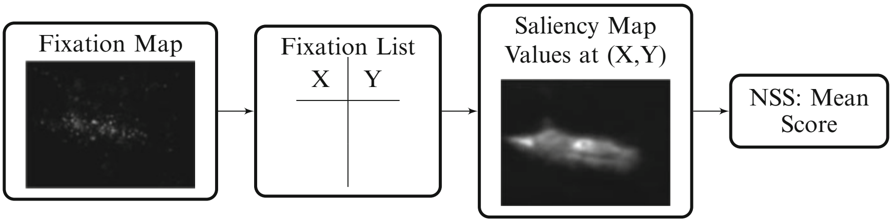


Fig. 12.3 NSS between saliency and fixation map

Description: The idea is to quantify the saliency map values at the eye fixation locations and to normalize it with the saliency map variance (Fig. 12.3):

$$NSS(p) = \frac{SM(p) - \mu_{SM}}{\sigma_{SM}} \tag{12.4}$$

where p is the location of one fixation and SM is the saliency map which is normalized to have a zero mean and unit standard deviation. Indeed, the NSS score should be decreased if the saliency map variance is important or if all values are globally similar (small difference between fixation values and mean) because it shows that the saliency model will not be very predictive, while it will precisely point a direction of interest if the variance is small or if the difference between fixation values and means is high.

The NSS score is the average of $NSS(p)$ for all fixations:

$$NSS = \frac{1}{N} * \sum_{p=1}^{N} NSS(p) \tag{12.5}$$

where N is the total number of eye fixations.

PF: Percentage of Fixations into the Salient Region (2006)

Authors: A. Torralba, A. Oliva, M. Castelhano and J. Henderson [15].

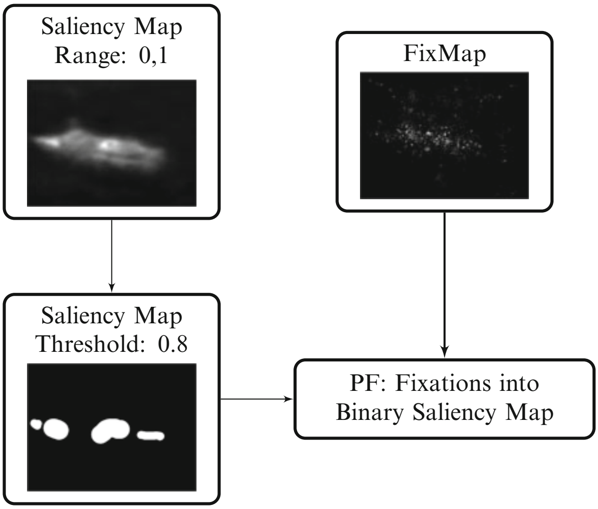


Fig. 12.4 PF between saliency and fixation map

Description: Its purpose is to measure the percentage of fixations into the salient region. In a first step, saliency maps are thresholded at $T = 0.8$ where the saliency is normalized between 0 and 1. The threshold is set so that the selected image region occupies a fixed proportion of the image size. In a second step, the percentage of fixations in this area is computed and called PF (Fig. 12.4).

P: Percentile (2008)

Authors: R. Peters and L. Itti [16].

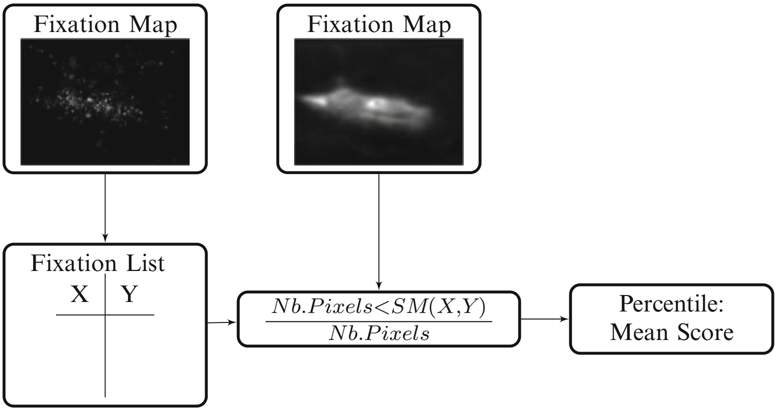


Fig. 12.5 P between saliency and fixation map

Description: The percentile metric is, for each pixel p on the eye fixation map, a ratio between the number of pixels in the saliency map with values smaller than the one corresponding to pixel p from the eye fixation map and the total number of pixels (Eq. 12.6) (Fig. 12.5):

$$P(p) = \frac{|x \in X : SM(x) < SM(p)|}{|SM|} \tag{12.6}$$

where X is the set of all pixels of the saliency map SM, p is the location of one eye fixation and $|SM|$ indicates the total number of pixels. Like in the case of NSS, the global percentile score is the average of $P(p)$ for all the eye fixations.

12.2.2 Distribution-Based Metrics: Focus on Saliency and Gaze Statistical Distributions

In the literature, there are two kinds of distribution-based metrics. Those which compute a similarity between two distributions and those which compute a dissimilarity. Moreover, some metrics which are not a distance are nonsymmetric. It means that by first considering the saliency map (SM) as the first input and secondly the fixation map (FM) as the first input, the results are not the same. This should be taken into account for the comparison. Two dissimilarity and three similarity metrics are proposed and described in the following subsections.

PCC: Pearson's Correlation Coefficient (2004)

Authors: N. Ouerhani, R. Von Wartburg, H. Hugli and R. Muri [17].

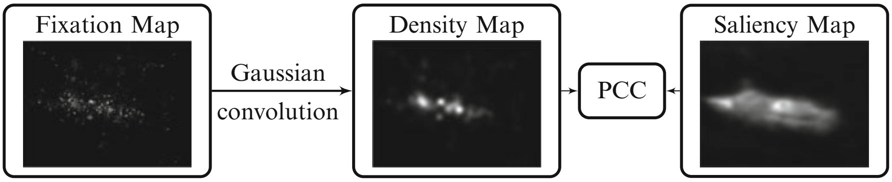


Fig. 12.6 Pearson's correlation coefficient between saliency and density map

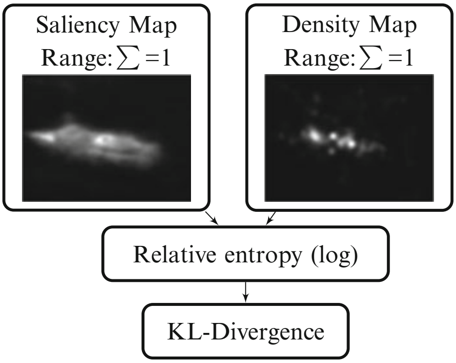
Description: Pearson's correlation coefficient also named linear correlation coefficient was first used in [17] as a metric. Other authors also used it such as in [18]. The linear CC output range is between −1 and 1. When the correlation value is close to −1 or 1, there is almost a perfect linear relationship between the two variables (Fig. 12.6):

$$CC = \frac{cov(SM, FM)}{\sigma_{SM} * \sigma_{FM}} \tag{12.7}$$

KLD: Kullback-Leibler Divergence (2004)

Authors: U. Rajashekar, L. Cormack and A. Bovik [19].

Fig. 12.7 KLD between saliency and density map



Description: The Kullback-Leibler divergence is a commonly used metric to estimate an overall dissimilarity between two distributions. Many authors like

[20] and [18] also used this metric to compare saliency maps with human eye fixations. The KLD is a measure of the information lost when the saliency map probability distribution (called *SM*) is used to approximate the human eye fixation map probability distribution (called *FM*) (Fig. 12.7):

$$KLD = \sum_{x=1}^{X} FM(x) * \log\left(\frac{FM(x)}{SM(x)+\epsilon} + \epsilon\right) \tag{12.8}$$

where X is the number of pixels and ϵ is a small constant to avoid log(0) and division by zero. *SM* and *FM* distributions are both normalized as in Eq. 12.9:

$$SM(x) = \frac{SM(x)}{\sum_{x=1}^{X} SM(x)+\epsilon} \qquad FM(x) = \frac{FM(x)}{\sum_{x=1}^{X} FM(x)+\epsilon} \tag{12.9}$$

When the two maps are strictly equal, the KLD value is zero.

SCC: Spearman's Correlation Coefficient (2011)

Authors: A. Toet [21].

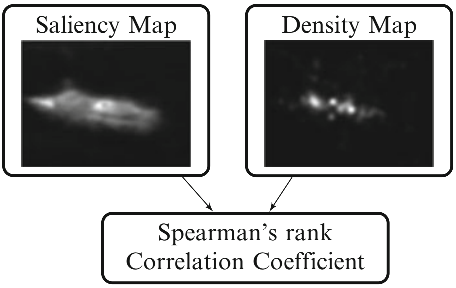


Fig. 12.8 Spearman's correlation coefficient between saliency and density map

Description: Spearman's rank correlation coefficient metric [21] is defined as the CC metric (Eq. 12.7) but on ranked variables. This can be understood as a non-linear correlation. Toets uses this metrics in [21] to evaluate 13 models (Fig. 12.8).

EMD: Earth Mover's Distance (2012)

Authors: T. Judd, F. Durand and A. Torralba [22].

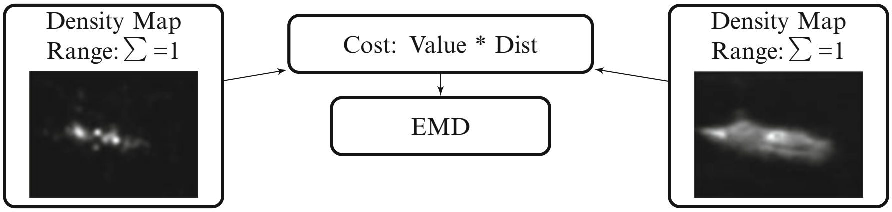


Fig. 12.9 EMD between saliency and density map

Description: Earth mover's distance metric is a measure of the distance between two probability distributions over a region (Fig. 12.9). Judd [22] used this metric in her benchmark which is now available online. She uses a fast implementation of EMD provided by Pele and Werman [23, 24], but without a threshold. It computes the minimal cost to transform the probability distribution of the saliency maps *SM* into one of the human eye fixations *FM*:

$$EMD = \left(\min_{f_{ij}} \sum_{i,j} f_{ij} d_{ij}\right) + \left|\sum_i FM_i - \sum_j SM_j\right| \max_{i,j} d_{ij}$$

$$s.t. f_{ij} \geq 0, \sum_j f_{ij} \leq FM_i, \sum_i f_{ij} \leq SM_j, \tag{12.10}$$

and

$$\sum_{i,j} f_{ij} = \min\left(\sum_i FM_i - \sum_j SM_j\right)$$

where each f_{ij} represents the amount transported from the i_{th} supply to the j_{th} demand. d_{ij} is the ground distance between bin i and bin j in the distribution. A larger EMD indicates a larger overall difference between the two distributions. An EMD of zero indicates that two distributions are the same.

S: Similarity (2012)

Authors: T. Judd, F. Durand and A. Torralba [22].

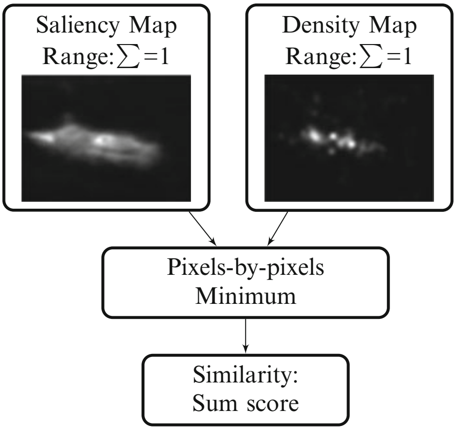


Fig. 12.10 S between saliency and density map

Description: The similarity metric [22] also uses the normalized probability distributions of the saliency map *SM* and human eye fixation map *FM* (Fig. 12.10). The similarity is the sum of the minimum values at each point in the distributions. Mathematically, the similarity between two maps *SM* and *FM* is:

$$S = \sum_{x=1}^{X} \min(SM(x), FM(x)) \tag{12.11}$$

where $\sum_{x=1}^{X} SM(x) = \sum_{x=1}^{X} FM(x) = 1$.

A similarity score of one indicates that the distributions are the same. A similarity score of zero indicates that they do not overlap at all and are completely different.

12.2.3 Location-Based Metrics: Focus on Location of Salient Regions at Gaze Positions

Location-based metrics are very popular to evaluate saliency maps. They are based on the notion of area under the receiver operating characteristic curve coming from the signal detection theory. Four main different implementations are available dealing with some limitations of the classical approach.

nAUC: Normalized Area Under the ROC Curve (2011)

Authors: Q. Zhao and C. Koch [25].

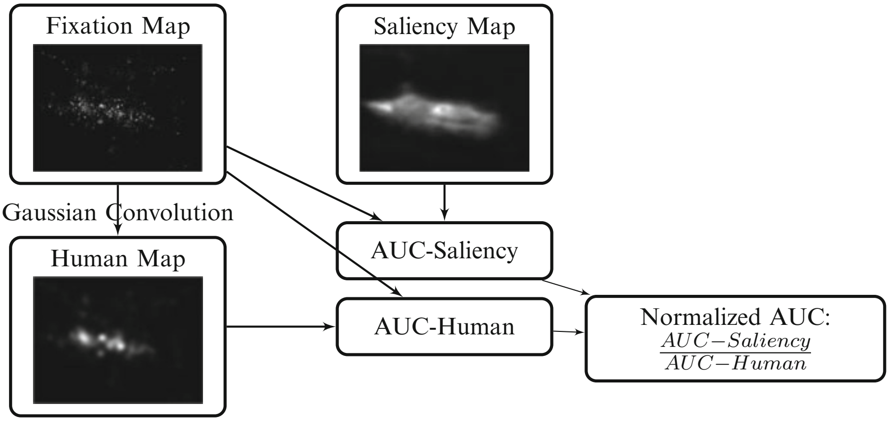


Fig. 12.11 nAUC between saliency and density map

Description: Zhao used a normalized AUC (Fig. 12.11). The idea is that saliency algorithms perform less well (on average) than the area under the ROC curve coming from inter-subject variability for each image. Zhao computes an ideal AUC by measuring how well the human fixations of one subject can be predicted by those of the other $n-1$ subjects, iterating over all n subjects and averaging the result with an upper limit of one. Finally, the AUC of the saliency map is normalized by this ideal AUC.

pAUC: Post-Processing for Area Under the ROC Curve (2011)

Authors: J. Li, L. Martin, A. Xiangjing and H. Hangen [11].

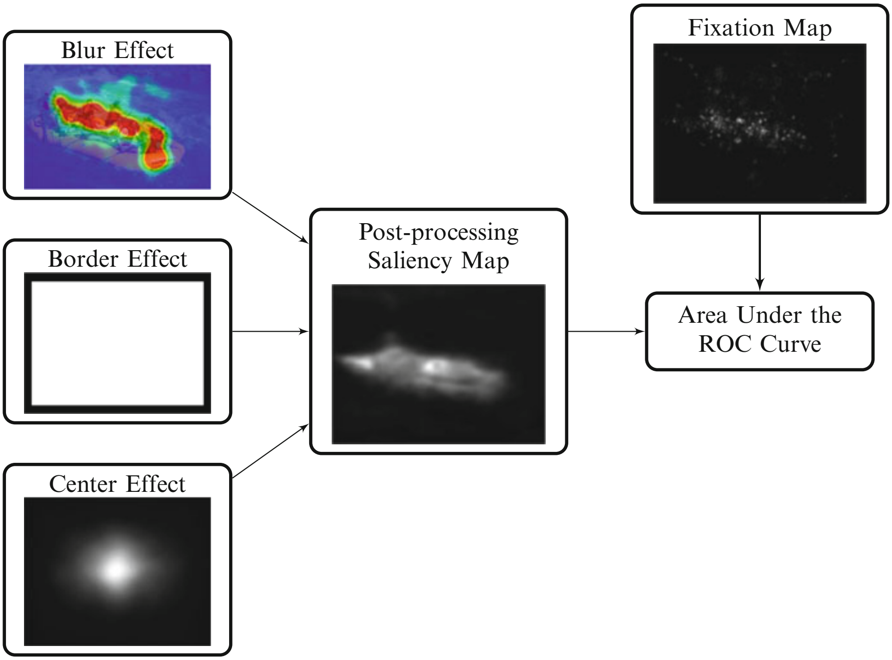


Fig. 12.12 pAUC between saliency and density map

Description: Li set the border cuts for all models to be of equal size and avoids in that way to artificially increase the AUC scores for the models which already do this preprocessing in comparison with those which do not. The border cut post-processing affecting the fairness during the assessment is thus eliminated (Fig. 12.12).

hAUC: Hit Rate for Area Under the ROC Curve (2012)

Authors: T. Judd, F. Durand and A. Torralba [22].

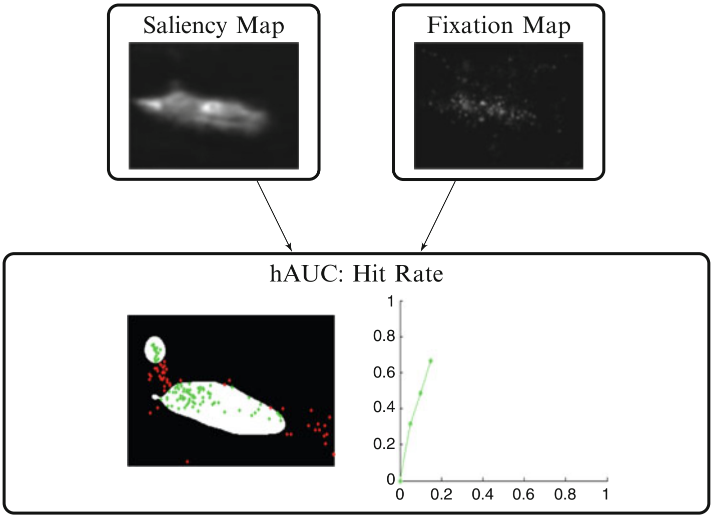


Fig. 12.13 hAUC between saliency and density map

Description: Judd proposed another version of AUC to validate saliency models. First, fixation pixels were counted once and the same number of random pixels is extracted from the saliency map. For one given threshold, saliency pixels can be seen as a classifier, with all points above threshold indicated as "fixation" and all points below threshold as "background".

For any particular value of the threshold, there is some fraction of the actual fixation points which are labelled as true positive (TP) and some fraction of points which were not fixated but labelled as false positive (FP). This operation is repeated one hundred times. Then the ROC curve can be drawn and the area under the curve (AUC) computed. An ideal score is one, while random classification provides 0.5 (Fig. 12.13).

sAUC: Shuffled Area Under the ROC Curve (2012)

Authors: A. Borji, D. Sihite and L. Itti [26].

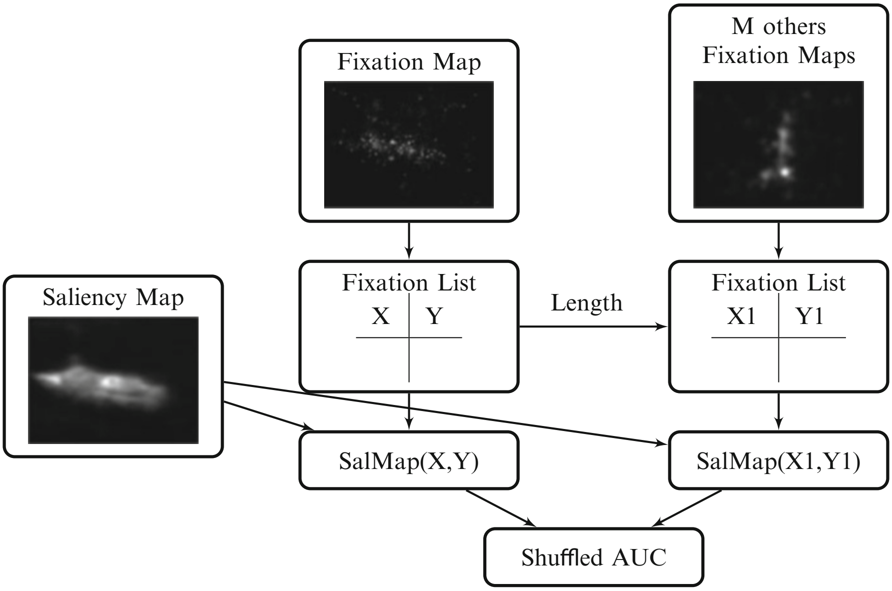


Fig. 12.14 sAUC between saliency and fixation map

Description: Borji applied to saliency map validation a suitable AUC metric called *shuffled* AUC (Fig. 12.14). In his classical AUC, saliency map values from random points from the image are addressed to create a binary mask. In the *shuffled* AUC metric, saliency values and fixations from another image (instead of random) of the same dataset are taken into account. In that way, the more or less centred distribution of the human fixations of the database is taken into account in the AUC computation. This point is important because the AUROC scores can dramatically increase if a saliency map is weighted by a centred Gaussian. Indeed, human eye fixations are rarely near the edges of general test images, and the amateur photographer often places salient objects in the image centre.

12.3 Discussions and Conclusions

There are a large variety of metrics in the literature which provide a score between saliency map and ground-truth data which have been processed into a bidimensional map. These metrics depend on the nature of the ground truth and what authors want to measure: amplitude, location, distribution or the three.

It shows the importance of choosing appropriate metrics for a validation framework. The authors need to clarify why they choose these metrics. Moreover, the framework validation needs a preliminary study to investigate the relevance of the chosen metric mix.

In this chapter, we focused only on approaches involving two bidimensional maps. Other metrics exist for comparing two scanpaths, using either distance-based methods (string edit technique or Mannan's distance) or vector-based methods. These metrics are described in [4]. They require taking into account a number of factors, such as the temporal dimension or the alignment procedure. To overcome these problems, most of saliency validation frameworks used two bidimensional maps.

12.4 Summary

- For object-based validation, all the metrics are based on the notion of TP/FP and TN/FN as F-score and weighted F-score.
- For eye tracking ground truth, there are dozens of metrics (amplitude based, location based, distribution based).

References

1. Winkler, S., & Ramanathan, S. (2013). Overview of eye tracking datasets. In *QoMEX*, Klagenfurt am Wörthersee (pp. 212–217).
2. Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207.
3. Borji, A., Cheng, M.-M., Jiang, H., & Li, J. (2014). Salient object detection: A survey. arXiv preprint arXiv:1411.5878.
4. Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods, 45*(1), 251–266.
5. Margolin, R., Zelnik-Manor, L., & Tal, A. (2014). How to evaluate foreground maps. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, Columbus (pp.248–255). IEEE.
6. Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami (pp. 1597–1604). IEEE.
7. Cheng, M.-M., Zhang, G.-X., Mitra, N. J., Huang, X., & Hu, S.-M. (2011). Global contrast based salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, Colorado Springs (pp. 409–416). IEEE.
8. Perazzi, F., Krahenbuhl, P., Pritch, Y., & Hornung, A. (2012). Saliency filters: Contrast based filtering for salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, Providence (pp. 733–740). IEEE.
9. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., & Shum, H.-Y. (2011). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(2), 353–367.

10. Cheng, M.-M., Warrell, J., Lin, W.-Y., Zheng, S., Vineet, V., & Crook, N. (2013). Efficient salient region detection with soft image abstraction. In *IEEE International Conference on Computer Vision (ICCV 2013)*, Sydney (pp. 1529–1536). IEEE.
11. Li, J., Levine, M., An, X., & He, H. (2011). Saliency detection based on frequency and spatial domain analyses. In *Proceedings of the British Machine Vision Conference* (pp. 86.1–86.11). BMVA Press. http://dx.doi.org/10.5244/C.25.86.
12. Borji, A., Sihite, D. N., & Itti, L. (2012). Salient object detection: A benchmark. In *Computer Vision–ECCV 2012*, Florence (pp. 414–429). Springer.
13. Borji, A. (2015). What is a salient object? A dataset and a baseline model for salient object detection. *IEEE Transactions on Image Processing, 24*(2), 742–756.
14. Peters, R. J., Iyer, A., Itti, L., & Koch, L. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45*(18), 2397–2416.
15. Antonio Torralba, M. C., Oliva, A., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review, 113*(4), 766–786.
16. Peters, R. J., & Itti, L. (2008). Applying computational tools to predict gaze direction in interactive visual environments. *ACM Transactions on Applied Perception (TAP), 5*(2), 9.
17. Ouerhani, N., Von Wartburg, R., Hugli, H., & Muri, R. (2004). Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis, 3*(1), 13–24.
18. Le Meur, O., Le Callet, P., Barba, D., et al. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research, 47*(19), 2483–2498.
19. Rajashekar, U., Cormack, L. K., & Bovik, A. C. (2004). Point-of-gaze analysis reveals visual search strategies. In *Proceedings of SPIE*, San Jose, USA (Vol. 5292, pp. 296–306).
20. Tatler, B. W., Baddeley, R. J., Gilchrist, I. D., et al. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*(5), 643–659.
21. Toet, A. (2011). Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(11), 2131–2146.
22. Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. MIT technical report.
23. Pele, O., & Werman, M. (2008). A linear time histogram metric for improved sift matching. In *Computer Vision–ECCV 2008*, Marseille (pp. 495–508). Springer.
24. Pele, O., & Werman, M. (2009). Fast and robust earth mover's distances. In *IEEE 12th International Conference on Computer Vision 2009*, Kyoto (pp. 460–467). IEEE.
25. Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision, 11*(3), 9.
26. Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing, 22*(1), 55–69.

Chapter 13
Study of Parameters Affecting Visual Saliency Assessment

Nicolas Riche

The computational modelling of visual attention has been developed and expanded considerably during the past 10 years. Many different saliency models are now available online (for still images and videos). At the same time, many popular image-video datasets with human gaze data or binary masks have been released to evaluate saliency models with commonly used evaluation metrics. The new challenges and future directions for this field are therefore to establish evaluation protocols and saliency benchmarks.

Although some evaluation studies (such as [1–3] and [4]) and online benchmarks (like [5] and [6]) have already been proposed and are major contributions, a key underlying issue is: *how can one fairly evaluate all these models?* In this chapter, we investigate this question with an evaluation, divided into four experiments, leading to the proposition of a new evaluation framework. Each experiment is based on an important aspect of visual saliency assessment in real-life images and is extended for videos in the validation framework. There are four questions that we will carefully consider:

1. What are the differences between eye fixations and manually segmented salient regions?
2. What is the relation between model performances and the properties (e.g. the size) of the salient regions into images?
3. What is the effect of saliency map post-processing?
4. Is one metric enough to evaluate a saliency model?

First of all, there are mainly two ground-truth categories to assess a saliency map: human eye fixations obtained using an eye tracker device and manually

N. Riche (✉)
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: nicolas.riche@umons.ac.be

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_13

segmented and labelled salient regions. In our study, we analyse the difference and the coherence between them. The second aspect of this study is about different categories of salient regions. Are saliency models equally efficient in predicting human gaze on three categories of salient regions: large, intermediate and small? This is an important issue as real-life objects and scenes contain a very wide range of object sizes. The third experiment is about saliency map post-processings. Which ones increase the score of a saliency map? Finally, various evaluation measures exist to compare saliency and ground-truth maps. We study the redundancy of these metrics and propose, among them, three metrics which should be used to obtain a complete assessment of saliency model performance.

Statistical analysis is used here to answer each of these four questions.

13.1 Experiment 1: Effects of Ground Truth

13.1.1 Goal

Nowadays, databases are coming with two ground truths: eye fixations and labelled objects. Some databases have the interest of providing both approaches for the same set of images. Some saliency models will better model eye fixations while others focus on object detection and segmentation and are assessed with region-based labelled objects. The main idea of this first experiment is to assess the coherence between the region-based and eye fixation-based ground truths.

13.1.2 Method

Database and Ground Truth: The database used here has been published by Jian Li et al. [7] and provides both region ground truth (human labelled) and eye fixation ground truth (collected with an eye tracker). In this experiment, we use the whole database containing 235 colour images.

Models: Twelve state-of-the-art models from a mix of eye tracking-based (80 %) and object detection-based (20 %) algorithms are used in this experiment. These models are detailed in the previous chapters and the taxonomy to present them is proposed by Borji's review paper [8] and used as a comparison feature in the previous chapters where models are sorted based on their mechanism to obtain saliency maps. We use a wide range of recently published saliency models. FSM model [9] represents the cognitive approach. SUN [10] and SDLF [11] are Bayesian models. AIM [12], DVA [13] and RARE [1] are into the information theory category. SR [14], PFT [15], QDCT [16], SSAFD [17] and FTSD [18] use a spectral analysis approach to compute their saliency map. Finally, AWS [19] which does not fit into Borji's taxonomy represents the *other models* category.

Metrics: In this study, the pAUC (post-processing for area under the ROC curve (2011) [7]) metric has been chosen. This metric can be applied to both eye tracking-based and region-based ground truths and mainly measures the eye fixation or region locations.

Kendall's W concordance measure is used for the statistical analysis. Kendall's W concordance measure [20] is an effect size measure. It defines how big the discordance between two distributions is. Indeed, while common significant tests only assess if there is enough evidence to determine whether the null hypothesis is likely between two or more groups, they do not provide information about the size of this effect. The effect size measures by how much the detected effect is significant in practice; in other words, it defines, in our case, how big the discordance between the region-based and eye fixation-based ground truths is.

It is defined in Eq. 13.1:

$$W = \frac{12 * S}{m^2 * (n^2 n)} \tag{13.1}$$

where n is the number of models and m is the number of metrics. So here $n = 12$ and $m = 2$ (pAUC on both ground truths). S, the sum of squared deviations, is defined as in Eq. 13.2:

$$S = \sum_{i=1}^{n} (R_i - \bar{R})^2. \tag{13.2}$$

where R_i is the ranking given to model i. A ranking as used here replaces the mean score of each model based on one metric by the assignment of labels (first, second, third, etc.). $\bar{R}$ is the mean value of those rankings.

Kendall's W concordance is a coefficient measuring the degree of agreement between metrics. The value ranges from 0 (no agreement between model ranks) to 1 (full agreement, same model ranking). Furthermore, some rules of thumb are provided to allow the researcher to interpret this measure as depicted in Table 13.1 [20].

However, in our study, the ranking range of 1–12 is small; therefore, higher thresholds are required to keep on the interpretation. That is why we decided to be much more selective than in Table 13.1: we interpret Kendall's coefficient as in

Table 13.1 Interpretation of Kendall's W coefficient

Kendall's W	Interpretation	Confidence in ranks
0.5	Moderate agreement	Fair
0.7	Strong agreement	High
0.9	Unusually strong agreement	Very high
1	Complete agreement	Very high

Table 13.2 Interpretation of Kendall's *W* coefficient on mean scores

W	Interpretation	Rank confidence
0.7	Moderate agreement	Fair
0.85	Strong agreement	High
0.93	Very strong agreement	High
0.98	Unusually strong agreement with 2 or 3 switched models	Very high
0.99	Unusually strong agreement with 1 or 2 switched models	Very high
1	Complete agreement	Very high

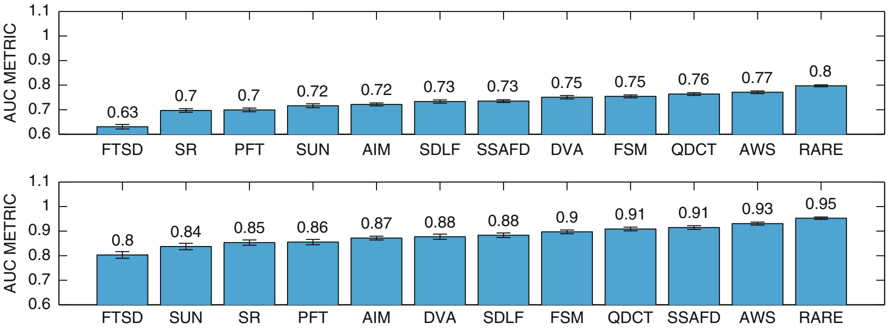


Fig. 13.1 Row 1: eye fixation mean score for all the models on the whole database with their standard deviations. Row 2: labelled region mean score for all the models on the whole database with their standard deviations. A higher pAUC means that the model is better

Table 13.2. Indeed this interpretation shows that $W = 0.98$ means that only two or three models are switched between the rankings.

13.1.3 Results

The mean results of pAUC metric for each model are computed in Fig. 13.1 for the entire database and both ground truths.

After this first score computation, a ranking-based statistical test is required. Considering our design, the 95 % confidence interval (CI) Friedman test allows to respond to the H_0 hypothesis: are the rankings of the individual results provided by the different models coherent between both ground-truth performance evaluations? As explained above, there is no specific effect size measure in the case of the 95 % CI Friedman test (only a binary response). Therefore, we use the presented Kendall's *W* concordance measure, which basically fulfils our needs (response between 0 and 1).

As shown in Table 13.3, although differences between eye fixation and region results are significant (Friedman test), Kendall's concordance between both ground truths is very good. This means that there is a difference between both rankings,

Table 13.3 Concordance based on Friedman test and Kendall's coefficient between eye fixation and region results

	Friedman test (p-value)	Kendall's concordance *W*
pAUC	≈0	0.82

but the size of this difference based on Kendall's *W* coefficient is relatively small. In other words, if models have good results with one ground truth, it is quite unlikely that these models completely fail with the other ground truths except due to statistical fluctuation. A saliency model which is good in predicting human eye fixations will remain good in predicting human-labelled regions and conversely.

These results depend on the experimental design. In our case, one database, 12 saliency models and one metric have been chosen. However, the same experiment was conducted in our paper [21] based on another metric (NSS) and leads towards exactly the same conclusion. These results are not presented in this section to avoid redundancy of information but to validate the interpretations.

13.2 Experiment 2: Effects of the Size of Salient Objects

13.2.1 Goal

In this experiment, we want to compare the effectiveness of the models on three different image categories (large, medium and small salient regions). In real-life images, all kinds of object sizes can be seen and saliency models which are tuned for a given object size are not suitable. It should be noted that this study is divided into two parts: First, the experiment is computed on saliency models based on eye tracking. Second, the same experiment is calculated on saliency models based on salient object detection.

13.2.2 Method

Database and Ground Truth: The same database as in experiment 1 is used [7] with both region-based and eye tracking-based ground truths. However, in this experiment, the whole database is not used. Only the first three categories are interesting for this study and therefore employed: 50 images with large salient regions, 80 with intermediate salient regions and 60 with small salient regions.

Models: In the first part of this experiment, nine state-of-the-art models from experiment 1 have been chosen. These are only eye tracking-based algorithms: FSM

[9], SUN [10], SDLF [11], AIM [12], DVA [13], RARE [1], SR [14], QDCT [16] and AWS [19].

In the second part of this experiment, nine salient object detection-based state-of-the-art models have been chosen: FTSD [18], SSOI [22], SMSI [2], SLMC [23], SDHAS [24], SDAIR [25], SDBM [26], SIM [27] and SDWT [28].

Metrics: As in the first study, the pAUC (post-processing for area under the ROC curve (2011) [7]) metric has been chosen for this experiment because it can be applied to both eye tracking-based and region-based ground truths and mainly measures the eye fixation or region locations. Kendall's W concordance measure is used for the statistical analysis.

13.2.3 Results

13.2.3.1 Models with Eye Tracking

Figure 13.2 shows the results for pAUC intro the three categories for eye tracking-based algorithms. The mean trend can be computed by a linear regression (black line on Fig. 13.2). The general trend which can be highlighted is that the small regions have higher score than medium and large regions. This observation is correct for all models. We can also pay attention to the SR model which significantly increases (in terms of score rank) for small regions.

To assess the coherence between categories, the same ranking-based statistical test is required as in experiment 1; however, in this case, it is applied to the

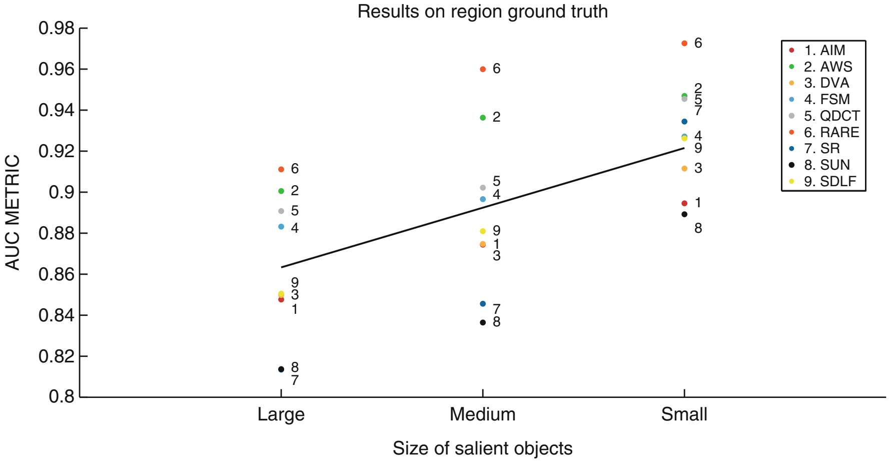


Fig. 13.2 Labelled region results on eye tracking-based algorithms on large, medium and small regions for pAUC

Table 13.4 Concordance based on Friedman test and Kendall's measure for large, medium and small regions

	Friedman test (p-value)	Kendall's concordance *W*
pAUC-labelled regions	$6 * 10^{-4}$	0.74

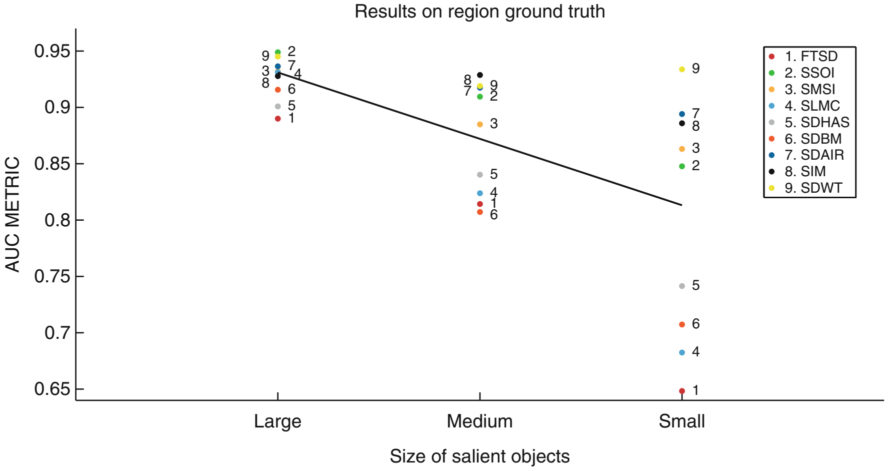


Fig. 13.3 Labelled region results on salient region detection algorithms on large, medium and small regions for pAUC

means of each of the three classes (large, medium and small). We use the averages because the number of images is different by categories. Kendall's *W* coefficient as used in experiment 1 shows us a smaller concordance. As shown in Table 13.4, the p-value is significant. It means that the ranks between models are statistically different between the three categories, but the size of this difference in terms of ranking is relatively small. Indeed, Kendall's concordance shows a moderate-strong agreement. In this experiment, the rankings are globally coherent (but less than between the two ground truths). So, the size of the salient region can have a stronger impact on our assessment than the chosen ground truth.

13.2.3.2 Models with Object Detection

Figure 13.3 shows the results on pAUC for the three categories for salient object detection-based algorithms. The mean tendency can be computed by a linear regression (black line on Fig. 13.3). The general trend which can be highlighted is the opposite of what we observed on Fig. 13.2. The large region has higher score than medium and small regions. This observation shows that most of saliency models are tuned to their ground truth (e.g. SOD-based models with the large binary masks and ET-based models with the small eye tracking distribution). It is correct

Table 13.5 Concordance based on Friedman test and Kendall's coefficient for large, medium and small regions

	Friedman test (p-value)	Kendall's concordance W
pAUC-labelled regions	$8 * 10^{-4}$	0.81

for almost all models. However, SDWT, for example, is different: its score is better with large salient regions than small ones, but its ranking is worse than both on medium regions. On the other hand, models with superpixels, like SDHAS, SDBM and SLMC, significantly decrease (in terms of score rank) for small regions.

To assess the coherence between categories, the same ranking-based statistical test is required as in the first part. We also use the average because the number of images is different depending on the categories. Kendall's W coefficient shows us a bigger concordance than in the first part. As shown in Table 13.5, the p-value is significant. It means that the ranks between models are statistically different between the three categories but the size of this difference in terms of ranking is relatively small. Indeed, Kendall's concordance shows a relatively strong agreement. In this experiment, the rankings are globally coherent (more than in the first part and approximately equal to the one between the two ground truths).

As mentioned for experiment 1, these results depend on the experimental design. In our case, one database, 18 saliency models divided in two groups and one metric have been chosen. However, the first part of this experiment was conducted again in our paper [21] based on another metric (NSS) and leads towards exactly the same conclusion. These results are not computed for SOD models. Indeed, NSS is not a metric for object labelling.

13.3 Experiment 3: Effects of Post-Processings

13.3.1 Goal

In this experiment, only databases with eye fixations will be used. The purpose is to investigate which post-processings increase the score of a saliency map. Indeed, there are three aspects which should be considered: the blurring, the border cut and the centre effects.

First, we study the blurring which is used to better correlate the noisy human eye movement data. Indeed, the saliency maps obtained from a model usually score lower than smoother versions of these maps. However, based on [5], there is an optimal Gaussian blur level for each model.

Then, we investigate the two other well-known problems for fair comparisons which are the centre bias and border effect. Centre bias means that a lot of fixations from natural image databases are located near the image centre because when taking pictures, the amateur photographer often places salient objects in the image centre.

The computational saliency models which include a centred Gaussian use the prior knowledge of working on natural images and increase their score on some metrics compared with other models without this information. Moreover, Zhang et al. [10] showed that metric scores are also corrupted by edge effects for the same reason. If we remove the edges of an image, metric scores usually increase as well. This is why a specific metric, called *sAUC*, has been designed to eliminate these undesirable effects. However, for other metrics (like NSS), these issues need to be taken into account.

These post-processing factors can dramatically influence some metric scores and affect the fairness of the validation. The main idea of this third experiment is to measure the impact of these factors on some saliency models.

13.3.2 Method

Database and Ground Truth: The database used here remains Jian Li's dataset [7] but only the eye fixation ground truth (collected with an eye tracker) will be employed. In this experiment, we use the whole database containing 235 images.

Models: For this experiment, six state-of-the-art models have been chosen. These are only eye tracking-based algorithms: FSM [9], SUN [10], AIM [12], DVA [13], RARE [1] and AWS [19].

Metrics: The NSS (normalized scanpath saliency (2005) [29]) metric has been chosen for this experiment. Kendall's *W* concordance measure is used for the statistical analysis.

13.3.3 Results

Figure 13.4 shows an example of smoothing effect for the six saliency models used in this experiment. To find this optimal blur width, we use the Y. Li's toolbox [30]. Some models such as FSM have already reached the optimal blur, while other models such as AIM, DVA and SUN increase their score with smoother maps.

For the six saliency models with optimal blur (SM), we first cut the edges (8 pixels at each border) of each saliency map. Second, we multiply the output of every saliency model by a centred Gaussian to observe their improvement.

Figure 13.5 illustrates how the post-processing factors impact the score of each model based on the NSS score. The general trend shows that all the scores increase. However, much depends on the saliency models.

Concerning the border cut, we observe that most of the saliency models such as AIM, FSM, DVA or RARE haven't improved their scores significantly. There are two reasons to this: some methods already remove edges into their mechanism or some selective models often have low score on the border. At the opposite, SUN

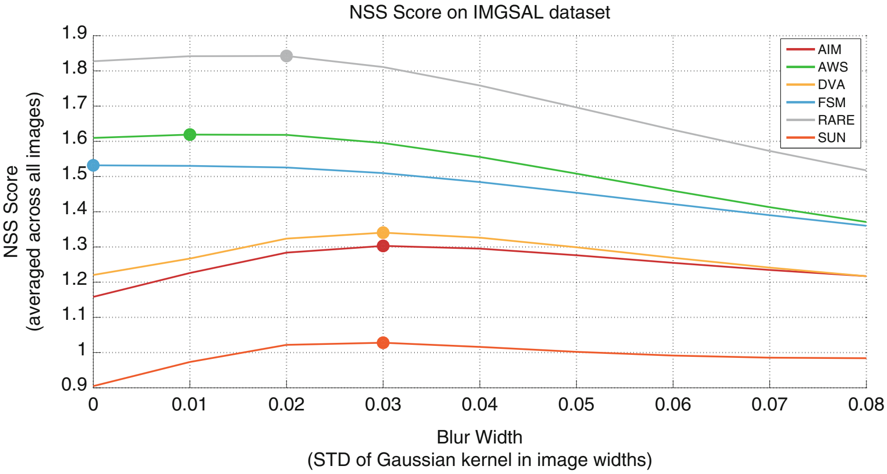


Fig. 13.4 Smoothing effect on six saliency models. The averaged NSS scores at all levels of blur widths are plotted and form a curve for each saliency model. The optimal blur is represented by a *dot*

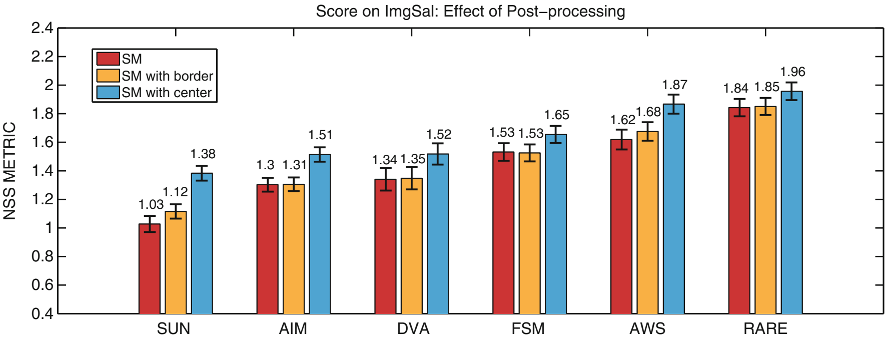


Fig. 13.5 Study of three post-processing factors: blurring, edge and centre effects. A higher NSS means that the model is better

improves its scores. It means that this model often has high values on its edges and needs to be more selective.

Concerning the 2D Gaussian centre, we can see that all models improve their score. These results confirm that many fixations are located near the image centre in Jian Li's database [7]. These measures can help quantify the centre bias of databases.

As mentioned above, these results depend on the experimental design. In our case, one database, six saliency models and one metric have been chosen. However, other correlated results from literature can be found in [5, 6, 10, 17], etc. They lead towards the same interpretation.

13.4 Experiment 4: Effects of Metrics

13.4.1 Goal

Due to the diversity of available metrics for eye fixation prediction assessment, several benchmarks were proposed. In 2011, Toets proposed in [4] to compare saliency models based on Spearman's rank correlation coefficient. In 2012, Borji built a benchmark [6] where three evaluation scores (PCC, NSS and sAUC) are used. Finally, Judd [5] proposed a platform using three different metrics: hAUC, S and EMD. Although these benchmarks are major contributions, none of those studies deeply discussed the relevance of their similarity metric mix.

The goal of this fourth experiment is twofold. First, it shows which metrics are close to each other. Second, it intends to reduce the dimensionality of the metrics we use and see which ones should be applied to do an efficient benchmark. Indeed, it is important to decide which metrics should be used together because they are complementary and which ones are useless to compute together because they will provide redundant information.

13.4.2 Method

Database and Ground Truth: The human eye fixation maps used are those in the database published by J. Li et al. [7] from experiment 1. This database provides eye fixation ground truth (collected with an eye tracker) for 235 colour images.

Models: In this experiment, the same twelve state-of-the-art models from experiment 1 have been chosen: FSM [9], SUN [10], SDLF [11], AIM [12], DVA [13], RARE [1], SR [14], PFT [15], QDCT [16],FTSD [18], SSAFD [17] and AWS [19].

Metrics: The 12 metrics presented in the previous chapter are used in this experiment. These metrics can be divided into three categories: value-based metrics which focus on saliency map values at eye gaze positions (NSS, P and PF), distribution-based metrics which focus on saliency and gaze statistical distributions (PCC, KLD, SCC, EMD and S) and location-based metrics which focus on location of salient regions at gaze positions (nAUC, pAUC, hAUC and sAUC). An average score by metric can thus be computed on the whole database for each model which leads to 12 different rankings of the 12 models, one for each comparison metric.

In the following, we will use the ranking between models and not their average score values. This is due to the fact that the output of the metrics can be very different in terms of range of score value and some of them should be maximized (correlation measures) while others should be minimized (divergence measures). Therefore, a direct score value comparison does not make a lot of sense. By contrast, the relative rank of the different models is a consistent measure common to all metrics and its range is here between 1 and 12 (respectively, from the best model to the weakest).

To compare model rank according to the different metrics, Kendall's W concordance measure [20] is used (as defined in Eq. 13.1 of experiment 1).

Kendall's W concordance is a coefficient measuring the degree of agreement between metrics. The value ranges from 0 (no agreement between model ranks) to 1 (full agreement, same model ranking). Furthermore, some rules of thumb are provided [20] to allow the researcher to interpret this measure as depicted in Table 13.2.

13.4.3 Results

13.4.3.1 Analysis of Consistency of Metrics

Intragroup Metrics: The concordance is computed between all metrics into the three categories: value-based (amplitude), location-based and distribution-based metrics (Table 13.6).

The concordance shows a moderate-strong agreement for location-based and distribution-based metrics. This means that these metrics provide some complementary information: they might provide different results for the same saliency map; thus, one of those metrics cannot just be ignored without a possible information loss about model ranking. However, one can see that the concordance between the amplitude metrics is high, which means that those measures are highly correlated and can therefore be summarized by a small subset of value-based metrics.

Intergroup Metrics: Contrary to the intragroup study that does not achieve enough concordance, the intergroup suggests that some metrics are very close as it is shown in Kendall's matrix of Fig. 13.6a. NSS, P, PCC and hAUC seem to be very close. On the opposite side, the KLD metric seems like an outlier in this matrix, and it is different from most of the other metrics in terms of model ranking.

To provide a better representation of the proximity in terms of model ranking among metrics, we apply, on Kendall's coefficient, a classical multidimensional scaling (MDS) technique which visualizes and explores similarities or dissimilarities in data. The results are displayed in Fig. 13.6b. In this representation, the x-axis (equivalent to a first eigenvector) is more important than the y-axis (equivalent to a second eigenvector). From the figure, one can see, for example, that PF and NSS are closer than PF and sAUC.

Table 13.6 Kendall's W coefficient of intragroup metrics

Group of metrics	W
Amplitude	0.9534
Distribution	0.7869
Location	0.8488

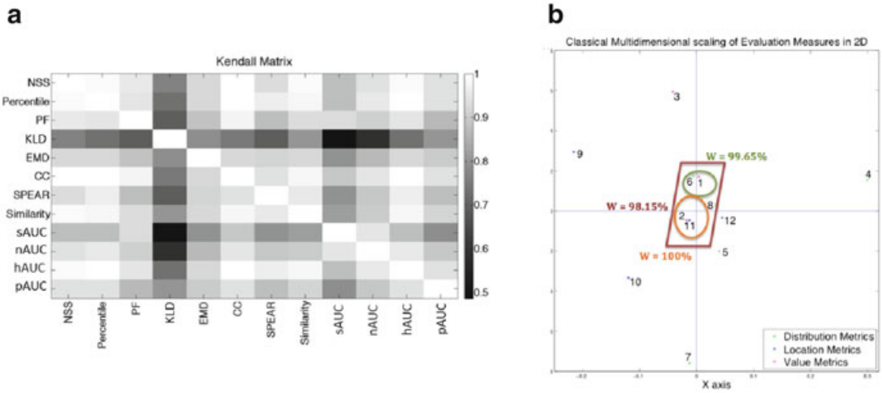


Fig. 13.6 Kendall's analysis. (**a**) Kendall's matrix on the 12 metrics. (**b**) Kendall's measure on group of metrics with classical multidimensional scaling of evaluation measures in 2D: 1. NSS / 2. P / 3. PF / 4. KLD / 5. EMD / 6. PCC / 7. SCC / 8. S / 9. sAUC / 10. nAUC / 11. hAUC / 12. pAUC

13.4.3.2 Study of the Dimensionality

Based on the representation of Fig. 13.6 and in order to reduce the dimensionality of metric space, we decide to use a concordance of 98 % as a threshold to fuse metrics (in terms of rank). By using this threshold, five metrics (NSS, P, PCC, S and hAUC) can be fused into a single metric called *cluster*. Indeed, as seen in Fig. 13.6b, the concordance between these metrics is 98.15 %. It means that only the rank of two or three couples of models has been inverted on the 12 models between these metrics. The ranking of *cluster* is defined as the mean ranking of all the metrics composing it.

For model validation, this *cluster* means that one measure from those included in this set is enough and the computation of the others inside this *cluster* is useless in terms of new information about model ranking. In this case, the five metrics can be summarized well enough by any of them.

To go further, a *global* metric which acts like the barycentre of all metrics is also computed as the mean of the ranking of all metrics.

The same study as in the first part of Sect. 13.4.3 is then applied but not on the same metrics. Indeed, we replace the five redundant metrics by the *cluster* metric and we add the *global* one. Kendall's matrix and the classical multidimensional scaling (MDS) technique are displayed in Fig. 13.7. We can observe that the *cluster* and *global* metrics are close. Moreover, along the x-axis (first eigenvector), the three metrics which cover most of the space are the cluster, sAUC and KLD.

These results depend on the experimental design. In our case, one database and 12 saliency models have been chosen. However, other saliency benchmarks exist online such as [5, 6] that use several metrics which lead towards the same interpretation.

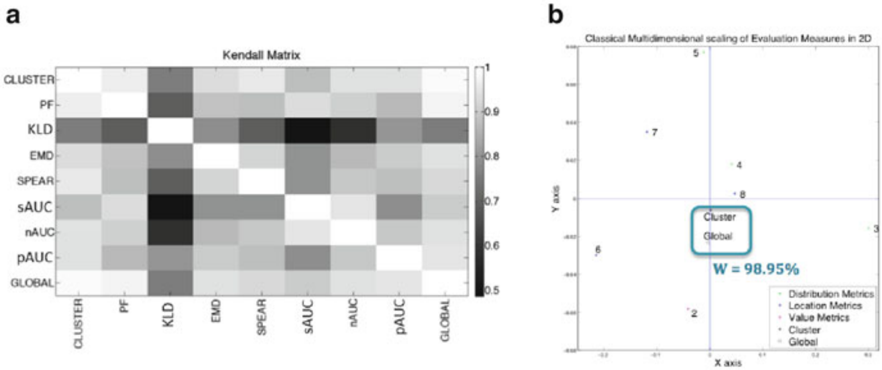


Fig. 13.7 Kendall's analysis. (**a**) Kendall's matrix on cluster, global and 7 metrics. (**b**) Kendall's measure on a group of metrics with classical multidimensional scaling of evaluation measures in 2D: 2. PF / 3. KLD / 4. EMD / 5. SCC / 6. sAUC / 7. nAUC / 8. pAUC

13.5 Conclusion

In conclusion, there are many parameters affecting visual saliency assessment. Four experiments investigate basic questions to fairly evaluate saliency maps with human gazes or labelled regions.

To build a validation framework, first, a database with ground truth needs to be chosen. Experiment 1 shows that there are significant differences between eye fixations and manually segmented salient region results, but the concordance between the rankings of models is strong. Moreover, the properties of the stimuli (e.g. in experiment 2, large, medium and small salient regions) are addressed with different degrees of accuracy by the saliency models. For eye tracking-based models, small salient regions are better detected than medium and large salient regions. With object detection, the exact opposite behaviour is observed. Therefore, the size of the salient region can have a stronger impact on our assessment than the chosen ground truth.

Consequently, for the validation framework which will be seen in the next chapter, three databases for still images have been chosen to have a large range of stimuli with only human eye fixation ground truth. Indeed, the purpose of RARE is only to find gaze distribution. Moreover, with experiment 1, if RARE is good in predicting human eye fixations, it will remain good in predicting human-labelled regions and conversely. Two databases for videos have been chosen with both ground truths. Indeed, the purpose of STRAP is to find gaze distribution but also to detect a salient object for application as seam carving.

Some metrics need to be chosen. For salient object detection, the gold standard F-measure is enough, but experiment 4 shows that one metric is not enough to evaluate the saliency model ranking on eye fixation data. The minimal set of similarity metrics which should be used is one of the metrics composing the cluster, sAUC and

KLD. The use of those three metrics is enough to cover most of the space (along the first eigenvector) and provide a fair ranking result.

For the validation framework, NSS has been selected to represent the *cluster*. This metric will be used with KLD which provides really complementary results and with sAUC which eliminates the effect of centred Gaussians. As only some models use centred Gaussians, eliminating this effect provides a fairer comparison.

Finally, state-of-the-art models must be selected. To be coherent, 18 eye tracking-based models have been used in the validation for still images and a mix of 9 eye tracking- and salient region detection-based models for videos. In terms of post-processing, experiment 3 shows that some factors such as centred bias, saliency map fuzziness and border cut have an important influence on the final result and can dramatically improve the score, especially for the centred bias. The optimal blur has been assigned on each model. For other parameters, in the validation framework, the ones given by authors have been kept. Indeed, with the chosen metric sAUC, the border cut and the centre Gaussian are not an advantage and a fair comparison can be done.

13.6 Summary

- Experiment 1 shows that the influence of the ground truth is not crucial: if models have good results with one ground truth, it is quite unlikely that these models completely fail with the other ground truths except due to statistical fluctuation.
- Experiment 2 shows that the properties of the stimuli (e.g. large, medium and small salient regions) are addressed with different degrees of accuracy by the saliency models. For eye tracking-based models, small salient regions are better detected than medium and large salient regions. With object detection, the exact opposite behaviour is observed. So the size of the salient region can have a stronger impact on our assessment than the chosen ground truth.
- Experiment 3 shows that several parameters such as centred bias, saliency map fuzziness and border cut have important influence on the final result. It is thus possible to optimize a model by choosing the best parameters.
- Experiment 4 shows that the minimal set of similarity metrics which should be used is (a) one of the metrics composing the cluster, (b) sAUC and (c) KLD. The use of those three metrics is enough to cover most of space and provide a fair ranking result.

References

1. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013). Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication, 28*(6), 642–658.

2. Vikram, T. N., Tscherepanow, M., & Wrede, B. (2012). A saliency map based on sampling an image into random rectangular regions of interest. *Pattern Recognition, 45*(9), 3114–3124.
3. Klein, D. A., & Frintrop, S. (2011). Center-surround divergence of feature statistics for salient object detection. In *IEEE International Conference on Computer Vision (ICCV 2011)*, Barcelona (pp. 2214–2219). IEEE.
4. Toet, A. (2011). Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(11), 2131–2146.
5. Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. MIT technical report.
6. Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing, 22*(1), 55–69.
7. Li, J., Levine, M., An, X., & He, H. (2011). Saliency detection based on frequency and spatial domain analyses. In *Proceedings of the British Machine Vision Conference* (pp. 86.1–86.11). BMVA. http://dx.doi.org/10.5244/C.25.86..
8. Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207.
9. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.
10. Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7), 32.
11. Antonio Torralba, M. C., Oliva, A., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review, 113*(4), 766–786.
12. Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. *Advances in Neural Information Processing Systems, 18*, 155–162.
13. Hou, X., & Zhang, L. (2008). Dynamic visual attention: searching for coding length increments. *NIPS, 5*, 7.
14. Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, Minneapolis.
15. Guo, C., Ma, Q., & Zhang, L. (2008). Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR 2008)*, Anchorage (pp. 1–8). IEEE.
16. Schauerte, B., & Stiefelhagen, R. (2012). Predicting human gaze using quaternion DCT image signature saliency and face detection. In *Proceedings of the 12th IEEE Workshop on the Applications of Computer Vision (WACV)/IEEE Winter Vision Meetings*, Breckenridge, Jan 2012 (pp. 9–11).
17. Li, J., Levine, M. D., An, X., Xu, X., & He, H. (2013). Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(4), 996–1010.
18. Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami (pp. 1597–1604). IEEE.
19. Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R., & Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision, 12*(6), 17.
20. Howell, D. (2012). *Statistical Methods for Psychology*. Belmont: Cengage Learning.
21. Riche, N., Duvinage, M., Mancas, M., Gosselin, B., & Dutoit, T. (2013). A study of parameters affecting visual saliency assessment. arXiv preprint arXiv:1307.5691.
22. Rahtu, E., Kannala, J., Salo, M., & Heikkilä, J. (2010). Segmenting salient objects from images and videos. In *Computer Vision–ECCV 2010*, Heraklion (pp. 366–379). Springer.

23. Xie, Y., Lu, H., & Yang, M.-H. (2013). Bayesian saliency via low and mid level cues. *IEEE Transactions on Image Processing, 22*(5), 1689–1698.
24. Fang, Y., Lin, W., Lee, B.-S., Lau, C.-T., Chen, Z., & Lin, C.-W. (2012). Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum. *IEEE Transactions on Multimedia, 14*(1), 187–198.
25. Fang, Y., Chen, Z., Lin, W., & Lin, C.-W. (2012). Saliency detection in the compressed domain for adaptive image retargeting. *IEEE Transactions on Image Processing, 21*(9), 3888–3901.
26. Xie, Y., & Lu, H. (2011). Visual saliency detection based on bayesian model. In *18th IEEE International Conference on Image Processing (ICIP 2011)*, Brussels (pp. 645–648). IEEE.
27. Margolin, R., Zelnik-Manor, L., & Tal, A. (2013). Saliency for image manipulation. *The Visual Computer, 29*(5), 381–392.
28. Imamoglu, N., Lin, W., & Fang, Y. (2013). A saliency detection model using low-level features based on wavelet transform. *IEEE Transactions on Multimedia, 15*(1), 96–105.
29. Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45*(18), 2397–2416.
30. Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The secrets of salient object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, Columbus (pp. 280–287). IEEE.

Chapter 14
Saliency Model Evaluation

Nicolas Riche

In this chapter, the validation framework will be applied on static and dynamic saliency models. The databases and metrics presented in the previous chapters will be employed, and the studies on the parameters influence have been taken into account as seen in the conclusion section. For each model, qualitative and quantitative results are detailed and explained. An additional comparative statistical analysis is applied on each quantitative result.

14.1 Validation of Saliency Models on Still Images

14.1.1 Qualitative Assessment for Still Image Models

Some qualitative results on synthetic patterns and selected images from the three static datasets for the validation framework are presented here. The goal of this section is to visually show results on one of the state-of-the-art saliency models called RARE [1] on simple and more complex images.

14.1.1.1 Synthetic Patterns

Psychophysical observations are synthetic stimuli showing a particular object (the target) among other objects (the distractors). All stimuli presented here have been widely used by the community [2, 3].

N. Riche (✉)
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: nicolas.riche@umons.ac.be

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_14

Nevertheless, RARE does not intend to fully explain human behaviour, and the dataset shown here is not large enough, and it has no eye-tracking data for an efficient comparison. The goal is to see if the global rarity and local contrast idea behind RARE make sense compared to human behaviour which will fixate the pop-out target. There are two parts in this section. First, eight synthetic patterns are selected for the specificity of their targets which are linked to RARE features: colour and orientation. In the second part, the selected targets are more complex. They are not necessarily directly linked to the features extracted by RARE.

In Fig. 14.1, RARE suitably reproduces pop-out phenomena related to colour and orientation targets. Indeed, the saliency is high (in red) on the targets. These results are expected due to the nature of the targets. For the colour/luminance differences,

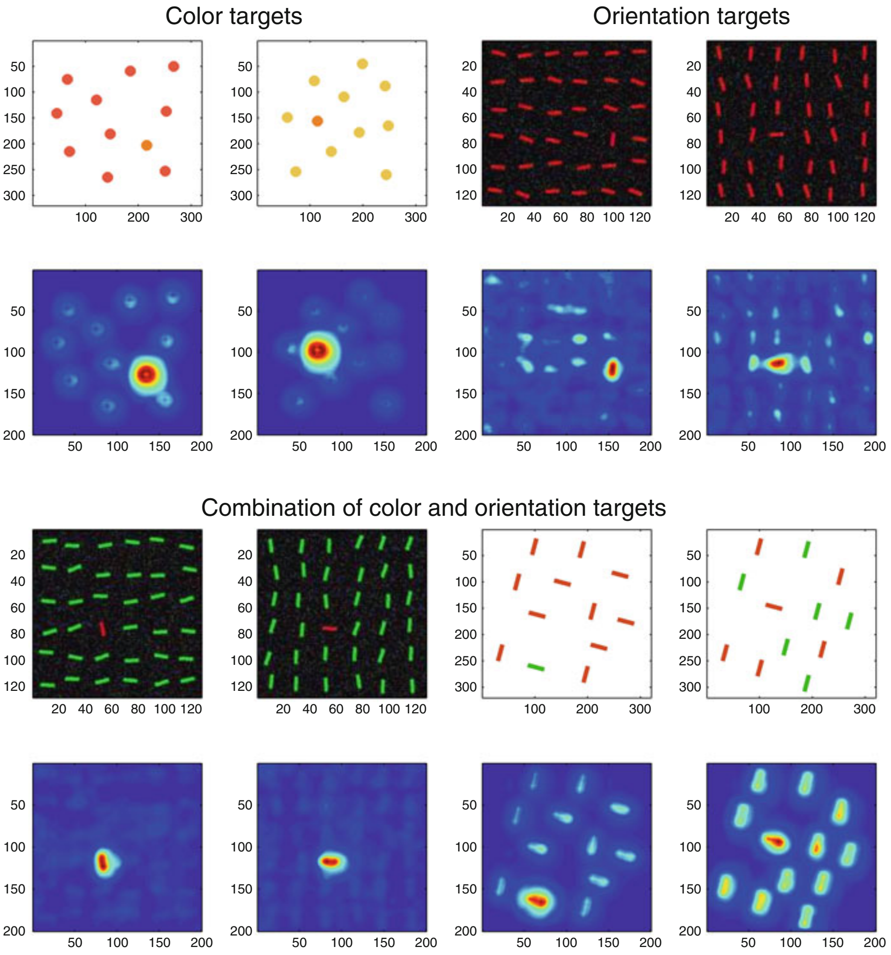


Fig. 14.1 Rows 1–2: Stimuli and RARE saliency maps for colour and orientation targets presented separately. Rows 3–4: Stimuli and related saliency maps for colour and orientation mixed targets. Globally, RARE works as expected

they are well detected even if the colour difference is not very important. This is due to the nature of the proposed model which is based on global rarity. Even if an object has a low contrast, but there are no other high contrast objects, it will be well highlighted. Concerning the combination of colour and orientation targets, it is interesting to see the influence of mixed targets or the heterogeneity of distractors. Indeed, the more distractors, the less selective the saliency map, even if the pop-out target is still detected as the maximum of the saliency map. This is again a consequence of the global rarity part of the algorithm. In Fig. 14.2, RARE points out all of the selected targets even if the features used here are more complex.

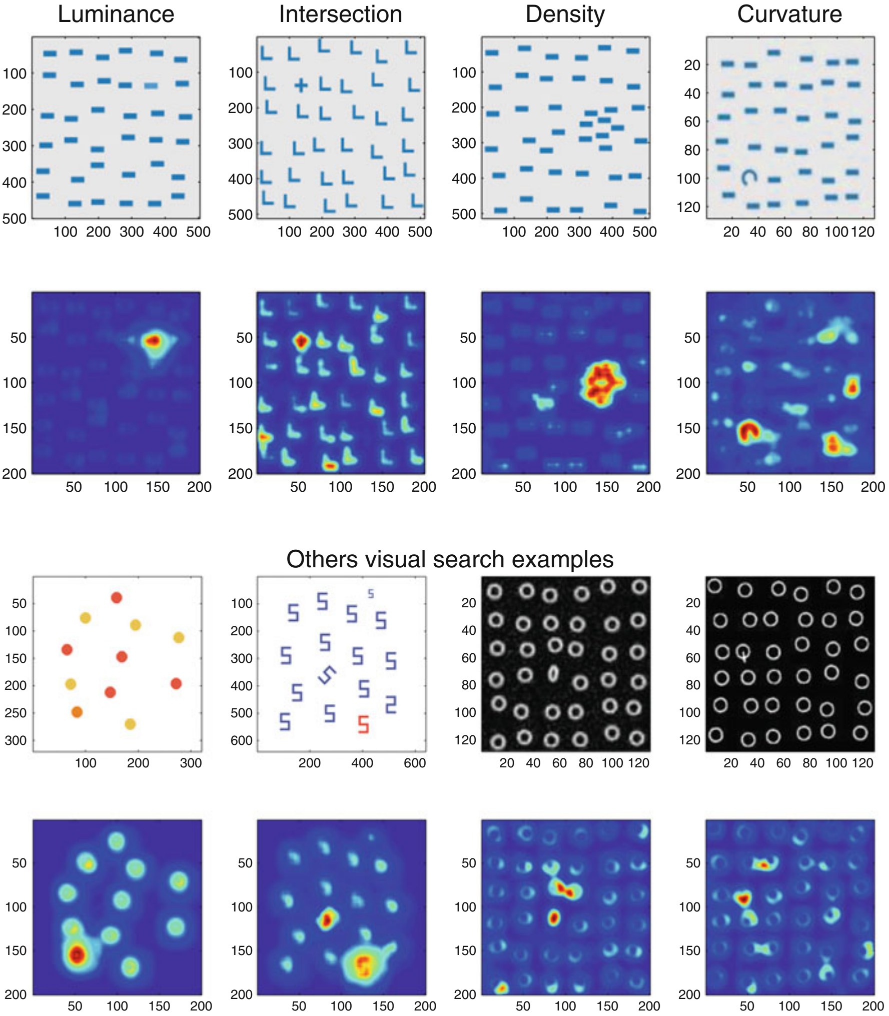


Fig. 14.2 Rows 1–2: Stimuli and RARE saliency maps for targets with different specificities. Rows 3–4: Stimuli and related saliency maps for synthetic patterns come from visual search task. Overall, RARE works slightly worse than in the first part. However, it also points out all targets

The selection of targets includes (1) luminance, (2) intersection and curvature, (3) density target and (4) visual search examples where all previous targets can be present. The saliency maps are noisier than in Fig. 14.1 but replicate the expected human behaviour. In addition to synthetic patterns, some qualitative results will be presented on selected images from the three datasets of the validation framework.

14.1.1.2 Toronto Database

Figure 14.3 displays selected images from the Toronto database (column 1). The eye- tracking results on these images which are superimposed on them (column 2) are compared to the results obtained from RARE (column 3) and the best state-of-the-art saliency models' results (following columns).

We first observe in Fig. 14.3 that three algorithms use a 2D-centred Gaussian as post-processing for their saliency maps, namely, GBVS, SERC and SKSE. There are many ways to introduce the 2D-centred Gaussian in a saliency model. Visually, this is clearly discernible within SERC and SKSE models. Then, we see that some methods, like AIM and GBVS, are less selective. This is specially the case with image 3 where the entire building is selected. Finally, RARE and AWS show similar results. They often find the salient distribution with more or less noise. For example, with image 1, these methods have some difficulties to find the salient area (the phone).

14.1.1.3 Kootstra Database

In Fig. 14.4, three images from the Kootstra database (column 1) with their eye-tracking data (column 2) are compared to RARE (column 3) and the best state-of-the-art saliency models (following columns).

In general, this database is more challenging than the Toronto database. Indeed, as shown in Fig. 14.4, there are a variety of challenges. On the first image, the background is very cluttered with repeating distractors. On the second one, there is

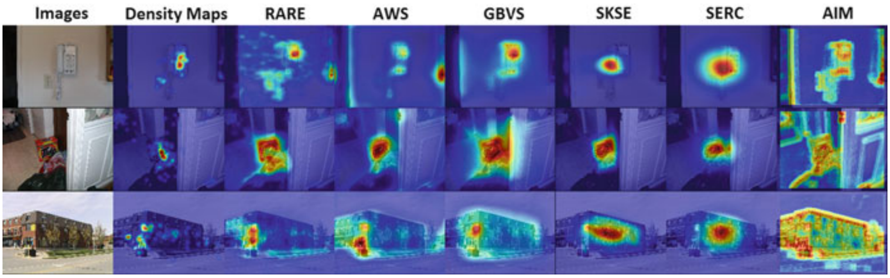


Fig. 14.3 Qualitative comparison of six models' (including the RARE algorithm) results with the eye-tracking ground truth (second column) on 3 images (rows) taken from the Toronto database

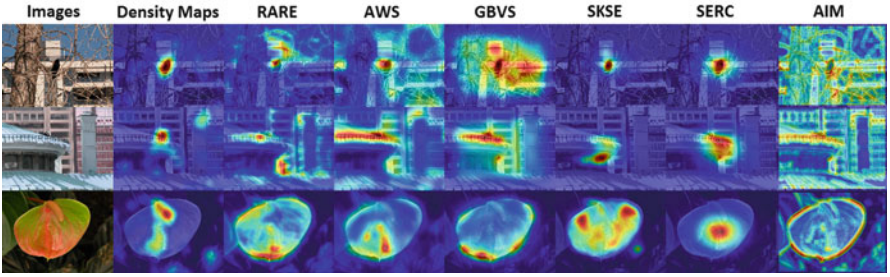


Fig. 14.4 Qualitative comparison of six models' (including the RARE algorithm) results with the eye-tracking ground truth (second column) on 3 images (rows) taken from the Kootstra database

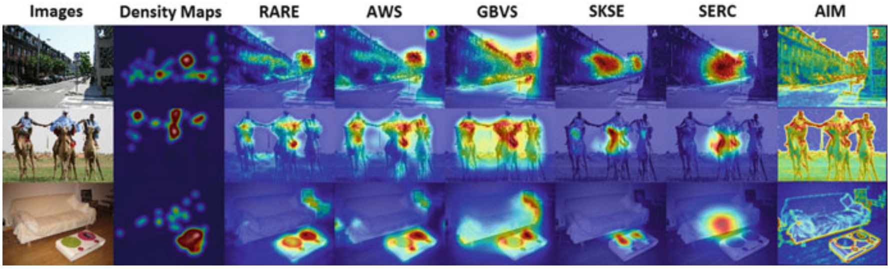


Fig. 14.5 Qualitative comparison of six models' (including the RARE algorithm) results with the eye-tracking ground truth (second column) on 3 images (rows) taken from the MIT database

no particular salient regions, only some buildings. Finally, on the last image, there is a large object which is displayed in most of image.

Visually, the same observations can be made regarding the algorithms. SERC and SKSE are still visual results close to a 2D-centred Gaussian. This approach is very efficient when there are no salient regions as in image 2. The SKSE distribution changes on the last image displaying a large object, but the salient area is not found. AIM and GBVS still have difficulties with the selectivity. In Fig. 14.4, this is the case with image 1 because the background is very complex. Finally, AWS and RARE find the salient area with more or less noise but still have difficulties with objects occupying a big part of the image.

14.1.1.4 MIT Database

Three images from the MIT database have been displayed in Fig. 14.5 (column 1). The eye- tracking results on these images (column 2) are compared to RARE (column 3) and the best saliency models (following columns).

In this database, the same characteristics can be seen for each model: centred Gaussian, selectivity and good detection. But it's interesting to watch more carefully image 2 where models fail. This is mainly due to the fact that here the bottom-up cues do not match with top-down information (faces, animals). This example also shows that purely bottom-up models are nowadays good enough to find most salient region and distribution. This can be improved by providing the top-down information.

In general, across all the databases, the behaviour of each model can be observed repeatedly.

14.1.2 Quantitative Validation

14.1.2.1 Experiment 1: Toronto Database

Figure 14.6 displays the RARE mean results along with their standard error compared to the other 18 saliency model results over the Toronto database. The graph shows how well saliency maps predict eye fixations under three metrics: sAUC, NSS and KLD. The models are displayed and sorted by metrics (from left to right).

RARE gives very competitive results on this dataset compared to the state-of-the-art models. It is the best performing model concerning the Toronto database relatively to NSS and KLD metrics and the third best performing one with sAUC metric. Based on the NSS metric, RARE is the only model which outperforms models with implicit centre bias (SKSE, GBVS and SERC). Besides, these models' performances are the worst with respect to the sAUC metric. However, even though RARE gives more accurate results compared to other models, this outperformance has not been proven significant. Statistical significance tests are required to verify this claim.

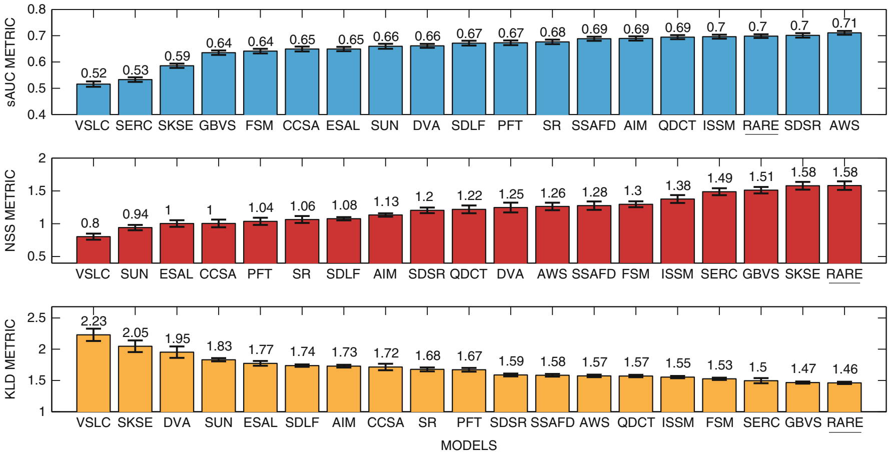


Fig. 14.6 Ranking saliency models over the Toronto database using three metrics: first row, mean and standard error of each model with sAUC; second row, with NSS; and third row, with KLD. For sAUC and NSS, higher is better, while for KLD, lower is better

For the statistical significance testing of mean scores between all models, we used a 95 % Confidence Interval (CI) Kruskal-Wallis test [4]. Indeed, preliminary to this statistical analysis, we checked by visual inspection if the metrics normality distribution assumption was met or not. Although the sAUC distributions of all models seem very close to the normal distribution, other metrics as KLD are clearly not normally distributed. We thus decided to use a Kruskal-Wallis statistical test that does not require normally distributed data.

Figure 14.7 gives a boxplot representation of each model for each metric and the results of the statistical test. The boxplot represents the data through their quartiles. The bottom and top of the bow are the first and third quartiles, and the band inside is the median. The whiskers in this case represent the lowest datum (still within 1.5 interquartile range) of the lower quartile and the highest datum (still within 1.5 interquartile range) of the upper quartile. Some outlier data can be represented with dots.

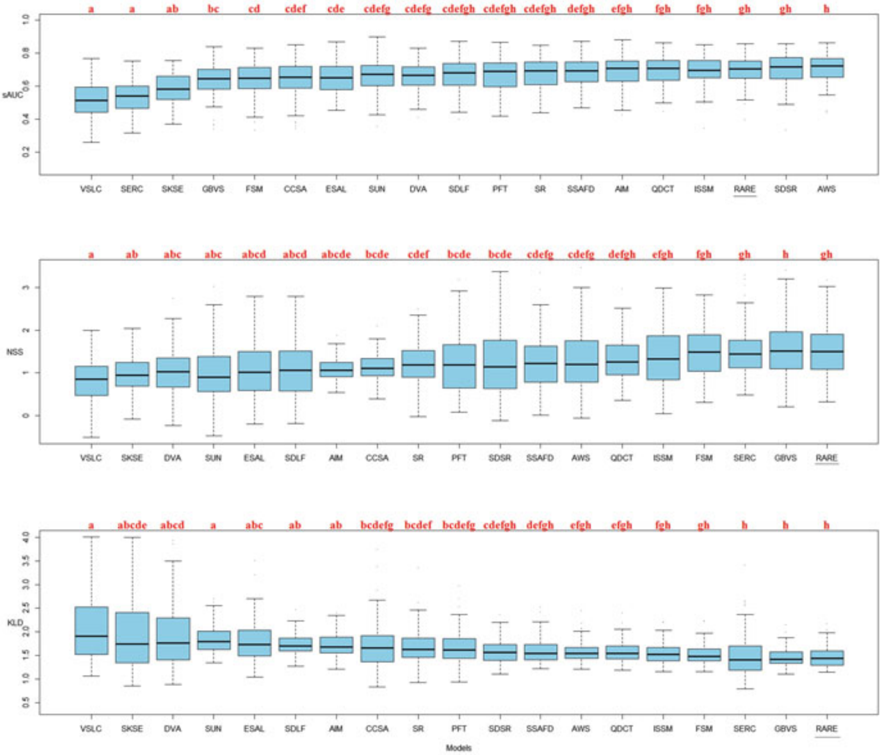


Fig. 14.7 Boxplot representation of statistical significance testing of mean scores between all models on the Toronto database. A 95 % CI Kruskal-Wallis test is used for each metric. The statistical results are given by the *red letters* above each model. If two models have the same *red letters*, the difference between them is not significant

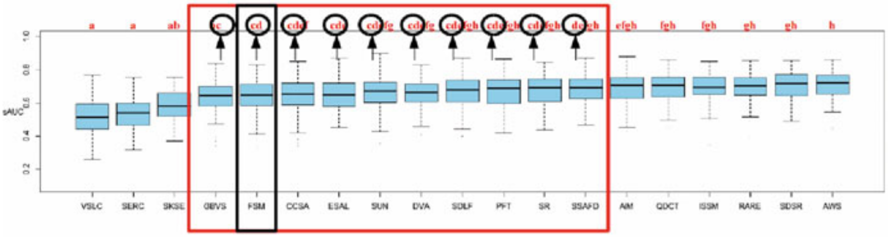


Fig. 14.8 Example of statistical results for the FSM method

The statistical results are given by the red letters above each model. For one metric, if two models have the same red letters, the difference between them is not significant. For example, in Fig. 14.8 based on sAUC, FSM method has two letters: c and d. It means that all the models with the same letters (from GBVS to SSAFD) are not significantly different. On the other hand, saliency models which have not the same letters are either significantly lower (from VSCL to SKSE) or higher (from AIM to AWS) than the FSM method.

In more general terms in Fig. 14.7, based on sAUC metric, AWS does not show a significantly better performance than SDLF (letter h). The best group of models (letter h) is composed of the top-ten models of Fig. 14.6 (sAUC), including RARE. Based on NSS metric, RARE does not significantly outperform QDCT (letter h). This group of models is composed of the top six ones of Fig. 14.6 (NSS). RARE outperforms significantly PFT but does not show a significantly better performance than SDSR (letter h) based on the KLD metric. The best group is therefore composed of the top nine models of Fig. 14.6 (KLD).

As a conclusion, RARE behaves well on this database composed of indoor and outdoor images. Based on the mean scores, it gives better results than the other models with respect to NSS and KLD and is always in the best group based on the statistical analysis.

14.1.2.2 Experiment 2: Kootstra Database

Figure 14.9 displays the RARE mean results along with their standard error compared to the other 18 saliency model results over the Kootstra database. Three metrics, sAUC, NSS and KLD, are used to show the saliency map performance. The models are displayed and sorted by metrics (from left to right).

RARE gives good results on this more challenging dataset compared to the state-of-the-art models. It is the best performing model in Kootstra on NSS metric and the second and fifth best performing one with the sAUC and KLD metrics, respectively. Based on the NSS metric, RARE, AWS and SSAFD outperform models with implicit centre bias (SKSE, GBVS and SERC). On the other side, these models have worst performance based on the sAUC metric.

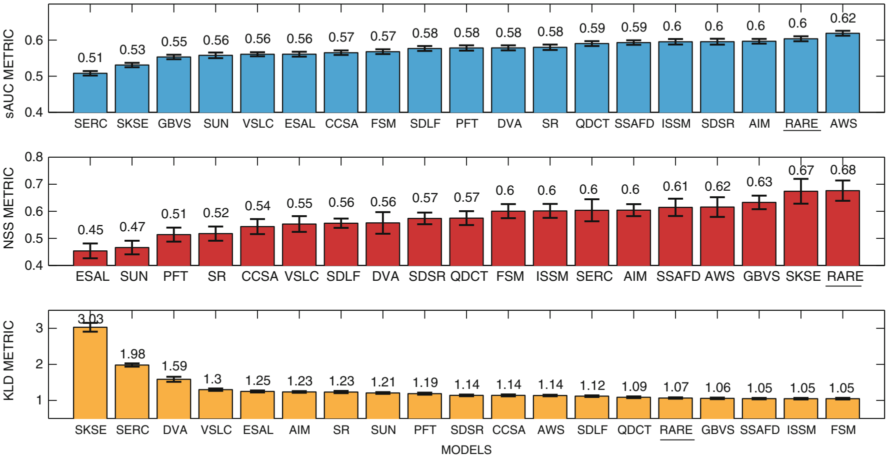


Fig. 14.9 Ranking saliency models over the Kootstra database using three metrics: first row, mean and standard error of each model with sAUC; second row, with NSS; and third row, with KLD. For sAUC and NSS, higher is better, while for KLD, lower is better

The same previously mentioned statistical test is performed in order to test the significance of the results. The results are shown in the boxplot of Fig. 14.10 where the red letters represent the results of each statistical test.

Based on sAUC, AWS outperforms significantly DVA but does not show a significantly better performance than PFT (letter g). RARE is part of this best group (letter g) composed of nine of the ten first models based on Fig. 14.9 (sAUC). With respect to NSS, RARE does not significantly outperform PFT (letter c). This best group is composed of almost all of the models. Only ESAL and SUN do not have the letter c and therefore are significantly lower than the other models. Based on the KLD, FSM does not show a significantly better performance than PFT (letter g). Only SKSE and SERC are significantly lower (letter a). RARE is in the first group composed of the top-eleven models based on Fig. 14.9 (KLD).

As a conclusion, RARE behaves well on this database composed of complex images. Based on the mean scores, it outperforms under one metric: NSS. Based on the statistical test, RARE always is in the best group. However, the first group always has a large number of models (nine, seventeen, eleven). This shows that all models don't perform well on this database.

14.1.2.3 Experiment 3: MIT Database

Figure 14.11 displays the RARE mean results along with their standard error compared to the other 18 saliency model results over the MIT database. Three metrics, sAUC, NSS and KLD, are used and sorted the models (from left to right).

RARE gives good results on this dataset with a lot of centre bias and top-down information (like faces) compared to other models, but it isn't the best performing

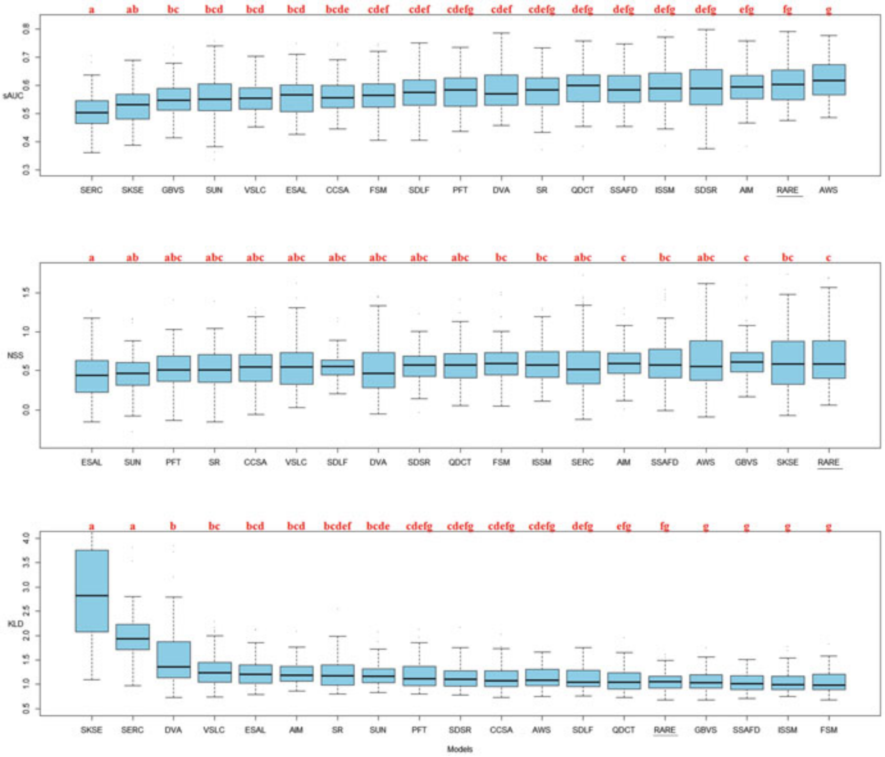


Fig. 14.10 Boxplot representation of statistical significance testing of mean scores between all models on the Kootstra database. A 95 % CI Kruskal-Wallis test is used for each metric. The statistical results are given by the *red letters* above each model. If two models have the same *red letters*, the difference between them is not significant

model on the MIT database. It is ranked third based on sAUC and KLD metrics and fourth with respect to NSS metric. Based on the NSS metric, all models with implicit centre bias (SKSE, GBVS and SERC) are the best performing ones. However, these model performances are the worst based on sAUC metric (Fig. 14.11).

To test the significant differences in the results of the means displayed in Fig. 14.12, an additional statistical 95 % CI Kruskal-Wallis test is required as previously mentioned. Figure 14.12 also gives the results of each statistical test by giving the red letters.

Based on sAUC metric, AWS does not show a significantly better performance than RARE (letter i). The models from SERC to SKSE are significantly lower than others models (letters a and b). RARE is in the best group composed of the top three models of Fig. 14.11 (sAUC). SKSE does not show a significantly better performance than GBVS (letter i) based on NSS metric. This is the first time that RARE is not part of the first group (letter i), but it is ranked in a second group (letter h) with the GBVS model. Indeed, GBVS does not significantly outperform

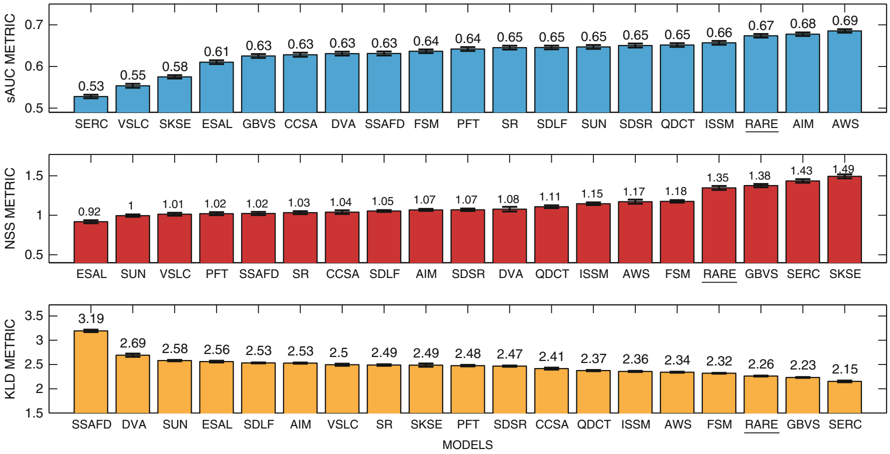


Fig. 14.11 Ranking saliency models over the MIT database using three metrics: first row, mean and standard error of each model with sAUC; second row, with NSS; and third row, with KLD. For sAUC and NSS, higher is better, while for KLD, lower is better

RARE, while RARE does not significantly outperforms FSM (third group, letter g). It is a good result because in the first group composed of the top three models based on Fig. 14.11 (NSS), all the models use implicit centred Gaussian. Based on the KLD metric, SERC does not show a significantly better performance than GBVS (letter i). RARE is ranked in the second group (letter h) with GBVS. However, contrary to the results with the NSS metric, RARE significantly outperforms FSM (third group, letter g). It is a good result for RARE because in the first group composed of the top-two models on Fig. 14.11 (KLD), all the models use implicit centred Gaussian.

As a conclusion, RARE behaves well on this database composed of images with centre bias and top-down information. Based on the statistical test, RARE is in the best group with respect to sAUC and among the second best performing based on NSS and KLD. However, the only models that significantly outperform RARE based on NSS and KLD are SKSE and SERC. These models use centred Gaussian which is an advantage on this database. We also observe larger significant differences between the models in this database (letter i instead of h, g or even c into others databases). This observation is due to the number of images (1003 images here instead of 100 images).

14.1.3 Multidimensional Scaling Analysis

To complete this comparison, the classical multidimensional scaling (MDS) technique has been applied, but this time, the distances of this MDS are computed

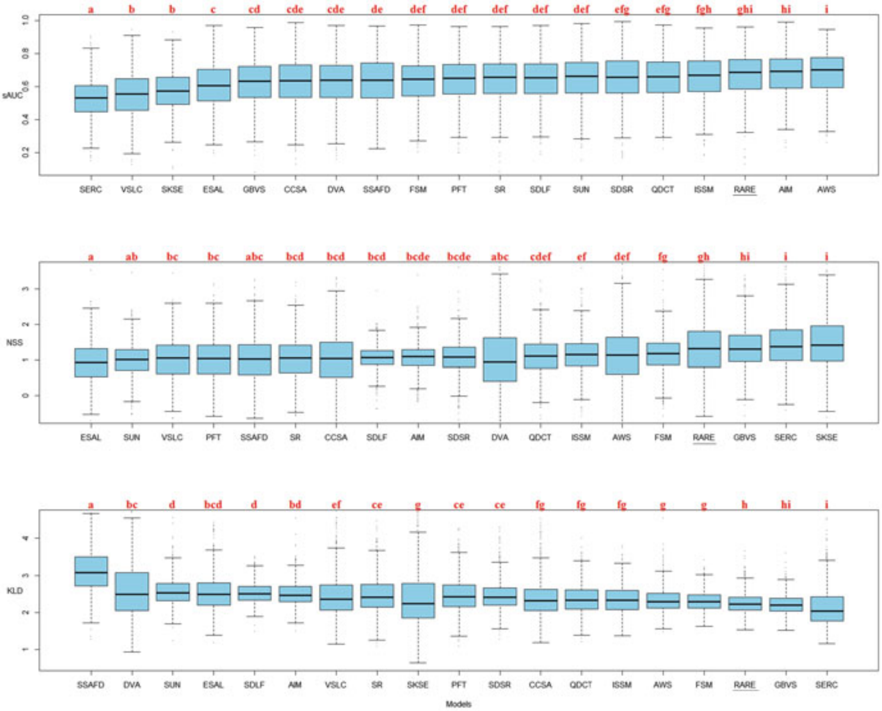


Fig. 14.12 Boxplot representation of statistical significance testing of mean scores between all models on the MIT database. A 95 % CI Kruskal-Wallis test is used for each metric. The statistical results are given by the *red letters* above each model. If two models have the same *red letters*, the difference between them is not significant

Table 14.1 Example of score concatenation for the calculation of a distance between two saliency models (AIM and SR) based on the scores

	Toronto database							Kootstra
	Img 1			Img 2			. . .	Img 1
	sAUC	NSS	KLD	SAUC	NSS	KLD		
AIM	0.7	1.07	1.7	0.67	1.1	1.63	. . .	. . .
SR	0.5	1.05	1.8	0.7	1.2	1.62		
D	0.2	0.02	0.1	0.03	0.1	0.01		. . .

from the scores obtained from the different metrics rather than the characteristics. Table 14.1 shows an example of how we calculate a distance between two saliency models (AIM and SR). For each model, the scores on different metrics and databases have been concatenated, and to build the distance matrix, the distances between each pair of models have been measured. The purpose is to visualize in 2D the similarity

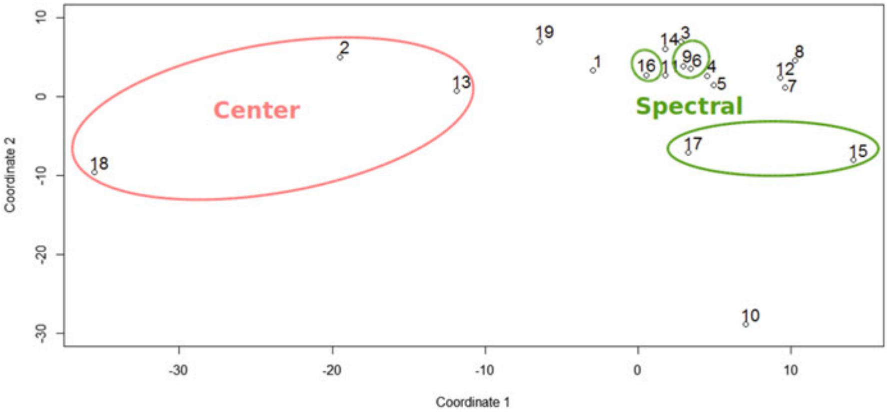


Fig. 14.13 Multidimensional scaling of nineteen eye- tracking saliency models based on score in 2D: 1. FSM / 2. GBVS / 3. CCSA / 4. AIM / 5. SDLF / 6. SR / 7. SUN / 8. DVA / 9. PFT / 10. SDSR / 11. VSLC / 12. ESAL / 13. SKSE / 14. AWS / 15. SSAFD / 16. ISSM / 17. QDCT / 18. SERC / 19. RARE

level between the models based on scores and compare the correspondence between the model characteristics and the scores.

We can see from Fig. 14.13 that on one side, saliency models with 2D-centred Gaussian bias (Models: 2,13 and 18) appear to have distances in the same range relatively to other models. On the other, side saliency models with spectral mechanism (Models: 6, 9, 15, 16 and 17) also seem to have distances in the same range.

Compared to the MDS presented in previous chapter, only the two observations explained above are found. We cannot see the impact of the stimuli characteristic (greyscale or colour) and of the approaches (local or global). It means that the model scores cannot be found from the characteristics (unlike the post processing).

14.2 Video Saliency Models Validation

14.2.1 Qualitative Assessment

Some qualitative results on selected frames from the two dynamic datasets for the validation framework are presented here. The goal of this section is to visually show results of one of the state-of-the-art video saliency model called STRAP on simple and more complex frames of videos. As for still images, the models, databases and metrics used for evaluation are described in the previous chapters.

14.2.1.1 ASCMN Database

Figure 14.14 shows some qualitative results on the ASCMN database. One can find on the first line 5 different example frames from each of the five categories of videos in the dataset (abnormal video, surveillance video, crowd video, video with a moving camera and videos with high motion contrast). The second line of the figure shows the eye-tracking density maps as heat maps on the original frames. The three following lines show results for the three best models, namely, STRAP, GBVS [5] and STVSM [6].

STRAP works well in all the situations, but it sometimes introduces peripheral noise. GBVS sometimes misses the main interesting regions in the frames, but when they are well detected, there is few noise, and the saliency map is well focused on these regions. STVSM performs well and also provides more focused results on part of the regions of interest. Overall, the three methods provide quite similar results.

14.2.1.2 SVS Database

On Fig. 14.15, 4 different sequences (Harbour, Tempete, Hall Monitor and Mother Daughter) have been selected to illustrate the results (columns). The two first sequences have lots of movements, while the two other sequences have a static

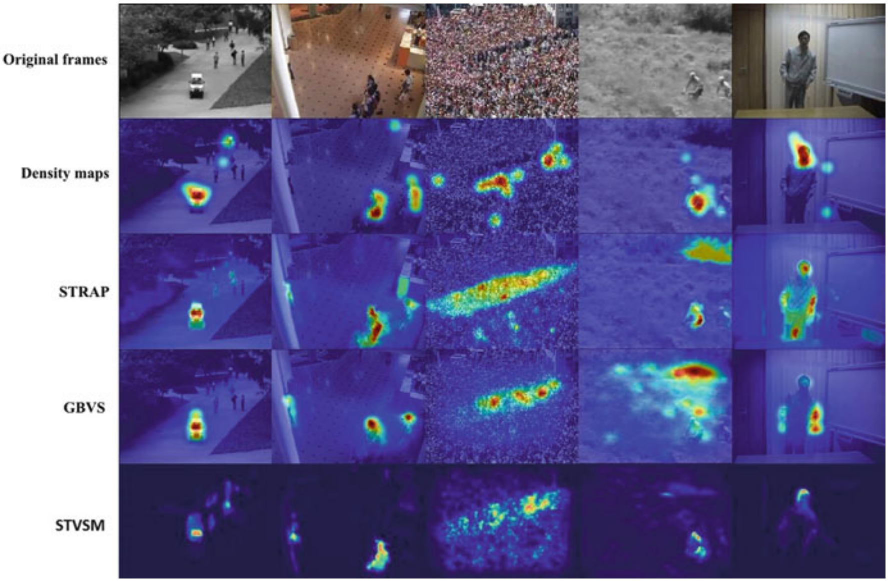


Fig. 14.14 Visual results of the best saliency models on ASCMN database: STRAP, GBVS [5], STVSM [6] compared with the eye-tracking density maps (row 2) for different original frames (row 1)

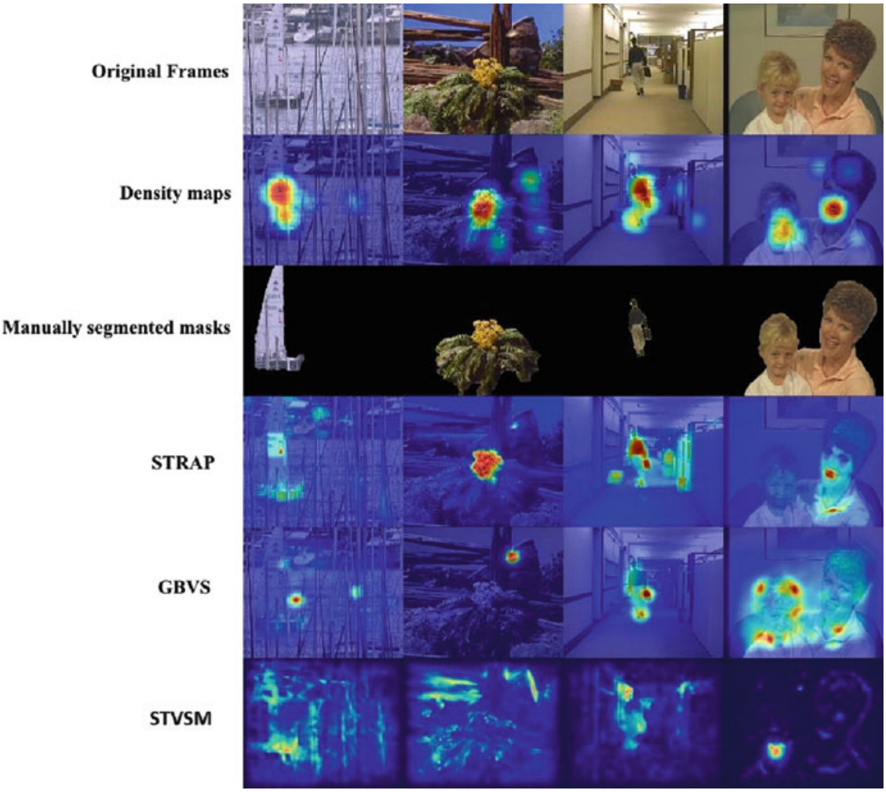


Fig. 14.15 Visual results of the best saliency models on the SVS database: STRAP, GBVS [5], STVSM [6] compared with the eye-tracking density maps (row 2), manually segmented masks (row 3) for different original frames (row 1)

camera with a salient moving object. The original frames can be seen on the first line, and the density maps, used as reference, are displayed on the second line. The third line shows the manually segmented masks which represent the second ground truth. The three following lines show again the best models STRAP, GBVS and STVSM.

It can be seen that the result for STRAP performs very well for Tempete and Hall Monitor. On Harbour and Mother Daughter, the salient objects are well identified, but the results could be more intense. Concerning the manually segmented masks, STRAP always hits inside, but sometimes the objects are not fully spotted.

GBVS is a very selective model which provides relatively good highlight of a small part of the salient objects. The most important object can be missed, and this very spotted approach can be a problem in the manually segmented masks detection assessment.

STVSM misses the important regions in videos containing a lot of motion and also has issues when the face detection algorithm fails. For the manually segmented masks, this approach seems less efficient.

14.2.2 Quantitative Validation

For quantitative assessment, three different experiments are carried out. First, in experiment 1, the ASCMN database is used to compare STRAP with the other models using the eye-tracking ground truth and three metrics. In experiment 2, STRAP is compared with state-of-the-art techniques based on the videos of the SVS database using the eye-tracking ground truth and three metrics. Finally, in experiment 3, STRAP is compared again with the state-of-the-art models on the SVS database, but this time on the manually segmented masks ground truth using one adapted metric.

14.2.2.1 Experiment 1: ASCMN Database

In this first experiment, we use the ASCMN database to assess the proposed STRAP model. Figure 14.16 displays the STRAP mean results along with their standard error compared to the other nine video saliency models over ASCMN database.

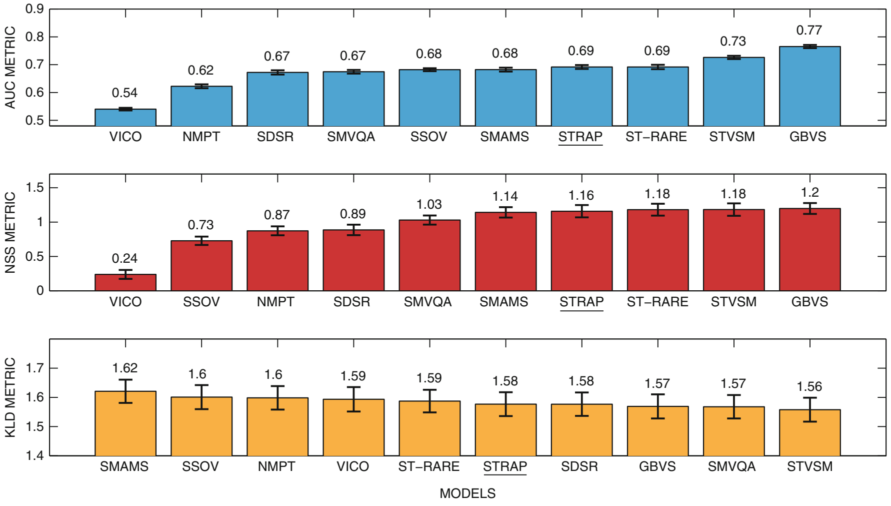


Fig. 14.16 Ranking saliency models over ASCMN database using three metrics: first row, mean and standard error of each model with sAUC; second row, with NSS; and third row, with KLD. For sAUC and NSS, higher is better, while for KLD, lower is better

The graph shows the performance of saliency maps under three metrics: AUC and NSS (higher score) and KLD (lower score). It is important to note that sAUC is not appropriate for the videos. Indeed, the centre bias is less present and appears only during the first second of a video [7]. Moreover, the sAUC metric is time consuming and therefore not suited to the videos. This is why we replace sAUC by a classical implementation of AUC which has been realized by S. Schroedl and is based on [8]. Therefore, we always have one metric for each category (location, distribution and amplitude).

STRAP gave good results on this specific dataset composed of five different kinds of movements which can be met in real-life scenarios. Compared with other models, it is not the best performing model in ASCMN on one metric. Indeed, STRAP is not specifically tuned to predict eye distribution but performs well compared to specific models as GBVS, STVSM and ST-RARE. It occupies the fourth rank on AUC and NSS metric and the fifth one with KLD metric.

However, to know which models outperform others, an additional statistical significance test is required. The same statistical significance 95 % CI Kruskal-Wallis test that is used for still images which do not require normally distributed data is applied. Figure 14.7 gives a boxplot representation of each model for each metric, and the results of each statistical test are given by the red letters (Fig. 14.17).

A general trend is that video saliency models are more significantly different than the image ones. This is also due to the fact that there are a lot of frames (implying more scores) during video. Based on AUC metric, GBVS shows a significantly better performance than other models (letter h). STRAP and ST-RARE are not significantly different (letter f) and are in the third group. Moreover, GBVS also shows a significantly better performance than other models (letter g) based on NSS metric. STRAP is in the second group as SMAMS or STVSM (letter f). It is more complicated based on the KLD metric. Indeed, STVSM does not show better result than STRAP (letter d). However, SMQVA shows better performance (letter e). STRAP is therefore in the second group of models (letter d) as STVSM, GBVS or SDSR.

As a conclusion, STRAP behaves well on this database composed of videos with different kind of motion. Based on the statistical test, STRAP is in the third group with AUC and the second on NSS and KLD.

14.2.2.2 Experiment 2: SVS Database Using the Eye Fixation Ground Truth

The videos in this database are very different compared to ASCMN. Indeed, there are only high-quality videos with very complex scenes with a lot of camera motion (zoom, tracking) or faces. In this case, the use of the temporal compensation has a very important role into this database.

Figure 14.18 displays the mean results along with their standard error of STRAP compared to the other nine video saliency models over eye tracking of SVS database.

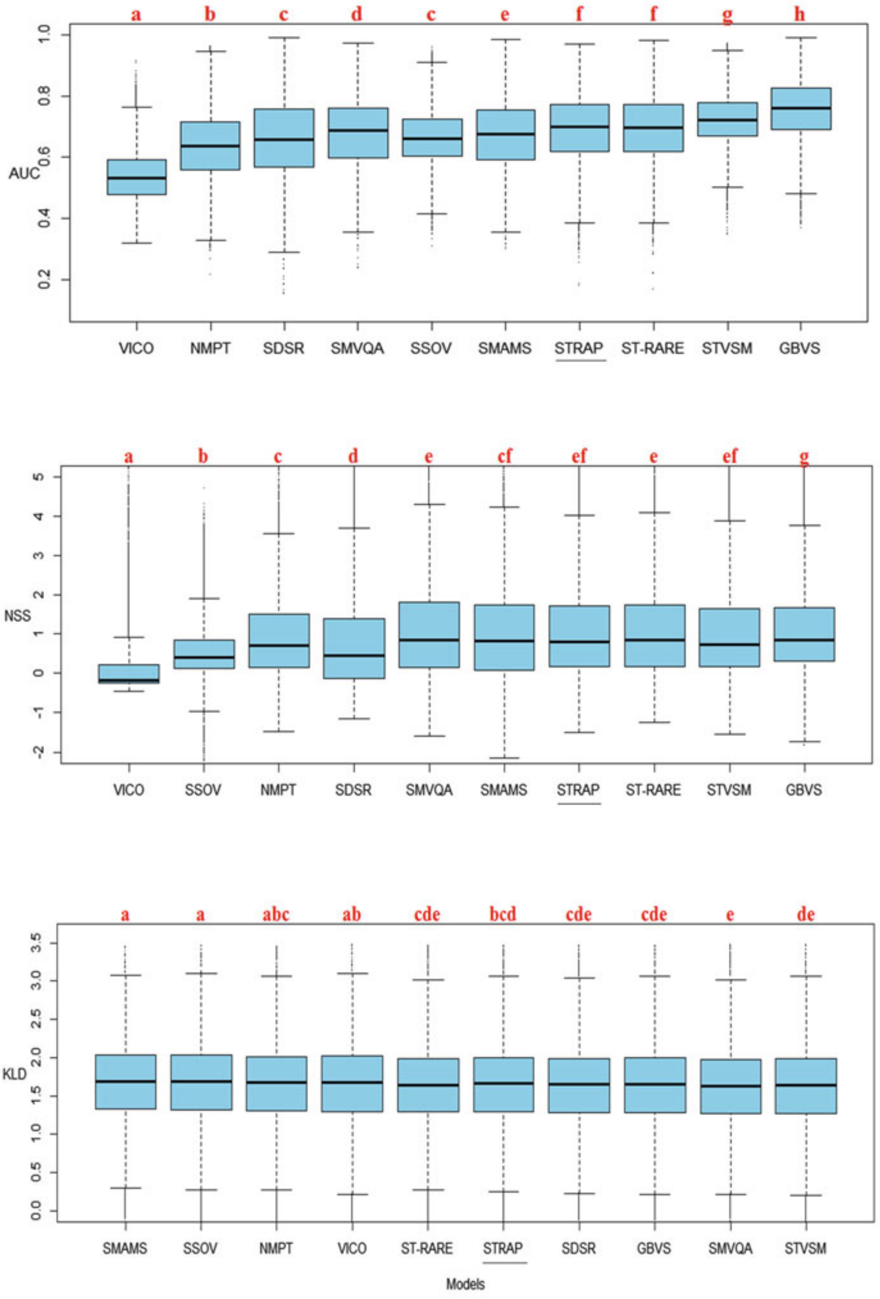


Fig. 14.17 Boxplot representation of statistical significance testing of mean scores between all models on ASCMN database. A 95 % CI Kruskal-Wallis test is used for each metric. The statistical results are given by the *red letters* above each model. If two models have the same *red letters*, the difference between them is not significant

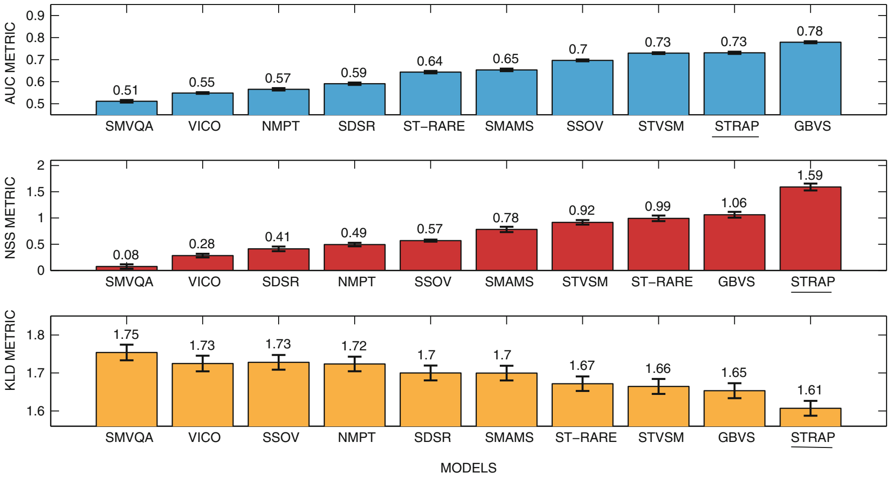


Fig. 14.18 Ranking saliency models over SVS database using three metrics: first row, mean and standard error of each model with AUC; second row, with NSS; and third row, with KLD. For AUC and NSS, higher is better, while for KLD, lower is better

The graph shows the performance of saliency maps under three metrics: AUC and NSS (higher score) and KLD (lower score). As in Experiment 1, AUC is used instead of sAUC. STRAP gives very good results on this dataset composed of high-quality videos with moving camera. Compared with other models, it is the best performing model in SVS on two metric: NSS and KLD. It ranks number two on AUC metric.

However, as in Experiment 1, to know which models outperform others, an additional statistical significance 95 % CI Kruskal-Wallis test is required. Figure 14.19 gives a boxplot representation of each model for each metric, and the results of each statistical test is given by the red letters.

Based on AUC metric, GBVS shows a significantly better performance than other models (letter h). STRAP and STVSM are not significantly different (letter g) and represent the second group. Based on NSS and KLD metrics, STRAP shows a significantly better performance than all the other models (respectively letter i and f).

As a conclusion, STRAP behaves very well on this database composed of complex videos with moving camera. STRAP is the best model with NSS and KLD based on the statistical test and in the second group based on AUC.

14.2.2.3 Experiment 3: SVS Database Using the Object-Oriented Ground Truth

Concerning the manually segmented masks validation into the YUV database, Fig. 14.20 displays the mean results along with their standard error of STRAP

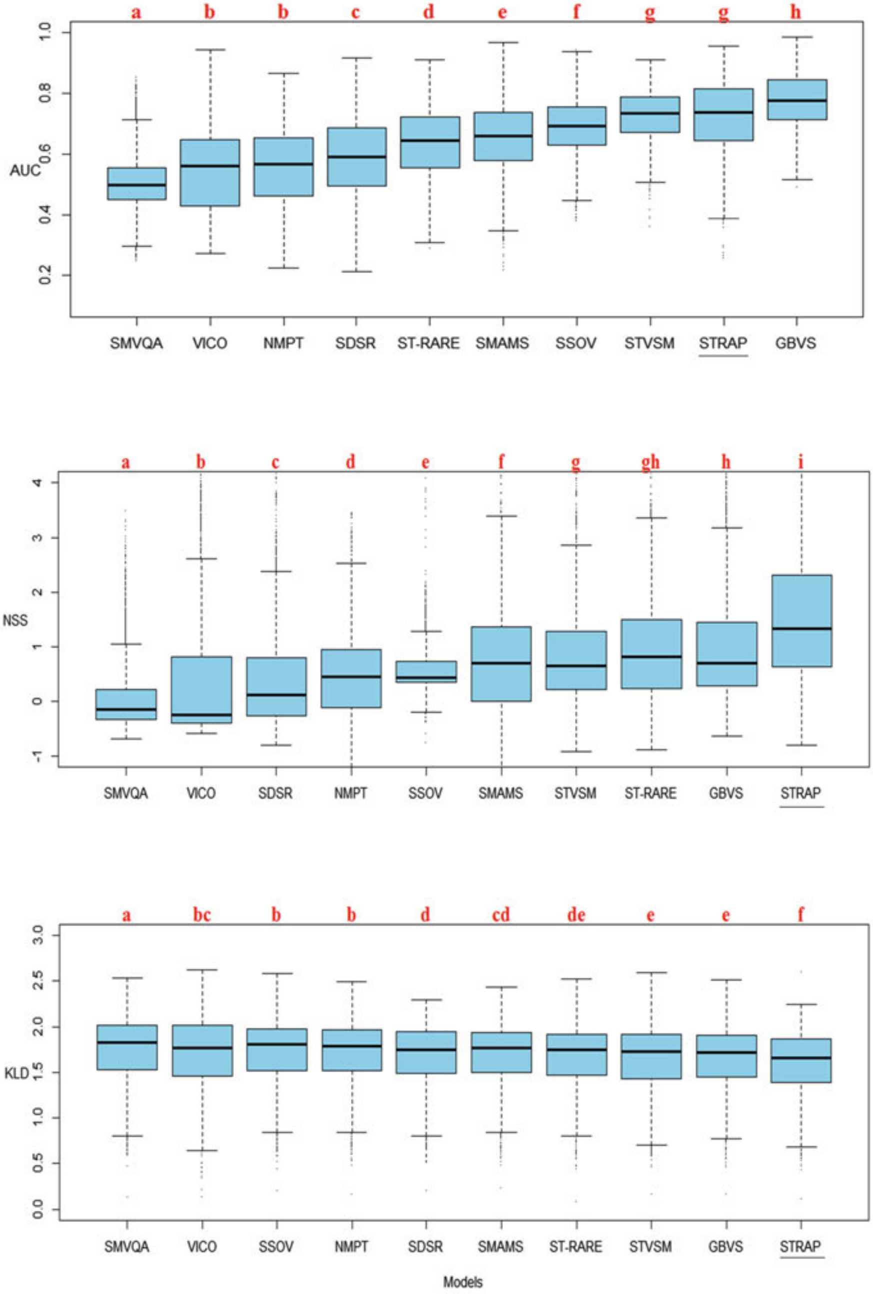


Fig. 14.19 Boxplot representation of statistical significance testing of mean scores on eye tracking between all models on SVS database. A 95 % CI Kruskal-Wallis test is used for each metric. The statistical results are given by the *red letters* above each model. If two models have the same *red letters*, the difference between them is not significant

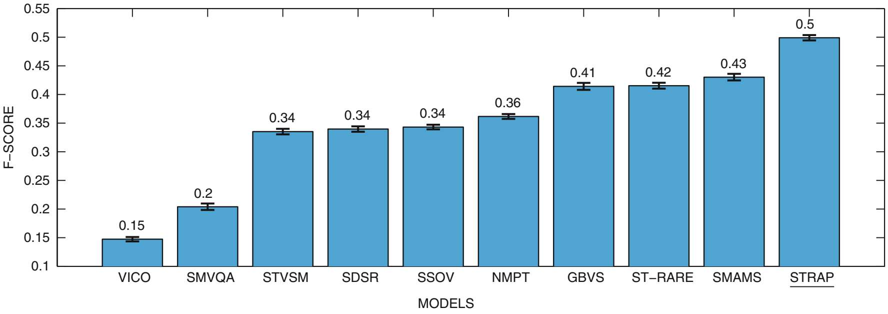


Fig. 14.20 Ranking saliency models over SVS database with binary masks using one metric. Mean and standard error of each model with F-score: higher is better

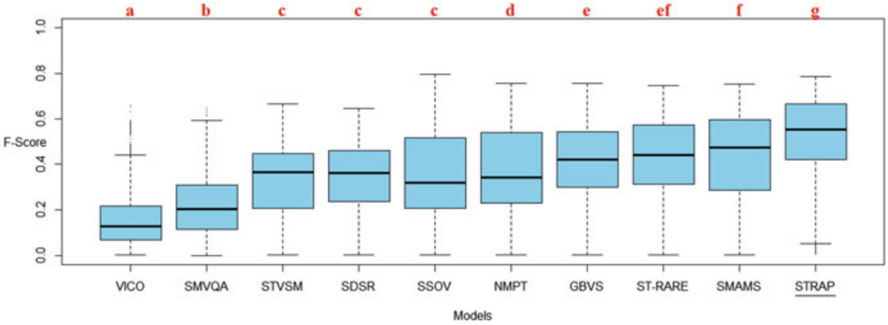


Fig. 14.21 Boxplot representation of statistical significance testing of mean score on salient object detection between all models on SVS database. A 95 % CI Kruskal-Wallis test is used. The statistical results are given by the *red letters* above each model. If two models have the same *red letters*, the difference between them is not significant

compared to the other nine video saliency models over binary masks of SVS database. The graph shows the performance of saliency maps under one metric: F-score (higher score is better). Compared with other models, it is the best performing model in SVS on this metric.

However, to know which models outperform others, an additional statistical significance 95 % CI Kruskal-Wallis test is required. Figure 14.19 gives a boxplot representation of each model for the metric and the results of each statistical test are given by the red letters.

Based on F-score metric, STRAP shows a significantly better performance than all the other models (letter g) (Fig. 14.21).

As a conclusion, STRAP performances are in line with the state-of-the-art models for eye fixations, but it definitely outperforms all the other models in case of object-oriented ground truth. The use of the segmentation module, while also supporting a little the eye fixation ground truth (see Fig. 14.15, row 2), has a very important role in object-oriented ground truth (see Fig. 14.15, row 3). The ability to handle both eye fixations and object prediction is a unique ability of STRAP.

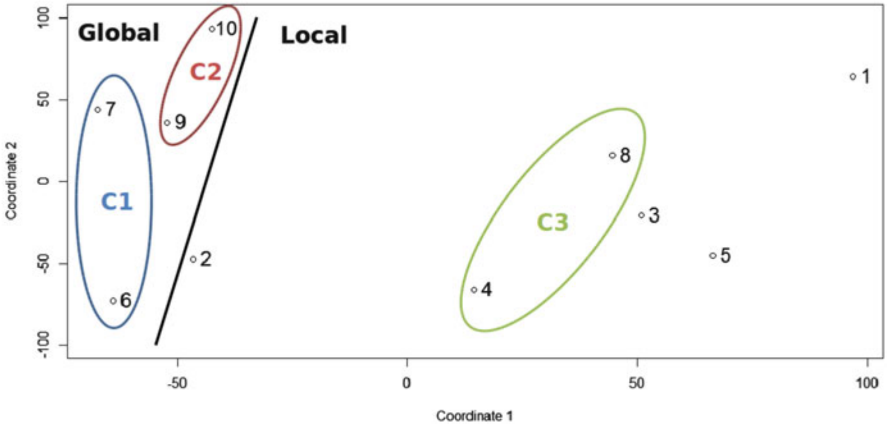


Fig. 14.22 Multidimensional scaling of ten saliency models for videos based on scores in 2D: 1. GBVS / 2. NMPT / 3. SSOV / 4. SDSR / 5. VICO / 6. SMVQA / 7. SMAMS / 8. STVSM / 9. ST-RARE / 10. STRAP

14.2.3 Multidimensional Scaling Analysis

To complete this assessment, as for still images, the classical multidimensional scaling (MDS) technique similar to the one exposed in Sect. 14.1.3 has been chosen. To build the distance matrix, the scores on different metrics and databases have also been concatenated, and distances between each pair of models have been measured for each model. The purpose is to visualize the similarity level between the dynamic models based on scores and compare the correspondance between the model characteristics and the scores obtained.

We can see from Fig. 14.22 that the first coordinate divides models in two clusters; on one side (right, models: 1, 3, 4, 5 and 8), saliency models with local approach appear to have distances in the same range relatively to other models. On the other side (left, models: 6, 7, 9 and 10), saliency models with global approach also seem to also have distances in the same range. Model 2 acts as an outlier.

Comparing with the MDS presented in the previous chapter, contrary to still images, most observations explained above are also found with the MDS based on video characteristics. Besides the fact that the first coordinates correspond to the approach characteristic, one another interesting observation is that three of the four clusters (from C1 to C3 on Fig. 14.22) can be observed from the characteristic graph. It means that for videos, there is a correlation between methods and scores. However, we do not see the impact of the stimuli characteristic (greyscale or colour). It shows that the model scores can be correlated with the chosen approach, while the way to exploit the stimuli (greyscale or colour) is less important.

14.3 Summary

- Validation of saliency models was made on 3 databases of eye tracking for still images and two databases of videos.
- Three metrics (sAUC, NSS, KLD) were used for validation as their information are complementary.
- One dataset for videos with pixel-wise object segmentation was used along with the F-measure metric.
- Additional statistical tests (Kruskal-Wallis test) showed that the different models' results on eye-tracking data are close between different saliency models, especially for static images.
- Multidimensional scaling of image and video models also shows that a few groups are enough to explain the different results.

References

1. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013). Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication, 28*(6), 642–658.
2. Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207.
3. Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R., & Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision, 12*(6), 17.
4. Chan, Y., & Walmsley, R. P. (1997). Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. *Physical Therapy, 77*(12), 1755–1761.
5. Harel, C. K. J., & Perona, P. (2006). Graph-based visual saliency. *Proceedings of Neural Information Processing Systems (NIPS)*, Vancouver.
6. Marat, S., Rahman, A., Pellerin, D., Guyader, N., & Houzet, D. (2013). Improving visual saliency by adding 'face feature map' and 'center bias'. *Cognitive Computation, 5*(1), 63–75.
7. Coutrot, A., & Guyader, N. (2014). How saliency, faces, and sound influence gaze in dynamic social scenesShort Title? *Journal of Vision, 14*(8), 5–5.
8. Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on machine learning*. ACM.

Part III
Evolution and Applications

Chapter 15
Object-Based Attention: Cognitive and Computational Perspectives

Anna Belardinelli

15.1 Introduction to Object-Based Attention

A much debated and still unresolved issue within the multidisciplinary research on visual attention concerns the question on the basic units of attention. In recent years the object-based account of attentional selectivity has gained evidence, in contrast to the spatial (location) account, yet it seems clear that the two strategies are very much likely to cooperate or being used for different purposes. Depending on the task and the situation at hand, indeed, selection can be location, feature, or object based. In Posner's cueing paradigm [46], attention is shifted to a location where nothing is yet present while during visual search target features or objects resembling the target capture attention. Since the postulation of the dual nature of vision [25, 62], space-based attention has been thought to pertain more to action (dorsal pathway), while object-based attention would be rather needed for object recognition and ventral processing. Again, the two systems have actually to interact for coherent behavior [30, 64].

In any case, the issue presents far-reaching consequences in diverse fields, since from the parsing of the perceivable scene into distinctly selectable units, higher-level concepts, action, and language can be bootstrapped. Even without going too far in considering complex levels of cognition, it is a daily experience for everyone that our attentional behavior is often object oriented: we look for something; we want to grasp or manipulate an object or to navigate an environment while avoiding obstacles. That is, every goal-oriented action or every perceptual understanding of a scene at some point has to come to terms with the selection of a discrete

A. Belardinelli (✉)
Cognitive Modeling, Department of Computer Science, University of Tübingen,
Sand 14, 72076, Tübingen, Germany
e-mail: belardinelli@informatik.uni-tuebingen.de

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_15

entity, may it be a completely formed object in a semantic, categorical sense or a candidate target, defined by a bundle of features of interest gathered in a single selectable token. This goal-driven perspective is particularly desirable, of course, also in the case of artificial systems, which need to instantiate the planning of their actions or select on what to focus their processing resources even before costly operations of object recognition have taken place. Moreover, compared to saliency computed on pixel-based feature contrast, human visual attention has proven to be able to span entire objects. In this respect computational approaches accounting for salient object selection present the advantage of helping scene understanding (e.g., by discarding the background and processing only foreground objects) and of allowing sparse representation and chunking while concurrently fostering timely action on the environment. Essentially, an object-based representation, even at the lowest level, is the basis for *situated vision* [47], a perspective on visual processing that embraces the different purposes of vision, from recognition to control of action. In this chapter, we will review some evidence for the capacity of our visual system to operate on an object basis, consider the implications of such an ability for cognitive systems, and explore how these concepts have inspired and have been implemented in artificial systems.

15.1.1 Evidences for Object-Based Processing of Visual Information

What is an object? This is a long-standing, haunting question for philosophers, psychologists, roboticists, and cognitive scientists alike. Even though everyone, just as for the term *attention*, has a fair understanding of what an object is, the distinctive properties making an object, determining the *objecthood*, are far less agreed on. Attentional units can range from amorphous bundles of colocated, grouped features, to proto-objects, to distinct, well-formed objects. Basis for each of these concepts is, nevertheless, the ability to segment the scene and group features that belong together according to some principle. The most evident and intuitive criteria for such perceptual grouping, as well known, were stated by the Gestalt movement already in the first decades of the last century [68]: elements similar for color, orientation, size, or direction of motion can be easily chunked together. Analogously, symmetry, parallelism, good continuation, and closure help segmenting unitary objects. The Gestalt laws provide a way to organize the perceptual input in grouped chunks and proceed from the primal sketch [39], identifying edges and intensity changes, upward toward structured entities comprised of surfaces and volumes. Simplified to the most basic terms, the problem of object segmentation can be reduced to fore-/background discrimination, yet this is not always univocally possible if no depth or light cues are present and often top-down biasing is necessary to recognize an object from ambiguous silhouettes (as in the case of the vase/face figure by Rubin).

For a more detailed investigation of Gestalt laws of perception, we refer the readers to texts on visual perception such as [8] or to the review by [66].

Assumed we have a way to extract discrete entities from the continuous visual input (in Sect. 15.1.2 definition and nature of such entities will be further scrutinized), what does it mean for selective attention? In which ways is cognition affected?

Even though for many decades space-based (and feature-based) accounts were the preferential approach in research on visual attention, object-based effects in selective attention have long been known to exist. We report here some of the most compelling studies (for thorough reviews see [55] and, more recently, [12]).

In the seminal work by [17], it was shown how subjects, when presented with two superimposed objects, could report more accurately two features pertaining to the same object rather than one feature of each object. This effect was termed the "same-object advantage." Stimuli contained, for example, a box crossed by a line, each of which had two variable dimensions (size and side of aperture for the box, texture and orientation for the line; see Fig. 15.1). The effect was reproduced in many different paradigms [18, 64], showing the advantage both in accuracy and in reaction time, within and between objects, with valid/invalid cues, with superimposed objects or flanker interference (objects grouped according to Gestalt criteria such as contours, motion, connectedness, good continuation) [55]. The critical point here is that a coherent object, with all of its properties and parts, can be attended without further effort in a single act of selection. The ability to focus on one of two superimposed stimuli was already demonstrated by [42] in a study where participants were able to attend to one of two overlaid videos and to ignore events in the other, with both videos running transparently on top of one another. A similar paradigm was used to show object effects at neural level in the well-known fMRI study by [44]. In this case photographs of a face and of a house were shown to subjects, and one of the two could be moving (oscillating along one axis). The areas sensitive to each of these elements are well and distinctly localized in our brain, namely, the fusiform face area for faces, the parahippocampal place area for places, and MT/MST for motion.

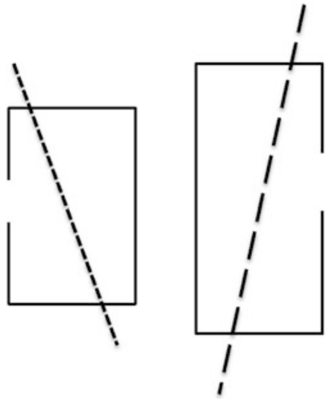


Fig. 15.1 Same-object advantage exemplification (Adapted from [17]): two possible stimuli used in the experiment. Two features of the line or of the box were reported more accurately and promptly than one feature of either object.

The authors thus were able to observe which area was most activated according to the task, i.e., to attend either to the face, to the house, or to the moving item. Results showed that both the area related to the target and the area pertaining to the second feature (motion) of the target were more activated than the one related to the object to be ignored. This phenomenon can be hardly explained just in terms of spatial-based or feature-based attention and is considered as neural evidence underpinning the "same-object advantage."

In general, object-based benefits for within-object saccades, working memory, and visual search have been demonstrated (see studies reviewed in [12]). Interestingly, even though there are many mutual interactions between segmentation and attention and the task also plays a role [16], perceptual objects do capture attention also in a purely bottom-up way. Kimchi and colleagues [34] showed how a target was more readily found in a visual array if it was placed inside a frame formed by a closed, symmetrical disposition of elementary objects rather than outside of it, even though the object-like form was not relevant to the task. When no object was present, but just the elementary items as distractors, performance was still better than when the object was present and the target was outside the perceptual object area.

Further important pieces of evidence of object-based representations come from studies on brain-lesioned patients, presenting conditions such as neglect, visual extinction, and simultanagnosia. Neglect is usually characterized by disregard of the contralesional visual hemifield and would hence be rather considered a space-based dysfunction, but, even so, it was shown that sometimes patients ignored the contralesional side of an object stimulus (i.e., the left side of a clock in the case of a right side lesion) even if this was entirely presented in the ipsilesional hemifield [15], suggesting the use in this case of an object-centered reference frame.

Visual extinction patients usually fail to detect one stimulus in the contralesional visual hemifield if this is presented concurrently with one stimulus on the other side. It would suggest an attention competition problem with the ipsilesional object winning the race and taking all the attentional resources, hence making patients unable to disengage from the stimulus in the hemifield relative to the non-damaged hemisphere. This effect was nevertheless attenuated when the two presented stimuli were manipulated in such a way that they could be perceptually grouped together [24].

Finally, simultanagnosia, part of the Balint's syndrome, causes the inability to perceive more than one object at a time. This makes scene perception and integration of single object parts in a larger compound object very difficult. This seems a merely object-based condition, is not lateralized, appears to be produced by a disrupted global Gestalt perception [28], and, again, can be alleviated by inducing a stronger perceptual grouping by connecting objects explicitly [30].

Lesion studies show that object coding is present at neural level, and this can happen even without focal attention. G. Humphreys [29] suggests the existence of a dual coding: a within-object representation (elements coded as a single object) and a between-object representation (elements coded independently). The first is needed for object recognition, while the second for action control; hence the two seem to pertain to different pathways while still existing in parallel. Primate single-cell studies involving tracing a line crossing a distractor line have also shown that object-based representations emerge as response enhancement as early as V1 [52].

Finally, assessed that object effects exist and can critically influence our representation of the world, which mechanisms can explain these effects? Chen [12] reports three main interpretations:

- Sensory enhancement: attention spans the entire object and in turn enhances its neural and perceptual representation, so that each of its features is processed more effectively with respect to unattended objects.
- Attentional prioritization: the scanning order in visual search is biased toward inspection of locations in an already selected object (suggesting that object effects do not emerge when location of the next target is known in advance, which is controversial).
- Attentional shifting: shifting within an object is less costly than between objects, since attention can spare the disengagement cost connected with shifting to another object representation.

15.1.2 Object Files and Proto-objects

Even if the Gestalt principles suggest important criteria to single objects out of multiple elements or to identify distinct textures, assigning them to different objects or to objects and background, a more basic question arises when these principles exploit a combination of features that need to be first linked together and assigned to the same entity. Moreover, a critical question for this chapter is which part of this processing is done pre-attentively and which part necessitates attention. From early sensing to conscious attentive perception, the flow is continuous and hierarchical, with different functional and neural layers acting on top of one another and bidirectionally interacting with one another. A first important distinction that can be driven is suggested by A. Clark [14]. At the level of early vision, features are computed in parallel in a pre-attentive way. In this case features are retinotopically arranged in maps, hence essentially location based. This representation is termed "feature-placing" and is very coarse and, importantly, nonconceptual and prelinguistic. The next layer is termed "proto-objects" (or in the words of [33] object files, " 'episodic' representations of real world objects"), is volatile (as in [49]), and constitutes the basis for attentional selection and object recognition. These entities can be indexed and selected for storage into visual working memory [47, 49] and fed into higher-level layers, such as that of "full-blooded," well-formed objects, both conceptually and linguistically determined.

Feature placing shares the idea stated in the Feature Integration Theory (FIT) [61] that different visual dimensions are computed in separate maps and they are then integrated in a single master map. An object with a unique property hence stands out ("pops out") because is the only one receiving contribution from the related feature map. On the other hand, a combination of features needs selective attention to be assessed across different maps, resulting in a less straightforward process. Still, an object is not the mere ensemble of its features, and location is not a feature as other phenomenological properties. The master map (or the saliency map in many attention models) can identify the co-occurrence of diverse features, but this is still defined by the overlapping in a precise physical location not by a more widely extended, yet connected, identifiable entity. The difference is akin to the one referred by [64] as grouped array versus spatially invariant representation. This leads to tackling the problem of "feature binding" in all its different aspects. Treisman [60] taxonomizes the binding problem into seven major categories: property (features), part (components), range (of different sensor values and cell firing), hierarchical (shape versus low-level features), conditional (interdepending properties), temporal (over successive states), and location ("what" to "where") binding. Most current models, referring to spatial attention, focus on the first two and combine specific feature maps by means of the same topographic organization, so that for each location the number of active, sensed features can be assessed. The binding in FIT is hence provided by the presence of features at the same location at one point in time, and attentive fixation glues them together. Hence, location is not a feature but the property, along with time, that permits the instantiation of object files. The temporal aspect is also very important in the definition of object units, since it allows for consideration of the dynamic aspects of perceptual objects. In the words of Rensink [49], "focused attention is needed to see change." In his *coherence theory*, he suggests a sketch of low-level vision based on three steps: a transduction stage, concerned with photo-reception at sensor level; a primary parallel processing stage computing image properties via linear filtering; and a secondary processing stage extracting proto-objects directly accessible to attention, which in turn provides coherence to selected objects in the form of a spatiotemporal unity (see Fig. 15.2).

15.1.3 The Role of Objects in Visual Search, Top-Down Attention, and Vision for Action

We are very goal-oriented beings, our survival used to depend on the action we undertook at any time, and the better the action planning, the higher the chances of getting through the challenges of our environment. This means that most of the time top-down control shapes our sensory-motor behavior in order to make informed decisions about the next move and to increase our situation awareness for responding to external solicitations. Visual search is one of the most classical experimental paradigms to study task-driven visual attention. Indeed, since the studies of Yarbus [72] and Buswell [10], it has been shown how substantial the role of the task can be in controlling the gaze wandering on some complex scene.

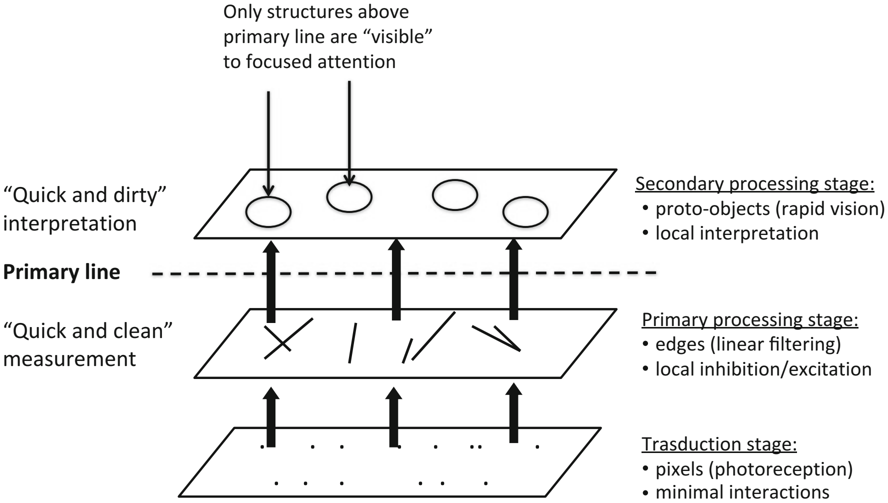

Only structures above
primary line are "visible"
to focused attention
"Quick and dirty"
interpretation
Primary line
"Quick and clean"
measurement
Secondary processing stage:
• proto-objects (rapid vision)
• local interpretation
Primary processing stage:
• edges (linear filtering)
• local inhibition/excitation
Trasduction stage:
• pixels (photoreception)
• minimal interactions

Fig. 15.2 Schematic representation (adapted) of the pre-attentive processing flow in the coherence theory of [49]. Only discrete proto-objects are accessible to attention processes via a *nexus*, a compact structure containing each object properties

Especially in a visual search task, the item to be targeted is more than an ensemble of features associated to the pixels it is composed of. The whole is more than just the sum of its features, something which seems supported by very recent studies. Naber et al. [41] conducted a study to disentangle the question whether grouped features elicit an object representation or if such a representation is necessary to bind features together. By showing ambiguous stimuli switching between a bound versus an unbound perception of the same features, it can be shown that detection of a probe and discrimination of features were both reported better when the bound perception was dominant. That objects are extracted pre-attentively was also shown in [70], where candidate targets in visual search are first segmented against a more or less cluttered background. That is, *objecthood* is also a guiding feature, even if not the one defining the target (as in the abovementioned case of [34]).

Something which is sometimes neglected is that visual search is not just a task but a whole visuomotor behavior. The visual system follows a precise strategy to bring the most promising target candidates within the fovea, with a sequence of saccades and fixations and possibly head movements. This is so because of the acuity drop in our peripheral view. In recent years, there has been an increasing number of studies investigating the role of object-based attention in gaze control. This has, indeed, proved to be prominent, especially when the subject has some kind of task [19, 26, 43], may it be explicit visual search or memory encoding for later recall. It appears clear that most fixations land within objects [43], in particular close to its center of gravity (COG), considered as the Preferred Viewing Location, especially in a search task condition (there still can be a small amount of undershooting, depending on the saccade direction). Einhäuser and colleagues [19] also showed that human fixations were better predicted by objects rather than by early saliency, even though objects gathering more salience within their boundaries were picked up for recall more often.

Apart from visual search, which, as said, involves an oculomotor behavior, object-based attention has been shown also to play a role in the control of action. In the context of the premotor theory of attention, indeed, it has been shown that some neurons discharge both during grasp execution and observation of graspable objects. Crucially, this happens only if the object is presented in size and shape compatible with the coded grip [51].

Finally, going back to the dualism space-/object-based attention, two models were proposed in the 1990s to explain how these perspectives may interact. The visual attention model (VAM) by [54] is a neurocognitive model reconciling vision for perception and vision for action in a single framework. Attention is here suggested to control segmentation, object recognition, and space-based motor action. "Visual chunks" are locally grouped and segmented as early as V1 in a stimulus-driven Gestalt manner, while attention processes act via top-down feedback from higher layers to produce global segmentation units as "object tokens." If just one "token" is produced, this can proceed to object recognition along the ventral pathway; otherwise the competition between multiple objects to be recognized requires where-based attention control from parietal areas to serially scan the possible targets.

The CODE Theory of Visual Attention [37] combines a previously proposed contour detector theory (CODE) [63] with the Theory of Visual Attention [9]. TVA, indeed, needs to act on objects or perceptual groups to account for within-object selection depending on feature and category evaluations. The CODE theory clusters locations by proximity assuming again that at early level (V1) location and identity are represented together while late processes separate the two (the magnocellular pathway, in the posterior parietal cortex, represents location, and the parvocellular pathway, ventral, represents identity). Again perceptual groups are produced acting on the analog representation of space (bottom-up) by applying top-down processes yielding a quasi-discrete representation.

In the next sections it will be discussed how such theories and concepts have been received in computer vision and inspired modeling in technical systems.

15.2 From Biological Inspiration to Modeling and Applications: Object-Based Attention in Artificial Systems

Although the modeling of visual attention for artificial systems has come a long way in the last 30 years, still most approaches are pixelwise, feature, and location based. The introduction of attentive processing was meant to allow a cognitive system (be it a camera or a robot) to rapidly focus on interesting objects and events, extracted from the wider visual scene. As such, pixel-based approaches need hence a further step to single out an object or a region for further processing. Moreover, when considering embodiment, vision is also inherently necessary for action control, usually exerted on objects. In many approaches, attentive selection is considered as a preliminary step for object detection and recognition. Indeed, even if most algorithms and systems for detection and recognition are basically scale and location invariant, they typically use pictures containing one object to recognize or, for detection, they often require a limited window to search within (e.g., [40, 65]). This window is usually slided over the whole picture, as such implying an exhaustive search. Attention is supposed to mitigate this computational burden, by directly selecting the regions where the searched object is likely to be found. In recent years multiple frameworks have been proposed, which try to move away from the selection of the most salient points toward the selection of salient objects or regions. The diverse approaches present in the literature can be basically classified according to when the segmentation operation is performed, namely, before or after saliency computation. What is of interest introducing these approaches is to stress again the problem of generically defining an object or the concept of *objecthood*, something which still remains quite elusive in vision science altogether. In most cases, computer vision approaches resort to feature-based definitions, considering an object as a homogeneous entity with respect to low-level features or to the criterion used for segmentation. In [2] authors focus on designing a measure of *objectness* independent of the object classes. They consider objects as "standalone things with a well-defined boundary and center, such as cows, cars, and telephones, as opposed to amorphous background stuff, such as sky, grass, and road," hence defined by closed boundaries and distinct, salient appearance with respect to the surroundings and global uniqueness. Elsewhere proto-objects or object candidates are formed as extended areas of saliency activation or segmented blobs with underlying features [45, 67, 69], as detailed in the next section.

15.3 Late Versus Early Segmentation: Salient Object Detection and Embodied Models

The question whether objects are computed pre- or post-attentively transfers also to the way current computational models intend object-based attention. In many cases, saliency is computed first at pixel level, and then saliency clusters are taken as proto-objects or regions containing a relevant object. In this case, saliency is considered causal of attention (at least, in a bottom-up way) or the best correlate of attention, assuming that object edges and shapes are mostly visually salient. Elazary and Itti [20] showed how, when applying the basic bottom-up model by [31] to a dataset of images containing objects spontaneously labeled by users, the most salient location fell well above chance in one of the annotated objects. The model has of course no concept of object, and it is possible that users chose the objects most easy to annotate, hence those well contrasted. Critically, when comparing pixel-based saliency to human performance, it has been proven the tendency to land saccades on the center of gravity (COG) of objects [21, 43]. It has been hence suggested that saliency has mostly an indirect, correlational relationship with the fixation distribution [19, 59]. This has initiated a debate, with Borji and colleagues [7] analyzing again the data of the abovementioned study of Einhäuser et al. [19]. In this latter study, objects – a map made of manually segmented objects – were shown to predict human fixations better than early saliency. In [7], the authors used a rich set of saliency models, with different parameter tuning and performance measures to show that most of these models indeed outperform the object model. Stoll et al. [56] picked up the question again considering not just segmented objects as predictors of human fixations but putting a higher weight on their COG (or PVL, Preferred Viewing Location). In this case the object model was performing as well as the best tested saliency model and even better when low-level saliency was manipulated while keeping the "objecthood" of displayed items intact.

Independently of where one stands in this debate, in many cases where attention serves as a spotlight (or as an object/background discrimination means) for more elaborated object detection and recognition techniques, quick, contrast-based saliency is the best choice. On the other hand, some consider attention, especially in task-driven settings, to be deployed on previously segmented units, saving saliency computation on the background and ranking sparse objects instead of the whole visual array. The image-based representation, dominating in the computational attention literature, does not take into account more high-level aspects of selection, connected to the *objecthood* of the target, whereas the cognitive relevance hypothesis postulates the role of the current task and related semantic and cognitive aspects in the selection of the saccade target [43].

In a way, late segmentation approaches consider object selection as a by-product of spatial attention and saliency extraction, while early segmentation approaches consider objects the only argument of attentive selection, as in Bundesen's model [9] or as for Rensink's proto-objects [49]. To put it in other words, models of the first type use saliency to perform object segmentation (e.g., [23]), while models of the second type use segmentation to compute object saliency. In general, late segmentation approaches are better suited for object detection and recognition, while early segmentation frameworks, more literally object based, serve better for visual search and, crucially, action/manipulation purposes in robotic settings. Both types of methods can be more or less biologically inspired, with the second ones usually paying more attention to a biologically plausible scan-path production.

15.3.1 Late Segmentation Approaches: Salient Object Detection

A wealth of saliency measures and techniques have been developed to overcome the issue of salient point location detection and to guide recognition by operating *salient object detection*. With the advancement of object detection/recognition techniques, the critical issue is learning objects during the training phase by presenting correctly segmented and labeled examples, since most learning is done supervisedly. Even if only the object of interest is present in the training image, this must be discriminated from the possibly more or less cluttered background. Pixel-based saliency offers in this case the advantage of being defined independently of image content, in an almost agnostic fashion as to the number of objects present, their features, or scale.

Rutishauser et al. [53] and Walther and Koch [67] extended the classic model by [31] to show how bottom-up saliency can be used to extract regions where salient objects are and use them to select, learn, and afterward retrieve the same objects. Proto-objects are formed by selecting the most salient location in the saliency map and by retrieving via feedback the feature and the scale that contributed the most to the saliency of the winning location. Segmentation of the proto-object occurs by selecting the region around the winning location which shares a similar level of activation. To perform object recognition, nevertheless, selected regions are further characterized by newly computed, more robust features, such as SIFT [38] or C2 coefficients in the HMAX model [50].

As mentioned above, a richer characterization of what makes a generic object, at least in computer vision terms, was proposed by Alexe and colleagues [2, 3]. The authors there argue that any object can be discriminated from the background according to one or more of these properties: a defined closed boundary, a locally different appearance, and a global uniqueness in the image. On this premise, a measure of "objectness" is built by considering multiscale saliency (based on spectral residual [27]) along with color contrast, edge density, and straddleness (considering superpixels straddling object boundaries). Still, this system is more a general object detector than an attention model. Instead of a saliency ranking of each object candidate, the output is a list of windows labeled with the likelihood that they contain an object.

Recent approaches have focused less on the abstract definition of what makes an object and more on salient object detection as an image segmentation problem, separating the (usually) one salient object on the foreground from the rest of the picture. These models often heavily rely on spectral information as a strong cue for boundaries or for features able to describe the global nature of objects as opposed to local contrast. In [36] the binary segmentation problem is solved by learning in a supervised fashion the optimal combination of local and global features describing an object. Images annotated by users are used to learn a conditional random field where the probability of a binary mask on the salient object in a given image depends conditionally on a combination of salient features – multiscale contrast (local), center-surround histograms (regional), and color spatial distribution (global) – and a pairwise feature, modeling the relationship between adjacent pixels. Achanta et al. [1], recognizing that classically defined saliency often fails in spreading over whole objects and consistently enhances border parts, use just color and luminance as features, since these are usually uniform within objects, at low frequencies. Saliency is computed as a center-surround operation where the surround is the whole averaged image (i.e., the DC component, as if the image was filtered by a Gaussian with infinite variance) and the center is a blurred version of the original image, filtered with a Gaussian with small variance, so to eliminate noise and textures but to keep object edges. That is, by optimally tuning the ratio between these variances, object extraction boils down to a difference-of-Gaussian operation, like a pass-band filter tuned to the bandwidth covering most salient objects.

Many other models have been proposed in this line of research, also trying to combine more "object-oriented features" (e.g., [11, 13, 32, 71]), and have been benchmarked in [6], but some limitations (such as biased datasets) have been pointed out in [5]. These techniques distance themselves from pure modeling of attention, since, to some extent, they mostly provide sophisticated image processing for object discovery. This nonetheless is paramount in many computer vision applications, specifically nowadays that object-based image classification is becoming imperative for many web-based services.

15.3.2 Early Segmentation Approaches

As the debate on objects grew in the human and primate attention community, computational models began also to strive for designing attention systems evaluating salience at a (proto-)object level. Such an approach would indeed present the advantage of delivering more accurately formed, discrete units to yield to higher-level cognition instead of amorphous locations. Moreover, in the case of complex scenes containing multiple objects, feature competition can be computed both within and between objects. One major caveat of this view is the critical reliance on a first segmentation stage, delivering object candidates, which hence are the more meaningful the more effective the scene parsing. The segmentation step depends also on the chosen features (and also on scale factors) and again rises the question of feature-homogeneous areas being objects or rather parts of objects, which, nevertheless, has not been completely solved and is not exclusively in the scope of technical systems.

One of the first models addressing object-based attention for computer vision was proposed by Sun and Fisher [57]. Their system for covert attention starts with a foveated version of the examined picture, computes feature multi-resolution pyramids as usual (color, intensity, orientation), and assumes that some perceptual grouping is computed on each layer by a given segmentation algorithm (which in the specific case was manual segmentation according to multiple Gestalt principles). A grouping, in the authors' view, is the basic unit of saliency computation and can represent altogether an object, a group of objects or features or a region, hierarchically arranged. Grouping saliency at any scale layer stems then from a combination of spatial, object, and feature saliency which also competes with the saliency of the surrounding. Competition (via winner take all) then proceeds from coarsest to finest level, also modulated by top-down biasing, in so achieving hierarchical selectivity. In a following work [58], the concept of grouping was replaced by “visual objects” (automatically segmented), and the model was able to produce overt shifts, hence adding one more competition for shifts within (attentional shifts, covert) and outside the current fixation region (gaze shifts, overt). A temporary inhibition of return mechanism prevents both type of shifts from staying on the last attended object.

The first proto-object-based embodied model was proposed in [45]. Vision for action was here a declared goal, since the model was meant to help a humanoid robot in both looking for objects in the scene (hence producing a scan path in a real scene) and in learning object appearance (blob components) by visually exploring the objects after grasping. Here, too, the image is first transformed to a foveated version via log-polar transformation. The following steps entail feature extraction by computing color contrast maps in a center-surround way (by means of difference of Gaussians applied to opponent color channels) and proto-object formation (by edge extraction and segmentation via watershed transform). This last operation delivers proto-objects in the form of closed areas of uniform color or uniform color gradient. Bottom-up and top-down saliencies are finally computed at object-level, proportionally to feature differences between the object and the surroundings or the target object, respectively. A sketch representation of their model is shown in Fig. 15.3. Inhibition of return was implemented basing on spatial position (in head-centered coordinates) and color, hence both spatial and object based.

A further approach based on earlier segmentation, but defined as region-based, was presented in [4]. This model builds on the consideration that pixel-based approaches cannot deliver shape or size information, while these object-level features are essential for top-down modulation, particularly in visual search. The color map is thus segmented in homogeneous regions which are further characterized

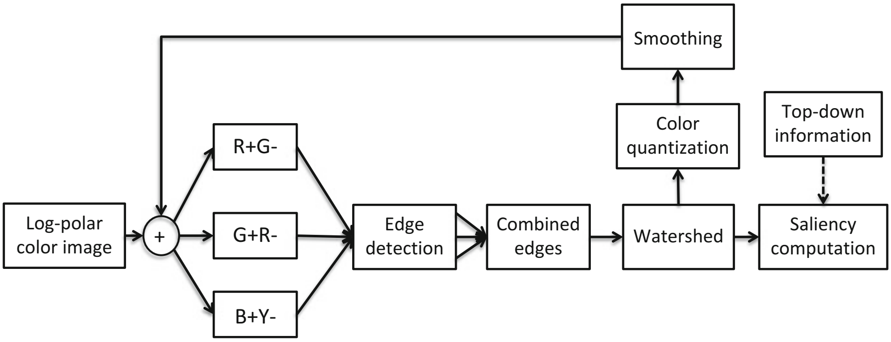


Fig. 15.3 The object-based model of Orabona et al. (adapted from [45])

in terms of color contrast, symmetry, size contrast, orientation, and eccentricity, all of which concur to determine the saliency of the region. These features are moreover not extracted just for the sake of saliency computation but can be handed to subsequent object-based processing.

An object-based model of attention aimed at scan-path generation in a visual search scenario has been put forward by Wischnewski and colleagues [69]. The model focuses on the selection of the landing item for the next saccade when looking for a target and relies on the Theory of Visual Attention (TVA) [9], which is intrinsically object based. TVA weight equation, indeed, according to which priority of each segmented object in the scene is computed, takes as argument entire objects (or proto-objects) whose feature similarity and pertinence with the target features are evaluated. In the proposed model feature extraction and proto-object segmentation are performed separately for static and dynamic features, imitating ventral and dorsal processing in the primate brain. After a fusion step, merging overlapping static and dynamic units, each proto-object is defined by its local features (mean color, orientation, intensity, motion direction, and motion energy) and its geometric features (size, shape, global orientation), computed after proto-object formation. The final attention priority map ranks candidate objects by assigning a weight, depending on its features and vicinity to the current point of fixation (see Fig. 15.4).

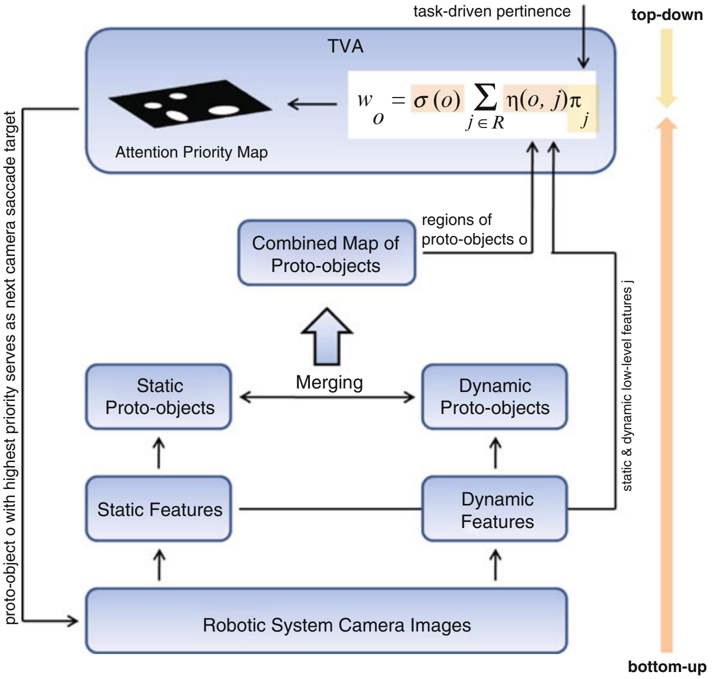

task-driven pertinence
top-down
TVA
$w_o = \sigma(o) \sum_{j \in R} \eta(o, j)\pi_j$
Attention Priority Map
regions of proto-objects o
Combined Map of Proto-objects
Merging
Static Proto-objects
Dynamic Proto-objects
Static Features
Dynamic Features
static & dynamic low-level features j
Robotic System Camera Images
proto-object o with highest priority serves as next camera saccade target
bottom-up

Fig. 15.4 The model proposed by [69]: static and dynamic objects are extracted separately and combined at intermediate level, while object priority is computer at the top of the architecture

In the scope of object discovery, similarly to [2], two approaches in between late and early segmentation have been proposed by [22] and [35]. In the first case, superpixel-based segmentation is performed in parallel to pixel-based saliency computation, basing on color and intensity channels. Afterward, segments collecting saliency above a certain adaptive threshold and spatially connected are merged together in an object hypothesis. This approach has been shown to work on cluttered images, besides salient object image databases. Leroy and colleagues [35], instead, first produce a multiscale segmentation by using decompositions with different numbers of superpixels on just color channels. These superpixels are then rated on the base of their rarity, according to the self-information of the superpixel color. Finally the saliency maps obtained in this way are merged, allowing both small and larger regions to emerge.

In general early segmentation approaches do not just parse the scene and rank extracted entities but deliver compact descriptions of these to be suitably processed by object recognition or action control modules. They lend themselves more easily to top-down biasing and visual search, higher-level abstraction, and implementation in embodied settings.

15.4 Summary and Perspectives

In this chapter it has been shown how attention can act by selecting discrete, extended units instead of or along with locations in space. Evidences have been collected at behavioral and neural level, demonstrating object effects as well as (proto-)object representations already at early levels in the visual system. Technical systems have started using attention to efficiently manage visual processing resources. To move toward high-level cognitive processes that can represent and interact with full-fledged objects, perceptually and possibly semantically coherent units need to be selected and characterized at a low and intermediate level. Depending on the specific application, different models have been proposed, basically considering saliency weighting either the cause or one effect of object segmentation. Even when not considering biologically inspired attention modeling, the need for a more abstract level going beyond the features of single pixels is demonstrated by the increasing use of superpixels [48], a segmentation technique tiling images in atomic regions, producing a more efficient representation both on a computational and a perceptual level.

The considerations above suggest that object-oriented approaches are a key to reduce the cognitive/computational load and to produce sparse and manageable representations upon which biological or artificial systems can concentrate their high-level reasoning capabilities.

As a final remark, most current models are feedforward with segmentation and selection coming before object recognition, while future improvement directions could envision feedback from upper levels iteratively top-down influencing and refining the segmentation and selection stages. The main issue in these systems is indeed how to attend to a meaningful object before this is fully recognized and, often, by relying just on low-level, “quick-and-dirty” bottom-up information.

References

1. Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). *Frequency-Tuned Salient Region Detection*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, 1597–1604.
2. Alexe, B., Deselaers, T., & Ferrari, V. (2010). What is an object? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, San Francisco, 73–80.
3. Alexe, B., Deselaers, T., & Ferrari, V. (2012). Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(11), 2189–2202.
4. Aziz, M., & Mertsching, B. (2008). Fast and robust generation of feature maps for region-based visual attention. *IEEE Transactions on Image Processing, 17*(5), 633–644.
5. Borji, A. (2015). What is a salient object? A dataset and a baseline model for salient object detection. *IEEE Transactions on Image Processing, 24*(2), 742–756.
6. Borji, A., Sihite, D., & Itti, L. (2012). Salient object detection: A benchmark. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Computer Vision – ECCV 2012* (Lecture notes in computer science, pp. 414–429). Berlin/Heidelberg: Springer.
7. Borji, A., Sihite, D. N., & Itti, L. (2013). Objects do not predict fixations better than early saliency: A re-analysis of einhäuser et al.,'s data. *Journal of Vision, 13*(10), 18.
8. Bruce, V., Georgeson, M. A., Green, P. R., & Georgeson, M. A. (2003). *Visual Perception: Physiology, Psychology and Ecology* (3rd ed.). Hove/New York: Psychology Press.
9. Bundesen, C. (1990). A theory of visual attention. *Psychological Review, 97*(4), 523–547.
10. Buswell, G. (1935). *How people look at pictures: A study of the psychology and perception in art*. Chicago: University of Chicago Press.
11. Chang, K.-Y., Liu, T.-L., Chen, H.-T., & Lai, S.-H. (2011). Fusing generic objectness and visual saliency for salient object detection. In *IEEE International Conference on Computer Vision (ICCV 2011)*, Barcelona (pp. 914–921).
12. Chen, Z. (2012). Object-based attention: A tutorial review. *Attention, Perception, & Psychophysics, 74*, 784–802.
13. Cheng, M.-M., Zhang, G.-X., Mitra, N. J., Huang, X., & Hu, S.-M. (2011). Global contrast based salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, Washington, DC (pp. 409–416). IEEE Computer Society.
14. Clark, A. (2004). Feature-placing and proto-objects. *Philosophical Psychology, 17*(4), 443–469.
15. Driver, J., Baylis, G. C., Goodrich, S. J., & Rafal, R. D. (1994). Axis-based neglect of visual shapes. *Neuropsychologia, 32*(11), 1353–1356.
16. Driver, J., Davis, G., Russell, C., Turatto, M., & Freeman, E. (2001). Segmentation, attention and phenomenal visual objects. *Cognition, 80*(1–2), 61–95.
17. Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology. General, 113*(4), 501–517.
18. Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology. General, 123*(2), 161–177.
19. Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *JOV, 8*(14), 18–18.
20. Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*(3), 3.
21. Foulsham, T., & Underwood, G. (2009). Does conspicuity enhance distraction? Saliency and eye landing position when searching for objects.*Quarterly Journal of Experimental Psychology, 62*(6), 1088–1098.

22. Frintrop, S., Garcia, G. M., & Cremers, A. B. (2014). A cognitive approach for object discovery. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR 2014)*, Stockholm (pp. 2329–2334). IEEE.
23. Ge, F., Wang, S., & Liu, T. (2006). Image-segmentation evaluation from the perspective of salient object extraction. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, Washington, DC (Vol. 1, pp. 1146–1153). IEEE Computer Society.
24. Gilchrist, I. D., Humphreys, G. W., & Riddoch, M. J. (1996). Grouping and extinction: Evidence for low-level modulation of visual selection. *Cognitive Neuropsychology, 13*(8), 1223–1249.
25. Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*(1), 20–25.
26. Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188–94.
27. Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Los Alamitos, CA (Vol. 0, pp. 1–8). IEEE Computer Society.
28. Huberle, E., & Karnath, H.-O. O. (2012). The role of temporo-parietal junction (TPJ) in global Gestalt perception. *Brain Structure & Function, 217*(3), 735–746.
29. Humphreys, G. W. (1998). Neural representation of objects in space: A dual coding account. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 353*(1373), 1341–1351.
30. Humphreys, G. W., & Riddoch, M. J. (1993). Interactions between object and space systems revealed through neuropsychology. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV* (pp. 143–162). Cambridge, MA: MIT.
31. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.
32. Jiang, H., Wang, J., Yuan, Z., Liu, T., & Zheng, N. (2011). Automatic salient object segmentation based on context and shape prior. In *BMVC*, Dundee (pp. 1–12).
33. Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24*(2), 175–219.
34. Kimchi, R., Yeshurun, Y., & Cohen-Savransky, A. (2007). Automatic, stimulus-driven attentional capture by objecthood. *Psychonomic Bulletin & Review, 14*(1), 166–172.
35. Leroy, J., Riche, N., Mancas, M., Gosselin, B., & Dutoit, T. (2014). Superrare: An object-oriented saliency algorithm based on superpixels rarity. In *IEEE International Conference on Robotics and Automation (ICRA 2014), Workshop "Robots in Homes and Industry: Where to Look First?"*, Hong Kong.
36. Liu, T., Sun, J., Zheng, N.-N., Tang, X., & Shum, H.-Y. (2007). Learning to detect a salient object. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, Minneapolis (pp. 1–8).
37. Logan, G. D. (1996). The CODE theory of visual attention: An integration of space-based and object-based attention. *Psychological Review, 103*(4), 603–649.
38. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.
39. Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.
40. Mutch, J., & Lowe, D. G. (2006). Multiclass object recognition with sparse, localized features. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, Washington, DC (Vol. 1, pp. 11–18). IEEE Computer Society.
41. Naber, M., Carlson, T. A., Verstraten, F. A. J., & Einhäuser, W. (2011). Perceptual benefits of objecthood. *JOV, 11*(4), 8–8.
42. Neisser, U. (1975). Selective looking: Attending to visually specified events. *Cognitive Psychology, 7*(4), 480–494.

43. Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*(8), 20.
44. O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature, 401*(6753), 584–587.
45. Orabona, F., Metta, G., & Sandini, G. (2005). Object-based visual attention: A model for a behaving robot. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005) – Workshops*, Washington, DC (Vol. 03, pp. 89–96). IEEE Computer Society.
46. Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology, 32*(1), 3–25.
47. Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80*(1–2), 127–158.
48. Ren, X., & Malik, J. (2003). Learning a classification model for segmentation. In *Proceedings of Ninth IEEE International Conference on Computer Vision, 2003*, Nice (pp. 10–17). IEEE.
49. Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition, 7*(1–3), 17–42.
50. Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*, 1019–1025.
51. Rizzolatti, G., & Luppino, G. (2001). The cortical motor system. *Neuron, 31*(6), 889–901.
52. Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature, 395*(6700), 376–381.
53. Rutishauser, U., Walther, D., Koch, C., & Perona, P. (2004). Is bottom-up attention useful for object recognition? In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, Washington, DC (Vol. 2, pp. II-37–II-44).
54. Schneider, W. X. (1995). VAM: A neuro-cognitive model for visual attention control of segmentation, object recognition, and space-based motor action. *Visual Cognition, 2*(2–3), 331–376.
55. Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80*(1–2), 1–46.
56. Stoll, J., Thrun, M., Nuthmann, A., & Einhäuser, W. (2015). Overt attention in natural scenes: Objects dominate features. *Vision Research, 107*(0), 36–48.
57. Sun, Y., & Fisher, R. (2003). Object-based visual attention for computer vision. *Artificial Intelligence, 146*(1), 77–123.
58. Sun, Y., Fisher, R., Wang, F., & Gomes, H. M. (2008). A computer vision model for visual-object-based attention and eye movements. *Computer Vision and Image Understanding, 112*(2), 126–142.
59. Tatler, B., Baddeley, R., & Gilchrist, I. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*(5), 643–659.
60. Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology, 6*(2), 171–178.
61. Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.
62. Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of Visual Behavior* (pp. 549–586). Cambridge, MA: MIT.
63. van Oeffelen, M. P., & Vos, P. G. (1982). Configurational effects on the enumeration of dots: Counting by groups. *Memory & Cognition, 10*(4), 396–404.
64. Vecera, S. P., & Farah, M. J. (1994). Does visual attention select objects or locations? *The Journal of Experimental Psychology: General, 123*(2), 146–160.
65. Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Los Alamitos (Vol. 1, pp. 511–518). IEEE.
66. Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin, 138*(6), 1172–1217.
67. Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*(9), 1395–1407.

68. Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. II. *Psychological Research, 4*(1), 301–350.
69. Wischnewski, M., Belardinelli, A., Schneider, W. X., & Steil, J. J. (2010). Where to look next? Combining static and dynamic proto-objects in a TVA-based model of visual attention. *Cognitive Computation, 2*(4), 326–343.
70. Wolfe, J., Oliva, A., Horowitz, T., Butcher, S., & Bompas, A. (2002). Segmentation of objects from backgrounds in visual search tasks. *Vision Research, 42*(28), 2985–3004.
71. Xie, Y., Lu, H., & Yang, M.-H. (2013). Bayesian saliency via low and mid level cues. *IEEE Transactions on Image Processing, 22*(5), 1689–1698.
72. Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum.

Chapter 16
Multimodal Saliency Models for Videos

Antoine Coutrot and Nathalie Guyader

16.1 Introduction

Over the past hundred years, *attention* – the focus on one aspect of the environment while ignoring others – has been one of the most studied topics within cognitive sciences and neurosciences. Researches aim at determining which part of signals perceived through different senses (e.g., sight, hearing, taste, smell, and touch) captures attention. The term of salient is used to characterize the part of the signal that attracts attention. Two types of attention can be distinguished: overt and covert attention. Overt attention is followed by eye movements, whereas covert attention concerns the displacement of attention without moving the eyes. In this field of research, most studies have been dedicated to overt visual attention.

Over the past 30 years, numerous computational saliency models have been proposed to model overt visual attention (see [1] for a review). Visual saliency mainly depends on two factors. The first one refers to bottom-up processes and is mostly driven by stimulus visual features [2, 3]; the latter refers to top-down processes and is driven by observer-related cues, such as task [4–6]. Most saliency models, also called visual attention models, simulate bottom-up processes to predict salient locations in visual scenes. These regions are supposed to attract attention and, hence, observers' gaze.

A. Coutrot (✉)
CoMPLEX, University College London, London, UK
e-mail: a.coutrot@ucl.ac.uk; acoutrot@gmail.com

N. Guyader
GIPSA-lab, Grenoble Alpes University, Grenoble, France
e-mail: nathalie.guyader@gipsa-lab.fr

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_16

The earliest saliency models were developed for static scenes using only static visual features such as luminance contrast, spatial frequencies, and orientations. They rapidly evolved to be adapted to dynamic scenes, adding motion amplitude as an additional feature [2, 7–9]. More recently, face detection has been also added to classical low-level feature extraction for scenes with faces [10, 11]. However, another feature, although ubiquitous in dynamic natural scenes, has been left aside: sound. Indeed, while clues for the existence of audiovisual interactions in attention are numerous, only few studies investigate the influence of sound on visual attention. Hence, when using eye-tracking and dynamic stimuli, authors rarely mention soundtracks or explicitly remove them, making participants look at videos without any sound, which is far from natural situations.

The goal of this chapter is to give an overview of visual saliency models and how they have been improved to take into account more features. We further discuss studies measuring the influence of sound on eye movements. Finally, we detail an audiovisual saliency model for the particular case of conversation scenes.

16.2 Visual Saliency Models

Since 1980, numerous visual attention models have been proposed [7–9, 12]. In a recent paper, Borji [1] proposes a classification of 65 models according to several types: cognitive models, Bayesian models, decision theoretic models, information theoretic models, graphical models, spectral analysis models, pattern classification models, and others.

In this chapter, we focus on cognitive models, i.e., models inspired by the biology of the human visual system. These models are based on the feature integration theory (FIT) proposed by Treisman and Gelade [3]. The FIT states that low-level visual features (edges, intensity, color, etc.) are extracted from the visual scene and combined into a master map to guide visual attention. The first conceptual model of this theory was proposed by Koch and Ullman [2] who introduced the definition of saliency map. This first model was interested in modeling neural processes. It decomposes a visual stimulus into several feature maps dedicated to specific visual features such as orientation, spatial frequencies, or intensity, see Fig. 16.1. Saliency models have then been generalized to dynamic scenes by adding in the decomposition a motion feature map [7–9]. In each map, the spatial locations that locally differ from their surroundings are emphasized (conspicuity maps). Then, the different feature maps are combined into a master saliency map that points out the spatial locations (regions) the most likely to attract the visual attention, and the gaze of observers. In fact, a close link between visual attention and eye movements is well established. The premotor theory of spatial attention stipulates that visual attention and oculomotor system share the same neural substrate [14]. This theory has been strengthened by neurophysiological experiments showing that intracranial subthreshold stimulation of some oculomotor brain areas results in enhanced visual

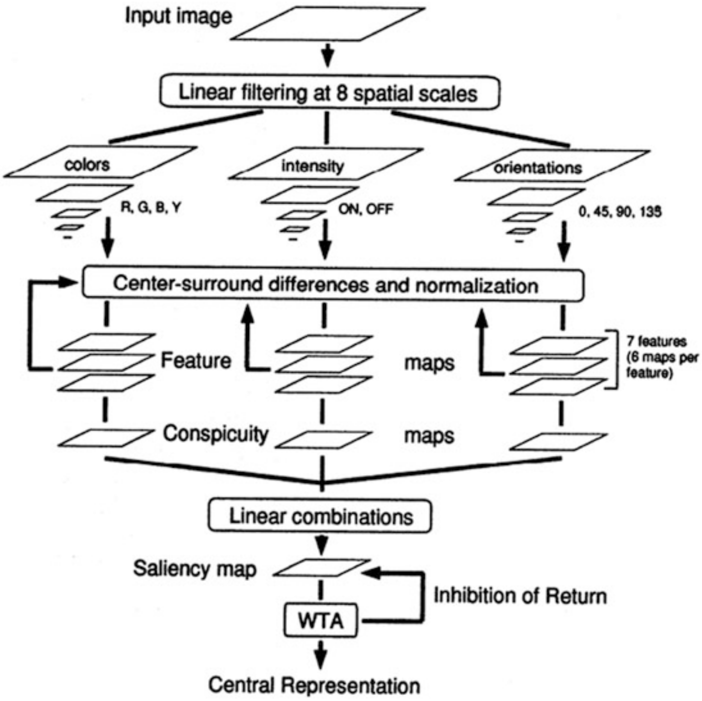


Fig. 16.1 Schematic diagram of the model proposed in Ref. [13]. Static visual features are computed at eight spatial scales, followed by center-surround differences to compute local spatial contrast in each feature. After competition for salience within each feature map, the latter are combined into a single "conspicuity map" for each feature type. The three conspicuity maps then are summed into the unique master saliency map. A winner-takes-all (WTA) mechanism spots the most salient location, an inhibition-of-return transiently inhibits it and directs attention to the next most salient location (Extracted from Ref. [13])

sensitivity at the corresponding retinotopic location [15]. Although some other studies suggest a greater separation of the two processes [16], the existence of a high correlation between eye movements and visual attention meets general consensus. This link between attention and eye movement allows authors to evaluate their visual attention models by comparing the predicted salient regions with the locations actually looked at by observers during an eye-tracking experiment.

Eye-tracking experiments are not only used to validate computational saliency models but also to better understand which factor or visual feature drives attention. For instance, many eye-tracking studies reported that observers tend to gaze more often at the center of the image than at the edges [17]. Several propositions have been made to explain this bias. Some are stimuli-related, like the photographer bias (one often places regions of interest at the center of the picture); others are inherent to the oculomotor system (motor bias) or to the observers' viewing strategy [18]. To take this tendency into account, several authors incorporate "center bias" as a feature in their models. Most saliency model simply models this bias with a two-dimensional Gaussian ([11, 19, 20]). However, recent studies show that taking into account more sophisticated oculomotor biases (such as saccade amplitude or saccade orientation) considerably improve the performance of the models [21].

All the aforementioned models are efficient to predict the most gazed at locations in various scenes. Yet these "classical" models cannot be generalized to many experimental contexts, since the social nature of visual perception is not taken into account [22]. Typical examples where they dramatically fail are visual scenes involving faces [23]. Despite their leading role in attention allocation, faces have rarely been considered in visual attention modeling. However, since the beginning of eye tracking, they have been known to attract gaze and capture visual attention more than any other visual feature [6, 24]. When present in a scene, faces invariably draw gaze, even if observers are explicitly asked to look at a competing object [25, 26]. Many studies have established that face perception is holistic [27–29] and pre-attentive [30, 31]. For all these reasons, more recently, visual saliency models combining face detection with classical low-level feature extraction have been developed and have significantly outperformed the classical ones [10, 11] especially for scenes with faces and people. In their paper, Rahman and colleagues decompose a scene into three maps dedicated to three different types of features. The static pathway extracts the texture information based on luminance. The dynamic pathway extracts information about object's motion against background. The face pathway extracts information about the presence and location of faces in the frames. Their model also integrates the center bias as a suitable modulation on the static and dynamic feature maps. This bias is not added on the face map because faces attract gaze independently of their location on a scene. Adding the center bias improves the model efficiency to predict eye movement. But above all, adding the face pathway greatly improves the model efficiency, making the full model (center bias modulation and face pathway) the most efficient model (Fig. 16.2).

16.3 Audiovisual Saliency Models

Hence, even if eye-tracking and modeling studies lead to a better understanding of which visual feature drive attention and eye movements, few studies have measured the influence of sound on eye movements. Yet hearing and sight constantly interact to perceive the surrounding world. Audiovisual illusions are certainly the most popular audiovisual interactions, like the McGurk effect, where mismatched acoustic and visual stimuli result in a perceptual shift: auditory /ba/ and visual /ga/ are audiovisually perceived as /da/ [32]. Another well-known audiovisual interaction is the help given by "lip reading" to understanding speech, even more when speech is produced in poor acoustical conditions or in a foreign language [33–35].

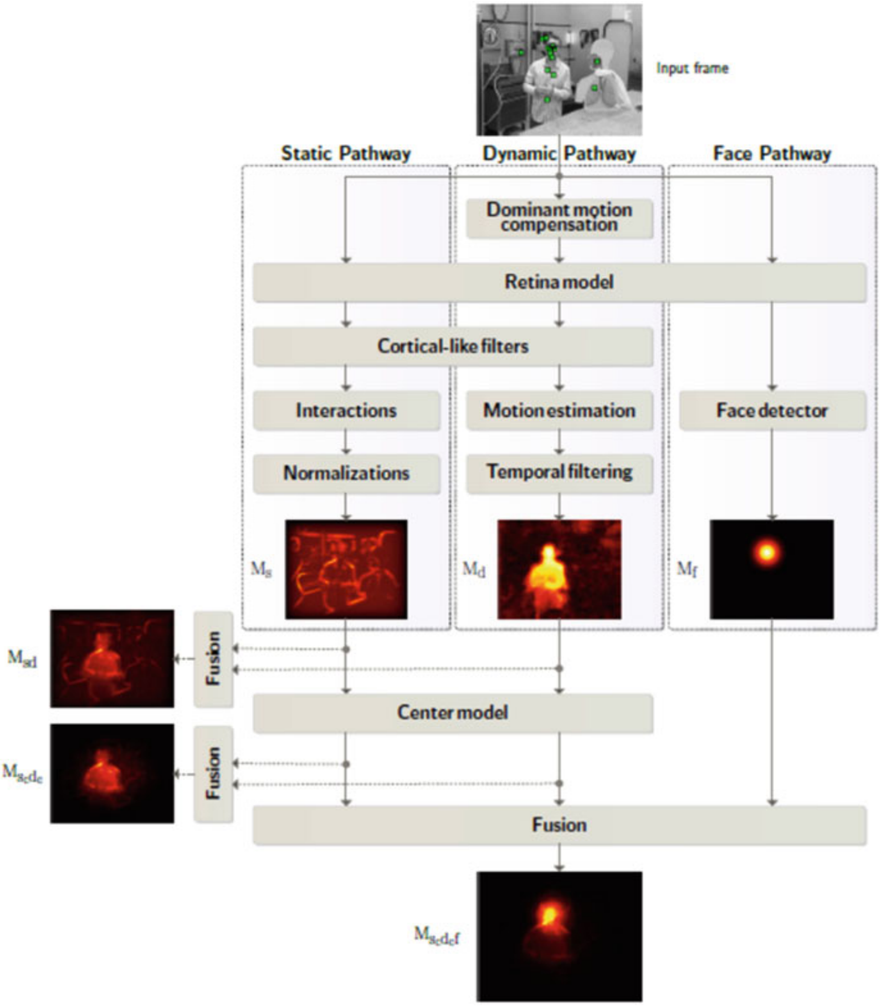


Fig. 16.2 Schematic diagram of the model proposed in Ref. [11]. Static, dynamic, and face features are computed in parallel pathways, each of them producing a saliency map: M_{s}, M_{d}, and M_{f}. The maps are then fused together either before or after applying the center model to take into account the influence of the center bias. M_{scdcf} is the final saliency model that combines all the three features with center bias (Extracted from Ref. [11])

Besides these perceptual phenomena, some studies measured the influences of competing visual and auditory stimuli on gaze shifts. Authors showed that speed and accuracy of eye movements in detection tasks were more efficient for congruent audiovisual stimulus compared to a mere visual or auditory unimodal stimulus [36–38]. Quigley et al. [39] presented static natural images and spatially localized (left, right, up, down) simple sounds. They compared eye movements of observers when viewing visual-only, auditory-only, or audiovisual stimuli. Results indicated that eye movements were spatially biased toward the regions of the scene corresponding to the sound sources.

However, spatial localization is not necessary to observe the influence of sound on visual attention. A study [40] showed that a nonspatial auditory signal improved spatial visual search. The correct mean reaction time was up to 4 s shorter (depending on the number of distractors) when a nonspatial beep was synchronized with the visual target change. After controlling alternative explanations of the so-called "pip and pop" phenomenon (an auditory "pip" makes the visual target pop out), authors proposed that the temporal information of the auditory signal directly interacted with the synchronous visual event. As a result, the visual target became more salient within its environment.

In a previous study, we investigated the effect of nonspatial auditory sounds (a monophonic soundtrack) on eye movements recoded while exploring video clips. We observed that during video exploration, gaze was impacted by the related soundtrack, even without spatial auditory information [41]. With the soundtrack, eye positions of participants were less dispersed and tended to go more away from the screen center, with larger saccades compared to a visual-only condition. Moreover, the observed locations significantly differed between the two conditions. These results indicate that the related soundtrack of a video impacts on its exploration. Hence, saliency models should take into account the sound information to be more efficient in predicting eye movements.

To go further, we tested whether this effect of sound on eye movements was stronger just after salient auditory events [42]. To automatically spot salient auditory events in the soundtrack, we used two audio saliency models: the discrete energy separation algorithm (DESA) and the energy model. If visual saliency models compute saliency maps in two spatial dimensions, these audio saliency models compute saliency curves with only one temporal dimension. Both audio saliency models provide a saliency time curve that was thresholded to extract the most salient events. We examined eye movement parameters just after these events rather than over all the video frames. We did not find any increased effect of sound after salient auditory events.

Altogether, our results indicate that if nonspatial auditory information does impact on eye movements, the exact auditory features capturing observers' attention remain unclear. In these first studies, the visual and auditory contents of videos were very diverse. However, previous studies showed that different types of sounds interact differently with visual information when viewing videos [43, 44].

Hence, using an eye-tracking experiment, we investigated whether the congruency between the visual and audio contents influenced eye movements of observers viewing videos [45, 20]. Observers watched videos belonging to four visual categories presenting different visual saliency distributions: landscapes, one moving object, several moving objects, and conversations. Videos were seen with their original soundtrack, with the soundtrack from another video belonging to the same visual category, or with the soundtrack from another video belonging to a different visual category. Videos and eye-tracking data are available online.[1] Recorded eye movements showed that sound has an impact on the several moving object category and even more on the conversation category. Unrelated soundtracks increased the variability between the locations gazed at by different observers. The effect was not constant across viewing time but appeared after 1 s of exploration. It seems logical that the auditory conditions impact on conversation and several moving objects rather than on one moving object and landscape categories, since the auditory information they convey is more rich and informative. We did not find any difference between unrelated soundtracks (from the same or different visual category). We hypothesized that unrelated soundtracks are not temporally correlated enough with the visual information to be bound to the visual information, preventing any further integration. In that case, observers might just filter out the unbound auditory information and focus on the sole visual stream.

Another research team has developed audiovisual saliency models in the context of video summarization [46–49]. For that purpose, they characterize each frame using three saliency curves. The first curve was obtained using a classical visual saliency model and by taking the average over the different features to produce one value per frame. The second curve averaged three auditory features: mean Teager energy, mean instant amplitude, and mean instant frequency. The last curve corresponded to textual saliency information. These three curves were further linearly combined to provide a master saliency curve. The local maxima of this curve corresponded to the frames latter chosen for the video summary. Note that their aim was to extract the most salient frames in a video and not to extract or predict the most salient locations in a frame.

Multimodal saliency models have also been proposed to control the overt attention of humanoid robots [50, 51]. Just as humans, robots have limited processing capabilities. To be able to interact with their environment in real time, multimodal saliency models have been integrated into their perception unit. In [50], the robot iCUB computes a traditional visual saliency map (based on intensity, color, motion) and an auditory spatial saliency map (based on binaural differences). These two maps are then combined into an audiovisual saliency map that controls the movements of the robot's eyes and neck. Since then, more sophisticated features such as face, emotion, expression, or speech recognition have been added to make robots even more "social" [52].

16.4 The Particular Case of Scenes of Conversations

Thus, what we hear has an impact on what we see. This is particularly true for speech and faces, which are known to strongly interact, as evidenced by the huge literature on audiovisual speech integration [35, 53, 54]. To investigate audiovisual integration, most of these studies presented talking faces to observers and measured

[1] http://antoinecoutrot.magix.net/public/databases.html

how visual or auditory modifications impacted on their eye movements or speech comprehension [55, 56, 57]. They identified the eyes and the mouth as two strong gaze attractors during audiovisual speech processing and showed that the degree to which gaze is directed toward the mouth depends on the difficulty of the speech identification task. Yet results emanating from experimental setups using isolated close-ups of faces might not be generally applied to the real world, where everything is continuously moving and embedded in a complex social and dynamic context. To address this issue, Võ et al. [58] eye-tracked participants watching videos of a pedestrian engaged in an interview. They showed that observers' gaze is dynamically directed to the eyes, the nose, or the mouth of the interviewee, according to depicted events (speech onsets, eye contact with the camera, quick movement of the head). The authors also found that removing the speech signal decreased the number of fixations on the pedestrian's face in favor of the scene background.

Nevertheless, in daily life, conversations are often made of several speakers embedded in a complex scene (objects, background), not only listening to what is being said but interacting dynamically. Thus, Foulsham and colleagues eye-tracked observers viewing video clips of people taking part in a decision-making task [59]. They showed that gazes followed the speech turn-taking, especially when the speaker had high social status. These results indicate that during dynamic face viewing, our visual system operates in a functional, information-seeking fashion. A few very recent papers quantified how the turn-taking affects the gaze of a noninvolved viewer of natural conversations [60, 61]. These studies presented conversations to participants with the related speech soundtracks or without any sound. They both showed that sound changed the timing of looks. With the related speech soundtracks, speakers were fixated more often and more quickly after they took the floor, leading to a greater attentional synchrony. All the previously reviewed studies reported behavioral and eye movement analyses, but did not quantify the relative contributions of faces (mute or talking) and of classical visual features to guide eye movements. Birmingham and Kingstone [23] showed static social scenes to observers and compared their eye positions to the corresponding low-level saliency maps (within the meaning of Itti and Koch [13]). The authors showed that saliency did not predict fixations better than chance and noticed that classical low-level saliency models do not account for the bias of observers to look at the eyes within static social scenes. However, this study did not use dynamic scenes for which motion is known to be highly predictive of fixations, much more than static visual features [62].

More recently, we analyzed how auditory conditions influence the eye movement parameters of participants viewing videos of conversations [20]. We compared the eye movements of participants watching movies either with the original speech soundtrack, with an unrelated speech soundtrack, with the noise of moving objects (abrupt onsets, e.g., falling cutlery), or with landscape continuous sound (slowly changing components, e.g., wind blowing). To analyze how auditory information modulates the relative predictive power of different visual features (faces, low-level static and dynamic visual saliencies, and center bias), the expectation-maximization algorithm (EM) was used. The EM algorithm is a statistical method widely used in statistics and machine learning that has been recently applied to visual attention modeling [19, 63, 64]. This method is a mixture model approach that uses participants' eye positions to estimate the relative contribution of different potential gaze-guiding features. We quantified the impact of sound on classical (saccade amplitudes, fixation durations, dispersion between eye positions) and less classical (distance between scan paths) eye movement parameters. Through experimental and statistical modeling results, we showed that regardless of the auditory condition, participants look more at faces, and especially at talking faces. However, with the original soundtrack, observers look even more at the speakers, following the speech turn-taking more closely.

These experimental results were incorporated in an audiovisual saliency model. Based on the saliency model proposed by Marat and colleagues [11], the proposed model decomposes an input video frame into several saliency maps: a center bias, a static saliency, a dynamic saliency, and a face/body map. Then a master saliency map was obtained through a weighted sum of the maps. Using two different statistical methods, the weights were obtained either using the EM algorithm [20] or the least absolute shrinkage and selection operator algorithm [65]. The Lasso is a regularized version of the least square method. Its major advantage is the sparsity imposed by the penalization constant, while the EM deals with all the parameters given as inputs and runs the risk of overfitting. A MATLAB implementation of the Lasso algorithm for saliency modeling is available online.[2] Since we previously found that speakers are more likely to attract gaze than other conversation partners, the model assigns to face and body maps different weights according to their talking-or-not status. Thanks to a speaker diarization algorithm, able to temporally distinguish talking conversation partners from silent ones, we proposed an audiovisual saliency model that increases the saliency of the speakers compared to the addressees (Fig. 16.3).

The speaker diarization algorithm used did not require training; it was based on two assumptions: (1) each speech turn-taking was separated by a silence and (2) speakers move more than other conversation partners [66, 67]. The speaker diarization algorithm relies on three stages. First, a voice activity detector extracts and appends the speech segments from the soundtrack, discarding silent segment. Second, an audio speaker clustering algorithm decides whether two successive speech segments were assigned to the same conversation partners or not. Each speech segment was described using the first 26 Mel frequency cepstral coefficients (MFCCs) on 10 ms intervals. A fixed size analysis window of 200 ms was centered on each sample s of the speech signal. We tested the hypothesis that a change occurred at sample s using the Bayesian information criterion (BIC) that compared two models: one that considers that the speech segment was pronounced by a single person and the second that considers that a turn-taking occurred. A difference in

[2] http://antoinecoutrot.magix.net/public/code.html

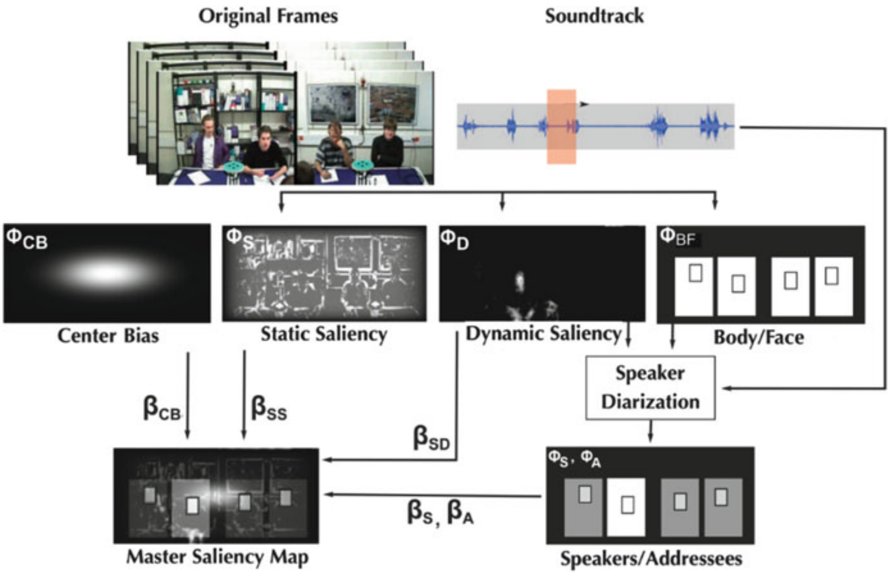


Fig. 16.3 Block diagram of our audiovisual saliency model. *Center bias*, *static and dynamic saliency*, *speakers, and addresses* (faces or bodies) maps are weighted with the $ß^{Lasso}$-estimated weights and merged into the audiovisual *master saliency* map (Extracted from Ref. [65])

the BIC value was computed for each sample, and a local maximum was extracted from each speech segment. The higher it was, the more likely a speaker transition occurred.

This step does not only consider the audio signal but also the visual signal. For each frame, we used the dynamic feature map, and for each speaker, we summed the pixels of the dynamic feature map contained in their corresponding face or body mask. Thus, we had the frame-by-frame evolution of the "activity" of each conversation partner. Then, we standardized these values and compared their mean over each speech sequence. For each conversation partner, the higher the modulus of the difference between two successive speech segments was, the more likely this person began or stopped moving. Finally, we standardized and added the audio and visual "transition probabilities" for each speech segment. If this combination was higher than an empirical threshold, the speech segments were said to be delivered by different speakers. Else, the speech segments were merged. Finally, to attribute each speech cluster to the right speaker, we used the same dynamic low-level saliency maps as described above. We summed the pixel values contained in each mask to get the activity of the corresponding conversation partner. We then averaged these activities over each speech cluster. The corresponding speech sequence was attributed to the most "active" conversation partner. Note that the face of speakers was more salient than their body.

16.5 Summary and Perspectives

This chapter gives an overview of how visual saliency models have been improved to take into account more features and better predict the locations that should attract eye fixations. We particularly focused on the few models that attempted to use audio information. We saw that in the absence of auditory spatial information (monophonic soundtrack), using audio and visual information to create two-dimensional saliency maps that emphasize salient locations within a scene is challenging. In the particular case of conversational videos, we showed that combining audio and visual information is efficient to automatically spot the speaker and, hence, reinforce its saliency compared to the attendees. Giving higher saliency value to the speakers greatly improves model's performances.

However, modeling audiovisual saliency for more general natural scenes remains an open problem. It may be important to adopt a relevant and biologically plausible metric for auditory saliency. One could, for instance, consider audio saliency models closer to the ones proposed for visual saliency. Kayser and colleagues [68] have proposed an audio saliency model to create a saliency map for auditory information. This model is similar to that proposed in visual attention by Itti and colleagues [13]. Yet, contrary to visual saliency model that may be validated by comparing the regions predicted as salient with actual eye positions, models of audio saliency are much more difficult to validate. In their research, Kaiser and colleagues tested their model by reproducing human judgments and by predicting the detectability of salient sounds embedded in noisy backgrounds.

Another key challenge concerns the fusion of auditory with visual saliency to create audiovisual saliency for dynamic scenes with various contents. The main difficulty would be to spatially match auditory and visual salient event in a complex dynamic scene. If the audio signal is at least stereophonic, 2D sound-source localization techniques could be applied [69]. Else, if the soundtrack is monophonic, multimodal signal processing techniques such as canonical correlation analysis, mutual information, or blind source separation could be harnessed and effectively applied to solve this issue. Another way to tackle this so-called cross-modal binding problem [70] would be to take advantage of the Bayesian framework. This paradigm allows to computationally model the effect of prior context on audiovisual integration. Here, context would mean the prior correlation between auditory and visual events: if these were less correlated in the past, it would make them less likely to bind in the future.

References

1. Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Patterns Analysis and Machine Intelligence, 35*(1), 185–207.
2. Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*, 219–227.

3. Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97–136.
4. Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision, 9*(3), 1–15.
5. Henderson, J. M., & Hollingworth, A. (1999). Eye movements during scene viewing: An overview. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (No. 12, pp. 269–290). Oxford: Elsevier Science.
6. Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum.
7. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.
8. Le Meur, O., Callet, P. L., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(5), 802–817.
9. Marat, S., Ho-Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., & Guérin-Dugué, A. (2009). Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision, 82*(3), 231–243.
10. Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing Systems, 20*, 241–248.
11. Marat, S., Rahman, A., Pellerin, D., Guyader, N., & Houzet, D. (2013). Improving visual saliency by adding 'face feature map' and 'center bias'. *Cognitive Computation, 5*(1), 63–75.
12. Tsotsos, J. K., Culhane, S. M., Yan Kei Wai, W., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence, 78*, 507–545.
13. Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*, 1489–1506.
14. Rizzolatti, G., Riggio, L., Dascola, I., & Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia, 25*(1, Part 1), 31–40.
15. Belopolsky, A. V., & Theeuwes, J. (2009). When are attention and saccade preparation dissociated? *Psychological Science, 20*(11), 1340–1347.
16. Klein, R. M. (1980). Does oculomotor readiness mediate cognitive control of visual attention? In R. S. Nickerson (Ed.), *Attention and performance viii* (pp. 259–276). Hillsdale: Lawrence Erlbaum.
17. Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14), 1–17.
18. Tseng, P. H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*(7), 4, pp. 1–16.
19. Gautier, J., & Le Meur, O. (2012). A time-dependent saliency model combining center and depth biases for 2D and 3D viewing conditions. *Cognitive Computation, 4*, 1–16.
20. Coutrot, A., & Guyader, N. (2014). How saliency, faces and sound influence gaze in dynamic social scenes. *Journal of Vision, 14*(8), 1–17.
21. Le Meur, O., & Liu, Z. (2015). Saccadic model of eye movements for free-viewing condition. *Vision Research*. doi:10.1016/j.visres.2014.12.026.
22. Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11*(5), 5, pp. 1–23.
23. Birmingham, E., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research, 49*, 2992–3000.
24. Buswell, G. T. (1935). *How people look at pictures: A study of the psychology of perception in art*. Chicago: University of Chicago Press.
25. Bindemann, M., Burton, A. M., Hooge, I. T. C., Jenkins, R., & de Haan, E. H. F. (2005). Faces retain attention. *Psychonomic Bulletin and Review, 12*(6), 1048–1053.

26. Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition, 13*(6), 657–665.
27. Boremanse, A., Norcia, A., & Rossion, B. (2013). An objective signature for visual binding of face parts in the human brain. *Journal of Vision, 13*(11), 6, pp. 1–18.
28. Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review, 105*(3), 482–498.
29. Hershler, O., & Hochstein, S. (2005). At first sight: A high-level pop out effect for faces. *Vision Research, 45*, 1707–1724.
30. Bindemann, M., Burton, A. M., Langton, S. R. H., Schweinberger, S. R., & Doherty, M. J. (2007). The control of attention to faces. *Journal of Vision, 7*(10), 15, pp. 1–8.
31. Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision, 10*(4), 16, pp. 1–17.
32. McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.
33. Gailey, L. (1987). *Psychological parameters of lip-reading skill in hearing by eye: The psychology of lip-reading*. Hillsdale: R. Dodd and B. Campbell.
34. Jeffers, J., & Barley, M. (1971). *Speechreading (lipreading)*. Springfield: Charles C. Thomas.
35. Summerfield, Q. (1987). *Some preliminaries to a comprehensive account of audio-visual speech perception*. Hillsdale: B. Dodd and R. Campbell.
36. Arndt, P. A., & Colonius, H. (2003). Two stages in crossmodal saccadic integration: Evidence from a visual-auditory focused attention task. *Experimental Brain Research, 150*, 417–426.
37. Corneil, B. D., VanWanrooij, M., Munoz, D. P., & Van Opstal, A. J. (2002). Auditory-visual interactions subserving goal-directed saccades in a complex scene. *Journal of Neurophysiology, 88*, 438–454.
38. McDonald, J. J., Teder-Sälejärvi, W. A., & Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature, 407*, 906–908.
39. Quigley, C., Onat, S., Harding, S., Cooke, M., & König, P. (2008). Audio-visual integration during overt visual attention. *Journal of Eye Movement Research, 1*(2), 1–17.
40. Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 34*(5), 1053–1065.
41. Coutrot, A., Guyader, N., Ionescu, G., & Caplier, A. (2012). Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research, 5*(4), 1–10.
42. Coutrot, A., Guyader, N., Ionescu, G., & Caplier, A. (2014). Video viewing: Do auditory salient events capture visual attention? *Annals of Telecommunications, 69*(1), 89–97.
43. Song, G., Pellerin, D., & Granjon, L. (2013). Different types of sounds influence gaze differently in videos. *Journal of Eye Movement Research, 6*(4), 1–13.
44. Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition, 118*(1), 75–83.
45. Coutrot, A., & Guyader, N. (2013). Toward the introduction of auditory information in dynamic visual attention models. In IEEE International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Paris, pp. 1–4.
46. Evangelopoulos, G., Zlatintsi, A., Skoumas, G., Rapantzikos, K., Potamianos, A., & Maragos, P. (2009). Video event detection and summarization using audio, visual and text saliency. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, pp. 3553–3556.
47. Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G., & Avrithis, Y. (2013). Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia, 15*(7), 1553–1568.
48. Rapantzikos, K., Evangelopoulos, G., Maragos, P., & Avrithis, Y. (2007). An audio-visual saliency model for movie summarization. In *IEEE international workshop on multimedia signal processing (MMSP)* (pp. 320–323). New York: Springer.
49. Zlatintsi, A., Maragos, P., Potamianos, A., & Evangelopoulos, G. (2012). A saliency-based approach to audio event detection and summarization. In European Signal Processing Conference (EUSIPCO 2012), Bucharest, pp. 1294–1298.

50. Ruesch, J., Lopes, M., Bernardino, A., Hörnstein, J., Santos-Victor, J., & Pfeifer, R. (2008). *Multimodal saliency-based bottom-up attention, a framework for the humanoid robot iCub* (pp. 962–967). Paper presented at the IEEE International Conference on Robotics and Automation, Pasadena.
51. Schauerte, B., Kühn, B., Kroschel, K., & Stiefelhagen, R. (2011). *Multimodal saliency-based attention for object-based scene analysis* (pp. 1173–1179). Paper presented at the International Conference on Intelligent Robots and Systems (IROS), IEEE/RSJ, San Francisco.
52. Zaraki, A., Mazzei, D., Giuliani, M., & De Rossi, D. (2014). Designing and evaluating a social gaze-control system for a humanoid robot. *IEEE Transactions on Human-Machine Systems, 44*(2), 157–168.
53. Bailly, G., Perrier, P., & Vatikiotis-Bateson, E. (2012). *Audiovisual speech processing*. Cambridge, UK: Cambridge University Press.
54. Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models of audiovisual fusion in speech perception. In R. Campbell, B. Dodd, & D. K. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 85–108). Hove, UK: Psychology Press.
55. Bailly, G., Raidt, S., & Elisei, F. (2010). Gaze, conversational agents, and face-to-face communication. *Speech Communication, 52*, 598–612.
56. Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics, 65*(4), 536–552.
57. Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisualspeech perception. *Perception & Psychophysics, 60*(6), 926–940.
58. Võ, M. L. H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision, 12*(13):3, 1–14
59. Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., & Kingstone, A. (2010). Gaze allocation in a dynamic situation: Effects of social status and speaking. *Cognition, 117*(3), 319–331.
60. Foulsham, T., & Sanderson, L. A. (2013). Look who's talking? Sound changes gaze behaviour in a dynamic social scene. *Visual Cognition, 21*(7), 922–944.
61. Hirvenkari, L., Ruusuvori, J., Saarinen, V. M., Kivioja, M., Peräkylä, A., & Hari, R. (2013). Influence of turn-taking in a two-person conversation on the gaze of a viewer. *PLoS One, 8*(8), 1–6.
62. Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2010). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation, 3*(1), 5–24.
63. Ho-Phuoc, T., Guyader, N., & Guérin-Dugué, A. (2010). A functional and statistical bottom-up saliency model to reveal the relative contributions of low-level visual guiding factors. *Cognitive Computation, 2*(4), 344–359.
64. Vincent, B. T., Baddeley, R. J., Correani, A., Troscianko, T., & Leonards, U. (2009). Do we look at lights? Using mixture modelling to distinguish between low- and high-level factors in natural image viewing. *Visual Cognition, 17*(6–7), 856–879.
65. Coutrot, A., & Guyader, N. (2015). *An efficient audiovisual saliency model to predict eye positions when looking at conversations*. Paper presented at the European Conference on Signal Processing (EUSIPCO), Nice.
66. Gebre, B G., Wittenburg, P., & Heskes, T. (2013). The gesturer is the speaker. *In IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, Vancouver, BC, pp. 3751–3755.
67. McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review, 92*(3), 350–371.
68. Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology, 15*(21), 1943–1947.
69. Deleforge, A., & Horaud, R. (2012). *2D sound-source localization on the binaural manifold*. Paper presented at the IEEE Workshop on Machine Learning for Signal Processing (MLSP), Satander.
70. Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*(4), 971–995.

Chapter 17
Toward 3D Visual Saliency Modeling

Leroy Julien and Nicolas Riche

17.1 Understanding 3D Saliency

For visual attention, depth perception is just as fundamental as the perception of texture or movement. Human vision is an extremely complex process which is intrinsically linked to the perception of depth. Indeed, we do not bear the same interest in objects if they are near or far, structured or disorganized, big or small, etc. Although these features can be extracted from an image, raw access to depth information greatly simplifies and increases the accuracy we can achieve using these characteristics. It is therefore essential to integrate this depth information if we want to be able to accurately model human visual attention. If the literature is very rich on computational models, they are in the vast majority dedicated to 2D image analysis. Concerning the 3D saliency, it is unfortunately meager. Even earlier than the appearance of well-known 2D saliency models, like Itti's model [1], authors have studied the subject of 3D and began to propose integrating 3D information into their models. This subject will remain underprivileged until the last 5 years. Indeed, a new enthusiasm has taken hold of the scientific community which takes advantages of the recent advances in 3D data acquisition. Three factors are likely to be considered:

L. Julien (✉)
TCTS Lab, Engineering Faculty of Mons (FPMS), University of Mons (UMONS),
31, Boulevard Dolez, B-7000 Mons, Belgium
e-mail: julien.leroy@umons.ac.be

N. Riche
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: nicolas.riche@umons.ac.be

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_17

1. The availability of 3D content visualization systems became increasingly democratic. 3DTV and 3D cinema have become essential in the media landscape. Indeed, research shows many new features between 3D and attentive behavior, such as cognitive overload-induced vision of 3D content. Understanding and modeling the careful visual process make sense when we want to improve the technical acquisition and visualization with 3D systems to enable the user to make the most of 3D media.
2. The availability of accurate and inexpensive depth sensors influenced the field of 3D image analysis and subsequently the 3D saliency. Previously, getting a disparity or depth map demanded extensive work, such as calibration or conversion of data for analysis with often expensive and poorly performing sensors. Currently, new sensors, such as Microsoft Kinect or Asus Xtion, make easy the use of depth information.
3. The modeling of human attention requires a validation step of algorithmic performances. In 3D, this step has long been difficult to perform by the lack of large databases that can be used. Nevertheless, as we said in the previous point, these new sensors have simplified the steps of acquiring and annotating 3D images for designing these databases dedicated to saliency. Coupled with monitoring systems as efficient binocular eye tracking system, it becomes easier to analyze the specific processes involved in 3D saliency but also to compare and validate the proposed models.

17.2 Why 3D Features for Attention?

The 2D feature extraction from videos can identify the relevant information within the (X,Y) plane. However, they show their limits when the information occurs on the Z (depth)-axis. As shown in Fig. 17.1, this is the case for motion feature extraction. Indeed, the relevant motion is poorly captured with 2D motion features as the main movement is along the Z-axis.

The (X,Y) motion is properly captured: the snow falling vertically (Y-axis) above the skier is detected (yellow vertical lines) and the snow moved by the skier on his right on the X-axis (blue horizontal lines). But the motion of the skier himself is not well described: the image shows several lines of different colors (X,Y directions) on the skier, while in reality he is coming toward the camera (Z-axis). This example shows that detection of the motion on the Z-axis would assign the skier with his real displacement. A better feature extraction will also enhance the attention model performance.

The availability of low-cost 3D sensors with active infrared illumination (as the Microsoft Kinect described in [2]) is an opportunity to easily extract scene depth (Z) information along with classical videos providing (X,Y) information.

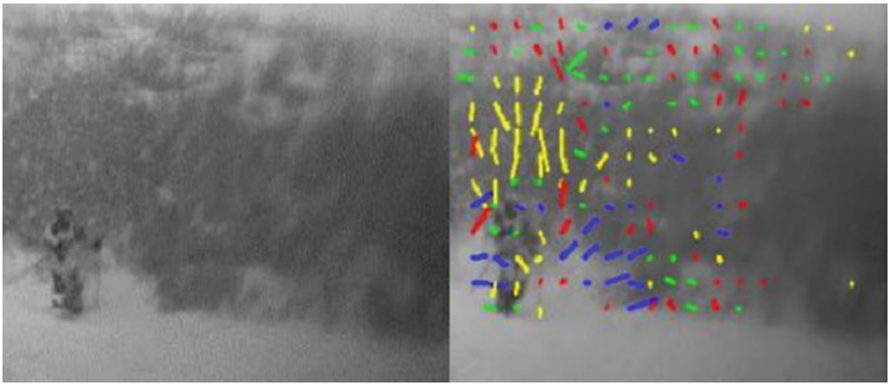


Fig. 17.1 A frame with a skier coming toward the camera (depth – Z-axis velocity): 2D motion features (optical flow for X and Y velocity)

17.3 When Using 3D Features

In Fig. 17.2 from SMAMS's model [3], the speed and direction saliency maps by using an RGB final saliency map are represented.

A red dominant means that the speed feature is the most interesting. A cyan dominant means that direction is the most important. A white blob (which is a mix of red and cyan) means that both speed and directions may attract attention. Here we used only 2D motion features on complex real scenes. In the first two images from the first row, there is a close scene with a frontal view. The other scenes contain wider and wider views with mostly top views.

On the second row, first and second images, we can see that people running toward the others are detected (1), and the person who is faster and with a different direction (2) is also highlighted. On the first row, third and fourth images, the two people walking against the main central flow (1) are well visible. It is also the case with some people having perpendicular directions (3). Finally, in the second row, third and fourth images, one person carried by the crowd (1) and a thrown object (2) are also well detected with a higher speed compared with the other moving objects. Nevertheless, the results are very poor for the first row, first and second image in Fig. 17.2. While the rapidly falling snow (Y-axis motion) is well detected (1) and the snow pushed by the skier (X-axis motion) on his right (2) is also detected, the skier himself (3) is not detected at all! The skier is the only moving object on the Z-axis; thus, it is very salient, but as only 2D features are extracted, he is not well detected. This scene comparison in 2D shows that the more the scene is wide and the camera has a top view, the less important the Z-axis motion is. Indeed, a top view will map most of the motion on the (X,Y) plane, and very small people doing gestures on Z (e.g., jumping) are almost not detectable in those configurations. An interesting conclusion is that while in video surveillance-like situations (wide field of view, almost top view) the knowledge of Z is important, for ambient intelligence

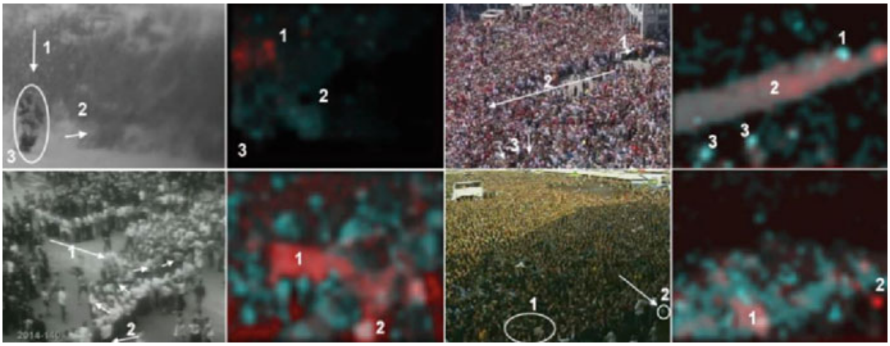


Fig. 17.2 First and third column: annotated frames. Second and fourth column: color saliency maps from SMAMS's model. A *red* dominant means that the speed feature is the most interesting. A *cyan* dominant means that direction is the important feature

and robot-like situations (smaller field of view, frontal view), the knowledge of the Z-axis is crucial. This is convenient, as the Kinect sensor horopter is between 25 cm and 6 m.

17.4 Chapter Organization

Saliency can use any input features. 3D saliency is based on the new 3D sensor output. This output is twofold: (1) the automatic recognition of people silhouette and skeleton which can provide high-level features about people behavior and (2) the 3D data output (RGB and depth maps or 3D point cloud) providing low-level 3D features.

This chapter first presents a model using high-level information about people. In a second part, several models using RGB/depth data or point cloud, more generic on low-level features, will be presented.

17.5 3D Saliency Model Based on High-Level Features

In this section, the design of a new intelligent system capable of selecting the most outstanding user from a group of people in a scene will be discussed. This ability to select a user to interact with is very important in natural interfaces and in emergency-related applications where several people can ask to communicate simultaneously. The proposed algorithm has three main steps: first, features are extracted from Kinect's sensor. In a second step, a contrast-based approach is applied, and, finally, those contrast-based feature maps are combined to focus the system attention on a specific user without complex rules.

17.5.1 Feature Extraction

The first step is to extract features from the observed people. For that purpose, we use the Kinect sensor for its ability to extract smooth depth maps in complex illumination conditions. Libraries as OpenNI (e.g., used in [4]) are available to detect human silhouettes and extract anatomical features from skeleton tracking.

Four features are extracted from the upper body part only as the legs are much less stable in our implementation. One of the four features is dynamic, namely, the **motion index**. It is computed as the mean variation of the same skeleton points between two frames in 3D (on X, Y, and Z). The barycenter point variation is extracted from the others (Eq. 17.2) in order to keep only the body relative motion which will describe an excitement degree or movement transition of the body without any assumption on the whole body speed:

$$D_{mk} = (\sum_{sk} |k_b - k_{sk}|)_t - (\sum_{sk} |k_b - k_{sk}|)_{t-1} \tag{17.1}$$

where $k = x, y, z$. The skeleton points are noted sk and the barycenter b.

$$MI = \sqrt{Dmx^2 + Dmy^2 + Dmz^2} \tag{17.2}$$

A second feature extracted from the upper body part is a static feature, namely, the **asymmetry index**. This feature is only computed on the X-axis by differencing the distances between the barycenter point and the right shoulder, elbow, and hand points with the left ones (Eq. 17.3). This index provides information about the symmetry of the upper body:

$$AI = \frac{\sum_{sk} |X_b - X_{sk_r}| - \sum_{sk} |X_b - X_{sk_l}|}{n_{sk}} \tag{17.3}$$

where n_{sk} is the number of skeleton points.

The third extracted feature is the **contraction index**. This index is the ratio between the maximal distance between skeleton points on X-axis and the maximal distance on the Y-axis (Eq. 17.4). This index tells us if the person is more or less contracted:

$$CI = \frac{|max(X) - min(X)|}{|max(Y) - max(Y)|} \tag{17.4}$$

The fourth and final feature is the **player height**. That one is simply computed by measuring the player barycenter Y coordinate.

After normalization, those four features provide a quite complete description about the level of excitement and the upper body configuration of each player.

17.5.2 Contrast-Based Mechanism

As stated in [5], a feature does not attract attention by itself: bright and dark, locally contrasted areas or not, and red or blue can equally attract human attention depending on their context. In the same way, motion can be as interesting as the lack of motion depending on the context. The main cue, which involves bottom-up attention, is the contrast and rarity of a feature in a given context.

The approach here follows the one in [6]. In our case, as the group of players can be small, the rarity computation is not relevant. Therefore, we only use the global contrast. Thus, the first step in this section is to calculate for the *i*th feature ($f_{i,k}$) a contrast between the different users *k*:

$$C_{i,k} = \sum_{j=1}^{N} \frac{|f_{i,k} - f_{i,j}|}{N-1} \tag{17.5}$$

where N is the number of users. Once all the contrasts for a given feature $C_{i,k}$ between each user and the others have been computed, they are ordered in ascending order $C_{i,k,o}$ with $o = [1 : N]$ from the maximum (o = 1) to the minimum (o = N). The difference between the two highest values is compared to a threshold T which decides if the contrast is large enough to be taken into account as in Eq. 17.6:

$$\begin{cases} \alpha = 0 \;\; if \;\; |C_{i,k,1} - C_{i,k,2}| < T \\ \alpha > 0 \;\; if \;\; |C_{i,k,1} - C_{i,k,2}| \geq T \end{cases} \tag{17.6}$$

17.5.3 Fusion

Only the features being the largest and passing this threshold T are merged with different weights (Eq. 17.7):

$$C_k = \sum_{i=1}^{H} \frac{C_{i,k} * W_i * \alpha}{H} \tag{17.7}$$

where H is the number of features and α is given in Eq. 17.6.

The weights W_i are initially set to be the same for all the four features which are used here. Then, the number of times a feature is contrasted enough for a given user ($\alpha > 0$), a counter is increased. The feature weight will be inversely proportional to its counter: if a feature i is often contrasted, its weight will be lower and lower, while a feature which is rarely contrasted enough will see its weight increased. This mechanism ensures a higher weight to novel behavior, while too repetitive behavior will be penalized. As an example, someone who will sit down for the first time (different height feature compared to the others), the height will have the maximum weight. If this person thinks that a different height is enough to attract the system attention, he will try again, but the more he tries again, the more the height feature weight will decrease as this behavior is no longer surprising. This approach allows the system to learn how much a feature is novel and provides higher weights to the most novel ones.

The contrast C_k represents the bottom-up saliency for each user k. Saliency will be higher for the people exhibiting the most contrasted features within a given frame. The process of bottom-up attention is summarized on Fig. 17.3 on a three-player scenario example. Each of the three players has its four features computed (in red for the asymmetry index, yellow for the contraction index, violet for the motion index, and green for the height). The contrast computation and threshold (Eqs. 17.5 and 17.6) are displayed in the second column. Finally, the contrasted feature combination (Eq. 17.7) is explained in the third and fourth columns.

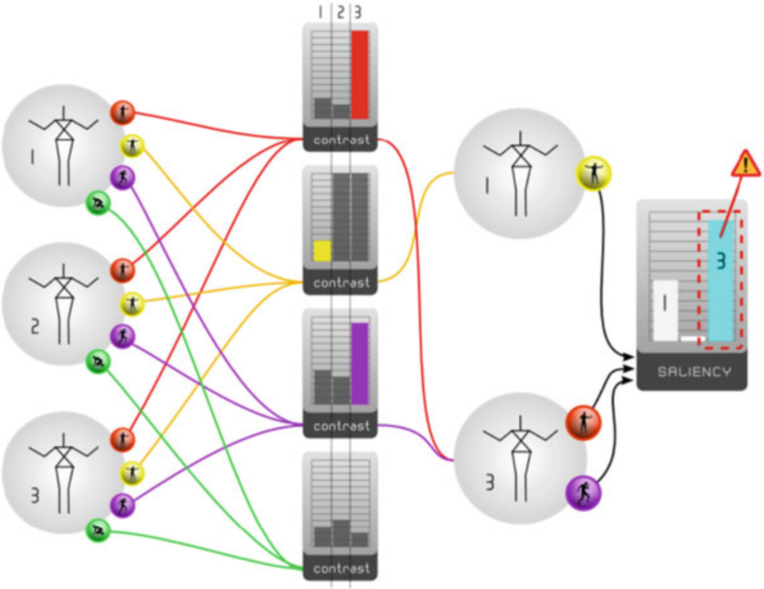


Fig. 17.3 Example of bottom-up saliency computation for three players. For each of the four behavioral features, a contrast is computed between the players. A threshold will eliminate features which are not contrasted enough between players (here the fourth feature in *green* is eliminated). The player having more contrasted features with higher weights will be selected as the most salient (here the third player)

17.6 3D Saliency Models Based On Low-Level Features

Several algorithms taking into account 3D information were already set up. However, this concept of 3D should be taken with caution. Under the aspect of integration of spatial information into a model of human attention, the concept of 3D is often used vaguely and should be refined. Indeed, it will be possible to find in the literature models that deal with 3D saliency, but they will be difficult to compare. Take on one side [7], the precursor model on the use of a fused depth map with a saliency map from a 2D analysis and on the other side [8], who proposed a saliency model on 3D meshes. These could be presented both as 3D saliency models, each using 3D information into their algorithm. However, the inherent nature of the data imposes a differentiation. The 3D data processed differs greatly; a mesh can hardly be equated with an organized disparity map. It will be necessary to distinguish between methods using depth information and methods using all the available 3D information. Disambiguation of 3D in saliency can be submitted via the concept of 3D imaging, 3D data, and their representation. When discussing a 3D image, the most common technique to reproduce the illusion of depth to an observer is stereoscopy. Stereoscopy brings together all the techniques used to reproduce a perception of depth from two planar images. Based on the acquisition of two slightly offset images similar to human eyes, it is possible to generate a disparity map. From this disparity by using epipolar geometry, we can estimate a depth map. Indeed, the disparity and depth are inversely related. As the distance from the cameras increases, the disparity decreases. It is then possible to allocate to each pixel of the image a depth information. We then obtain a depth map where 3D information is represented by the triple (i,j,d), where i,j are coordinates in the image plane and d is the depth of the pixel. This is typically what RGBD sensors will get; often a depth map is also calibrated with a 2D color image. Let us now take a 3D model generated by any 3D modeling software. The 3D image or volume will consist of a set of vertex spatially represented by Cartesian coordinates.

We will do the following distinctions between different classes of saliency models:

1. The ones that we will call “2.5D.” These methods are based on the use of spatial information using disparity or depth map. These methods take as input depth image or stereoscopic images. All these models have a step in which to calculate the final saliency map; 2D visual features and 2D saliency maps are estimated.
2. The 3D methods. These methods are based on all available spatial information and geometry. These methods are heavily based on 3D geometry to extract salient information. These methods apply to 3D reconstructed objects and scenes or 3D modeled scenes. The result is a 3D saliency map.

This chapter discusses 3D salience as a whole. We begin with a review of the methods related to the modeling of salience in 3D. For this review, we will make a distinction between types of models presented based on their input data on which the algorithm is applied. Indeed, it is necessary to make the separation between the methods using depth information and methods using pure 3D structures. After this review, we will discuss this classification, which can be too simplistic, and present other classification possibilities better able to take into account all the intricacies related to 3D salience. Finally, we will present a new model of salience in 3D with two particularities to process large amounts of 3D data regardless of their type (mesh, cloud point, or RGBD data) and integrate color texture information. Indeed, as the review will show, models capable of handling information on structures such as mesh or large point clouds systematically droping the color information process large amounts of 3D data regardless of their type: meshes, point clouds, or RGBD data.

17.7 3D Saliency: A Review

In this section, we will follow the proposed classification and present the most representative models of these categories. The first category of algorithms focuses on the integration of depth usually as a disparity map in their saliency model.

17.7.1 2.5D Models or Depth-Based Saliency

17.7.1.1 Depth and Disparity-Based Models

Maki et al. [7] is one of the first models incorporating a stereo disparity map as a pre-attentive characteristic which will be added to the information flow and motion detection. A depth target mask, which corresponds to the depth conspicuity map in the saliency-based model of attention, is computed. This research, however, is aimed at the integration of a saliency mechanism as an element of selection of a moving part, and no validation of the algorithm performance with a human reference has been performed.

Ouerhani and Hugli [9] propose an extension of the well-known Itti model based on the analysis in center-surround and extend with the introduction of a depth map. Their analyses also involved other features related to depth as the use of gradient or curvature but reject their use because of noise on the data.

Jost et al. [10] is to our knowledge one of the first researches on the impact that depth information can have on attention and especially with an objective comparison with the computational counterpart. The authors demonstrate through two simple experiments the potential impact of the use of depth. First, they show with random-dot stereograms (RDS), so pure disparity maps, that depth perception influences our attention. They draw conclusions as the objects with the greatest disparity attract first fixations. The generalization of this conclusion is that elements with a large disparity are more easily perceived, attract the earlier fixations, and so are more salient. Second, they validate their observation with an objective comparison of human attention map, acquired with an eye-tracker, a saliency map. Their conclusion is that the introduction of the depth drastically increases the measured performance. Although the experimental context can be discussed, given the metrics and the small size of the database operated, this study was drawing the basic conclusions on the idea and the impact of the integration of depth in a saliency mechanism. The depth is an important characteristic capable of optimizing the similarity of saliency algorithms with human attention.

In [11], the authors offer a very similar approach to previous authors and realize an interesting comparison of the performance of their model through several validations. Their model is thus divided on the basis of its features: grayscale, color, and depth. The analysis is interesting because they show not only the interest of color-based features but also the impact of depth on the results while discussing performance according to the nature of the analyzed scene.

In [12], the authors study the possibilities of using laser data for attention mechanisms. They propose a model (BILAS) based on that of Koch & Ullman but including here as input depth and reflectance images from a laser sensor.

In [13], the authors are interested in the use of attention mechanism exploiting 3D features to assist in the segmentation step preceding robotic tasks such as grasp and object manipulation. For this, they study several 3D features as a surface height, orientation, and relative area occluded edges and merge them with 2D information (color, orientation, etc.) through a probabilistic approach.

In [14], the authors focus on the exploitation of the depth map as support for extracting information related to motion. As they point out, very few models operate depths of the data, and to our knowledge they will be the first to integrate depth data from the stream of depth camera to constrain their model. Their premise is the limitation of movement of characterization possibilities in an RGB image; if one can easily define this movement in an XY plane, it becomes complex along a depth axis. A better feature extraction thus becomes trivial with the depth of information and will improve the performance of their attention model.

In [15], the authors suggest to study the differences of visual attention behavior when the depth is involved. For this, the authors propose a first substantial database containing 600 fixation measures obtained on pairs of 2D and 3D images. These data come from a Kinect camera. The authors want to measure differences in fixation between 2D and 3D images and the impact that the introduction of depth data can have on well-known 2D saliency model performance. The authors exhibit a set of priors related to the depth that are consistent with the attention process. Depth cues modulate visual saliency to a greater extent at farther depth ranges. Furthermore, humans fixate preferentially at closer depth ranges. A few interesting objects account for majority of the fixations, and this behavior is consistent across both 2D and 3D. They also found that the relationship between depth and saliency is nonlinear and characteristic for low and high depth-of-field scenes. The additional depth information led to an increased difference of fixation distribution between 2D and 3D version, especially when there are multiple salient stimuli located in different depth planes. Using their framework and approach on various models of 2D saliencies, the authors obtained a significant increase in the algorithm performance.

In [16], the authors focus on the extension of a model based on the contrast saliency, allowing it to integrate the depth through a disparity map from a set of stereo images. The authors also show increased performance of their model by introducing the disparity data. They also offer a large database of 1,000 stereoscopic images to validate their method.

In [17], the authors deal with saliency for 3D stereoscopic images. An interesting point of their approach is the comparison of the integration possibilities of depth either as a weighting element or through the establishment of a depth saliency map. The last method seems to give better results. The authors propose a detailed analysis of several 3D saliency models and define a classification of the models based on hox depth information that is integrated. Finally, they propose an adaptable framework for the existing 2D saliency model. The depth saliency map and 2D saliency map from a generic 2D saliency model are merged to provide the final saliency map. Extensive validation is provided for the various models.

In [18], the authors are interested in the role of depth in situations of competing saliencies due to appearance, depth-induced blur, and center bias. They propose a new saliency model by integrating first depth contrast, then many other features like color histogram, contour compactness, dimensionality, etc. They create a feature vector of 82 elements that are fused by a learning algorithm (SVM). Their approach shows that 3D saliency outperforms the other 2D saliency models.

In [19], the authors propose a new model of saliency based on the depth. They propose not to use the depth measurements as another channel of an image but by explicitly constructing 3D layout and shape features from depth measurements. The main idea is that humans use coplanarity to guide their assessment of saliency. Their method is based on fitting planes to the 3D points of the depth image, allowing to associate each pixel with the dominant plane that contains it. Therefore, they penalize points which lie on different depth planes and compute a dissimilarity measure between patches (locally adjacent pixels) to create a saliency map. The authors demonstrate the application of their algorithm on the segmentation of objects on RGBD data. To validate their approach, the authors have made a new dataset for depth-based saliency, including pixel-level ground truth segmentation of salient objects.

In [20], the authors are interested in a particular element related to attention and for which they will introduce a new feature map called the "depth-of-field map." The idea is that the depth-of-field map functions work similarly to the depth-of-field effect of human vision by enhancing the saliency of the regions near the point of gaze in the direction of depth and reducing the gaze movements between regions widely separated from each other. Their model is based on that of Itti and Ozeki which they are adding their own constraint based on depth of field.

In [21], the authors had two major contributions: their primary objective is to offer a wide enough RGBD database to be a real benchmark for 3D saliency; secondly, they proposed a model of saliency based on the depth that does not treat as an independent feature as in many models but simultaneously takes account of depth and appearance information from multiple layers. They based their approach on low-level feature contrast, mid-level region grouping, and high-level prior enhancement. Thanks to their large database, they carry out a quantitative analysis of their method against other well-known 2D augmented to 3D saliency models.

17.7.1.2 Stereoscopic-Based Models

In [22], the authors propose a saliency model in the context of the stereovision. Their model is based on a biological approach and highlights the problems of binocular vision that have a direct impact on the attention as the concepts of binocular rivalry. Their model is based on an existing model, the selective tuning model, which extends naturally as they demonstrate to the binocular vision.

In [23], the authors are interested in stereoscopic vision and involvement that it can have on the design of a saliency model, based on a biological modeling of the human attentive process. Indeed, if we consider the binocular nature, the source of stimuli is double and redundant. This issue is entitled the Attentional Stereo Correspondence Problem (ASCP). The authors propose a model of attention based on the depth that tends to consider this issue, proposing a model close to a model with relevant psychophysical characteristic of attention in depth.

17.7.2 3D Structure-Based Models

17.7.2.1 Mesh-Based Saliency

In [8], the authors introduce for the first time the concept of mesh saliency as a measure of importance for regions of a mesh structure. Their mesh saliency is defined in a scale-dependent manner using a center-surround operator on Gaussian-weighted mean curvatures. The model is based on the assumption that for a 3D mesh, geometry is the largest contributor to saliency. Their method estimates the saliency in terms of mean curvature with a mechanism of center-surround.

In [24], the authors propose a new method for extracting salient critical points of a mesh combining saliency mesh with Morse theory. Their method is based on a center-surround mechanism but also Gaussian-weighted average of the scalar of vertices. It offers an extension of the previous model using, as a weighting element, a bilateral filter rather than an absolute difference in weighted Gaussian.

In [25], the authors propose a variation of the model based on the difference of Gaussian for the extraction of salient points for the purpose of correspondence between various views of an object. Their method is based on measuring the displacement of vertices with respect to their original position after the various filters.

In [26], the authors propose a 3D object retrieval method based on the extraction of salient points in 3D. Their method of extracting salient points is based on classification by an SVM of the low characteristic histogram for each vertex of the model. The characteristic used is the absolute value of the curvature filtered by a Gaussian. With this classification, each vertex is defined as salient or not with a confidence score. For a 2D projection, the method generates a two-dimensional map of salient points that will be used to perform the signature recovery object.

In [27], the authors have an approach based on the definition of an information channel between a set of views and polygons of an object. It is this mutual information channel expressed by the Jensen-Shannon divergence [28] which allows them to firstly define a measure of similarity between views and secondly to extract the saliency of the 3D object. The idea is to express the way in which the polygons are perceived as a function of a set of viewpoints. For this, they express the saliency of a polygon as the average variation in the difference of Jensen-Shannon between this polygon and its neighbors. Based on these characteristics, the salient points are extracted with a classifier that detects points that have a combination of high curvature and low entropy values.

In [29], the authors propose an extension of their previously proposed method where they use now for the characterization of surfaces the absolute values of Gaussian curvature and Besl-Jain surface curvature characterization as the low-level surface properties.

In [30],the authors propose a new method for the detection of regions of interest on surfaces. Their method is local and global. It also takes into account the distance to the foci of attention. They use this method to determine the best possible view for a 3D object. Their method is based on a local approach and is based on the calculation of a descriptor on each vertex. Besides this local approach, the method integrates two characteristics that are the extremity detection and definition of patch association to represent the fact that regions of interest are the look that attracts more specific points.

In [31],the authors propose a rarity model on two levels: local and global. Indeed, they make the observation that all the models presented before them were based only on a local analysis of mesh, but a human observer also had a global vision. They thus introduce global saliency calculated on the mesh. The local part is based on the calculation of a heightmap to encode local structure. The global rarity is achieved using the same characteristic but where the comparison to a vertex is no longer local but global. To reduce the necessary calculation time, the authors use clustering to group the vertices with similar properties.

17.7.2.2 Point Cloud-Based Saliency

In [32], the authors propose one of the first models of saliency that apply specifically to point clouds. A major interest of this approach is the extraction of geometric features for each point of the cloud based on its neighborhood. The final saliency map is the composition of two intermediate maps: the first one is obtained based on what the authors call the local surface property (LSP) based in particular on normal surfaces or curvature. A second map is generated based on the distance of the points in the camera. Both maps are then linearly combined to produce the final saliency map.

In [33], the authors propose a 3D object detection framework based on the saliency. Although the system is intended to extract in the 3D environment salient objects, the employed attention mechanism is only based on 2D color image from an RGBD sensor. The interesting contribution is the idea of inhibition of return mechanisms (IOR) that inhibit the currently attended region in 3D.

In [34], the authors propose for the first time a saliency model specifically designed to handle large unorganized point clouds. The proposed method is based on the concept of global "distinctiveness." Their method employs the use of a descriptor for each point of the cloud. The authors define the simplified point feature histogram (SPFH), variation of fast feature point histogram (FPFH). A dissimilarity measure based on chi-square will be used to estimate the saliency.

17.7.3 Discussion

The classification of different methods solely based on their input data is a bit simplistic. Indeed, a system based on an RGBD sensor, although providing a depth map, may very well be converted into 3D point cloud by knowing the intrinsic properties of the sensor. Models designed for point cloud can be adapted to depth map processing and vice versa. For models based on meshes, data can be converted into a point cloud based on tessellation, opening the possibility to use a method for point cloud. There is thus a possibility of conversion and interoperability between methods.

An interesting models classification was proposed by [17] based on how the spatial information is integrated. Three categories are made on integrating the spatial aspect:

1. Depth-weighting models. This type of models (e.g., [7, 35]) does not contain any depth map-based feature-extraction processes. Apart from detecting the salient areas by using 2D visual features, these models share a same step in which depth information is used as the weighting factor of the 2D saliency. The saliency of each location (e.g., pixel, target, or depth plane) in the scene is directly related to its depth. Both 2D scene and depth map are taken as input.
2. Depth saliency models. The models (e.g., [13] and [9]) in this category take depth saliency as additional information. This type of models relies on the existence of "depth saliency maps." Depth features are first extracted from the depth map to create additional feature maps, which are then used to generate the depth saliency maps. These depth saliency maps are finally combined with 2D saliency maps (e.g., from 2D visual attention models using color, orientation, or intensity) by using a saliency map pooling strategy to obtain a final 3D saliency map. This type of model also takes the 2D scene and the depth map as input.
3. Stereovision models. Instead of directly using a depth map, this type of models (e.g., [22]) takes into account the mechanisms of the stereoscopic perception

in the HVS. Bruce and Tsotsos extend the 2D models that use a visual pyramid processing architecture [36] by adding neuronal units for modeling the stereovision. Images from both views are taken as input, from which 2D visual features can be considered. In addition, the model takes into account the conflicts between two eyes resulting from occlusions or large disparities.

The classification proposed here by [17] was proposed on a set of 2.5D model; but it could be extended to all models which extract geometrical characteristics from the input data. The 3D features are indeed not only limited to the depth map; richer features can be extracted from 3D data such as mesh.

The idea of depth saliency and depth weighting is interesting because of the abstract data type on which we work. The concept of depth saliency can be extended and be generalized to the extraction of salient geometric features related to spatial structure regardless of the data. We can therefore speak of spatial saliency. The depth weighting also introduced another idea for classification. Weighting by the depth map is fundamentally linked to the sensor and its field of vision. It therefore depends on the viewpoint. Inversely, if we apply this idea to methods on mesh, the methods are not viewpoint dependent.

This distinction between models linked or not with a point of view is fundamental. In particular, it raises issues related to the validation and the link with the human visual system. Could we still consider a saliency model, an algorithm running on a mesh processing 3D volumes independently from the viewpoint? Such a system operates basically beyond human visual capabilities.

The classification we propose is to make the following distinctions:

1. 2.5D saliency that processes 3D data but is dependent on the viewpoint
2. 3D saliency that processes 3D structure as a whole and is not dependent of the viewpoint

This separation makes abstraction of the data type being processed while taking into account the nature of the data dependency or not to a point of view.

17.8 SuperRare3D: A New Model of Point Cloud 3D Saliency Based on Supervoxel Rarity

We propose a novel object-oriented algorithm of bottom-up attention dedicated to analyze colored point clouds. This model builds on the one proposed in [36]. One contribution is the use of a rarity-based approach not based on superpixels as in [36] but on supervoxels. Supervoxels consist of an over-segmentation of a point cloud in regions with comparable sizes and characteristics (in terms of color and other 3D features). More details on supervoxels and the method used here are provided in the next sections. Our approach has four major interests:

1. Supervoxels let us reduce the amount of processing and allow our method to work on organized or unorganized clouds. Thus, it can analyze point clouds or even fused point clouds coming from various sensors.
2. Supervoxels allow us to have an object-oriented approach in the 3D space.
3. Supervoxel multilevel decomposition allows us to maintain detection performance regardless of the size of the salient objects present in the data.
4. This approach provides a bottom-up 3D saliency map which is viewer independent. It is then possible to add viewer-dependent top-down information as a viewer-dependent centered Gaussian and depth information. In our paper, we only use the centered Gaussian that all the other models also use to remain fair in our comparison.

Our method only uses one feature of the point cloud: the color. Other features like supervoxel orientation or other specific 3D features will be taken into account in future work. As the color feature is the only one we use, this approach is subject to the influence of the choice of the color representation. To provide a first solution to this influence, we propose to fuse the saliency maps computed on several color spaces. Our algorithm can be divided into three major stages: (1) supervoxel decomposition, (2) supervoxel rarity-based saliency mechanism, and (3) fusion. We present in the following subsections the three main steps of our algorithm (Fig. 17.4).

17.8.1 Supervoxel Cloud Segmentation

The superpixels are the result of over-segmentation of an image into regions of pixels having similar perceptual properties. This is a step commonly used in computer vision as a preprocessing stage to reduce the amount of information to be processed while still minimizing the loss of information.

We build our system, on the same idea by using supervoxels instead of processing at the point level. We use the voxel cloud connectivity segmentation method (VCCS) [38] that extracts supervoxels from an organized or unorganized point cloud. The supervoxels can replace the structure of the voxel-based original point cloud by a set of atomic regions that capture the local redundancy of information. They provide a convenient way to summarize the point cloud and thus greatly reduce the complexity of the following analysis process. But if there is a major difference between the size of supervoxels and the size of the salient object to be detected, this one can be merged with a nearby supervoxel and its information is lost Fig. 17.5. To avoid this situation, the rarity mechanism is applied to different levels of supervoxel decomposition so that at some level of detail the salient object is well captured. At this point the pathway of the algorithm is split between the different levels of supervoxel decomposition. This separation is made to capture all the information of salient objects by adjusting the size of supervoxels. Indeed, like shown in Fig. 17.5,

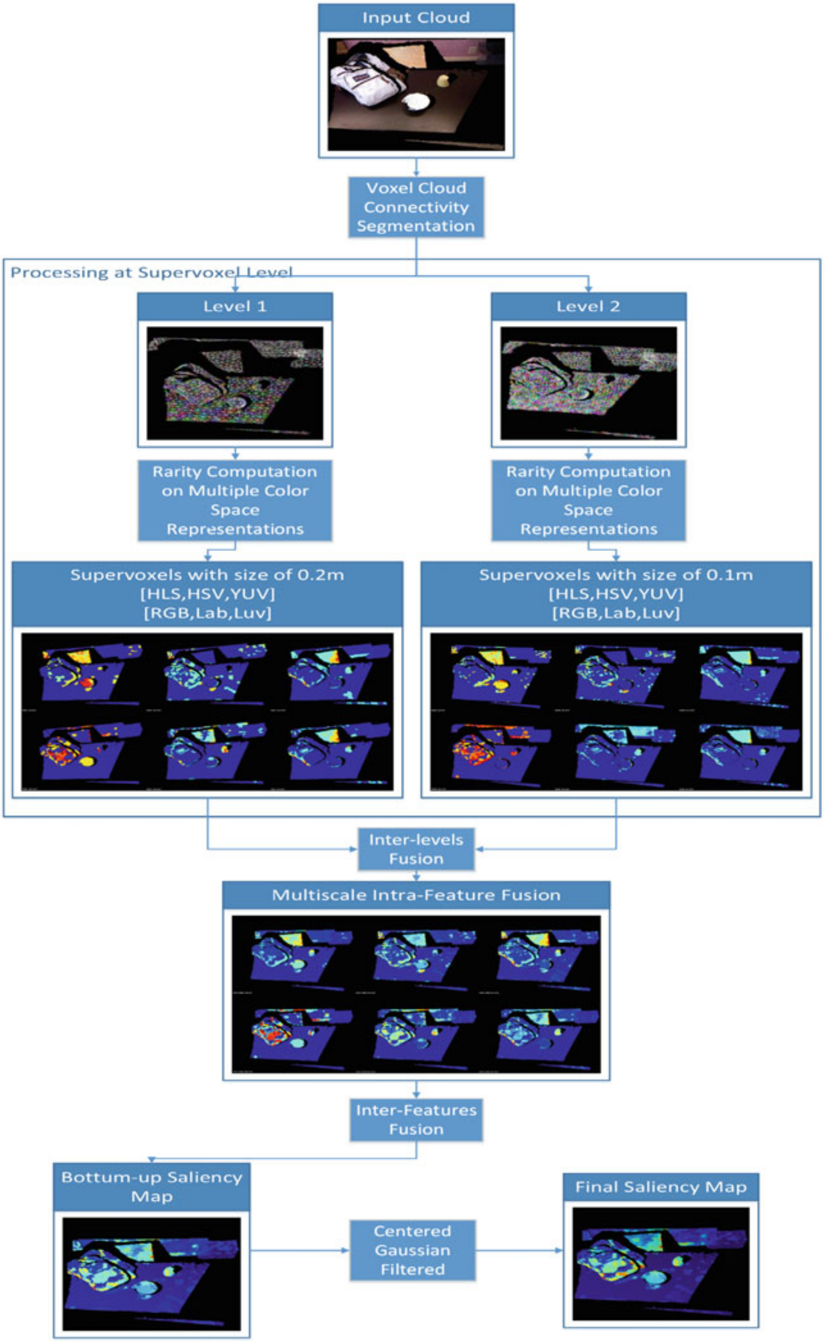

Input Cloud
Voxel Cloud Connectivity Segmentation
Processing at Supervoxel Level
Level 1
Level 2
Rarity Computation on Multiple Color Space Representations
Rarity Computation on Multiple Color Space Representations
Supervoxels with size of 0.2m [HLS,HSV,YUV] [RGB,Lab,Luv]
Supervoxels with size of 0.1m [HLS,HSV,YUV] [RGB,Lab,Luv]
Inter-levels Fusion
Multiscale Intra-Feature Fusion
Inter-Features Fusion
Bottum-up Saliency Map
Centered Gaussian Filtered
Final Saliency Map

Fig. 17.4 Our method is divided in three major steps: (1) multiscale supervoxel decomposition, (2) color rarity applied on multiple color spaces, and (3) inter-level and inter-feature fusion. A top-down centered Gaussian can be used to simulate the human centric preference [37]

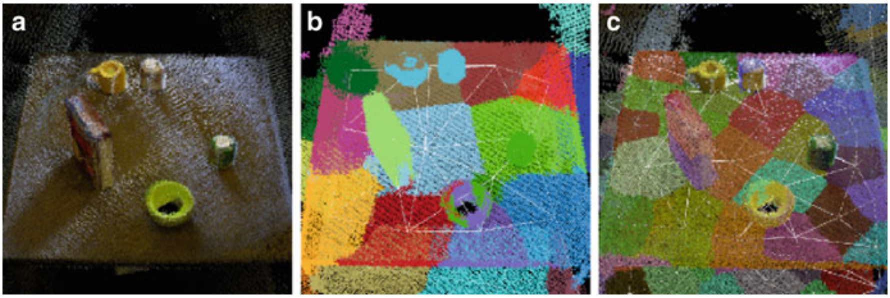


Fig. 17.5 (**a**) Example of table point cloud with a box and cups, (**b**) supervoxel segmentation using VCCS with a seed size of 0.5 m, (**c**) supervoxel segmentation using VCCS with a seed size of 0.25 m. The size of the supervoxels is essential to extract the information of all the objects. If the seed is too large, like in (**b**), we see the object being absorbed in an adjacent supervoxel, losing the information for the rarity mechanism

if a supervoxel is too large, it may not stick properly to an object and it is seen disappearing into an adjacent supervoxel. To remedy this, the algorithm works on several levels in parallel that will then be merged into a final saliency map, to maintain both the information of large objects and smaller ones, while refining the segmentation of salient regions.

17.8.2 Rarity-Based Saliency

The rarity mechanism consists, for each supervoxel vector, to compute the cross-scale occurrence probability of each of the N supervoxels. At each color component i, a rarity value is obtained by the self-information of the occurrence probabilities of the supervoxel as shown in Eq. (17.8). P_i is the occurrence probability of each supervoxel Sv_i value in respect with the empirical probability distribution represented by the histogram within the ith color channel:

$$Rarity(Sv_i)_i = -log(P_i/N) \tag{17.8}$$

Then, the self-information is used to represent the attention score for the supervoxel region. This mechanism provides higher scores for rare regions. The rarity value falls between 0 (all the supervoxels are the same) and 1 (one supervoxel is different from all the others).

17.8.2.1 Intra- and Inter-supervoxel Level Fusion

The rarity maps obtained from the rarity mechanism on each color channel (in this case, we select six color space representations: HSV, HLS, YUV, RGB, Lab, Luv) are first intra-color combined. In our example, we empirically select a two-level decomposition using supervoxel seed of 0.05 m and 0.02 m for balance between accuracy and computation time. A fusion between same color rarity maps is achieved at each decomposition level by using the fusion method proposed in Itti et al. [40]. The idea is to provide a higher weight to the map which has important peaks compared to its mean (Eq. 17.9):

$$S = \sum_{i=1}^{N} EC_i * map_i \tag{17.9}$$

where EC_i is the efficiency coefficient for each channel and is computed as in Eq. 17.10:

$$EC_i = (max_i - mean_i)^2 \tag{17.10}$$

These coefficients let us sort the different maps (map_i) based on each map efficiency coefficient EC_i. Each map is then multiplied by a fixed weight defined as $i = 1 \ldots K$ where K is the number of maps to mix (here $K = 3$) and i the rank of the sorted maps as shown in the first line of Eq. 17.11. T is an empirical threshold defined in [41]:

$$\forall i \in [1, K] \begin{cases} saliency_i = 0 & \text{if} \frac{EC_i}{EC_K} < T \\ saliency_i = \frac{i}{K} * map_i & \text{if} \frac{EC_i}{EC_K} \geqslant T \end{cases} \tag{17.11}$$

At the end of this first fusion process, the model provides one saliency map per color space representation. The second fusion step, an inter-color feature fusion between each map coming from the different color space representation, is achieved using the same method as the one explained for the inter-decomposition level fusion (Eq. 17.9).

17.8.3 Color Space Influence

Our method estimates saliency using the rarity only on color feature. The accuracy of this feature is very important and our method is strongly influenced by the choice of the color space representation. If we observe independently saliency maps for the different color modes, we can see that the performance is highly dependent on the color space, ranging from excellent to poor, but in all cases at least one map provides good performance. For this reason, we have chosen to apply the rarity on several color spaces and merge the different rarity maps.

17.8.4 Final Saliency Map

Finally, in this case, we work with an organized point cloud; we apply a Gaussian-centered filter to represent the central preference that people exhibit in images [37]. In the case of object avoidance, this centered human preference makes also sense in the context of robotics as one wants to correct the path of a robot to avoid collisions with objects in front of it.

17.9 Validation

17.9.1 Database

The database that we used to validate our method was published by [19]. It has 80 shots obtained using a Microsoft Kinect sensor mounted on a Willow Garage PR2 robot. The database consists of RGB images, depth maps, and point clouds associated with pixel-level ground truth segmentation masks. The 80 scenes are very complex both in terms of number and shape of objects, colors, and illumination but also in terms of depth differences. Indeed, there are a lot of objects which have little depth difference with those objects.

17.9.2 Metric

Several measures like the area under the ROC curve (AUROC) and the precision-recall curve have been suggested to evaluate the accuracy of salient object detection maps. However, as shown in [42], these most commonly used measures do not always provide a reliable evaluation. The authors start by identifying three causes of inaccurate evaluation: (1) interpolation flaw, (2) dependency flaw, and (3) equal-importance flaw. By amending these three assumptions, they propose a new reliable measure called $F_\beta^w - measure$ and defined in Eq. 17.12:

$$F_\beta^w = (1+\beta^2)\frac{Precision^w * Recall^w}{\beta^2 * Precision^w + Recall^w} \tag{17.12}$$

with

$$Precision^w = \frac{TP^w}{TP^w + FP^w}$$
$$Recall^w = \frac{TP^w}{TP^w + FN^w}.$$

where TP = true positives, FP = false positives, FN = false negatives.

The weight *w* has been chosen to resolve the flaws. This metric provides better evaluation than previous measures. We will use this new method to validate SuperRare3D on the database in order to be as fair and precise as possible.

17.9.3 Method

We made the validation of our SuperRare3D model (called SR3D) in two steps. First, we computed a 2D saliency map as a view of the 3D saliency map (2D projection). We compared SR3D to five other depth-extended (2.5D) models. The weighted F-measure is used to compare SR3D with 2.5D saliency methods given a pre-segmented ground truth. Models of visual attention can be split in two main categories based on their purpose. The first category of models aims to predict the human eye gaze distribution. The second category focuses on finding interesting objects. Our model fits in the second category and intends to segment complex scenes into an object hierarchy based on the objects of interest. Some of them are extended to use depth feature maps (called further in this paper 2.5D models). Those models are the ones also used to asses our method in this section.

In [9], the authors aim at an extension of the visual attention model with the integration of depth in the computational model built around conspicuity and saliency maps. This model is an extension of center-surround 2D saliency with depth proposed by [1]. In [19], the method constructs 3D layout and shape features from depth measurement that they integrate with image-based saliency. This method is an extension of center-surround 2D saliency with depth proposed by [43].

17.9.4 Results

Our full-3D model (SR3D) provides a 3D viewpoint-independent saliency map of any kind of organized or unorganized point cloud. Figure 17.6 shows two examples of results for two different types of point clouds. The first column shows the results from a single Kinect point cloud. The second column, an example of result on a point cloud obtained using a co-calibrated laser scanner, is displayed. First row shows the input colored point clouds and second row the full-3D bottom-up viewpoint-independent saliency maps. This figure shows two crucial advantages of the proposed model over any existing 2D or 2.5D saliency model: (1) the ability to work on any kind of structured or unstructured point cloud and (2) the ability to provide viewpoint-free 3D saliency maps which might be adapted to any given viewpoint (Figs. 17.7 and 17.8).

Figure 17.9 shows the results of the validation. Concerning the comparison with the 2.5D models, SR3D outperforms all the other models. However, like shown on figure, this performance difference is not statistically significant.

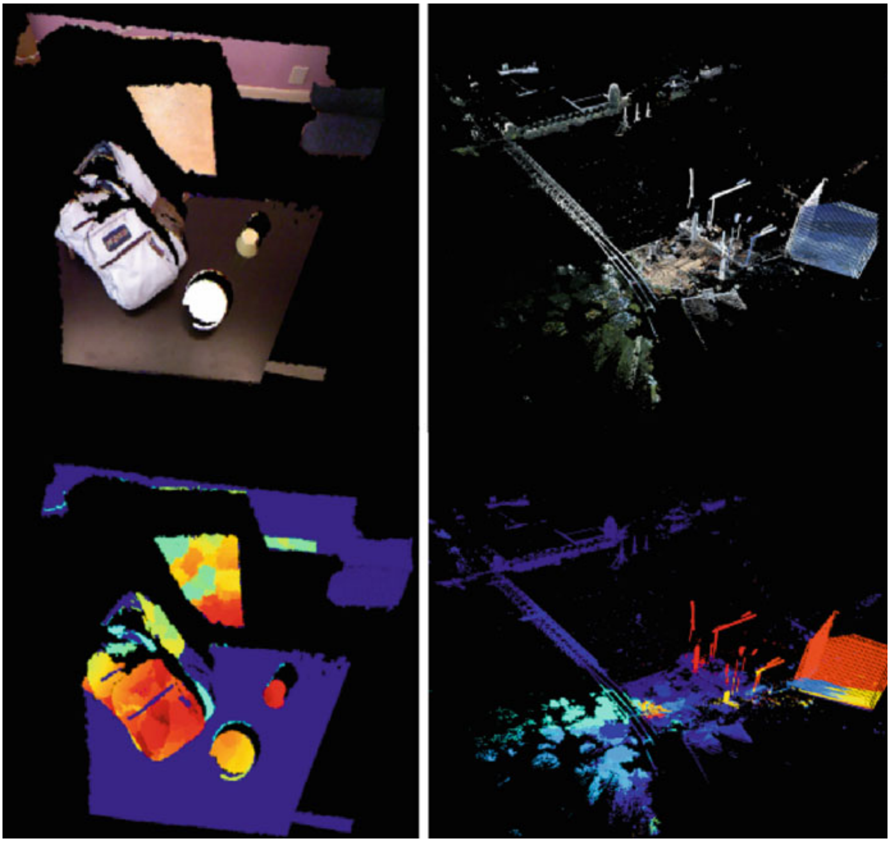


Fig. 17.6 Examples of results obtained with our method on three different point clouds: (*left*) a Kinect cloud (organized, 307,200 points, three levels of decomposition with 92, 32, and 13 supervoxels); (*right*) a point cloud recorded using a Riegl VZ-400 and a co-calibrated Canon 1000D camera with 10 megapixels [39] (unorganized, 5,976,977 points, one level of decomposition with 271 supervoxels)

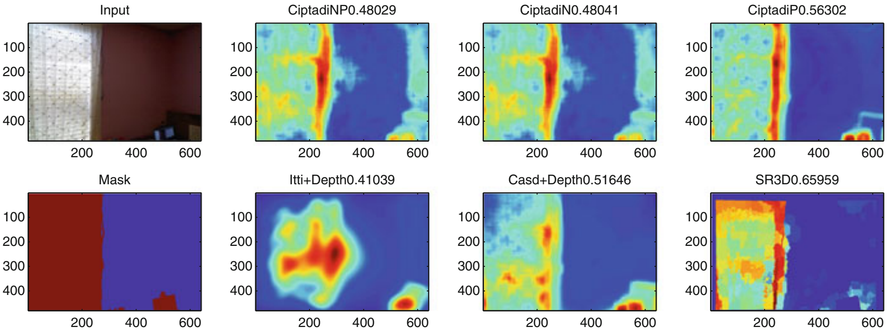


Fig. 17.7 Qualitative comparison of our model

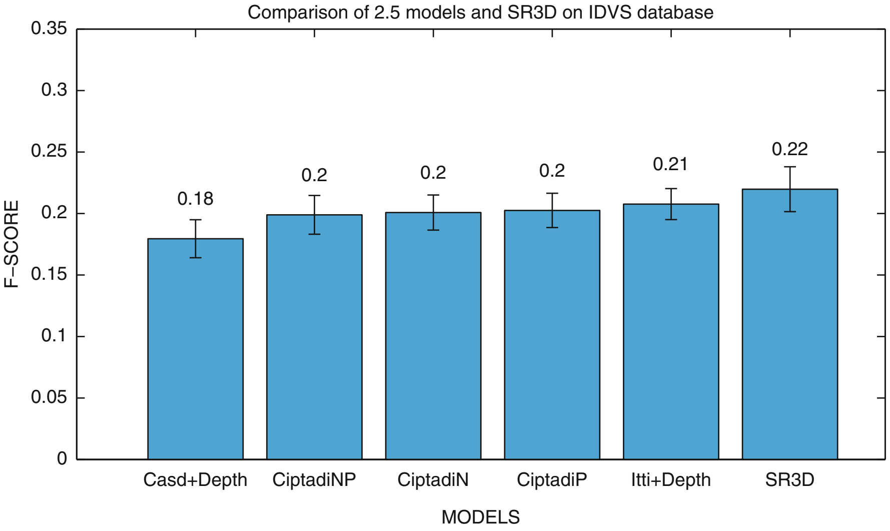


Fig. 17.8 Quantitative comparison of our model with five state-of-the-art 2.5 saliency models from the database [19]. SR3D outperforms with the other models

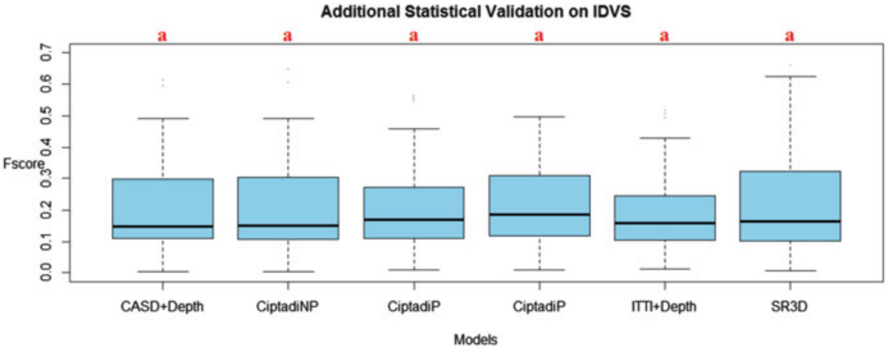


Fig. 17.9 If our model outperforms the others, however, it is not significantly above

17.10 Summary

3D is a fundamental element of the human vision system, and it is as much for visual attention mechanisms. If the study and integration of 3D features in the design of computational models of attention began early, it is only in recent years they have really grown. 3D attention can by applied on high-level features as extracted from human silhouettes or it can also be applied on low-level features as depth maps or 3D features. This 3D information plays an essential role in a 3D attention mechanism whatsoever in both on bottom-up and top-down. For bottom-up, the use of spatial information not only weight conventional salience map for giving importance to the regions according to their proximity but also by extracting 3D features, could improve significantly the performance of a saliency model. For the top-down, 3D data offer many opportunities to extract information on the environment, the scene, or the objects, leading to a more detailed or semantic analysis of the environment to constrain the saliency. In this chapter, we made a review of multiple methods of "3D saliency." Indeed, it is necessary to distinguish between models based on their dependence on a point of view; this is what prompted us to redefine our vision of saliency models by classifying them according to the notions of 2.5D and 3D. Given this classification, we have proposed a new model of salience based on rarity, effective in both categories, capable of handling large amounts of 3D data while taking into account the color information. Surprisingly few models are interested in full 3D that integrates this color information. Our model is efficient. 3D saliency is an early field but now reemerges, thanks to the appearance of numerous 3D sensors. This area is rich and complex and still offers many challenges in modeling human attention.

References

1. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*, 1254–1259.
2. Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE MultiMedia, 19*(2), 4–10.
3. Mancas, M., Riche, N., Leroy, J., & Gosselin, B. (2011). Abnormal motion selection in crowds using bottom-up saliency. In *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP 2011)*, Brussels (pp. 229–232). IEEE.
4. Villaroman, N., Rowe, D., & Swan, B. (2011). Teaching natural user interaction using openni and the microsoft kinect sensor. In *Proceedings of the 2011 Conference on Information Technology Education*, West Point (pp. 227–232). ACM.
5. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013). Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication, 28*(6), 642–658.
6. Mancas, M., Riche, N., Leroy, J., Gosselin, B., & Dutoit, T. (2011). Toward a social attentive machine. In *AAAI Fall Symposium: Robot-Human Teamwork in Dynamic Adverse Environment*, Arlington.
7. Maki, A., Nordlund, P., & Eklundh, J.-O. (1996). A computational model of depth-based attention. In *Proceedings of 13th International Conference on Pattern Recognition* (Vol. 4, no. 1, pp. 132–141).
8. Lee, C. H., Varshney, A., & Jacobs, D. W. (2005). Mesh saliency. *ACM Transactions on Graphics, 24*(3), 659.
9. Ouerhani, N., & Hugli, H. (2000). Computing visual attention from scene depth. *Proceedings of the 15th International Conference on Pattern Recognition (ICPR-2000)*, Barcelona (Vol. 1, pp. 375–378).
10. Jost, T., Ouerhani, N., & Wartburg, R. (2004). Contribution of depth to visual attention: Comparison of a computer model and human. In *Proceedings Early cognitive vision workshop, 3D Modelling*.
11. Hügli, H., Jost, T., & Ouerhani, N. (2005). Model performance for visual attention in real 3D color scenes. In *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach* (pp. 469–478).
12. Frintrop, S., Rome, E., Nüchter, A., & Surmann, H. (2005). A bimodal laser-based attention system. *Computer Vision and Image Understanding, 100*(1–2)(Special issue), 124–151.

13. Potapova, E., Zillich, M., & Vincze, M. (2011). Learning what matters: Combining probabilistic models of 2D and 3D saliency cues. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6962*, 132–142.
14. Riche, N., Mancas, M., Gosselin, B., & Dutoit, T. (2011). 3D Saliency for abnormal motion selection: The role of the depth map. In *Computer Vision Systems* (pp. 143–152). Berlin/Heidelberg: Springer.
15. Lang, C., Nguyen, T. V., Katti, H., Yadati, K., Kankanhalli, M., & Yan, S. (2012). Depth matters: Influence of depth cues on visual saliency. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7573*, 101–115.
16. Niu, Y., Geng, Y., Li, X., & Liu, F. (2013). Leveraging stereopsis for saliency analysis. *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
17. Wang, J., Da Silva, M. P., Le Callet, P., & Ricordel, V. (2013). Computational model of stereoscopic 3D visual saliency. *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 22*(6), 2151–2165.
18. Desingh, K., Krishna, K. M., Rajan, D., & Jawahar, C. (2013). Depth really matters: Improving visual salient region detection with depth. In *Proceedings of the British Machine Vision Conference 2013*, Bristol (pp. 98.1–98.11).
19. Ciptadi, A., Hermans, T., & Rehg, J. (2013). An in depth view of saliency. *Proceedings of the British Machine Vision Conference 2013*, Bristol (pp. 112.1–112.11).
20. Ogawa, T., Ozeki, M., Oka, N. (2014). A visual attention model using depth information from the point of gaze. *ii.is.kit.ac.jp* (pp. 125–130).
21. Peng, H., Li, B., Xiong, W., Hu, W., & Ji, R. (2014). RGBD salient object detection: A benchmark and algorithms. In *ECCV*, Zurich (no. 1, pp. 92–109).
22. Bruce, N. D. B., & Tsotsos, J. K. (2005). An attentional framework for stereo vision. *The 2nd Canadian Conference on Computer and Robot Vision (CRV'05)*, Victoria.
23. Bruce, N. D. B., & Tsotsos, J. K. (2012). Attention in Stereo Vision: Implications for Computational. In *Developing and Applying Biologically-Inspired Vision Systems: Interdisciplinary Concepts: Interdisciplinary Concepts* (pp. 65).
24. Liu, Y. S., Liu, M., Kihara, D., & Ramani, K. (2007). Salient critical points for meshes. In *Proceedings of the 2007 ACM symposium on solid and physical modeling*. ACM.
25. Castellani, U., Cristani, M., Fantoni, S., & Murino, V. (2008). Sparse points matching by combining 3D mesh saliency with statistical descriptors. *Computer Graphics Forum, 27*(2), 643–652.
26. Atmosukarto, I., & Shapiro, L. G. (2008). A salient-point signature for 3D object retrieval. In *Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval (MIR 2008)*, Vancouver (p. 208)
27. Feixas, M., Sbert, M., & González, F. (2009). A unified information-theoretic framework for viewpoint selection and mesh saliency. *ACM Transactions on Applied Perception, 6*(1), 1–23.
28. Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory, 37*(1), 145–151.
29. Atmosukarto, I., & Shapiro, L. G. (2010). 3D object retrieval using salient views. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR 2010)*, Philadelphia (p. 73)
30. Leifman, G., Shtrom, E., & Tal, A. (2012). Surface regions of interest for viewpoint selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, Providence (pp. 414–421).
31. Wu, J., Shen, X., Zhu, W., & Liu, L. (2013). Mesh saliency with global rarity. *Graphical Models, 75*(5), 255–264.
32. Akman, O., & Jonker, P. (2010). Computing saliency map from spatial information in point cloud data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6474*(PART 1), 290–299.

33. Garcia, G. M., & Frintrop, S. (2013). A computational framework for attentional 3D object detection. In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 2984–2989).
34. Shtrom, E., Leifman, G., & Tal, A. (2013). Saliency detection in large point sets. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2013)*, Sydney (pp. 3591–3598).
35. Zhang, Y., Jiang, G., Yu, M., & Chen, K. (2010). Stereoscopic visual attention model for 3D video.In *Advances in Multimedia Modeling*, Chongging (pp. 314–324).
36. Leroy, J., Riche, N., Mancas, M., Gosselin, B., & Dutoit, T. (2014). SuperRare: An object-oriented saliency algorithm based on superpixels rarity. *IEEE International Conference on Robotics and Automation* (ICRA 2014).
37. Judd, T., Durand, F., & Torralba, A. (2012). *A benchmark of computational models of saliency to predict human fixations*.
38. Papon, J., Abramov, A., Schoeler, M., & Worgotter, F. (2013).Voxel cloud connectivity segmentation – Supervoxels for point clouds. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Portland (pp. 2027–2034).
39. Dorit, B., Jan, E., HamidReza, H., & Andreas, N. Robotic 3D Scan Repository. http://kos.informatik.uni-osnabrueck.de/3Dscans/.
40. Itti, L., & Koch, C. (1999). A comparison of feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging, 10*, 161–169.
41. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013). Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication, 28*(6), 642–658.
42. Margolin, R., Zelnik-Manor, L., & Tal, A. (2014). How to evaluate foreground maps? In *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus.
43. Goferman, S., Zelnik-Manor, L., & Tal, A. (2012). Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(10), 1915–1926.

Chapter 18
Applications of Saliency Models

Matei Mancas and Olivier Le Meur

18.1 Attention Modeling: A Very Wide Spectrum of Applications

In engineering, the automatically computed output of a saliency model is called a saliency map. Saliency maps can be computed on still images (Chap. 9), videos (Chap. 10), audio signal (Chap. 16), and even 3D data (Chap. 17). Those maps provide for each pixel in an image or video frame each voxel on a 3D model or at a given time position in an audio file the probability to be attended by human gaze. They include bottom-up information using low-level features directly extracted from the signal or they can also include top-down information related to memory or emotions.

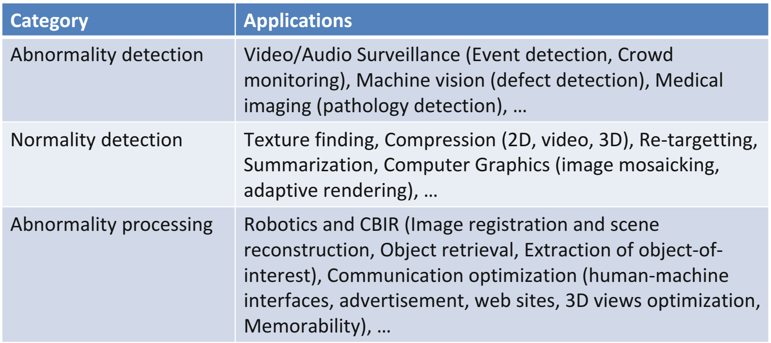
The applications of saliency maps are numerous and they can occur in many domains. For some applications, like in advertising or interface optimization, the saliency maps and their analysis are the final goal, while for others (compression, object recognition, etc.) saliency maps are not a goal per se, but they act like informational filters to improve the efficiency of other techniques.

While an exhaustive list of saliency map applications would be difficult to provide and to structure, we propose in this chapter a taxonomy composed of three categories. Different application domains can be split within those three categories (Fig. 18.1).

M. Mancas (✉)
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: matei.mancas@umons.ac.be

O. Le Meur
IRISA, University of Rennes, Campus Universitaire de Beaulieu, 35042, Rennes, France

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_18



Category	Applications
Abnormality detection	Video/Audio Surveillance (Event detection, Crowd monitoring), Machine vision (defect detection), Medical imaging (pathology detection), …
Normality detection	Texture finding, Compression (2D, video, 3D), Re-targetting, Summarization, Computer Graphics (image mosaicking, adaptive rendering), …
Abnormality processing	Robotics and CBIR (Image registration and scene reconstruction, Object retrieval, Extraction of object-of-interest), Communication optimization (human-machine interfaces, advertisement, web sites, 3D views optimization, Memorability), …

Fig. 18.1 A three-class taxonomy for saliency applications

Basically, attention maps provide cues about the surprising parts of a signal. A first category of applications directly takes advantage of the detection of those surprising, thus abnormal, areas in the signal. We will call this class of applications "abnormality detection." Surveillance and event detection are examples of application domains in this category.

A second category will focus more on the opposite of the first one: as the attention maps provide us with an idea about the surprising parts of the signal, one can deduce where is the normal (homogeneous, repetitive, usual, etc.) signal. We will call this category "normality modeling." The main application domains are in signal compression or retargeting.

Finally, the third application category is related to the surprising parts of the signal but will go further than a simple detection. This application family will be called "abnormality processing," and it will need to compare and further process the most salient regions. Domains such as robotics, object retrieval, or interface optimization can be found in this category.

In the rest of the chapter we will follow this taxonomy which has the advantage to group dozens of applications into only three categories (Fig. 18.1). The review of the applications of attention modeling has the ambition to be as exhaustive as possible by listing all the known applications. Nevertheless, if examples and references are provided for each application, those references are not necessarily exhaustive.

In each of the three categories, some applications will be listed and others further detailed. Moreover, some of other chapters such as Chaps. 16, 17, 19, 20, or 21 are dedicated to specific saliency map applications.

18.2 Applications Based on Abnormality Detection

In this section the main focus is on the first category of applications within our taxonomy: the applications which use the detection of the areas having the higher saliency scores. Those areas correspond with events, defects, pathologies, or social-related interactions in real-life applications.

18.2.1 Video Surveillance

Video surveillance encompasses a sheer number of applications which can benefit from attention modeling. Here we provide some non-exhaustive examples of the use of attention in this context.

An interesting European research project, called "SeaRise," focused on saliency and video surveillance in real-life situation. The main purpose of the project is to develop a trinocular active cognitive vision system called "Smart-Eyes" which first detects abnormal motion using saliency models and then focuses on the detected area for tracking and categorization of salient events [1]. Saliency models were developed to take into account spatial but also motion information [2]. Additional long-term information was used to distinguish usual paths or motion from abnormal motion (Fig. 18.2).

Other authors took into account the concept of "usual motion" either by using accumulation of motion features in given regions which provide a "normality" of the motion in those regions [3] or using more complex systems as hidden Markov models (HMMs) to predict future normal motion [4]. Crowd monitoring is an important topic in surveillance as it is very difficult to detect quickly and in an automatic way a suspect behavior. Some approaches rely on motion rarity [5] or motion textures [6] while [7] working on human gaze modeling on crowd videos. A benchmark of several models on a dataset which also includes crowd videos is available in [8].

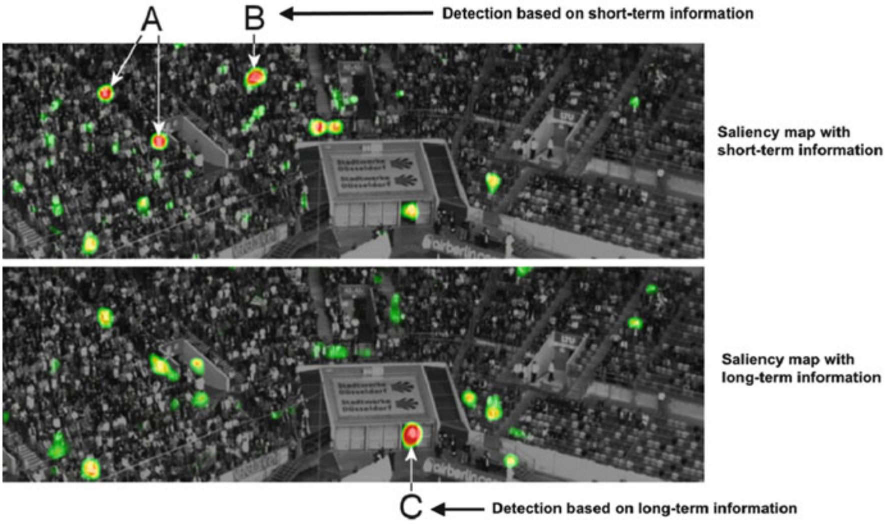

A
B
Detection based on short-term information
Saliency map with short-term information
Saliency map with long-term information
C
Detection based on long-term information

Fig. 18.2 An example of the use of long-term information. While the *regions A* and *B* on *top* contain a high amount of motion and are very salient when only short-term information is taken into account, when "usual" paths are taken into account, the region C becomes much more interesting (Adapted from [2])

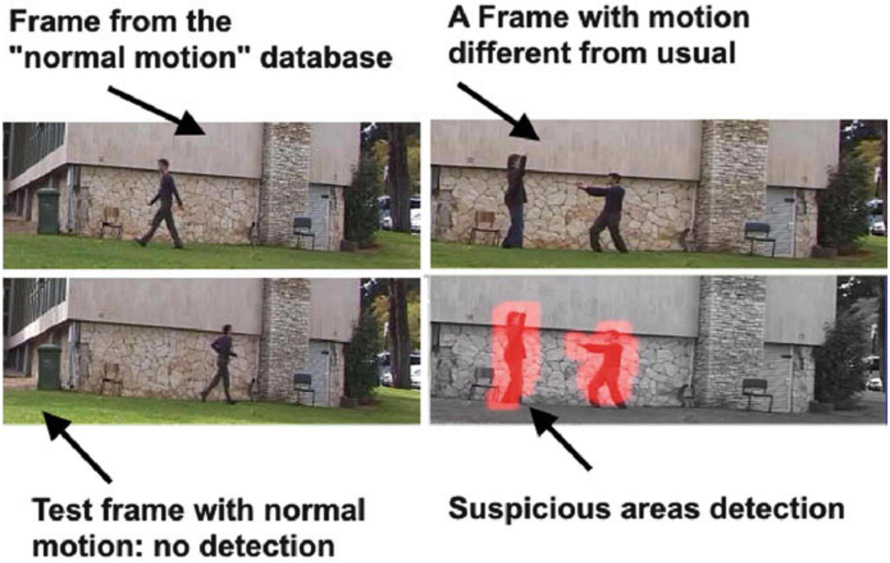

Frame from the "normal motion" database
A Frame with motion different from usual
Test frame with normal motion: no detection
Suspicious areas detection

Fig. 18.3 *Top left*: a frame from a normal motion dataset. *Bottom left*: a test frame containing similar motion. *Top right*: a frame with motion different from the dataset. *Bottom right*: suspicious motion detection and localization (Adapted from [9])

While abnormal motion has been mostly used for crowd scenes, some authors like in [9] provide models which work on any general scene containing motion (Fig. 18.3). An issue with this model is that it is very computationally expensive as it takes into account not only video patches but also the relative position of those patches. In addition video datasets are needed to learn normal motion before being able to detect any suspect behavior.

18.2.1.1 A Discussion

Attention in video surveillance is a prolific domain with lots of recent references. The search for abnormal events is one of the most important "quests" in the domain. While saliency is currently not a mainstream idea in video surveillance, there are good chances for it to become an important axis of research in the domain in the next years.

18.2.2 Audio Surveillance

Audio surveillance is a domain which is much less investigated compared to video surveillance. Nevertheless, microphones could also be added to surveillance

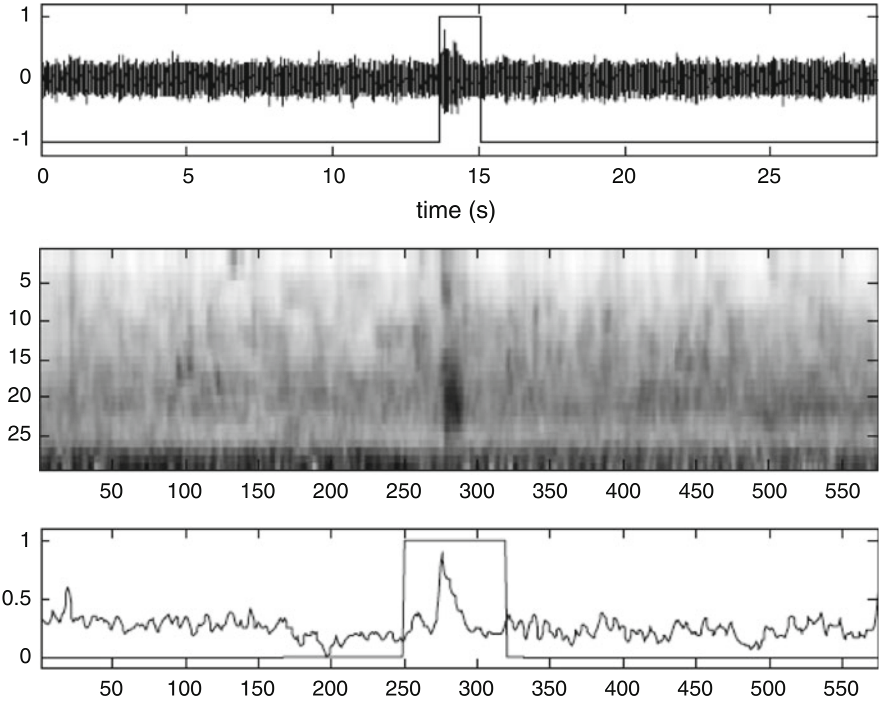


Fig. 18.4 *Top*: audio signal with reference segmentation. *Middle*: a space-time representation of the signal (cochleogram). *Bottom*: attention peak detection (Adapted from [10])

cameras. While a human operator still can watch several screens in the same time to monitor data from the cameras, it is almost impossible for him to listen to several audio sources simultaneously. In this latter case, automatic methods for audio event detection are crucial.

For instance, some saliency models were used [10, 11] to spot unusual sounds in classical contextual sounds like a gunshot in the middle of a metro station audio ambiance (see Fig. 18.4). The idea is to automatically select the camera corresponding with the microphone where the unusual audio event is detected. A normalized environment adaptive audio attention model based on space and audio clues was also proposed in [12].

18.2.2.1 A Discussion

Compared to video surveillance, audio surveillance is a smaller investigation field by itself. Moreover, there are very few audio models existing. The use of both audio and video saliency is only achieved in the fields of robotics or social interactions (see the Chaps. 16 and 21). Nevertheless, attention models have a real interest in the domain. While this application should stay rather limited in a short-term perspective, there is a lot of potential at a long-term perspective.

18.2.3 Machine Vision: Defect Detection

Machine vision is the application of computer vision to industry and manufacturing. One of the applications of machine vision is the automatic inspection of manufactured goods. Machine vision systems perform precise tasks such as counting objects on a conveyor, reading serial numbers, and searching for surface defects. These systems are preferred for repetitive high-speed tasks, and they are sometimes used to complement human's work which provides a finer perception over a short period of time and which is much more flexible in classification and adaptation to new defects.

In [13], machine vision was applied first to automatic fruit grading. Automatic quality inspection of fresh fruits by machine vision is a challenge not only due to their largely varying physical appearances but also because of the need to decrease the cost, time, and error of inspection introduced by human experts. Figure 18.5 shows the results using a global rarity-based model. As the apple is the main object of the scene, local contrast is not needed and the use of global rarity alone is the best approach once a preprocessing step which eliminates the apple contours and background is achieved. The results in the middle are promising, but some regions that are neither contour nor defective have also high attention scores. These "false positives" are mainly due to illumination artifacts or to the presence of stem or calyx regions which are quite similar to defects.

An "atlas" is used to provide the algorithm with images containing healthy apples. If uneven illumination and shadows often occur on healthy apple images,

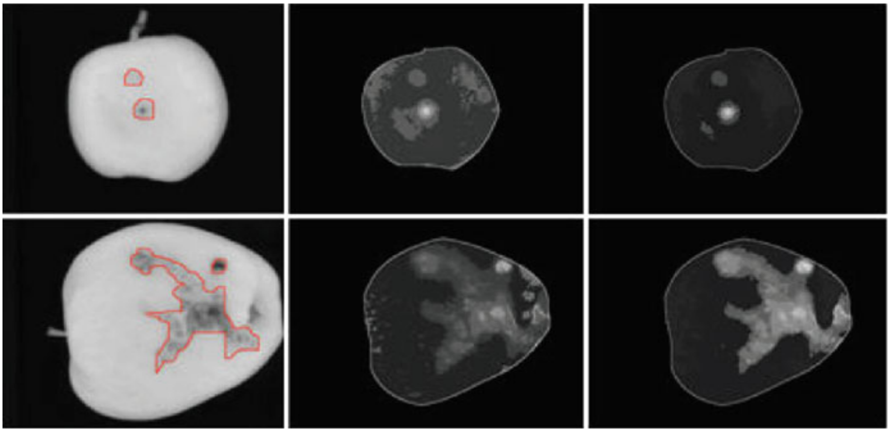


Fig. 18.5 *Left*: initial apple images. *Middle*: global rarity-based saliency maps after preprocessing. *Right*: global rarity saliency maps using an "atlas" or dataset of healthy apples (Adapted from [13])

they will be also found in the atlas; thus, even if these artifacts are rare within the initial image, they will be less rare if the entire atlas is taken into account. On the contrary, the defective skin will be even more rare as it never occurs within the atlas, but only on the test image. The results of the use of healthy apple examples are visible in the right column of Fig. 18.5. The defects become in this case more visible and noise due to illumination is eliminated.

In [9], in addition to video surveillance, their model can also apply to static images using or not an additional atlas. Figure 18.6 shows the results of defect

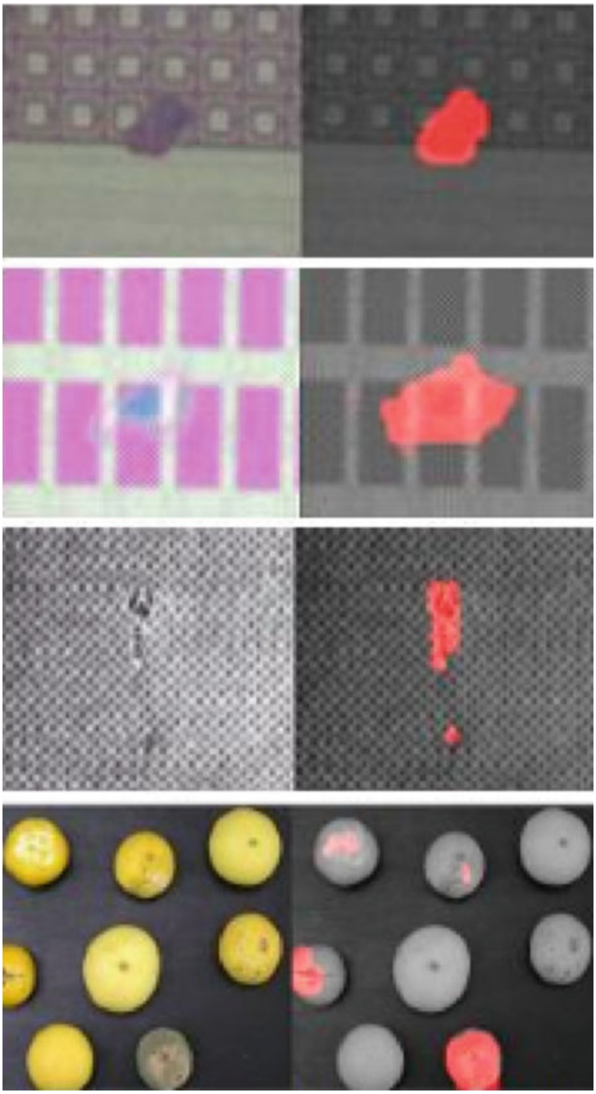


Fig. 18.6 *Left column*: initial images containing defects. *Right column*: defected located using examples of images with no defects (Adapted from [9])

detection using small atlases (sometimes only one image with no defect is enough to be able to find the defect afterward).

Saliency models are applied for defect detection on a wide variety of applications such as the semiconductor manufacturing and electronic production [14], metallic surfaces [15] or wafer defects [16], etc.

18.2.3.1 A Discussion

Defect detection and saliency modeling is a niche application field which is rapidly developing with recent references. As the image-based saliency models become reliable, these fields of application should grow in the next years.

18.2.4 Medical Imaging: Pathology Detection

Pathologies, such as tumors, might be considered also as defects with respect to the healthy tissues which are considered as normal. In [11], head and neck tumors are taken into account. The two main features characterizing those tumors are that (1) they are located close to the throat and (2) they induce an asymmetry in the neck tissues relative to the throat.

The first feature is taken into account by computing the log-polar image of the computed tomography scanner (CT scan) slices (Fig. 18.7b). The logarithmic approach gives much more importance to the areas which are located around the center point, which is here the throat.

The second feature is taken into account by computing for each gray level (reduced to only 16 in the paper) the ratio between the pixels on the right side and left side of the image. This provides, for each image, 16 coefficients of symmetry which are close to 0 if the gray level is symmetric or close to 1 or -1 if there is an asymmetry toward one or the other side. In Fig. 18.7c each column represents the symmetry coefficients for one slice: there are 16 values on the Y-axis, while the X-axis represents the number of slices in the CT scan volume.

A rarity-based attention approach is applied on each line to provide the result in Fig. 18.7d. Indeed if a gray-level symmetry is rare (unusual) in the context of the other slices of the CT scan volume, this means that the gray level is abnormally asymmetric for the given slice. Abnormal asymmetries are thus detected in Fig. 18.7d. This result, once projected on the X-axis, shows pics for the slices having a high probability of containing tumors (Fig. 18.7e).

As in previous sections, the use of a set of additional healthy slices called “atlas” can increase the efficiency of the algorithm, and it can provide a level of normality. An atlas is used in Fig. 18.7f at the left of the vertical line. This atlas provides a reference for a threshold of the final result (Fig. 18.7h). The approach in [11] is able to detect the slices in a CT scan which might contain tumors.

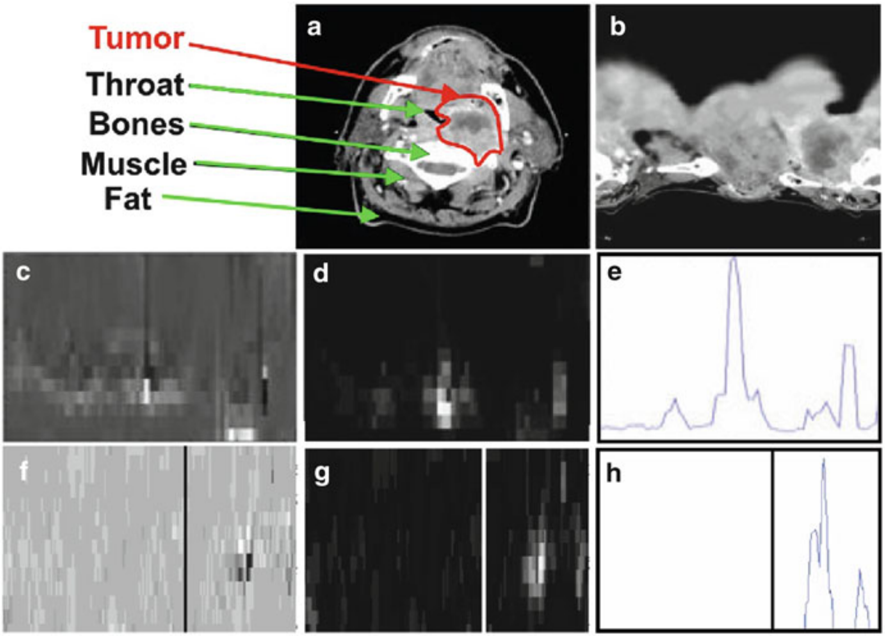


Fig. 18.7 (**a**) Initial CT SCAN slice example with annotated tumor and body parts. (**b**) Log-polar representation of image (**a**) centered on the throat. (**c**) *16 gray-level* symmetry coefficients (*Y-axis*) for all the slices into the 3D CT scan volume (*X-axis*). (**d**) Rarity-based attention computed on each line (a *gray-level* symmetry coefficient is rare if it is different from the coefficients of the same *gray level* in the other slices of the 3D CT scan volume). Clear coefficients are the ones which are unusually asymmetric. (**e**) Vertical projection of (**d**) which shows pics of abnormality for some slices which might contain tumors. (**f**) The same as (**c**) but using also an atlas with healthy CT scan slices (*left* of the *vertical line*). (**g**) Rarity-based attention computed on (**f**). (**h**) Slices possibly containing tumors after thresholding based on the maximum attention level in the atlas where no tumors are present (Adapted from [11])

Many other papers introduce saliency models as promising approaches in improving existing medical imaging techniques. In [17] a rarity-based approach is also used on magnetic resonance imaging (MRI) images. In [18], they use saliency to improve medical image registration. In [19] bright lesion detection and classification in color retinal images are based on saliency models. Several saliency models are tested on different image modalities in [20].

18.2.4.1 A Discussion

While some years ago the interaction between attention models and medical imaging was sparse, more and more publications make use of saliency models. With the improvement of saliency models for still images and with the arrival of 3D models of attention, there is a real development potential in the medical domain at middle term.

18.2.5 Expressive Gestures and Social Abilities Based on Saliency

Gestures are an important part of nonverbal communication. They are extensively used in robotics but also in the study of the expressiveness and communication or in human-computer interactions (HCI). Although there are few references in the domain, some papers used saliency models to investigate the role of gestures. In [21] and [22] close gestures are analyzed using dynamic saliency models to show how changes in gestures are more interesting than repetitive gestures. In [23] it is shown that the important moments in gestures which are detected by an attention model are close to what several users provided as manual annotation.

The computation of saliency which is common to several points of view and several images or videos can be used to exhibit the common interesting objects for several people/agents or robots [24]. Human gaze is also a very important social cue which will instinctively push others to gaze in the same direction. This joint attention and saliency modeling can be used together for robot-to-robot communication like in [25]. Moreover, saliency models can help in refining the estimation of a viewer gaze (Fig. 18.8) by proposing a set of salient areas close to the estimated gaze point [26, 27].

In robotics, several references also use saliency models to introduce the notion of gestures like in [28, 29]. The use of saliency models in human-robot interaction is evaluated in [30], and pointing gestures are related to saliency measures in [31]. Research on how to manage the point of focus of a robot using important objects and habituation is described in [32]. A set of interesting projects using attention models in robotics are also described in [33].

For avatar synthesis, a more natural behavior can be inferred by using attention models [34] which will direct the avatar attention on events which are of interest for humans. Other references can be found in this domain like [35] or [36].

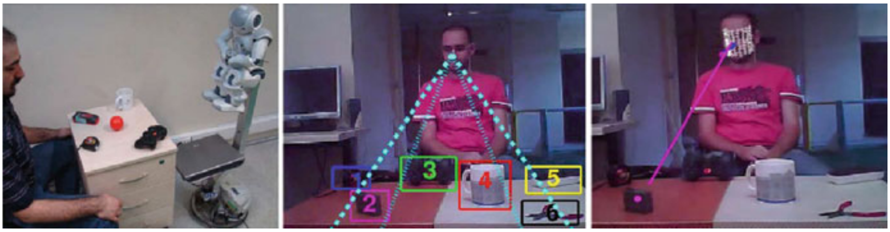


Fig. 18.8 *Left image*: experimental setup. *Middle image*: several possibilities from the face direction system and the important objects detected using a saliency models. *Right image*: closest salient object to the face direction gaze is selected (Adapted from [27])

18.2.5.1 A Discussion

Social signal processing in human-machine or human-human interaction is a growing research field. For the moment, the efforts of integrating saliency models are sparse and quite rare due to the fact that the community is not yet aware of what some models can bring and also to the fact that audiovisual and video models are not yet mature enough to act in very complex and dynamic scenes containing a lot of top-down information such as the social scenes. With the fact that attention is a filter which brings real-world signals toward conscience and awareness, social interaction extensively uses signals which aim to attract others' attention. Gesture saliency modeling, multimodal saliency, and co-saliency or joint attention are all very important points which should bring a lot to the field in the next years.

18.2.6 Attention-Based Computer Graphics

In computer graphics, saliency maps can help in rendering with less details the areas with lower saliency and with more details the areas which are more salient (Fig. 18.9). The idea is close to the one of compression but applied to rendering in computer graphics [37]. Other attention-based rendering techniques can be found in [38]. In addition to rendering, other computer graphics techniques as the meshes for 3D models can also be taken into account by saliency models like in [39].

Tone mapping can also take advantage from saliency models. Tone mapping is a technique for mapping a set of colors to another to approximate the appearance of high-dynamic-range (HDR) images in image which has a more limited dynamic range. In [40] they use visual attention for tone mapping on HDR images. In [41], the authors showed that tone mapping has a real influence on human perception which can be a disadvantage (e.g., in the case of compression) or an advantage (in the case of artistic or computer graphic applications).

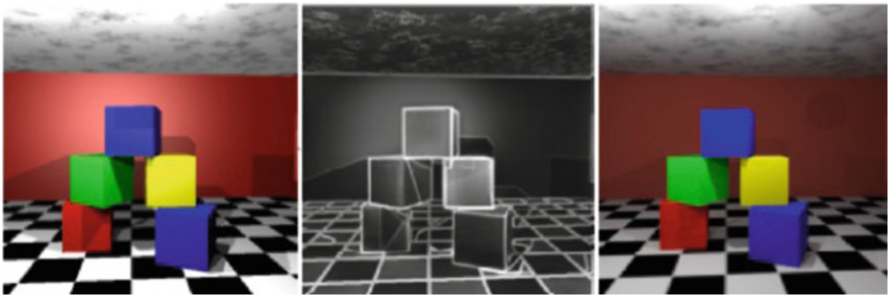


Fig. 18.9 *Left*: initial full rendering (see *shadows* and *lights* in the back). *Middle*: saliency map. *Right*: attention-based rendering (Adapted from [37])

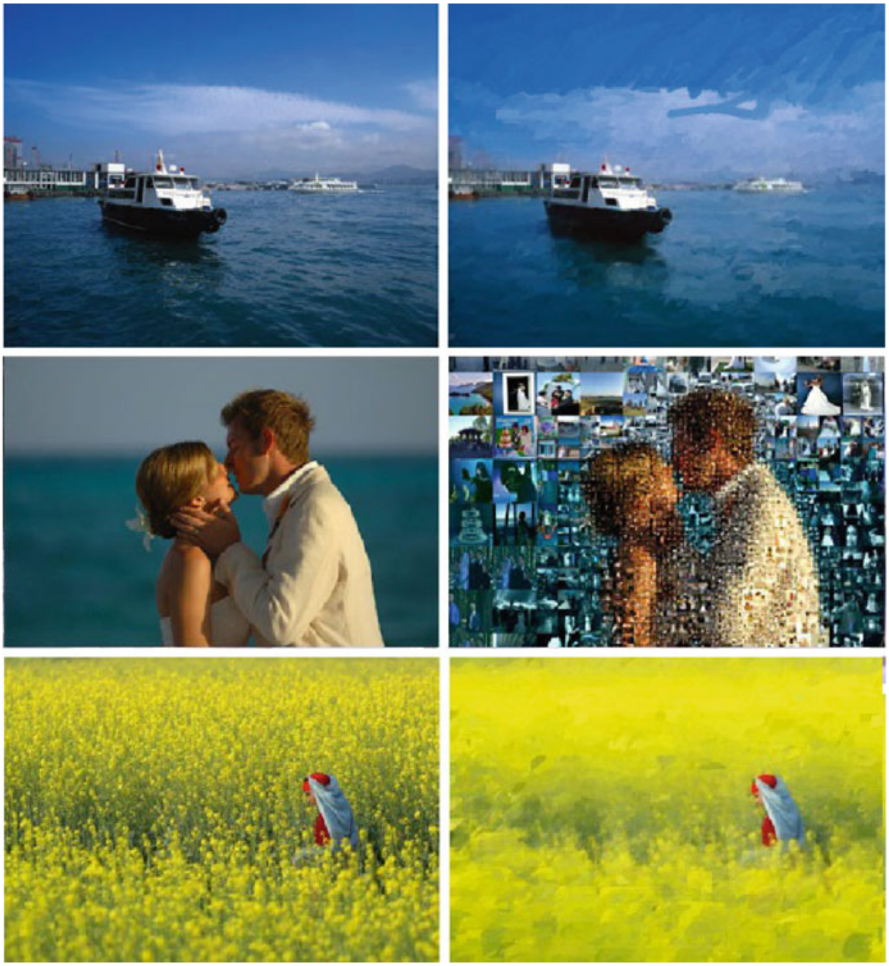


Fig. 18.10 *Left column*: initial images. *Right column*: attention-based artistic effect application (Adapted from [42])

In [42] several artistic effects are applied based on a saliency model. The authors show that the use of saliency maps helps the main objects to remain less altered and more visible (Fig. 18.10). Other references as [43] can be found in this domain which provides automatically interesting perceptually aware artistic effects.

Other papers related to attentive art deal with saliency-based aesthetics: [44], [45]. The saliency models take as input the segmented image and an order of importance of each segment. This input helps the algorithm to adapt the parameters of color, orientation and sharpness to change the image in order to stick to the proposed regions order. Those new images are considered as more aesthetic. In [46] the authors compute the image aesthetics based on saliency models.

18.2.6.1 A Discussion

The use of saliency models for artistic purposes or for computer graphics is a new axis of research, and one can find a lot of new references in this domain. This direction of research will probably grow especially in 3D rendering.

18.2.7 Attention-Based Quality Metric

Assessing the quality of an image or video sequence is a complex process, involving the visual perception as well as the visual attention. It is actually wrong to think that all areas of the picture or video sequence are accurately inspected during a quality assessment task. People preferentially and unconsciously focus on regions of interest. For these types of regions, our sensitivity to distortions might be significantly increased compared to non-salient regions. Even though we are aware of this, very few IQM (image quality metric) or VQM (video quality metric) approaches take this property into account. Therefore, it seems natural to use saliency maps to give more importance to distortion occurring on the salient part.

18.2.7.1 Saliency-Based Quality Metrics

For most of saliency-based metrics [47–51], the use of saliency map consists in modifying the pooling strategy. Quality metrics are composed of several stages. The last one is called the pooling which aims at computing the final quality score from a 2D distortion (or error) map. The degree of saliency of a given pixel can be used as a weight, giving more or less importance to the error occurring on this pixel location.

The difference between these methods concerns the way the weights are defined. As presented in Ninassi et al. [48], different methods to compute the weights can be used:

$$\begin{aligned}
w_0(x,y,t) &= 1 \\
w_1(x,y,t) &= SM_n(x,y,t) \\
w_2(x,y,t) &= 1 + SM_n(x,y,t) \\
w_3(x,y,t) &= SM(x,y,t) \\
w_4(x,y,t) &= 1 + SM(x,y,t) \\
w_5(x,y,t) &= SM_b(x,y,t) \\
w_6(x,y,t) &= 1 + SM_b(x,y,t)
\end{aligned} \tag{18.1}$$

where $SM(x, y, t)$ is the unnormalized human saliency map, $SM_n(x, y, t)$ is the human saliency map normalized in the range [0, 1], and $SM_b(x, y, t)$ is a binarized human saliency map. The weighting function w_0 is the baseline quality metrics in which the pooling is not modified. The functions w_1, w_3, and w_5 give more importance to the salient areas than the others. Indeed, the offset value of 1 in the weighting functions w_2, w_4, and w_6 allows us to take into account distortions appearing also on the non-salient areas.

The use of saliency map in the pooling stage provides contrasted results. In [48], the use of saliency map does not improve the performance of the quality metric. On the other hand, Akamine and Farias [51] showed that the performance of very simple metrics (PSNR and MSE) has been improved by the use of saliency information. However, for the SSIM metric [52], the saliency does not allow to improve the metric performance. In addition, they showed that the performance improvement depends both on the saliency model used to generate the saliency map and on the distortion type (white noise, JPEG distortions). More details on video quality and saliency can be found in Chap. 20.

18.2.7.2 Quality in Stereoscopic 3D Images

The conflicting vergence and accommodation cues are widely accepted to be a main cause of visual discomfort in stereoscopic viewing [53]. In addition, fast salient object motion has also been proposed as a cause for viewing discomfort [54]. In both cases, the use of efficient saliency maps is very useful in improving the viewing comfort.

Attention models have been first used to assess the viewing discomfort as in [55–57]. In this case, saliency maps are compared with the disparity (depth) maps to provide objective metrics for discomfort. A second use of saliency models is in the enhancement of the viewer comfort as in [58, 59] where the blurring of the image is done according to the saliency of the areas.

18.2.7.3 A Discussion

Many authors working in this field consider that visual attention is important for assessing the visual quality of images. However, there are still a number of open issues as demonstrated by [48, 51]. New strategies to incorporate visual attention into quality metrics as well as a better understanding of the interactions between saliency and distortion need to be addressed.

The development of the research in 3D stereoscopic viewing comfort opens a new research avenue to saliency models in the near future.

18.3 Applications Based on Normality Detection

In this section we focus on the second category of applications from our taxonomy: applications using the prior knowledge of the locations having the lowest saliency scores. Those areas correspond with repeating and less informative regions; thus, they can be easily compressed or cropped, for example (Fig. 18.11).

18.3.1 Attention-Based Texture

In [60], the authors show that saliency models, which are based on the global information of the image, can be related directly to the homogeneity of the texture. The more the image is complex and unique, the less the saliency is high. The natural tendency of saliency models to only focus on important signal while discarding the usual and repetitive one is very useful in the case of image texture. Indeed, low saliency is a synonym of highly homogeneous textures or colors. More recently, other authors proposed to use an attention model as a regularity metric for textures [61, 62].

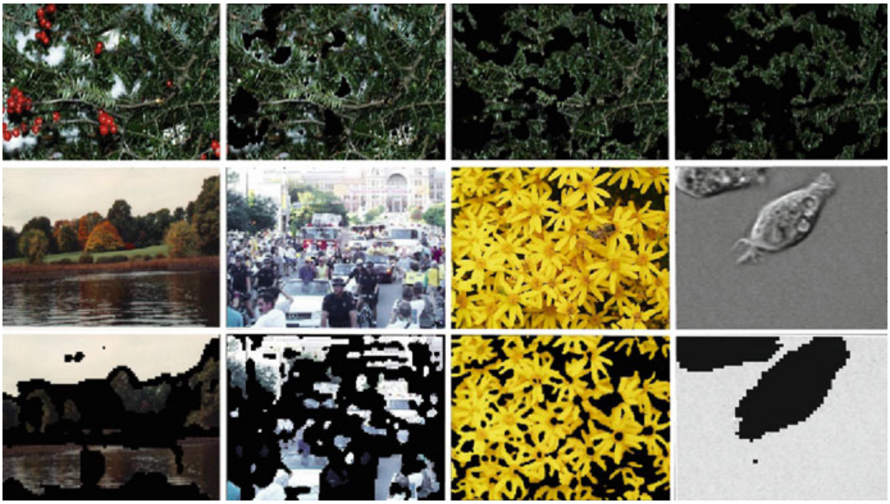


Fig. 18.11 *First line* from left to *right*: initial image, the most salient eliminated, the medium salient regions eliminated, and only the less salient regions remain-the texture is more and more homogeneous. *Second line*: four examples of images. *Third line*: medium salient regions eliminated-texture regions detected (Adapted from [60])

18.3.1.1 A Discussion

Very few papers deal with saliency models and texture regularity, even if this is a promising research field. Recent publications should make the texture segmentation community more aware about the potential of saliency models in texture segmentation and feature extraction. One of the applications of texture regularity detection is in image compression which will be further discussed in the following section.

18.3.2 Attention-Based Compression

Video compression is the process of converting a signal into a format that takes up less storage space or transmission bandwidth. It can thus be considered as a coding scheme that reduces bits of information representing the original signal (audio, images, videos).

Since the late 1990s techniques based on attention have been introduced in the field of image and video coding [63, 64]. Attention can be used to select the less interesting areas in images or videos and compress them or to transmit the most salient parts first during the data transfer from a server to a client.

The classical compression methods tend to distribute the coding resources evenly in an image. On the contrary, attention-based methods encode visually salient regions with high priority while treating less interesting regions with low priority (Fig. 18.12). The aim of these methods is to achieve compression without significant degradation of perceived quality.

Although there is currently no unified taxonomy, we have divided the methods into indirect and direct methods, the latter being the most commonly studied.

A first family of compression methods can be called "interactive." Early approaches relied on eye-tracking devices to monitor human attention focus [64].

With such devices which are able to follow the focus of the observer, encoding continuously and efficiently the images is natural. Indeed, observers usually do not even notice any degradation of the frames they watch. However, these techniques are neither practical (because of the use of the eye-tracking device) nor general (because they are restricted to a single viewer).

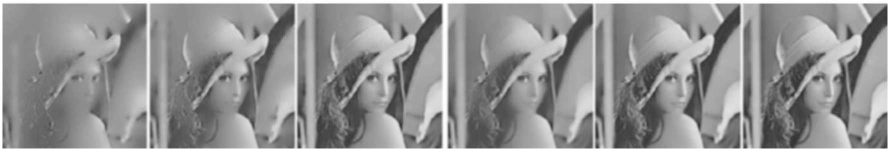


Fig. 18.12 Distortions introduced by general compression methods (*three first images* on the *left*) compared to saliency-based compression (*three last images* on the *right*), at three different compression levels (Adapted from [65])

Attempts to automatize this approach by using attention-based methods are very complex as top-down information is very important, and if clear salient objects are not present in a frame, people gaze can be very different. Even in the case where progresses in attention modeling are achieved, it is not possible to have a reliable model of human gaze in case where there is no specific salient object in the frame and where the viewers' gaze has naturally a very high dispersion.

18.3.2.1 Indirect Approaches

Indirect compression consists of modifying the source image to be coded while keeping the same coding scheme. Such methods are thus generally driven by a saliency map-based methods.

In [66], a saliency map for each frame of a video sequence is computed and a smoothing filter is applied to all non-salient regions. Smoothing leads to higher spatial correlation, a better prediction efficiency of the encoder, and therefore a reduced bit rate of the encoded video.

Another method combines both top-down and bottom-up information, using a wavelet decomposition for multiscale analysis [67]. Bit rate gains ranging from 15 % to 64 % for MPEG-1 videos and from 10.4 % to 28.3 % for MPEG-4 are reported.

An indirect approach based on their attention model is proposed by [68]. An anisotropic prefiltering of the images or frames is achieved keeping highly salient regions with a good resolution while low-pass filtering the regions with less important details (Fig. 18.13).

In [69], the depth based on the level of blur of the regions in an image is also taken into account: closer areas should be thus less compressed than objects which might be located far from the camera.

The main advantage of indirect approaches is that they are easy to set up because the coding scheme remains the same. The intelligence of the algorithm is applied as a preprocessing step, while standard coding algorithms are used afterward. This fact also let people to easily quantify the gain in terms of compression as the main compression algorithm can be used directly on the image or on the saliency preprocessed image. However, one possible problem is that the degradation of less salient zones can become strong. Selective blurring can lead to artifacts and distortions in low-saliency regions [70].

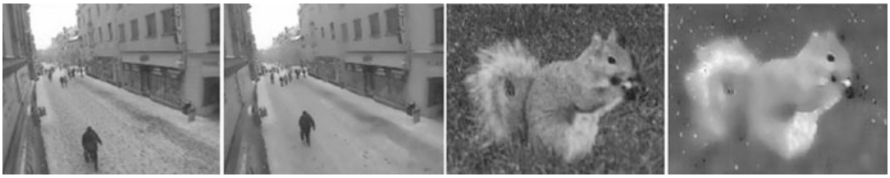


Fig. 18.13 Two pairs of images (original and anisotropic filtered) (Adapted from [68])

18.3.2.2 Direct Approaches

Recent works on modeling visual attention (Le Meur, Itti, Parkhurst, Chauvin, etc.) paved the way to efficient compression applications that modify the heart of the coding scheme to enhance the perceived quality. In this case some modifications to the saliency map are generally necessary to dedicate it directly to coding. The saliency maps will not only be used in the preprocessing step but also in the entire compression algorithm.

An extension of [66] uses a similar neurobiological model of visual attention to generate a saliency map [70]. The most salient locations are used to generate a so-called guidance map. The latter is used to guide the bit allocation through quantization parameter (QP) tuning by constrained global optimization. Considering its efficiency at achieving compression while preserving visual quality and the general nature of the algorithm, the authors suggest that it might be integrated in general-purpose video codecs. Future work in this direction should include a study of possible artifacts in the low-bit rate regions of the compressed video, which may themselves become salient and attract human attention. Another possible issue pointed out in [70] is that the attention model does not always predict accurately where people look at. For example, high-speed motion increases saliency, but regions with lower motion can attract more attention (e.g., a person running on the sidewalk, while cars are going faster).

Other approaches with lower computational complexity have been investigated, and in particular two methods using the spectrum of the images: the spectral residual [71] and the phase spectrum of quaternion Fourier transform [72]. The goal of both approaches is to suppress spectral elements corresponding to frequently occurring features.

The phase spectrum of quaternion Fourier transform (PQFT) is an extension of the phase spectrum of Fourier transform (PFT) to quaternions incorporating interframe motion. The latter method derives from the property of the Fourier transform that the phase information specifies the location each of the sinusoidal components resides within the image. Thus, the locations with less periodicity or less homogeneity in an image create the so-called proto-objects in the reconstruction of the image's phase spectrum, which indicates where the object candidates are located. A multi-resolution wavelet foveation filter suppressing coefficients corresponding to the background is then applied.

These Fourier-based approaches have two main drawbacks linked to the properties of the Fourier transform. First, if an object occupied most of the image, only its boundaries will be detected, unless resampling is used (at the expense of a blurring of the boundaries). Second, an image with a smooth object in front of a textured background will lead to the background being detected (saliency reversal).

Using the bit allocation model of [70], a scheme for attention video compression has been suggested by [73]. This method is based on learning feature integration algorithm, with a relevance vector machine architecture, incorporating visual saliency propagation (using motion vectors), to save computational time. This architecture is based on thresholding of mutual information between successive frames for flagging frames requiring recomputation of saliency.

Recently, attention-based image compression patents like [74] has been accepted, which also show that compression algorithms are more and more efficient in real-life applications and become close to reach the market.

18.3.2.3 A Discussion

For images and videos, the expectations from the saliency models were very high. In the first step not all these expectations were met. As the saliency models are not perfect and the classical compression already includes some cognitive elements, the compression factor given the information quality decrease is not optimal. Current saliency-based compression algorithms are mainly suitable for applications where the too high compression of some areas (which creates artifacts catching human attention) is not an issue like in video surveillance. Indeed, in video surveillance the perceived quality of background regions is not important if the foreground is not degraded. However, current work shows an enhancement of the techniques from which some become close to market as recent patents like [74] demonstrate.

Future developments in the direction of 3D compression seem very interesting, and new research avenues should be shortly opened in that direction. Indeed, a simple MS Kinect One device records RGB, depth, and infrared images at almost two gigabytes per second. Devices able to provide 3D images or point clouds all need efficient ways of compression to cope with the huge amount of data they deliver.

18.3.3 Attention-Based Retargeting

Compression aims in reducing the amount of data in a signal. A usual approach consists in modifying the coding rate, but other approaches can also reduce the amount of data in the signal by cropping or resizing the signal. An obvious idea which drastically compresses an image is of course to decrease its size. This size decrease can be brutal (zoom on a region and the rest of the image is discarded) or softer (the resolution of the context of the region of interest is decreased but not fully discarded). The first approach will of course be more efficient from a compression point of view, but it will fully discard the context of the regions of interest which can be disturbing.

The direct image cropping will be called here "perceptual zoom," while the second approach which still keeps some context information around the region of interest will be called "anisotropic resolution." Both approaches provide image retargeting. Retargeting is the process of resizing images while minimizing visual distortion and keeping at best the salient content.

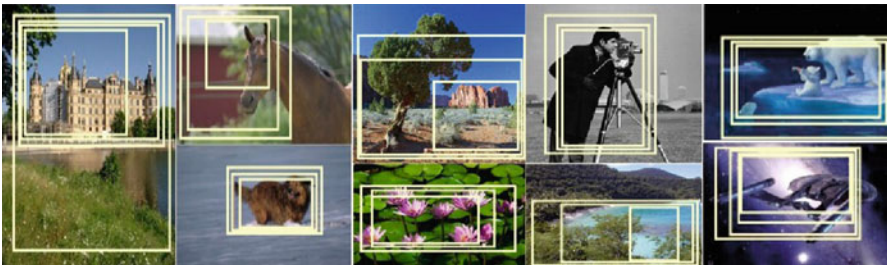


Fig. 18.14 Examples of images along with *rectangles* providing different attention-based automatic zooms. After a saliency map [76] is computed and low-pass filtered, several threshold values are used to extract the bounding boxes of the more interesting areas. Depending on this threshold, the zoom is more or less precise/important

18.3.3.1 Spatiotemporal Visual Data Repurposing: Perceptual Zoom

Human beings are naturally able to perceive interesting areas of an image. Zooming in images should therefore focus on such regions of interest.

Image manipulation programs provide tools to manually draw these rectangles of interest, but the process can be automated with the help of attention algorithms. Interestingly, such techniques can also be used for real-time spatiotemporal image broadcast [75].

Figure 18.14 shows several perceptual zooms depending on a parameter which will threshold the smoothed saliency map from [76].

The authors in [77] use Itti algorithm to compute the saliency map [78] that serves as a basis to automatically delineate a rectangular cropping window. A fast greedy algorithm was developed to optimize the window that has to encompass most of the saliency while remaining sufficiently small.

The self-adaptive image cropping for small displays [79] is based on an Itti and Koch bottom-up attention algorithm but also on top-down considerations as face detection, skin color, etc. According to a given threshold, the region is either kept or eliminated.

In [80], the authors start by segmenting the image into several regions, for which saliency is calculated to provide a global saliency map. The regions are classified according to their attractiveness, which allows to present image regions on small-size screens and to browse in big-size images.

A completely automatic solution to create thumbnails according to the saliency distribution or the cover rate is presented by [81]. The size of the thumbnail can be fixed and centered on the saliency map global maximum or adapted to certain parameters such as the saliency distribution. The gaze fixation predicted by a winner-take-all algorithm can thus be used, and the search for the thumbnail location ends when a given percentage of the total image saliency is reached.

An algorithm proposed in [82] starts by adaptively partitioning the input image into a number of strips according to the combined saliency map, which contains both gradient information and visual saliency to measure significant regions and is also used to guide the sampling process when scaling image strips.

A video retargeting method based on a spatiotemporal saliency model is described in [83]. Based on a spatiotemporal saliency map, a salient object detection method is used to locate salient object regions in the video. Finally, cropping and uniform scaling operations are performed on the basis of salient object regions to generate the retargeted video.

A hybrid framework of video retargeting with a domain-enhanced spatiotemporal grid optimization can be found in [84]. First, they combine visual attention with higher-level features. Second, they build a semantic importance map representing the spatial importance and temporal continuity, which is incorporated with a 3D rectilinear grid scale plate to map frames to a target display, thereby keeping the aspect ratio of semantically salient objects as well as the perceptual coherency.

The methods of intelligent perceptual zooming based on saliency algorithms become more and more interesting with the advances in saliency map computation in terms of both real-time and spatiotemporal cue integration. Even big companies as Google [85] become more and more involved in developing applications based on perceptual zooms. The idea is to generalize the perceptual zoom for images and videos and keep the temporal coherence of the zoomed image even in case of objects of interest which might brutally appear in the image far from the previous zoom area.

18.3.3.2 Spatiotemporal Resolution Decrease for Uninteresting Regions: Anisotropic Resolution

Perceptual zoom does not always preserve the image structure. For example, Fig. 18.14 shows that the smallest zoom on the left image only comprises part of the castle, which is likely to attract attention. In this case the zoom loses the structure and context of the original image. To keep the image structure when retargeting, two methods are described in this section: warping and seam carving. These methods may cause nonlinear visual distortions on several regions of the image [86], but they provide enough contextual information to let the viewer understand the main structures. When adapted to videos, those techniques are also easier to stabilize as the context is more present than for the perceptual zoom.

Warping

Warping is an operation that maps a position in a source image to a position in a target image by a spatial transformation. This transformation could be a simple scaling transformation [87].

Nonhomogeneous content-driven video retargeting [88] proposes a real-time retargeting algorithm for videos. Spatial saliency, face detection, and motion detection are computed to provide a saliency matrix. An optimized mapping is computed with a sparse linear system of equations which takes into account some constraints such as importance modeling, boundary substitutions, and spatial and time continuity.

A retargeting method based on global energy optimization is detailed in [89]. Some content-aware methods only preserve high-energy pixels, which only achieve local optimization. They calculate an energy map which depends on the static saliency and face detection. The optimal new size of each pixel is computed by linear programming.

The same group proposes a retargeting approach that combines a uniform sampling and a structure-aware image representation [90]. The image is decomposed with a curve-edge grid, which is determined by using a carving graph such that each image pixel corresponds to a vertex in the graph. A weight is assigned to each vertex connection (only vertical direction) which depends on an energy map using saliency region and face detection. The paths with high-connection weight sums in the graph are selected, and the target image is generated by uniformly sampling the pixels within the grids.

A warping method which uses the grid mesh of quads to retarget the images (Fig. 18.15) is defined in [91]. The method determines an optimal scaling factor for regions with high content importance as well as for regions with homogeneous content which will be distorted. A significance map is computed based on the product of the gradient and the saliency measure which characterizes the visual attractiveness of each pixel. The regions are deformed according to the significance map. A global optimizing process is used repetitively to minimize the quad deformation and grid bending.

Another approach is a patch-based retargeting scheme [92] with an extended significance measurement to preserve shapes of both visually salient objects and structure lines while minimizing visual distortions. In the proposed scheme, a similarity transformation constraint is used to force visually salient contents to undergo as-rigid-as-possible deformation, while an optimization process is performed to smoothly propagate distortions. These processes enable to yield more pleasing content-aware warping and retargeting.

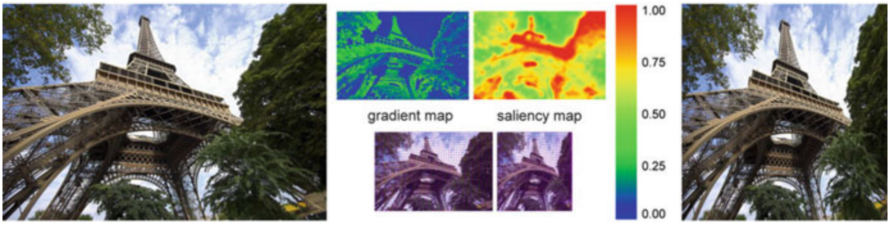


Fig. 18.15 The original image (*left*) is deformed by a grid mesh structure to be fit in the required size (*right*). The scaling and stretching depend on the gradient and saliency map (Adapted from http://graphics.csie.ncku.edu.tw/Image_Resizing/)

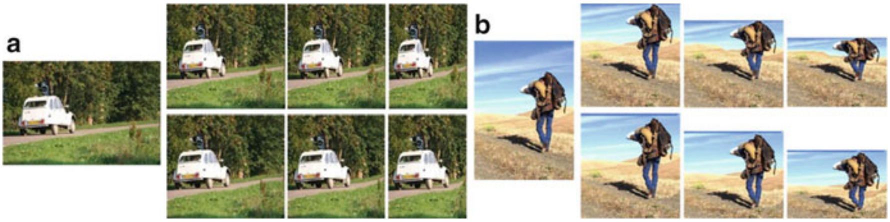


Fig. 18.16 The original images (*A* and *B*) and for each one seam removal (*vertical* seams for *A* and *horizontal* seams for *B*) using gradient (*top row*) and using a saliency map (*bottom row*) (Adapted from http://cilabs.kaist.ac.kr)

Seam Carving

Seam carving [93] allows to retarget the image, thanks to an energy function which defines the pixels' importance. The most classical energy function is the gradient map, but other functions can be used such as entropy, histograms of oriented gradients, or saliency maps [94]. Low-energy pixels are connected together to make a seam path. The seam paths cross vertically and horizontally the image and are removed. Dynamic programming is used to calculate the optimal seams. The image is readjusted by shifting pixels to compensate the disappeared seams. The process is repeated as often as required to reach the expected sizes.

Figure 18.16 shows an example of seam carving: the original images (A and B) are reduced either by discarding vertical or horizontal seams. On the top row, the classical gradient is used as the energy map, while saliency maps of [95] are used for the bottom row. Depending on the energy map, shapes as well as aspect ratio distortions can cause anisotropic stretching [75]. Even if saliency maps most of the time work better than a simple gradient, they are not perfect, and the results can be very different depending on the method used.

For spatiotemporal images, [96] propose to remove 2D seam manifolds from 3D space-time volumes by replacing a dynamic programming method with graph cut optimization to find the optimal seams. A forward energy criterion is presented which improves the visual quality of the retargeted images. Indeed, the seam-carving method removes the seams with the least amount of energy and might introduce energy into the images due to previously nonadjacent neighbors becoming neighbors. The optimal seam is the one which introduces a minimum amount of energy.

A saliency-based spatiotemporal seam-carving approach with much better spatiotemporal continuity than [96] is proposed by [97]. The spatial saliency maps are computed on each frame, but they are averaged over and history of frames in order to smooth the maps from a temporal point of view. Moreover, the seams are temporally discontinuous providing only the appearance of a continuous seam which helps in keeping both spatial and temporal coherence.

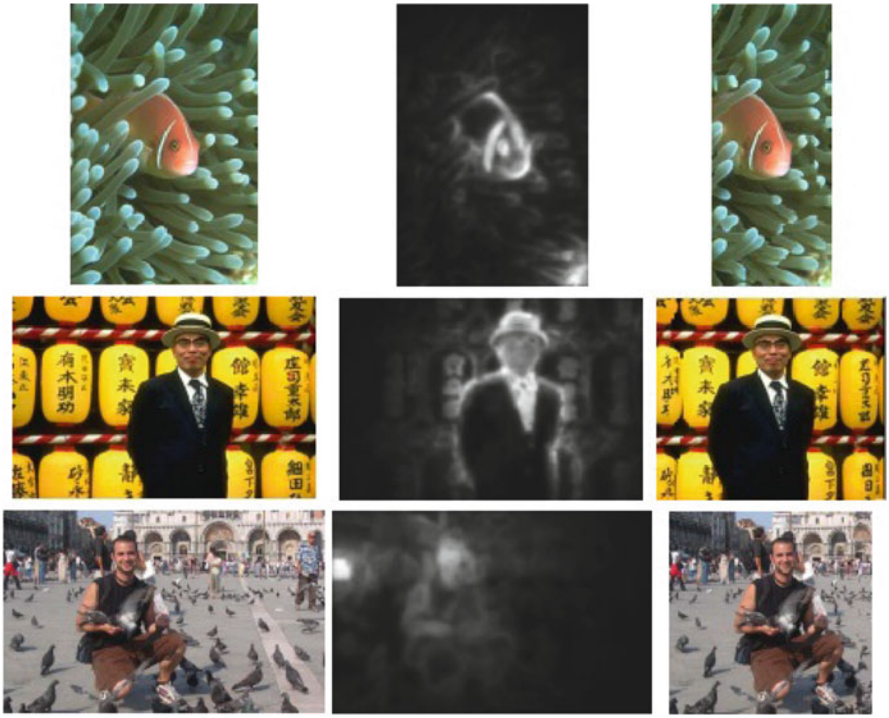


Fig. 18.17 *Left*: original images. *Middle*: saliency maps. *Right*: retargeted images (Adapted from [98])

In [98], the authors describe a saliency map which takes more into account the context and proposes to apply it to seam carving. The idea leads to good results as shown in Fig. 18.17.

In [99] the authors used attention algorithms for video retargeting based on seam carving. An efficient spatiotemporal grouping is done to determine the temporal rate of reduction depending on the content, to suppress groups of isolated seams, to identify spatiotemporal groups of seams, and to approximate by constant segments the number of seams for each group while keeping the total sum of seams constant. Problems of geometric distortion, anachronism, and length of summary have been also addressed.

Interestingly, recent papers as [100] propose to mix seam-carving and warping techniques. Firstly, based on the importance partition with the saliency map, they apply a weighted seam-carving approach to make the seams distributed dispersedly in the important regions. Then they propose content-aware image distance (CAID) to assess the deformation caused by removing seams. The weighted seam carving will stop with a fixed threshold to guarantee little visual image quality degradation. Finally, the grid-based warping is utilized to achieve the final size with a global optimization model, since warping tends to avoid discontinuity artifacts of an important region and typically make the distortion distribution of unimportant region more coherently.

18.3.3.3 Attention-Based Summarization

Summarization of images or videos is a term which is similar to retargeting. It might be based on cropping (closer to the first retargeting family) [101]. It might also be closer to the second family based on carving as in [102]. The main purpose is to provide a relevant summary of a video or an image.

In [103] the authors used video summarization to provide a mash-up of several videos into a unique pleasant video containing the important sequences of all the concatenated videos. This approach shows the possible extension of the notion of summarization from a single image or video document to a whole archive of documents. This application has common points with Sect. 18.2.6 and image mosaics. In [104] the authors proposed to make an intelligent collage based on saliency maps (Fig. 18.18). This approach also led to a patent [105] on this topic.

18.3.3.4 A Discussion

While the use of saliency maps for classical compression does not bring the expected improvements when using the nowadays state-of-the-art models, the retargeting methods (perceptual zooms, warping, or seam carving) can benefit a lot from saliency methods. Automatic attention computation is based on the use of context (contrast, rarity, surprise in a given spatial and/or temporal context). These models can highly improve retargeting methods and preserve objects of interest while

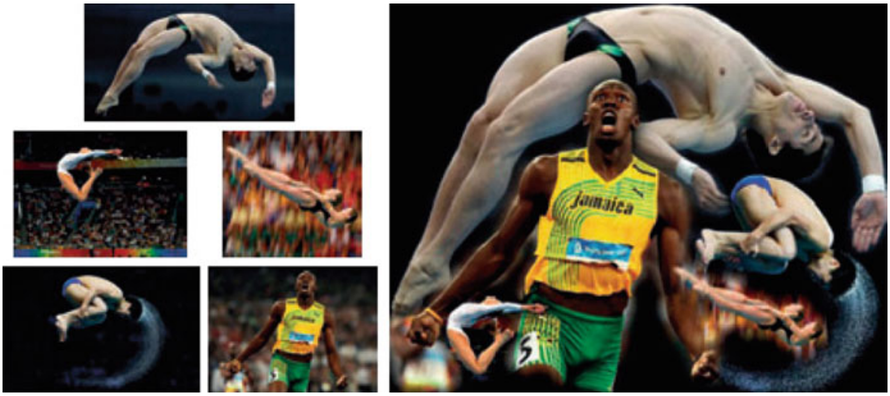


Fig. 18.18 *Left*: images to be summarized. *Right*: final attention-based collage (Adapted from [104])

also keeping the minimum of context information. Industrial applications begin to rise with the enhancement of the saliency models both in terms of accuracy and computational efficiency.

18.3.4 Watermarking and Security

Watermarking consists of hiding data in an image with a minimal visual altering of this image.

An idea is to hide data in the most interesting areas of the image which are computed based on a saliency model [106] to get more robust watermarks. Indeed, the high frequencies of the watermark are less easy to notice if hidden within other high-frequency areas which is generally the case for salient regions.

Another idea is, on the opposite, to hide data in the less salient regions as those regions have a lower probability to be noticed [107]. This assumption is true if the background is cluttered (grass, trees, complex buildings, etc.) as watermarks are easier to hide in high-frequency areas. The saliency-based watermarking is capable of hiding lower injected-watermark energy onto more sensitive regions and higher energy onto the less perceptually significant regions in the image [108]. The use of saliency models helps to get better visual quality of the watermarked image and an improved robustness of the watermark.

18.3.4.1 A Discussion

Watermarking only uses the saliency model as a filter to select the areas where information should be inserted. They could be inserted depending on the image entropy in the most salient or less salient areas.

18.3.5 Attention-Based Advertising Insertion

In [109] one can find an interesting description of attention-based advertising insertion. The first approach is called linear advertising, while the second is called nonlinear.

Linear advertising will insert content-related ad clips into less intrusive temporal positions. In [110], the authors propose a two-step approach. The first step aims in selecting an ad which is related to the current content. In the paper this was done using text mining. The second-step goal is to find the moment which has low spatiotemporal saliency to insert the ad in a less intrusive way. Figure 18.19 shows the main scheme of the system.

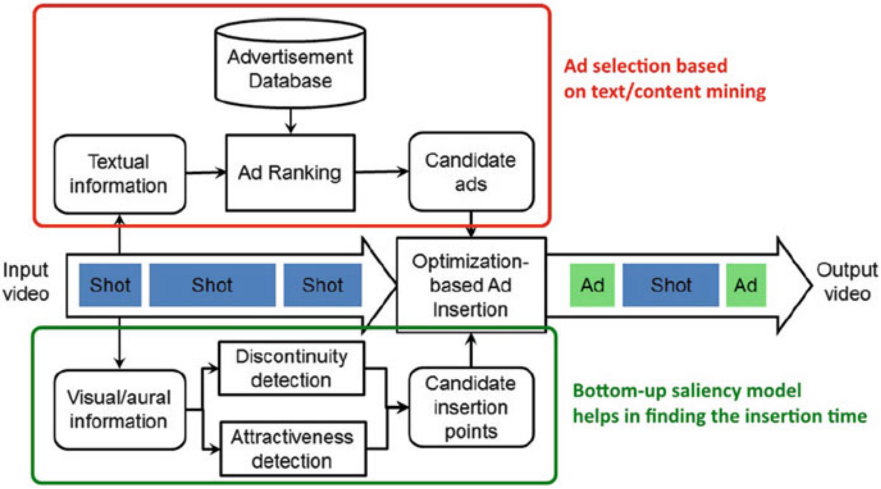


Fig. 18.19 Linear ad insertion at the less salient moments (Adapted from [110])

Another approach is the nonlinear one [111]. In this case, there are also two steps. The first consists in finding the right location in the frame where the ad should be inserted. This step uses the saliency map to locate an area close to the most interesting one. The second step will produce color harmonization of the ad to be less intrusive when projected onto the frame. Finally, the harmonized ad is projected close to the most interesting area. This will lead to a very noticeable ad (close to salient regions), while the color harmonization reduces its intrusiveness.

18.3.5.1 A Discussion

The way an ad is shown to viewers is of utmost importance in the way they perceive and remember the message. The linear and nonlinear approaches might find their way in the audiovisual production and broadcasting if the attention models become more efficient and the systems can really be real time.

18.4 Applications Based on Abnormality Processing

The third category of our attention-based application taxonomy concerns abnormality processing. Some applications go further than the use of the simple detection of the areas of interest. They use comparisons between the areas, relative positioning, and other operations on the saliency maps. Application domains such as robotics or advertisement highly benefit from this category of applications.

18.4.1 Attention-Based Robotics, Object Recognition, and Registration

Robotics is a very large domain of application with various needs. As robotics aims at mimicking human reactions, the field aggregates several techniques of which some can be used in other domains as robotics. We describe here rapidly three research axes where robots can take advantage from saliency models: (1) image registration and landmark extraction, (2) object recognition, and (3) robot action guidance. We only provide a rapid view about those research actions here as they are explained more in detail in “Chap. 21.”

18.4.1.1 Image Registration and Landmarks

An important need of a robot is to know where it is located. For this aim, the robot can use the data from its sensors to find landmarks (salient feature extraction) and register images taken at different times (salient feature comparison) to build a model of the scene. The general process of real-time building of a view of the scene is called simultaneous localization and mapping (SLAM). The use of RGB cameras using or not the depth information is called visual SLAM. Saliency models can help a lot in the extraction of more stable landmarks from images which can be more robustly compared [112]. A detailed review of attentive SLAM can be found in “Chap. 21.”

Saliency maps are also used in other domains as for medical image registration [18], lunar images, and crater impact detection [113] or on 3D object registration [114].

All those techniques imply first the computation of saliency maps, but the results are not used directly: they need to be further processed (like extraction of regions of interest and their comparison).

18.4.1.2 Object Recognition

Another important need of robots after they establish the scene is to recognize the objects which are present in this scene and which might be interesting to interact with. To recognize objects two steps are needed. First of all, the robot needs to detect the object in a scene. For this goal saliency models can help a lot as they can provide information about proto-objects [115] or area objectness [116]. For more details on proto-object and object detection, see Chap. 15.

Once objects are detected, they need to be recognized. In this area the main approach is to (1) extract features (SIFT, SURF, or any others) from the object, (2) filter the features based on a saliency map, and (3) perform the recognition based on a classifier (such as a SVM or others). Papers like [117] or [118] apply this technique which let a computer drastically decrease the number of needed key points to perform the object recognition. Further details can be found in Chap. 19 which is focusing on this approach.

Another approach was used in [119] or [120]. Here the features which are mostly present in the searched object and not present in the surroundings are learned, and this learning phase provides a new set of weights for bottom-up attention models. In this way, the features which are the most discriminant in the searched object will get the higher response in the final saliency map. The bottom-up model is in that way tuned by top-down information on the discriminant features learned from the searched object.

A third approach can be found in [121] where relative position of salient points (called cliques) is used for image recognition. More details on this approach can be found in the Chap. 8.

18.4.1.3 Action Guidance

Once robots know where they are (attentive visual SLAM) and they also recognize objects around them (attentive object recognition), they need to decide what to do next. One of the decisions they need to make is to know where to look next, and this decision is obviously taken based on visual attention. Several robots implement multimodal attention like the iCub robot in Fig. 18.20. They combine visual and audio saliency in an ego-sphere, and this is used to point the gaze on the next location. More details about robots and gaze can be found in Chap. 21 and an interesting survey on attention for interactive robots can be found in [122].

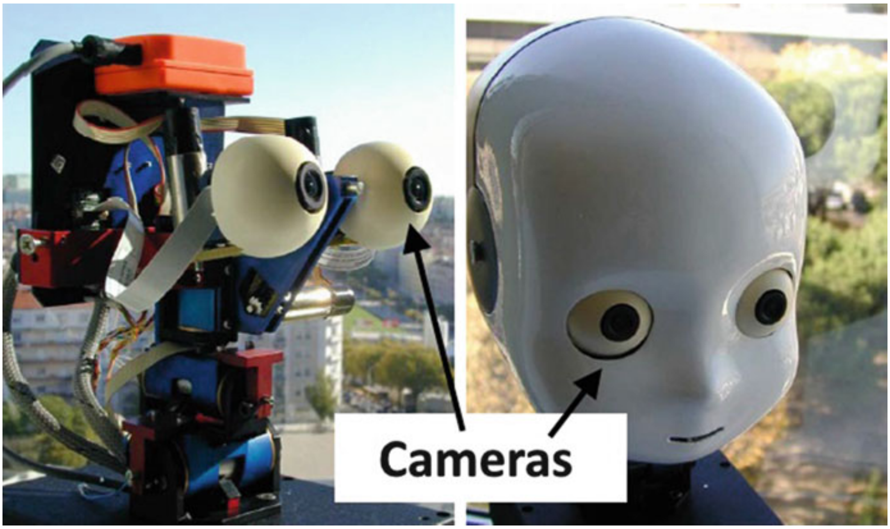


Fig. 18.20 iCUB robot head. The robot implements a multimodal saliency system (Adapted from [123])

The social interactions also include gestures as the pointing gestures which are important top-down factors. The use of gesture direction is used in [29, 124] to detect the object of utmost interest and to learn to the system where to look.

Robots are embodied agents, but other agents like virtual agents can implement attention models [34]. In [125] an attentive system based on high-level features (people skeleton extracted using an RGB-D camera) is described. More details on this approach can be found in "Chap. 17."

18.4.1.4 A Discussion

Robotics, especially humanoid robotics, is a very complete field of research. Even if we restrain the domain to electrical engineering and computer science, there is still an impressive list of topics necessary to build a convincing robot. Here we focused on the use of attention models in robots and related fields which give us the three main axes of research. The advances in those topics are huge, but still it is difficult to have a realistic social robot capable of naturally interacting and adapting to novel unusual situations. This is one of the big challenges to which attention modeling might bring a solution in the future years.

18.4.2 Attention-Based Marketing and Communication Optimization

Marketing optimization can be applied to a large amount of practical cases such as Web sites, advertisement, product placement in supermarkets, signages, 2D and 3D object placement in galleries, etc. All these application cases can benefit from attention maps themselves but also from regions of interest comparison and further analysis of the attention maps. Moreover, attention only tells if people will notice the important message in an ad, but not if they remember it. Thus, the "memorability" of an image is an important topic where attention models can help.

This section is structured in three subsections: we investigate the use of saliency maps in (1) Web sites and ad optimization, (2) image memorability, and (3) 3D object best viewpoint calculation.

18.4.2.1 Attention-Based Web Sites and Advertisement Optimization

Among the different applications of automatic saliency computation, the marketing and communication optimization is probably one of the closest to market. As it is possible to predict an image attention map, which is a map of the probability that people attend each pixel of the image, it is possible to predict where people are likely to look on a marketing material like an advertisement or a Web site.

Attracting customer attention is the first step of the process of attracting people interest, inducing desire and need for the product, and finally pushing the client to action as described in the AIDA pyramid [126].

It is important to stress the fact that attention alone is not enough to push a potential client to action, but at least it is a key step toward this goal.

There are already two main techniques which are able to provide information about people attention on marketing material. The first one uses eye-tracking studies on marketing material like in [127]. This approach is very accurate as the precise gaze location of the users on a Web site/advertisement is measured. The drawbacks of this approach are in the time needed to conduct the study, the price, and the fact that only finished or almost finished documents can be tested. Another drawback is that long fixations do not mean that this area is necessarily very salient: it might only mean that it is difficult to understand, and people spend a lot of time in trying to figure out what this area is about.

Figure 18.21 shows an example of a heat map computed from the average eye-tracking data from several users on a Web site page.

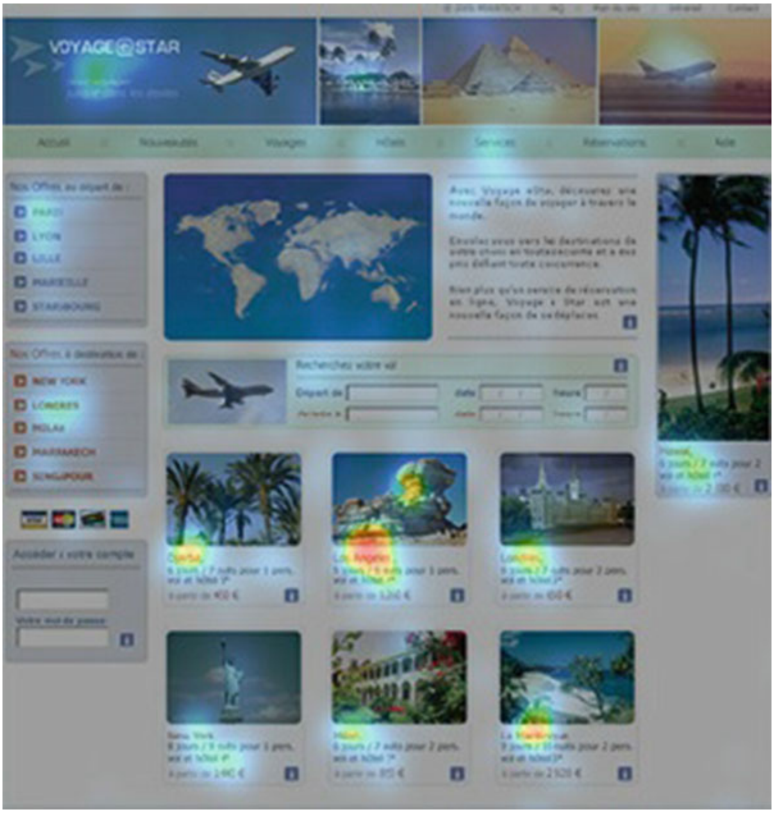
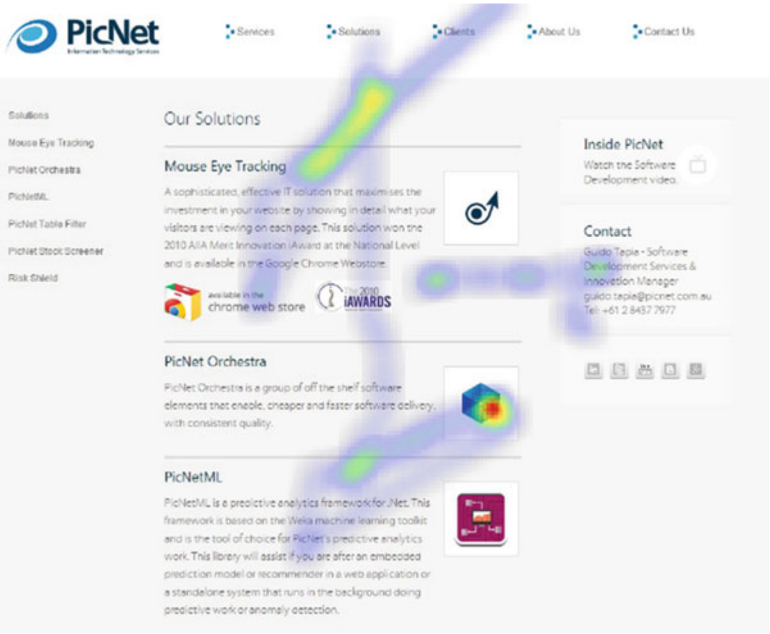


Fig. 18.21 Eye tracking on a Web site: gaze heat map overlay on the initial image from Miratech (Extracted from [127])



Our Solutions

Mouse Eye Tracking

A sophisticated, effective IT solution that maximises the investment in your website by showing in detail what your visitors are viewing on each page. This solution won the 2010 AIIA Merit Innovation iAward at the National Level and is available in the Google Chrome Webstore.

PicNet Orchestra

PicNet Orchestra is a group of off the shelf software elements that enable, cheaper and faster software delivery, with consistent quality.

PicNetML

PicNetML is a predictive analytics framework for .Net. This framework is based on the Weka machine learning toolkit and is the tool of choice for PicNet's predictive analytics work. This library will assist if you are after an embedded prediction model or recommender in a web application or a standalone system that runs in the background doing predictive work or anomaly detection.

Inside PicNet

Watch the Software Development video.

Contact

Guido Tapia - Software Development Services & Innovation Manager
guido.tapia@picnet.com.au
Tel: +61 2 8437 7977

Fig. 18.22 Mouse tracking on a Web site: a gaze heat map overlay on the initial image from PicNet (Extracted from [128])

Another technique uses mouse tracking and is also used on marketing material like in [128] or [129]. There are two ways of using mouse tracking: either with no special indication like in [128] or by telling people to locate their mouse where they look [130]. The second version is precise but only used for research purposes and provides a quite good approximation of the user gaze [131] (more than 80 % of the eye tracking). In the case of [128], the accuracy is much lower. The advantage of mouse-tracking techniques is that they are cheaper than eye tracking, and more users are available via Web sites, while eye tracking needs the user to be physically present in front of an eye-tracking device. The study time varies but could be a little shorter than by using eye tracking.

Figure 18.22 shows an example of a heat map computed from the average mouse-tracking data from several users on a Web site page.

If no special indication is provided like in [128], the result is less accurate than eye tracking which is due to the fact that the mouse pointer never exactly focuses on the object of interest at least for visibility reasons. Also the mouse motion does not always follow the eye motion. However, the mass effect of the number of users which can be much higher than in the case of eye tracking can partly compensate this issue.

Finally, the predictive method which is the main focus here uses automatic saliency maps. This approach is much faster than eye-tracking tests (seconds versus days) and also much cheaper (around ten times cheaper). The prices are in the same

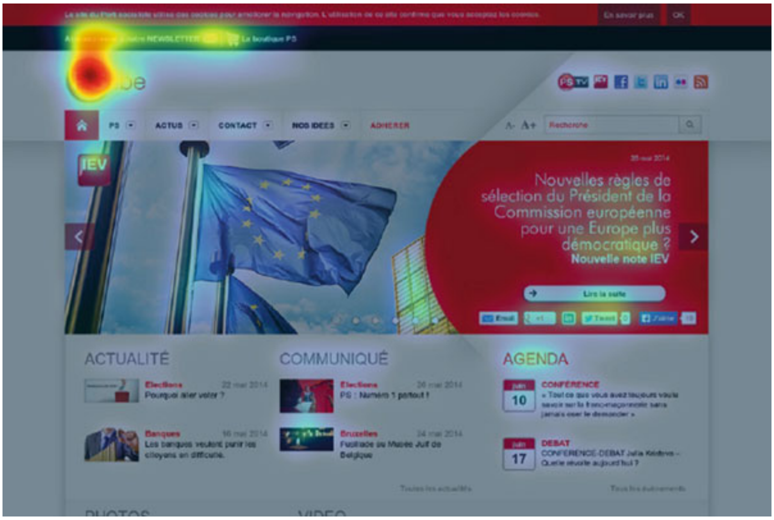

ACTUS
CONTACT
NOS IDEES
ADHERER
IEV
Nouvelles règles de sélection du Président de la Commission européenne pour une Europe plus démocratique ?
Nouvelle note IEV
Lire la suite
ACTUALITÉ
COMMUNIQUÉ
AGENDA
10
17

Fig. 18.23 Attention on a Web site: a gaze heat map overlay on the initial image from EyeQuant

range as the ones of mouse tracking. The results are less good than eye tracking, and they are equivalent to mouse tracking following proprietary studies such as [132]. The predictive methods can be achieved in real time and used at any time in the creative process: while humans (both using eye and mouse tracking) will be disturbed by unfinished documents, an automatic algorithm will not. This means that eye tracking or mouse tracking can mainly be achieved once the document is almost finalized (which might be too late for important changes), while the use of automatic algorithms can be used in real time and provides several feedback loops during the creative process. Another advantage of the predictive method is that it is image based only and it is possible to screen any kind of Web site or advertisement including the ones of concurrent companies.

Figure 18.23 shows an example of a heat map automatically computed on a Web site page using the automatic saliency maps from EyeQuant [133].

As the approach of saliency and marketing is one of the closest to the market, several companies based on the use of saliency in marketing were set up.

Feng-GUI [134] is an Israeli company mainly focusing on Web pages and advertising optimization even if the algorithm is also capable to analyze video sequences. Among the bottom-up features they use, one can find color, orientation, density and contrast, intensity, size, and weight and intersections. The top-down features are face detection, text detection, and skin detection. They also use context information about the type of the document (natural image, Web site, billboard, advertisement) which probably corresponds to different probability densities depending on the kind of document as in [130]. The main targeted applications are Web pages and advertising optimization even if the algorithm is also capable to analyze video sequences.

AttentionWizard [135] is a US company mainly focusing on Web pages. There are few hints on the used algorithm, but it uses bottom-up features like color differences, contrast, density, brightness and intensity, edges and intersections, length and width, curves, and line orientations. Top-down features include face detection, skin color, and text (especially big text) detection.

3M VAS [136] is the only big international player in this field. Very few details are given on the used algorithm, but it is also capable to provide video saliency. The main difference with the other competitors is in the customer segments with a much wider range of possible applications. They provide attention maps for Web page optimization but also advertisement with static images or videos, packaging, or in-store merchandising.

EyeQuant [133] is a German company specialized in Web site optimization. Their algorithm uses extensive eye-tracking tests to train the algorithm and make it closer to real eye tracking for a given task. They also can modify the saliency map if the viewer is involved or not in a task or if he simply goes through the page by modifying their algorithm weights. Finally, they provide a cue on "visual clarity" which seems to be related to a study on the image entropy. Other newcomers as Ittention (http://www.ittention.com) are coming on this market which shows a growing interest in the topic.

All those companies claim around 90 % accuracy for the first 3/5 viewing seconds [132]. They base their claim on different comparisons between their algorithm and several existing databases using several ROC metrics. They always compare the results with the maximum ROC score obtained by the human users. Nevertheless, for real-life images and for given tasks and emotion-based communication, this accuracy dramatically drops but still remains usable.

In addition to those four companies, another approach of using saliency models is proposed by a US company called EyePredict [137]. The main idea is to test a maximum of a combination of product catalog and propose the configuration which best optimizes a given product visibility.

18.4.2.2 Predicting Memorability of Pictures

The study of image memorability in computer science is a recent topic [138–141]. From those first attempts, it appears that it is possible to predict the degree of an image's memorability quite well. In this section, we present the concept of memorability of pictures and the computational models predicting the extent to which a picture is memorable.

Humans have an amazing visual memory. Only a few seconds is enough to memorize an image [142]. However, not all images are equally memorable. Some are very easy to memorize and to recall, whereas the memorization task appears to be much more difficult for other pictures. Isola et al. [138] was the first paper to build a large dataset of pictures associated to their own memorability score. The score varies between 0 and 1. 0 indicates that the picture is not memorable at all, while 1 indicates the highest score of memorability. The memorability has been quantified by performing a visual memory game. Six hundred sixty-five participants were involved in the test to score the memorability of 2222 images. This dataset is freely available on the author's Web site.

From this large amount of data, authors in [138] investigated the contributions of different factors and envisioned the first computational model for predicting the memorability scores.

An interesting step which followed was the use of the temporal context in memorability: indeed when seeing a lot of desert images, if a single image of forest appears, that one will be very memorable [141].

Memorability and Saliency Models

As mentioned earlier, Isola et al. [138] were the first to propose a computation model for predicting the memorability score of an image. Authors used a mixture of several low-level features which have been automatically extracted. A support vector regression classifier was used to infer the relationship between those features and the memorability. The best result was achieved by mixing together GIST [143], SIFT [144], HOG [145], SSIM [146], and pixel histograms (PH).

In [140], the authors proposed to go one step further by considering saliency-based features, namely, the saliency coverage and the visibility of structure. The saliency coverage which describes the spatial computational saliency density distribution could be approximated by the mean of the normalized saliency maps (computed by the RARE model [147]). A low coverage would indicate that there is at least one salient region in the image. A high coverage may indicate that there is nothing in the scene visually important as most of the pixels are attended. The second feature related to the visibility of structure is obtained by applying a low-pass filter several times on images with kernels of increasing sizes like in Gaussian pyramids (see [140] for more details). By using saliency-based features, the performance in term of linear correlation increases by 2 % while reducing the number of features required to perform the learning (86 % less features).

In the same year, the work in [138] was extended by [148] who proposed an attention-driven spatial pooling strategy. Instead of considering all the features (SIFT, HOG, etc.) with an equal contribution, the idea is to emphasize features of salient areas. This saliency-based pooling strategy improves the memorability prediction. Two levels of saliency were used: a bottom-up saliency and an object-level saliency. A linear correlation coefficient of 0.47 was obtained.

In [141], the context of the displayed images is studied and the influence of the viewing context is shown. Figure 18.24 shows the difference of memorability score function of the scene categories when images of those categories are shown surrounded of other images from the same category (in blue) or of different categories (in red). The memorability score dramatically increases when an image is shown surrounded of images from other scene categories.

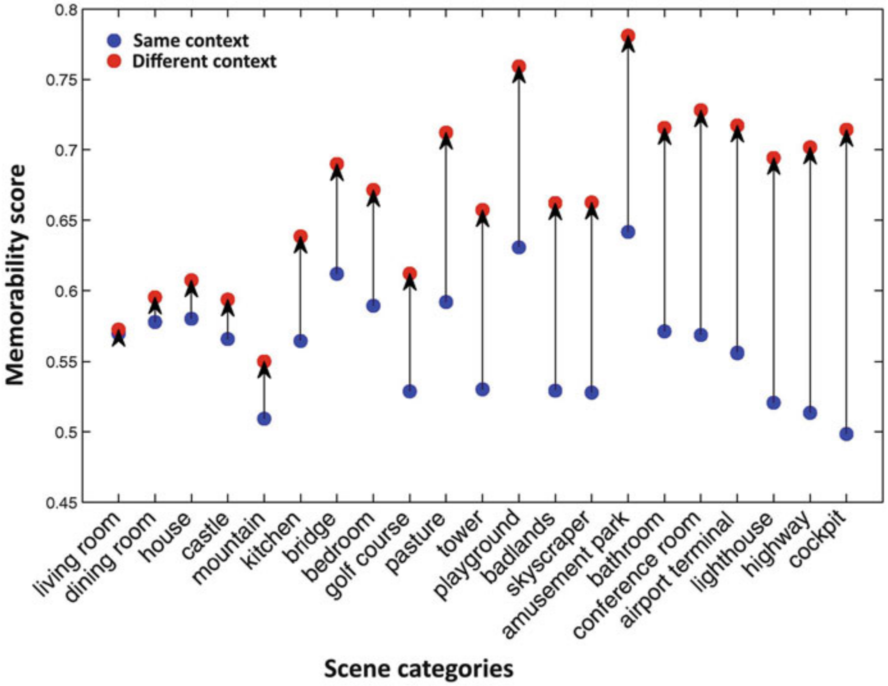


Fig. 18.24 Memorability vs. scene categories. In *blue*, the images are shown in the middle of images of the same category. In *Red*, the images are shown in the context of other scene categories (Adapted from [141])

18.4.2.3 Best Viewpoint

With more and more 3D objects which are created, manipulated, sold, or even printed, 3D saliency is a very promising future research direction. The main idea is to compute the saliency score of each view of a 3D model: the best viewpoint is the one where the total object saliency is maximized [149]. Mesh saliency was introduced based on adapting to the mesh structure concepts for 2D saliency [39]. The notion of viewpoint and mesh simplification is also related through the use of mesh saliency [150].

As the 3D approach for the best viewpoint is quite novel, it is not obvious to validate if the best viewpoint proposed by the algorithm is the one which people would select. The authors in [151] proposed a Web-based solution for viewpoint validation based on votes. Those votes are projected to a 3D heat map which is used to choose the best viewpoint (Fig. 18.25).

While the best viewpoint application can be used for computer graphics or even 3D mesh compression, marketing is one of the targets of this research topic: more and more 3D objects are shown even on the Internet, and the question of how to display them in an optimal way is very interesting in marketing.

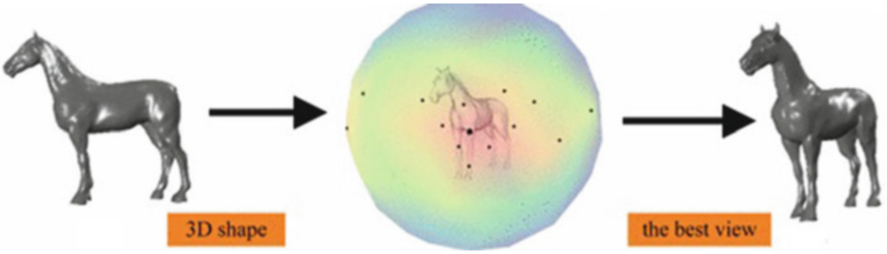


Fig. 18.25 Viewpoint evaluation: a Web-based evaluation can be done on a lot of viewers who vote for their best view. A 3D heat map of their votes can be projected on a sphere around the object, and the maximum of this heat map represents the best view (Adapted from [151])

18.4.2.4 A Discussion

The marketing optimization application of the automatic saliency algorithms has a promising future with already existing companies making money from the idea alone.

However, even if the results are very promising, there is room for a lot of improvement. More and more top-down information must be added to classical bottom-up attention to make the result fit with more precise user categories.

An issue which might be stressed is the banner blindness [152] which consists in ignoring the areas where the presence of an advertising is detected. Even if those advertisements are ignored, they are actually seen [153], and if one of them is of interest for the user, he will for sure attend to it.

In addition, features linked to image memorability (see Sect. 18.4.2.2) will also be taken into account. Indeed, in [154] features related to memorability seem to provide better visibility to advertisements like higher gray-level contrast or a smaller number of salient components, with all components close to the center of the creative and the major component consistent with the rule of third.

Finally, as more and more objects can be represented in 3D to be better visualized or even directly sold as 3D models for 3D printing, for example, the use of saliency on those models is a new research avenue. The automatic computation of the best viewpoint can provide interesting insights for 3D object visualization.

18.4.3 Attention-Based Focus or Symmetry

Saliency maps provide areas of interests or key points. By comparing those key points and their relative position, higher-level characteristics of the image can be found. For example, in [155] the author shows how a symmetry axis can be found using a saliency model (Fig. 18.26). Using again the comparison between patches in the image, it is also possible to find the vanishing points [156]. Using again the same approach, the autofocus of camera can be controlled [157].

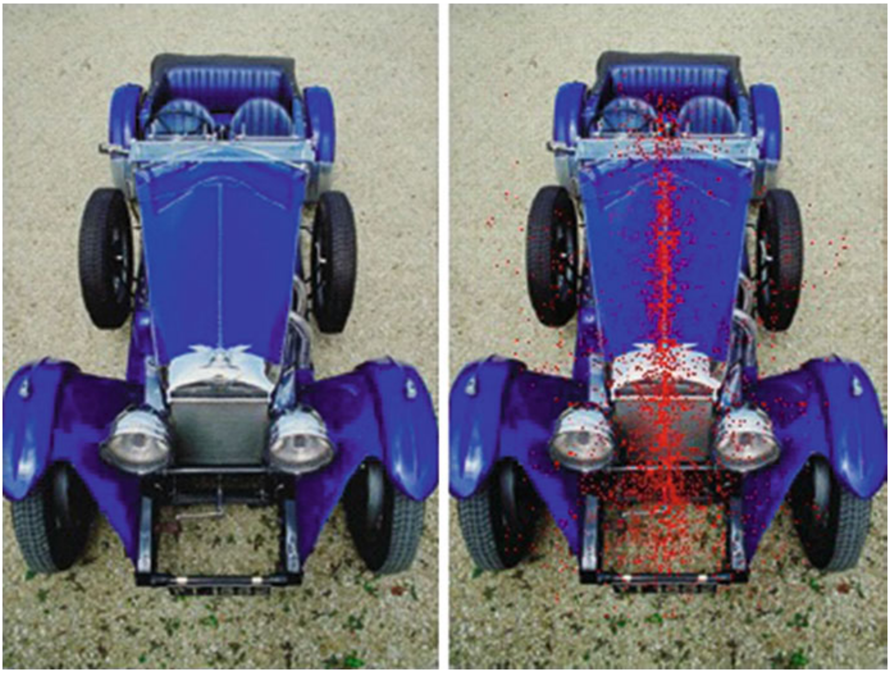


Fig. 18.26 Comparison between different regions of the image and object symmetry detection (Adapted from [155])

18.4.3.1 A Discussion

This work led to numerous applications. Several patents like [158] or [159] show that the technology becomes mature enough to be integrated in consumer electronics.

18.5 Conclusion

During the last two decades, significant progresses have been made in the area of visual attention. Although that the picture is much clearer, there are still a number of hurdles to overcome. For instance, the eye-tracking datasets used for evaluating the performance of computational models are more or less corrupted by biases. Among them, the central bias, which is the tendency of observers to look near the screen center, is probably the most important [160]. The central bias, which is extremely difficult to cancel or to remove, is a fundamental flaw which can significantly undermine conclusions of some studies and models' performance. Also other evaluation frameworks like the ones using segmented objects and even application-driven validation [161] will improve validation of the saliency models for real-life applications.

Regarding the applications, we decided to build a taxonomy made of three big categories:

- Abnormality detection: use the most salient areas' detection.
- Normality detection: use the less salient areas' detection.
- Abnormality processing: compare and further process the most salient areas.

These categories let us simplify and classify a very long list of applications which can benefit from attention models. We are just at the early stages of the use of saliency maps into computer vision applications. Nevertheless, the number of already existing applications shows a promising avenue for saliency models in improving existing applications and for the creation of new applications. Indeed, several factors are nowadays turning saliency computation from labs to industry:

- The models' accuracy drastically increased in two decades both concerning bottom-up saliency and top-down information and learning. The results of the recent models are way better than the first results in 1998.
- The models working both on videos and images are more and more numerous and provide more and more realistic results. New models including audio signals and 3D data are released and are expected to provide convincing results in the near future.
- The combined enhancement of computing hardware and algorithm optimization led to real-time or almost real-time good-quality saliency computation.

While some industry already began to use attention maps (marketing), others (TV, multimedia) come now to the use of such algorithms. Video surveillance and video summarization will also come into the game of using saliency maps shortly. This move from labs to industry will further encourage research on the topic toward understanding human attention, memory, and human motivation. New models both using bottom-up and more and more top-down information will appear. Moreover, more validation techniques, mainly application-driven, should be available in the next years to convince industry to use more attention modeling in their applications.

References

1. Fraunhofer (2011). Searise eu project. http://cordis.europa.eu/project/rcn/85425_en.html.
2. Bruce, N. D., & Kornprobst, P. (2009). On the role of context in probabilistic models of visual saliency. In *Proceedings of the 16th IEEE International Conference on Image Processing (ICIP 2009)*, Cairo (pp. 3089–3092). IEEE.
3. Mancas, M., & Gosselin, B. (2010). Dense crowd analysis through bottom-up and top-down attention. In *Proceedings of the Brain Inspired Cognitive Systems (BICS)*, Shenyang.
4. Jouneau, E., & Carincotte, C. (2011). Particle-based tracking model for automatic anomaly detection. In *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP 2011)*, Brussels (pp. 513–516). IEEE.

5. Mancas, M., Riche, N., Leroy, J., & Gosselin, B. (2011). Abnormal motion selection in crowds using bottom-up saliency. In *18th IEEE International Conference on Image Processing (ICIP 2011)*, Brussels (pp. 229–232). IEEE.
6. Li, W., Mahadevan, V., & Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(1), 18–32.
7. Jiang, M., Xu, J., & Zhao, Q. (2014). Saliency in crowd. In *Computer Vision–ECCV 2014*, Zurich (pp. 17–32). Springer.
8. Riche, N., Mancas, M., Culibrk, D., Crnojevic, V., Gosselin, B., & Dutoit, T. (2013). Dynamic saliency models and human attention: A comparative study on videos. In *Computer Vision–ACCV 2012*, Daejeon (pp. 586–598). Springer.
9. Boiman, O., & Irani, M. (2007). Detecting irregularities in images and in video. *International Journal of Computer Vision, 74*(1), 17–31.
10. Couvreur, L., Bettens, F., Hancq, J., & Mancas, M. (2007). Normalized auditory attention levels for automatic audio surveillance. In *International Conference on Safety and Security Engineering*, Malta.
11. Mancas, M., Couvreur, L., Gosselin, B., & Macq, B. et al. (2007). Computational attention for event detection. In *Proceedings of Fifth International Conference on Computer Vision Systems*, Bielefeld.
12. Hu, R., Hang, B., Ma, Y., & Dong, S. (2010). A bottom-up audio attention model for surveillance. In *IEEE International Conference on Multimedia and Expo (ICME 2010)*, Singapore (pp. 564–567). IEEE.
13. Mancas, M., Unay, B., Gosselin, B., & Macq, D. (2007). Computational attention for defect localisation. In *Proceedings of ICVS Workshop on Computational Attention & Applications*, Bielefeld.
14. Bai, X., Fang, Y., Lin, W., Wang, L., & Ju, B. F. (2014). Saliency-based defect detection in industrial images by using phase spectrum. *IEEE Transactions on Industrial Informatics, 10*(4), 2135–2145.
15. Bonnin-Pascual, F., & Ortiz, A. (2014). A probabilistic approach for defect detection based on saliency mechanisms. In *IEEE Emerging Technology and Factory Automation (ETFA 2014)*, Barcelona (pp. 1–4). IEEE.
16. Mishne, G., & Cohen, I. (2014). Multi-channel wafer defect detection using diffusion maps. In *IEEE 28th Convention of Electrical & Electronics Engineers in Israel (IEEEI 2014)*, Eilat (pp. 1–5). IEEE.
17. Alpert, S., & Kisilev, P. (2014). Unsupervised detection of abnormalities in medical images using salient features. In *SPIE medical imaging* (pp. 903 416–903 416). Bellingham: International Society for Optics and Photonics.
18. Shiwei, Y., Tingzhu, H., Xiaoyun, L., & Wufan, C. (2013). Partial mutual information based medical image registration guided by saliency maps. *Chinese Journal of Scientific Instrument, 6*, 002.
19. Deepak, K. S., Chakravarty, A., & Sivaswamy, J. et al. (2013). Visual saliency based bright lesion detection and discrimination in retinal images. In *Proceedings of the IEEE 10th International Symposium on Biomedical Imaging (ISBI 2013)*, San Francisco (pp. 1436–1439). IEEE.
20. Jampani, V., Sivaswamy, J., & Vaidya, V. et al. (2012). Assessment of computational visual attention models on medical images. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, Mumbai (p. 80). ACM.
21. Pirri, F., Pizzoli, M., & Mancas, M. (2012). Human-motion saliency in complex scenes. In *Gesture and sign language in human-computer interaction and embodied communication* (pp. 81–92). Berlin: Springer.
22. Mancas, M., Pirri, F., & Pizzoli, M. (2011). Human-motion saliency in multi-motion scenes and in close interaction. In *Proceedings of Gesture Workshop*, Athens.

23. Mancas, M., Glowinski, D., Volpe, G., Coletta, P., & Camurri, A. (2010). Gesture saliency: A context-aware analysis. In *Gesture in embodied communication and human-computer interaction* (pp. 146–157). Berlin/Heidelberg: Springer.
24. Chen, H. T. (2010). Preattentive co-saliency detection. In *Proceedings of the 17th IEEE International Conference on Image Processing (ICIP 2010)*, Hong Kong (pp. 1117–1120). IEEE.
25. Yucel, Z., Salah, A. A., Meriçli, C., & Meriçli, T. (2009). Joint visual attention modeling for naturally interacting robotic agents. In *Proceedings of the 24th International Symposium on Computer and Information Sciences, 2009. ISCIS 2009*, Guzelyurt (pp. 242–247). IEEE.
26. Yucel, Z., & Salah, A. A. (2009). Resolution of focus of attention using gaze direction estimation and saliency computation. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009)*, Amsterdam (pp. 1–6). IEEE.
27. Yucel, Z., Salah, A. A., Meriçli, Ç., Meriçli, T., Valenti, R., & Gevers, T. (2013). Joint attention by gaze interpolation and saliency. *IEEE Transactions on Cybernetics, 43*(3), 829–842.
28. Sugiyama, O., Kanda, T., Imai, M., Ishiguro, H., & Hagita, N. (2005). Three-layered draw-attention model for humanoid robots with gestures and verbal cues. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005.(IROS 2005)*, Edmonton (pp. 2423–2428). IEEE.
29. Schauerte, B., & Stiefelhagen, R. (2014). "look at this!" learning to guide visual saliency in human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014)*, Chicago (pp. 995–1002). IEEE.
30. Schillaci, G., Bodiroža, S., & Hafner, V. V. (2013). Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *International Journal of Social Robotics, 5*(1), 139–152.
31. Clair, A. S., Mead, R., & Matarić, M. J. (2011). Investigating the effects of visual saliency on deictic gesture production by a humanoid robot. In *RO-MAN, 2011 IEEE*, Atlanta (pp. 210–216). IEEE.
32. Zaraki, A., Mazzei, D., Lazzeri, N., Pieroni, M., & De Rossi, D. (2013). Preliminary implementation of context-aware attention system for humanoid robots. In *Biomimetic and biohybrid systems* (pp. 457–459). Heidelberg: Springer.
33. Balkenius, C., Gulz, A., Haake, M., & Johansson, B. (2013). Intelligent, socially oriented technology: Projects by teams of master level students in cognitive science and engineering: Anthology of master level course papers (p. 154). Lund University Cognitive Studies.
34. Itti, L., Dhavale, N., & Pighin, F. (2004). Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Optical Science and Technology, SPIE's 48th Annual Meeting*, San Diego (pp. 64–78). International Society for Optics and Photonics, San Diego, US.
35. Avila-Contreras, C., Medina, O., Jaime, K., & Ramos, F. (2014). An agent cognitive model for visual attention and response to novelty. In *Agent and multi-agent systems: Technologies and applications* (pp. 27–36). Berlin/New York: Springer.
36. Picot, A., Bailly, G., Elisei, F., & Raidt, S. (2007). Scrutinizing natural scenes: Controlling the gaze of an embodied conversational agent. In *Intelligent virtual agents* (pp. 272–282). Berlin/Heidelberg: Springer.
37. Longhurst, P., Debattista, K., & Chalmers, A. (2006). A GPU based saliency map for high-fidelity selective rendering. In *Proceedings of the 4th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*, Cape Town (pp. 21–29). ACM.
38. McNamara, A., Mania, K., Koulieris, G., & Itti, L. (2014). Attention-aware rendering, mobile graphics and games. In *ACM SIGGRAPH 2014 Courses*, Vancouver (p. 6).
39. Lee, C. H., Varshney, A., & Jacobs, D. W. (2005). Mesh saliency. In *ACM transactions on graphics (TOG)* (Vol. 24, pp. 659–666). New York: ACM.

40. Li, Z., & Zheng, J. (2014). Visual-salience-based tone mapping for high dynamic range images. *IEEE Transactions on Industrial Electronics, 61*(12), 7076–7082.
41. Narwaria, M., Da Silva, M. P., Le Callet, P., & Pepion, R. (2014). Tone mapping based HDR compression: Does it affect visual experience? *Signal Processing: Image Communication, 29*(2), 257–273.
42. Margolin, R., Zelnik-Manor, L., & Tal, A. (2013). Saliency for image manipulation. *The Visual Computer, 29*(5), 381–392.
43. Gai, M., & Wang, G. (2015). Artistic low poly rendering for images. *The visual computer* (pp. 1–10). Heidelberg: Springer.
44. Wong, L. K., & Low, K. L. (2009). Saliency-enhanced image aesthetics class prediction. In *Proceedings of 16th IEEE International Conference on Image Processing (ICIP 2009)*, Cairo (pp. 997–1000). IEEE.
45. Wong, L. K., & Low, K. L. (2011). Saliency retargeting: An approach to enhance image aesthetics. In *IEEE Workshop on Applications of Computer Vision (WACV 2011)*, Kona (pp. 73–80). IEEE.
46. Zhou, Y., Tan, Y., & Li, G. (2014). Computational aesthetic measurement of photographs based on multi-features with saliency. In *Intelligent computing theory* (pp. 357–366). Cham: Springer.
47. Ninassi, A., Le Meur, O., Le Callet, P., & Barbba, D. (2007). Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. In *IEEE International Conference on Image Processing, 2007. ICIP 2007*, San Antonio (Vol. 2, pp. II-169–II-172). doi:10.1109/ICIP.2007.4379119.
48. Ninassi, A., Le Meur, O., Le Callet, P., & Barba, D. (2009). Considering temporal variations of spatial visual distortions in video quality assessment. *IEEE Journal of Selected Topics in Signal Processing, Special Issue On Visual Media Quality Assessment, 3*(2), 253–265.
49. Liu, H., & Heynderickx, I. (2011). Visual attention in objective image quality assessment: based on eye-tracking data. *IEEE Transactions on Circuits and Systems for Video Technology, 21*(7), 971–982.
50. Guo, A., Zhao, D., Ans Shaohui, L., Fan, X., & Gao, W. (2011). Visual attention based image quality assessment. In *IEEE International Conference on Image Processing*, Brussels (pp. 3297–3300).
51. Akamine, W. Y. L., & Farias, M. C. Q. (2012). Incorporating visual attention models into image quality metrics. In *VPQM*, Scottsdale.
52. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.
53. Park, J., Lee, S., & Bovik, A. C. (2014). 3d visual discomfort prediction: Vergence, foveation, and the physiological optics of accommodation. *IEEE Journal of Selected Topics in Signal Processing, 8*(3), 415–427.
54. Lee, S. I., Jung, Y. J., Sohn, H., Ro, Y. M., & Park, H. W. (2011, February). Visual discomfort induced by fast salient object motion in stereoscopic video. In *IS&T/SPIE Electronic Imaging* (pp. 786305–786305). International Society for Optics and Photonics.
55. Sohn, H., Jung, Y. J., Lee, S. I., Park, H. W., & Ro, Y. M. (2011). Attention model-based visual comfort assessment for stereoscopic depth perception. In *17th International Conference on Digital Signal Processing (DSP 2011)*, Corfu Island (pp. 1–6). IEEE.
56. Du, S. P., Masia, B., Hu, S. M., & Gutierrez, D. (2013). A metric of visual comfort for stereoscopic motion. *ACM Transactions on Graphics (TOG), 32*(6), 222.
57. Jiang, Q., Wang, S., & Shao, F. (2015). An objective visual comfort prediction metric of stereoscopic images based on stereoscopic saliency model. *Industrial Electronics and Engineering, 93*, 263.
58. Jung, C., Cao, L., Liu, H., & Kim, J. (2015). Visual comfort enhancement in stereoscopic 3D images using saliency-adaptive nonlinear disparity mapping. *Displays, 40*, 17–23.

59. Chang, C. H., Liang, C. K., & Chuang, Y. Y. (2011). Content-aware display adaptation and interactive editing for stereoscopic images. *IEEE Transactions on Multimedia, 13*(4), 589–601.
60. Mancas, M., Mancas-Thillou, C., Gosselin, B., Macq, B. M. et al. (2006). A rarity-based visual attention map-application to texture description. In *ICIP*, Atlanta (pp. 445–448).
61. Varadarajan, S., & Karam, L. J. (2013). A no-reference perceptual texture regularity metric. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver (pp. 1894–1898). IEEE.
62. Varadarajan, S., & Karam, L. J. (2014). Effect of texture regularity on perceptual quality of compressed textures. In *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Chandler.
63. Maeder, A. J., Diederich, J., & Niebur, E. (1996). Limiting human perception for image sequences. In B. E. Rogowitz, & J. P. Allebach (Eds.), *Society of Photo-Optical Instrumentation Engineers (SPIE). Conference Series*, San Diego (Vol. 2657, pp. 330–337).
64. Kortum, P., & Geisler, W. (1996). Implementation of a foveated image coding system for image bandwidth reduction. In *Human Vision and Electronic Imaging, SPIE Proceedings*, San Francisco (pp. 350–360).
65. Yu, S. X., & Lisin, D. A. (2009). Image compression based on visual saliency at individual scales. In *International Symposium on Visual Computing*, Las Vegas (pp. 157–166).
66. Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing, 13*(10), 1304–1318.
67. Tsapatsoulis, N., Rapantzikos, K., & Pattichis, C. (2007). An embedded saliency map estimator scheme: Application to video encoding. *International Journal of Neural Systems, 17*(4), 1–16. http://www.image.ece.ntua.gr/publications.php.
68. Mancas, M., Gosselin, B., & Macq, B. (2007). Perceptual image representation. *Journal on Image Video Process, 2007*, 3–3. doi:http://dx.doi.org/10.1155/2007/98181.
69. Khanna, M. T., Rai, K., Chaudhury, S., & Lall, B. (2015). Perceptual depth preserving saliency based image compression. In *Proceedings of the 2nd International Conference on Perception and Machine Intelligence*, Kolkata (pp. 218–223). ACM.
70. Li, Z., Qin, S., & Itti, L. (2011). Visual attention guided bit allocation in video compression. *Image and Vision Computing, 29*(1), 1–14. doi:10.1016/j.imavis.2010.07.001. http://www.sciencedirect.com/science/article/pii/S0262885610001083
71. Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR'07*, Minneapolis (pp. 1–8). doi:10.1109/CVPR.2007.383267.
72. Guo, C., & Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans Image Process, 19*(1), 185–198. doi:10.1109/TIP.2009.2030969. http://dx.doi.org/10.1109/TIP.2009.2030969
73. Gupta, R., & Chaudhury, S. (2011). A scheme for attentional video compression. *Pattern Recognition and Machine Intelligence, 6744*, 458–465.
74. Zund, F., Pritch, Y., Hornung, A. S., & Gross, T. (2013), Content-aware image compression method. U.S. *Patent App. 13/802,165*.
75. Chamaret, C., Le Meur, O., Guillotel, P., & Chevet, J. C. (2010). How to measure the relevance of a retargeting approach?. In *Workshop Media Retargeting ECCV 2010*, Crete (pp. 1–14). http://hal.inria.fr/inria-00539234/en/
76. Mancas, M. (2009). Relative influence of bottom-up and top-down attention. In *Attention in cognitive systems* (Lecture notes in computer science, Vol. 5395). Berlin/Heidelberg: Springer.
77. Suh, B., Ling, H., Bederson, B. B., & Jacobs, D. W. (2003). Automatic thumbnail cropping and its effectiveness. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology (UIST)*, Vancouver (pp. 95–104).
78. Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194–203.

79. Ciocca, G., Cusano, C., Gasparini, F., & Schettini, R. (2007). Self-adaptive image cropping for small displays. *IEEE Transactions on Consumer Electronics, 53*(4), 1622–1627.
80. Liu, H., Jiang, S., Huang, Q., Xu, C., & Gao, W. (2007). Region-based visual attention analysis with its application in image browsing on small displays. In *ACM multimedia*, Augsburg (pp. 305–308).
81. Le Meur, O., Le Callet, P., & Barba, D. (2007). Construction d'images miniatures avec recadrage automatique basé sur un modéle perceptuel bio-inspiré. In *Traitement du signal, 24*(5), 323–335.
82. Zhu, T., Wang, W., Liu, P., & Xie, Y. (2011). Saliency-based adaptive scaling for image retargeting. In *Seventh International Conference on Computational Intelligence and Security (CIS 2011)*, New Orleans (pp. 1201–1205). doi:10.1109/CIS.2011.266.
83. Du, H., Liu, Z., Wang, J., Mei, L., & He, Y. (2014). Video retargeting based on spatiotemporal saliency model. In J. J. J. H. Park, Y. Pan, C. S. Kim, & Y. Yang (Eds.), *Future information technology* (Lecture notes in electrical engineering, Vol. 309, pp. 397–402). Berlin/Heidelberg: Springer. doi:10.1007/978-3-642-55038-6_61. http://dx.doi.org/10.1007/978-3-642-55038-6_61
84. Wang, J., Xu, M., He, X., Lu, H., & Hoang, D. (2014). A hybrid domain enhanced framework for video retargeting with spatial-temporal importance and 3d grid optimization. *Signal Processing, 94*(0), 33–47. doi:http://dx.doi.org/10.1016/j.sigpro.2013.06.007. http://www.sciencedirect.com/science/article/pii/S0165168413002259
85. Grundmann, M., & Kwatra, V. (2014). Methods and systems for video retargeting using motion saliency. http://www.google.com/patents/US20140044404, U.S. *Patent App. 14/058,411*.
86. Zhou, Lu, L., & Bovik., A. (2003). Foveation scalable video coding with automatic fixation selection. *IEEE Transactions on Image Processing, 12*(2), 243–254. doi:10.1109/TIP.2003.809015.
87. Liu, F., & Gleicher, M. (2005). Automatic image retargeting with fisheye-view warping. In *Proceedings of User Interface Software Technologies (UIST)*, Williamsburg. http://graphics.cs.wisc.edu/Papers/2005/LG05
88. Wolf, L., Guttmann, M., & Cohen-Or, D. (2007). Non-homogeneous content-driven video-retargeting. In *Proceedings of the Eleventh IEEE International Conference on Computer Vision (ICCV 2007)*, Rio de Janeiro.
89. Ren, T., Liu, Y., & Wu, G. (2009). Image retargeting using multi-map constrained region warping. In *ACM Multimedia*, Beijing (pp. 853–856).
90. Ren, T., Liu, Y., & Wu, G. (2010). Rapid image retargeting based on curve-edge grid representation. In *ICIP*, Hong Kong (pp. 869–872).
91. Wang, Y. S., Tai, C. L., Sorkine, O., & Lee, T. Y. (2008). Optimized scale-and-stretch for image resizing. *ACM Transactions on Graphics, 27*(5), 118.
92. Lin, S. S., Yeh, I. C., Lin, C. H., & Lee, T. Y. (2013). Patch-based image warping for content-aware retargeting. *IEEE Transactions on Multimedia, 15*(2), 359–368. doi:10.1109/TMM.2012.2228475.
93. Avidan, S., & Shamir, A. (2007). Seam carving for content-aware image resizing. *ACM Transactions on Graphics, 26*(3), 10.
94. Vaquero, D., Turk, M., Pulli, K., Tico, M., & Gelf, N. (2010). A survey of image retargeting techniques. SPIE Optical Engineering + Applications. International Society for Optics and Photonics.
95. Wonjun, K., Chanho, J., & Changick, K. (2011). Spatiotemporal saliency detection and its applications in static and dynamic scenes. *IEEE Transactions on Circuits and Systems for Video Technology, 21*(4), 10.
96. Rubinstein, M., Shamir, A., & Avidan, S. (2008). Improved seam carving for video retargeting. *ACM Transactions on Graphics (SIGGRAPH), 27*(3), 1–9.
97. Grundmann, M., Kwatra, V., Han, M., & Essa, I. (2010). Discontinuous seam-carving for video retargeting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco (pp. 569–576). doi:10.1109/CVPR.2010.5540165.

98. Goferman, S., Zelnik-Manor, L., & Tal, A. (2012). Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(10), 1915–1926.
99. Decombas, M., Dufaux, F., & Pesquet-Popescu, B. (2013). Spatio-temporal grouping with constraint for seam carving in video summary application. In *18th International Conference on Digital Signal Processing (DSP 2013)*, Santorini (pp. 1–8). doi:10.1109/ICDSP.2013.6622744.
100. Wu, L., Cao, L., Xu, M., & Wang, J. (2014). A hybrid image retargeting approach via combining seam carving and grid warping. *Journal of Multimedia, 9*(4). http://ojs.academypublisher.com/index.php/jmm/article/view/jmm0904483492
101. Ejaz, N., Mehmood, I., Sajjad, M., & Baik, S. W. (2014). Video summarization by employing visual saliency in a sufficient content change method. *International Journal of Computer Theory and Engineering, 6*(1), 26.
102. Dong, W., Zhou, N., Lee, T. Y., Wu, F., Kong, Y., & Zhang, X. (2014). Summarization-based image resizing by intelligent object carving. *IEEE Transactions on Visualization and Computer Graphics, 20*(1), 1–1.
103. Zhang, L., Xia, Y., Mao, K., Ma, H., & Shan, Z. (2015). An effective video summarization framework toward handheld devices. *IEEE Transactions on Industrial Electronics, 62*(2), 1309–1316.
104. Goferman, S., Tal, A., & Zelnik-Manor, L. (2010, May). Puzzle-like Collage. In *Computer Graphics Forum* (Vol. 29, No. 2, pp. 459–468). Blackwell Publishing Ltd.
105. Tal, A., Zelnik-Manor, L., & Goferman, S. (2014). Automated collage formation from photographic images. U.S. *Patent 8,693,780*.
106. Agarwal, C., Bose, A., Maiti, S., Islam, N., & Sarkar, S. K. (2013). Enhanced data hiding method using dwt based on saliency model. In *IEEE International Conference on Signal Processing, Computing and Control (ISPCC 2013)*, Shimla (pp. 1–6). IEEE.
107. Basu, A., Talukdar, S., Sengupta, N., Kar, A., Chakraborty, S. L., & Sarkar, S. K. (2015). On the implementation of a saliency based digital watermarking. In *Information systems design and intelligent applications* (pp. 447–455). Berlin/New York: Springer.
108. Niu, Y., Kyan, M., Ma, L., Beghdadi, A., & Krishnan, S. (2011). A visual saliency modulated just noticeable distortion profile for image watermarking. In *19th European Signal Processing Conference, 2011*, Barcelona (pp. 2039–2043). IEEE.
109. Li, J., & Gao, W. (2014). *Visual saliency computation: A machine learning perspective* (Vol. 8408). Cham: Springer.
110. Mei, T., Hua, X. S., Yang, L., & Li, S. (2007). Videosense: towards effective online video advertising. In *Proceedings of the 15th International Conference on Multimedia*, Augsburg (pp. 1075–1084). ACM
111. Chang, C. H., Hsieh, K. Y., Chiang, M. C., & Wu, J. L. (2010). Virtual spotlighted advertising for tennis videos. *Journal of visual communication and image representation, 21*(7), 595–612.
112. Frintrop, S., & Jensfelt, P. (2008). Attentional landmarks and active gaze control for visual slam. *IEEE Transactions on Robotics, 24*(5), 1054–1065.
113. Chen, H. Z., Jing, N., Wang, J., Chen, Y. G., & Chen, L. (2014). A novel saliency detection method for lunar remote sensing images. *Geoscience and Remote Sensing Letters, IEEE, 11*(1), 24–28.
114. Zhao, Y., Liu, Y., Song, R., & Zhang, M. (2012). Extended non-local means filter for surface saliency detection. In *19th IEEE International Conference on Image Processing (ICIP 2012)*, Orlando (pp. 633–636). IEEE.
115. Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*(9), 1395–1407.
116. Alexe, B., Deselaers, T., & Ferrari, V. (2010). What is an object? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, San Francisco (pp. 73–80). IEEE.
117. Zdziarski, Z., & Dahyot, R. (2012). Feature selection using visual saliency for content-based image retrieval. In *Signals and Systems Conference (ISSC 2012)*, IET Irish, Rostock (pp. 1–6). IET.

118. Awad, D., Courboulay, V., & Revel, A. (2012). Saliency filtering of sift detectors: Application to cbir. In *Advanced concepts for intelligent vision systems* (pp. 290–300). Berlin/New York: Springer.
119. Navalpakkam, V., & Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, New York (Vol. 2, pp. 2049–2056). IEEE.
120. Frintrop, S., Backer, G., & Rome, E. (2005). Goal-directed search with a top-down modulated computational attention system. In *Pattern recognition* (pp. 117–124). Berlin/New York: Springer.
121. Stentiford, F., & Bamidele, A. (2010). Image recognition using maximal cliques of interest points. In *17th IEEE International Conference on Image Processing (ICIP 2010)*, Hong Kong (pp. 1121–1124). IEEE.
122. Ferreira, J. F., & Dias, J. (2014). Attentional mechanisms for socially interactive robots–a survey. *IEEE Transactions on Autonomous Mental Development, 6*(2), 110–125.
123. Beira, R., Lopes, M., Praga, M., Santos-Victor, J., Bernardino, A., Metta, G., Becchi, F., & Saltarén, R. (2006). Design of the robot-cub (ICUB). head. In *IEEE International Conference on Robotics and Automation, 2006. ICRA 2006. Proceedings 2006*, Orlando (pp. 94–100). IEEE.
124. Schauerte, B., Richarz, J., Fink, G. et al. (2010). Saliency-based identification and recognition of pointed-at objects. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)*, Taipei (pp. 4638–4643). IEEE.
125. Mancas, M., Madhkour, R. B., De Beul, D., Leroy, J., Riche, N., Rybarczyk, Y. P., & Zajéga, F. (2011). Kinact: A saliency-based social game. In *Proceedings of the 7th International Summer Workshop on Multimodal Interfaces eNTERFACE11*, Pilsen (Vol. 8). Citeseer.
126. Russell, C. P. (1921). How to write a sales-making letter. *Printers' Ink*.
127. Miratech website proposes eye-tacking experiments for marketing material. http://miratech.fr/
128. Crazy egg website proposes mouse-tacking experiments for marketing material. http://www.crazyegg.com/
129. Picnet website proposes mouse-tacking experiments for marketing material. http://met.picnet.com.au/
130. Mancas, M. (2009). Relative influence of bottom-up and top-down attention. In *Attention in cognitive systems* (pp. 212–226). Berlin/Heidelberg: Springer.
131. Chen, M. C., Anderson, J. R., & Sohn, M. H. (2001). What can a mouse cursor tell us more?: Correlation of eye/mouse movements on web browsing. In *CHI'01 extended abstracts on Human factors in computing systems*, Seattle (pp. 281–282). ACM.
132. Page containing the 3m vas studies showing algorithm accuracy in general and in a marketing framework. http://solutions.3m.com/wps/portal/3M/en_US/VAS-NA/VAS/eye-tracking-software/eye-tracking-studies/.
133. Eyequant website proposes automatic saliency maps for marketing material. http://www.eyequant.com/.
134. Feng gui website proposes automatic saliency maps for marketing material. http://www.feng-gui.com/.
135. Attention wizzard website proposes automatic saliency maps for marketing material. https://www.attentionwizard.com/.
136. 3M vas website proposes automatic saliency maps for marketing material. http://solutions.3m.com/wps/portal/3M/en_US/VAS-NA/VAS/.
137. Eye predict website proposes automatic saliency models for web galleries. http://eye-predict.com/.
138. Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, Colorado Springs (pp. 145–152).
139. Khosla, A., Xiao, J., Torralba, A., & Oliva, A. (2012). Memorability of image regions. In *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe.

140. Mancas, M., & Le Meur, O. (2013). Memorability of natural scene: The role of attention. In *ICIP*.
141. Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research, 116*, 165–178.
142. Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology, 25*, 207–222.
143. Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145–175.
144. Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, New York (Vol. 2, pp. 2169–2178). IEEE.
145. Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR 2005)*, San Diego (Vol. 1, pp. 886–893). IEEE.
146. Shechtman, E., & Irani, M. (2007). Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, Minneapolis (pp. 1–8). IEEE.
147. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013). Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication, 28*(6), 642–658. doi:http://dx.doi.org/10.1016/j.image.2013.03.009.
148. Bora Celikkale, B., Erdem, A., & Erdem, E. (2013). Visual attention-driven spatial pooling for image memorability. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2013)*, Portland (pp. 1–8). IEEE.
149. Takahashi, S., Fujishiro, I., Takeshima, Y., & Nishita, T. (2005). A feature-driven approach to locating optimal viewpoints for volume visualization. In *Visualization (VIS 2005). IEEE*, Minneapolis (pp. 495–502).
150. Castelló, P., Chover, M., Sbert, M., & Feixas, M. (2014). Reducing complexity in polygonal meshes with view-based saliency. *Computer Aided Geometric Design, 31*(6), 279–293.
151. Liu, H., Zhang, L., & Huang, H. (2012). Web-image driven best views of 3d shapes. *The Visual Computer, 28*(3), 279–287.
152. Benway, J. P., & Lane, D. M. (1998). Banner blindness: Web searchers often miss "obvious" links. *Internetworking, ITG Newsletter*.
153. Bayles, M. (2000). Just how "blind" are we to advertising banners on the web. *Usability News, 2*(2), 520–541.
154. Azimi, J., Zhang, R., Zhou, Y., Navalpakkam, V., Mao, J., & Fern, X. (2012). The impact of visual appearance on user response in online display advertising. In *Proceedings of the 21st international conference companion on World Wide Web*, Lyon (pp. 457–458). ACM.
155. Stentiford, F. (2005). Attention based symmetry detection in colour images. In *IEEE 7th Workshop on Multimedia Signal Processing, 2005*, Shanghai (pp. 1–4). IEEE.
156. Stentiford, F. (2006). Attention-based vanishing point detection. In *IEEE International Conference on Image Processing, 2006*, Atlanta (pp. 417–420). IEEE.
157. Shilston, R., & Stentiford, F. (2006). An attention based focus control system. In *IEEE International Conference on Image Processing, 2006*, Atlanta (pp. 425–428). IEEE.
158. Shilston, R. T., & Stentiford, F. W. (2011). Method for focus control. U.S. *Patent 8,040,428*.
159. Stentiford, F. W. (2012). Image analysis relating to extracting three dimensional information from a two dimensional image. U.S. *Patent 8,135,210*.
160. Tatler, B. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*.
161. Awad, D., Mancas, M., Riche, N., Courboulay, V., & Revel, A. A. (2015). A CBIR-based evaluation framework for visual attention models. Signal Processing Conference (EUSIPCO), 2015 23rd European. IEEE.

Chapter 19
Attentive Content-Based Image Retrieval

Dounia Awad, Vincent Courboulay, and Arnaud Revel

19.1 Introduction

Imagine you are a few years in the future and you have just received your brand-new “autonomous robot” car. This car is able to recognize the road signs and detect passengers, pedestrians, and other cars on the road. At every moment, this car has to deal with a large amount of visual data. How can it manage so many information?

Typically, the kind of image the computer embedded into the car has to deal with is similar to the one given in Fig. 19.1a. For instance, let us consider this image as a classical VGA image (640 wide and 480 high) coded in RGB (24 bits per pixel). If we only consider the informational aspect of the problem, the number of different images that can be generated with this format is $(640 \times 480)^{2^{24}} \approx 10^{100000000}$…which is obviously unmanageable! Consequently, it is necessary to reduce the complexity of the image by detecting regularities and structures into the image.

By detecting the edges of the objects, the data to store is reduced to a few values (a boolean edge/no-edge value per pixel). Nevertheless, an important part of the properties (localization, shape) of the objects in the image is kept (see Fig. 19.1b).

In image processing, many methods have been proposed. Among them, some take inspiration from human visual system, and many studies have been done to

D. Awad (✉)
L3i Laboratory, University of La Rochelle, La Rochelle, France
Vision Laboratory, CINTAL-University of Algavre, Campus Gambelas, 2000 Faro, Portugal
e-mail: dounia.awad@gmail.com

V. Courboulay • A. Revel
L3i Laboratory, University of La Rochelle, La Rochelle, France

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_19

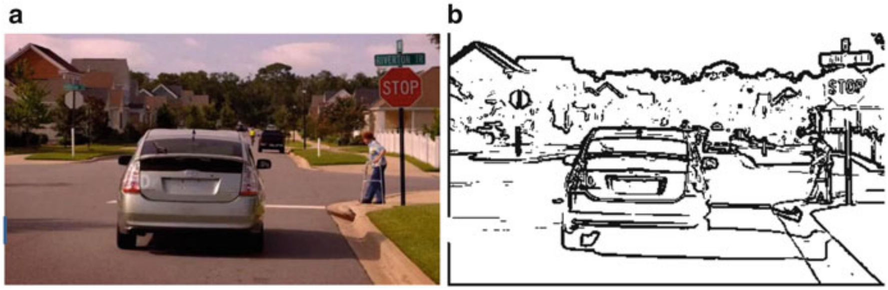


Fig. 19.1 (**a**) Example image adapted from autoGoer.com. (**b**) Edge detection on the previous image

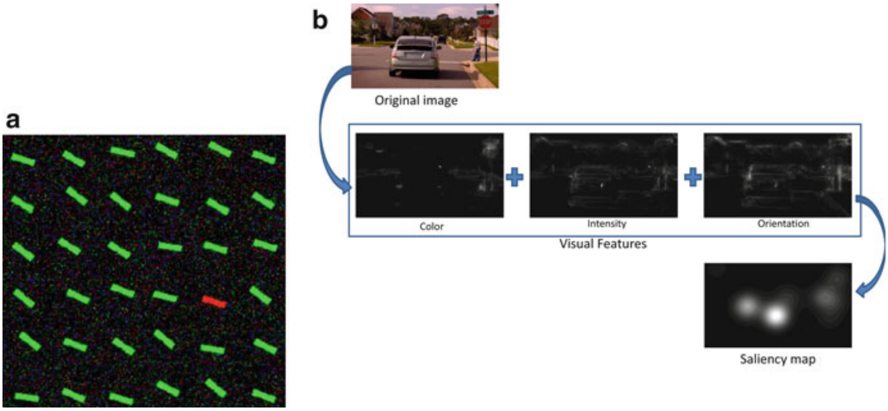


Fig. 19.2 (**a**) Pop-up effect and (**b**) classical saliency model

explain this "visual selection" mechanism. According to Walther [1], our capacity to identify and understand of our environment, known as "human perception," consists of two parts:

- Attention which selects information based on the saliency[1] in the image itself (bottom-up) and on prior knowledge about the scenes (top-down) [2, 3]
- Object recognition which permits to infer the presence of an object or members of an object category in this image

Considering the first part, it was found in neurobiology [4] that the visual system is able to perform parallel processing of the image coming from the retina to detect given features such as the edges, colors, textures, etc. In psychology [5], it has also been shown that such feature maps would be involved in a "pop-up" phenomenon that would be a part of the attentional system: a red rectangle lost in the middle of many green rectangles is detected almost instantaneously (see Fig. 19.2a).

[1] According to the Collins dictionary, the saliency is the quality of being prominent, conspicuous, or striking.

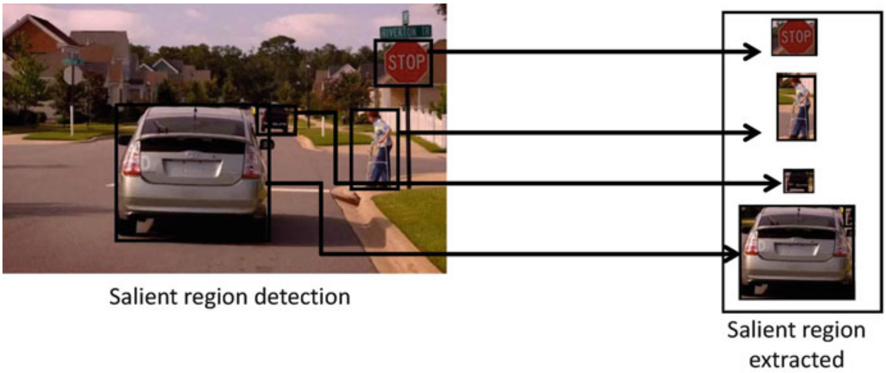


Fig. 19.3 Detection of objects in an image

Following Itti's [3] pioneer work, several computational models have been proposed to perform "attentional" selection (see Fig. 19.2b). Applied to an image such as our example, they can detect regions of interest which consist of "interesting" subparts of the image (ideally, objects in the image – see Fig. 19.3).

Once detected, these subparts of the image must be recognized. In fact, it is quite a difficult problem and many methods try to tackle it [6]. Among them, some are based on the idea that an image can easily be recognized if an image of the same kind has already been recognized. The principle is then to provide an image database in which every image is stored together with a given label corresponding to the object it contains. To recognize the subpart of the image, it is necessary to compare this subpart with all the images in the database (see Fig. 19.4): this kind of technique is thus known as content-based image retrieval (CBIR).

In this chapter, we propose to merge the two processes into a single one: the first part consists of an attentional system which selects the interesting information in the image. The second part is a CBIR-like system that intends to recognize the elements in the image. The idea is to investigate what saliency models can bring to the CBIR community.

19.1.1 CBIR State of the Art

The domain of "content-based image retrieval (CBIR)" is considered as one of the most challenging domains in computer vision, and it has been an active and fast-advancing research area over the last years. Most retrieval methods are based on extracting points of interest using interest point detectors [7] as Harris and Harris-Laplace (see Fig. 19.5) and described it by multidimensional feature vectors using SIFT descriptors [8]. The set of these feature vectors is known as bag of features [9]. Although these approaches have demonstrated a high efficiency, some weakness

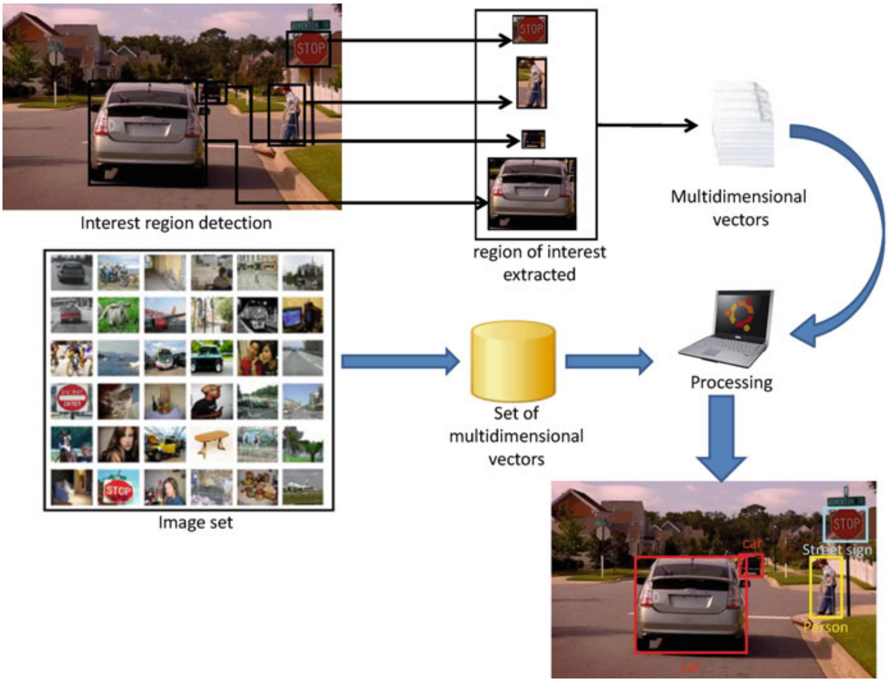

Interest region detection
region of interest extracted
Multidimensional vectors
Image set
Set of multidimensional vectors
Processing
car
Street sign
Person
car

Fig. 19.4 Overall schema of a generic CBIR algorithm

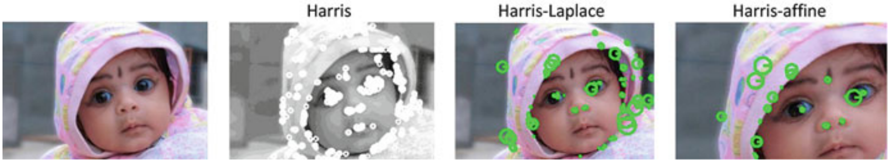


Fig. 19.5 Example of interest points detected by Harris algorithm and its variants

may be mentioned [10]. The first limitation is represented in the interest point detectors. Most of these detectors are based on geometric forms as corners, blobs, or junctions and consider that the interest of the image is correlated with the presence of such features.

Furthermore, some studies [11] have demonstrated that these detectors were not designed to detect the most pertinent regions for object recognition. Moreover, although SIFT shows a high efficiency, scalability remains an important problem due to the large number of features generated for each image [12]: many of them are outliers [13].

An alternative way for extracting regions of interest is derived from visual attention domain. This domain had been investigated intensely in the last years and many models had been proposed (see Table 19.1) [6]. In this chapter, we focus

Table 19.1 A taxonomy of visual attention models (Adapted from [6])

Class	Year	Models	Description
Cognitive models	1985	Koch and Ullman [14]	Their algorithm was inspired by Feature integration models. An important contribution of the work is the WTA network
	1998	Itti et al. [15]	Its model is considered as a derivation of Koch and Ullman algorithm. this model serves a basis for many research group. Itti introduced the image pyramids for the feature computations
	2004	Le Meur et al. [16]	Approach for bottom-up saliency based on the structure of the human visual system. Contrast sensitivity function, perceptual decomposition, visual masking, center surround interaction are some of the feature implemented in his model
	2005	Fintrop [17]	In their model, they separate the intensity feature computation into on-off and off-on computation instead of combining them in a single map
Bayesian models	2003	Torralba [18] and Oliva et al. [19]	They proposed a Bayesian framework for visual search tasks. Bottom-up saliency is derived from their formulation as $1/p(f\|f_G)$ where f_G represents a global feature that summarizes the probability density of the target object in the scene
	2005	Itti and Baldi [20]	They defined surprising stimuli as those which significantly change beliefs of an observer. This is modeled in a bayesian framework by computing the KL divergence between posterior and prior beliefs
Decision theoretic models	2004	Gao and Vasconcelos [21]	They argued that for recognition, salient features are those that best distinguish a class of interest from all other visual classes. They then defined top-down attention as classification with minimal expected error
	2007	Gu et al. [22]	An activation map was first computed by extracting primary visual features and detecting meaningful objects from the scene. A retinal filter is used after to generate the region of interest
Information theoretic models	2005	Bruce and Tsotsos [23]	They proposed the AIM model (Attention based on Information Maximization) which uses Shannon's self-information measure for calculating saliency of image regions. Saliency of a local image region is the information that region conveys relative to its surrounding

(continued)

Table 19.1 (continued)

Class	Year	Models	Description
Graphical models	2002	Salah et al. [24]	They proposed an approach for attention based on Observable Markov Model (OMM). Regions visited by a fovea treated as states of OMM. An inhibition of return allows the fovea to focus on the other position in the image
	2005	Rao et al. [25]	They built a model based on assumptions that the goal of the visual system is to know what is where and that visual processing happens sequentially. In this model, attention emerges as the inference in a Bayesian graphical model which implements interaction between ventral and dorsal areas
	2007	Liu et al. [26]	They proposed a set of novel features and adopted a Conditional Random Field to combine these feature for salient object detection on their regional saliency dataset
Spectral analysis models	2007	Hou and Zhang [27]	Their model based on the idea that similarities imply redundancies that propose that statistical singularities in the spectrum may be responsible for anomalous regions in the image, where proto-objects become conspicuous
Pattern classification models	2007	Petters and Itti [28]	They trained a simple regression classifier to capture the task-dependent association between a given scene and preferred location to gaze at while human subjects were playing video games
	2009	Kienzle et al. [29]	They built non-parametric bottom-up approaches for learning attention directly from human eye tracking data
Other models	2002	Ramstrom and Christensen [30]	They introduced a saliency measure using multiple cues based on game theory concepts. Feature maps are integrated using a scale pyramid where the nodes are subject to trading on a market and the outcome of the trading represents the saliency

on bottom-up visual attention models [15, 31]. The objective of these models is to extract the regions that attract our interest without any prior knowledge about the scene or image. These regions are called salient and they are selected according to some discriminative features such as color, orientation, and intensity. To check how much these regions are salient, the evaluation consists of comparing the saliency map generated by visual attention models with a heatmap processed from eye-tracking experiments [31, 32]. These experiments consist of displaying an image a few seconds in order to capture enough eye fixations to build a statistical heatmap. According to Perreira Da Silva [33], this evaluation is complex and suffers from known biases, such as the semantic bias. Indeed, one can gaze to meaningful regions which are not necessarily salient.

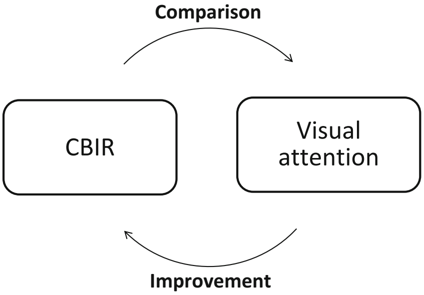


Fig. 19.6 Attention and CBIR self-improvement relationship

Recently, many works have been proposed to combine these domains, given what we called "attentive content-based image retrieval (Attentive C.B.I.R)." This idea was introduced earlier in [34], who indicated that object recognition in human perception consists of two steps: "attentional process selects the region of interest and complex object recognition process restricted to these regions." Based on this definition, Walther [1] proposed an algorithm for image matching: his algorithm detects SIFT keypoints inside the attentional regions. These regions determine a search area, whereas the matching is on SIFT keypoints. This approach was successful since they used very complex objects and those which do not change a viewpoint. Others as Frintrop and Jenselft [35] applied directly SIFT descriptors to the attention regions. They applied their approach on robot localization. Although this approach achieved an improvement in the detection rate for indoor environment, it fails in the outdoor environment and open areas. In this chapter, we will focus on the usefulness of attentive CBIR toward both communities: CBIR and visual attention. We hypothesize that attention can improve object recognition systems in query run time and information quality (Fig. 19.6) since these models generate salient regions on large scales, considering the context information. This property of attentional models generates fewer salient points regardless interest point detector. These detectors extract regions of interest on small scales, resulting several hundreds or thousands of points. This idea was presented previously by Frintrop [36] who indicated that the task of object recognition becomes easier if an attentional mechanism first cued the processing on regions of potential interest because of two reasons. First, it reduces the search space and results in lower computational complexity. Second, most recognition and classification methods work best if the object occupies a dominant portion of the image.

This chapter will be organized as follows. In Sect. 19.2, we will present a brief overview of our attentive CBIR system and the related algorithms used in this system. more detail about the system and experiment will be presented in Sect. 19.3.

19.2 System Architecture and Results

19.2.1 Bottom-Up Attention Model

In this section, we will present the bottom-up attentional systems that model human selective attention. These models aim to determine the most relevant parts within the large amount of visual data. Their basis is originally adapted from psychological theories like "feature integration theory" [5, 37] and "guided search model" [38] where they stated which visual features are important and how they are combined to direct human attention over pop-out and conjunction search task. Three features have been used based on these theories in computational models of attention: intensity (or intensity contrast, or luminance contrast), color, and orientation. The first complete implementation and verification of an attention model was proposed by Itti et al. [15] and was applied to synthetic as well as natural scenes. Its main idea was to compute features and to fuse their saliencies in a representation which is usually called saliency map. This algorithm is considered one of the famous central representation models that encode attention in a 2D topographic map. This map serves as reference for allocating attention though various mechanisms (winner takes all, inhibition of return, etc.).

Perreira Da Silva et al. [31] proposed a new hybrid model (called PVAS) based on the classical algorithm proposed by Itti [15], in which the first part of architecture relies on the extraction of three conspicuity maps (color, intensity, orientation) based on low-level characteristic computation.

Perreira Da Silva et al. [39] propose to substitute the second part of Itti's model by an optimal competitive approach: a prey/predator system (Fig. 19.7). We applied the same optimal parameter used by Perreira Da Silva [39] in our evaluation method. The output of this algorithm is a saliency map $S(I, t)$ computed by a temporal average of the focalization computed through a sliding temporal window. Hence, Perreira Da Silva and Itti models were chosen to test our evaluation method based on attentive CBIR system. In the following, we will present the CBIR approach used in our attentive CBIR system.

19.2.2 Object Recognition

As mentioned in Sect. 19.1, content-based image retrieval has seen considerable progress over the past years. Many challenges have been proposed to test the efficiency and robustness of the recognition methods. One of the most popular challenges is the Visual Object Classes Challenge [40]. VOC was proposed for the first time in 2005 with one objective: recognizing objects from a number of visual object classes in realistic scenes [9]. Since then, it has been organized every year and integrates new constraints in order to provide a standardized database to the researchers.

In 2005, 12 algorithms have been proposed to compete for winning the challenge; it is interesting to mention that all algorithms were based on local feature detection. We propose a taxonomy in Table 19.2. Finally, INRIA-Zhang appeared to be

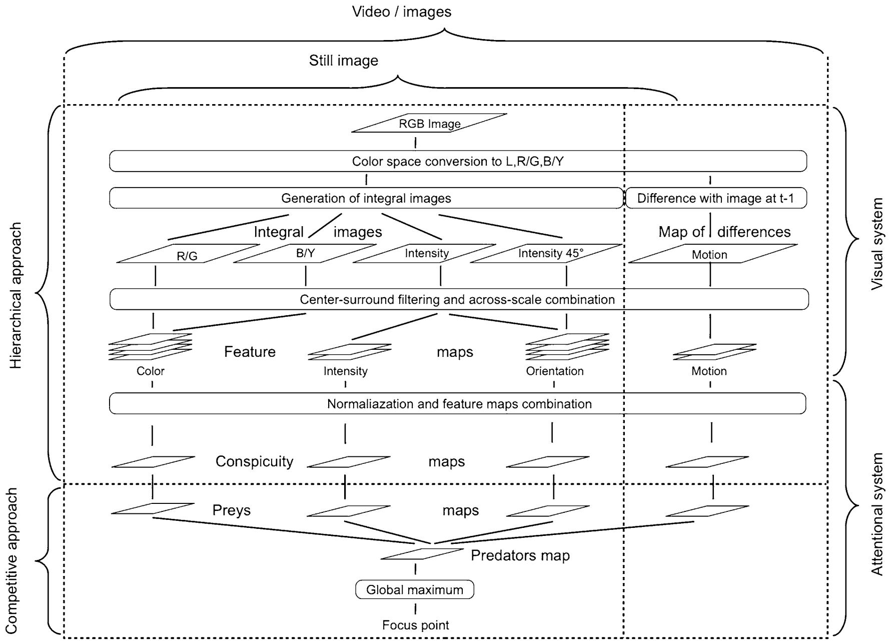


Fig. 19.7 Architecture of the PVAS computational model of attention

Table 19.2 Taxonomy of methods proposed in VOC2005 (Adapted from [9])

Category	Description
Distribution of local image feature	Images are represented by probability distributions over the set of descriptors, basing on two methods. Bags of words [41] in which an image is represented by a histogram over a dictionary recording the presence of each word. An alternative way is based on kernel as Bhattacharyya kernel [42]. Finally, the model is learned using classification methods as SVM
Recognition of individual local feature	In this approach, interest point detectors are used to focus the attention on a small number of local patches. Then each patch in each image is associated with a binary label. Vectors are built by grouping these labels. A parametric model of the probability that the patch belongs to a class is built. Finally the output is the posterior class probability for a patch feature vector
Recognition based on segmented regions	This method combines the features extracted from the image and the regions obtained by an image segmentation algorithm. The self-organizing maps (SOM) [43] are defined on the different feature spaces that were used to classify the descriptors resulting from the segmented regions and the whole image
Classifications by detection	It extracts patches in an image using interest point detectors. A code book is built using a clustering method. A new object class is detected using a matching method. Then a hypothesis on accept and refusal is defined

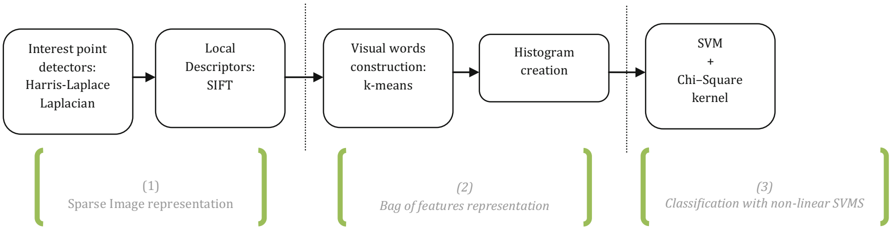

Interest point detectors: Harris-Laplace Laplacian
Local Descriptors: SIFT
Visual words construction: k-means
Histogram creation
SVM + Chi–Square kernel
(1) Sparse Image representation
(2) Bag of features representation
(3) Classification with non-linear SVMS

Fig. 19.8 Architecture of Zhang algorithm

the most efficient method. We decide to take it as the reference algorithm for object recognition. The algorithm shown in Fig. 19.8 consists of extracting an invariant image representation and classifying this data with nonlinear support vector machines (SVM) with an χ^2 kernel.

19.2.3 Results

To validate our hypothesis, we implemented our approach and evaluated it on the VOC 2005 database. The VOC challenge proposed two image subsets, the subset S_1 with selected images and another subset S_2 with Google image randomly selected. Thus, our approach can be performed independently during learning and for the test process. We evaluate the binary classification using receiver operating characteristic (ROC) curve [44]. With this curve, it is easy to observe the trade-off between two measures: proportion of false positives plotted on the x-axis and true positives plotted on the y-axis.

In Fig. 19.9, some ROC curves are shown. These curves present the results of our evaluation method, for two computational attention models: Itti and Perreira Da Silva. The idea, here, is to develop two attentive CBIR models and to test their efficiency:

- P/P+Zhang: this system represents the combination of Perreira Da Silva models with Zhang nominal algorithm.
- Itti+ Zhang: this system represents the combination of Itti models with Zhang's nominal implementation.

Finally, three curves were drawn, representing our implementation of Zhang algorithm and two attentive CBIR models.

Furthermore, we tested also the usefulness of attentive CBIR toward object recognition domain. Our test consists on using Perreira Da Silva system as a mask to select among all the SIFT keypoints only those which are the most salient. Results are shown in Fig. 19.10 representing, respectively, our implementation of Zhang

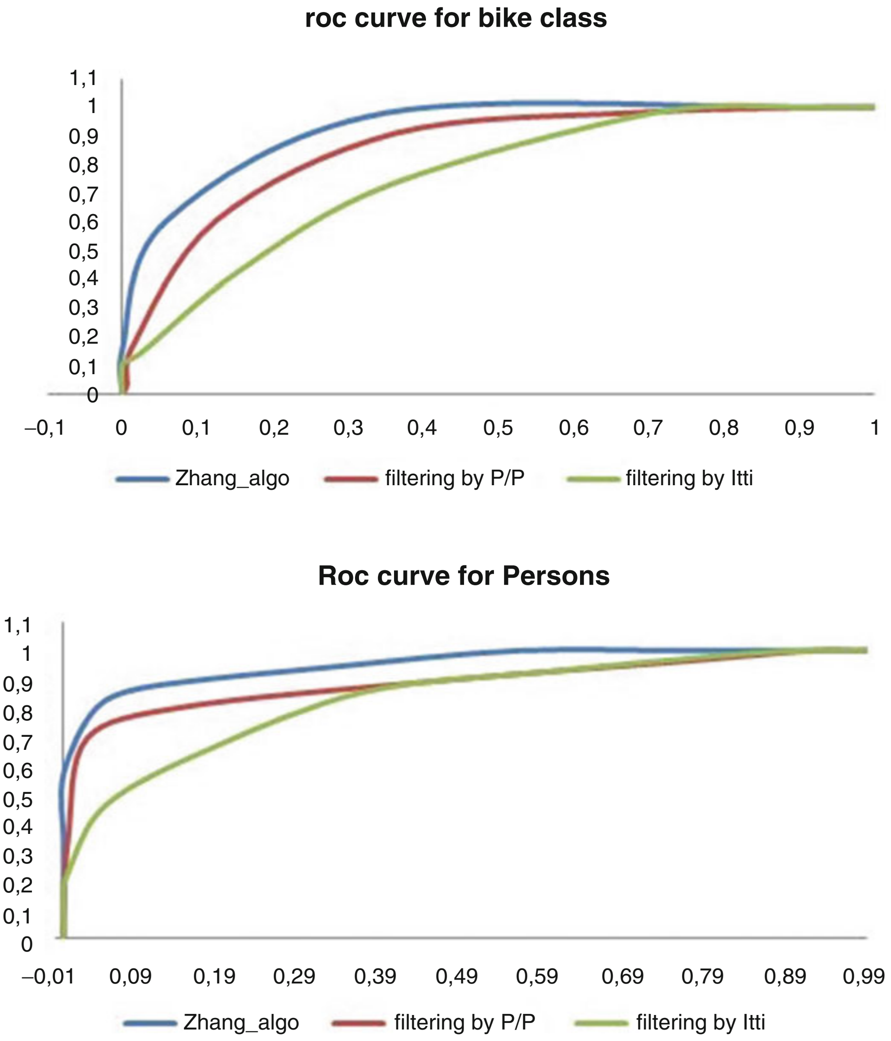


Fig. 19.9 ROC curve with and without our filter approach for two different classes

algorithm without filtering and with several filtering. For reason of clarity, we don't present the tests we did exhaustively. We selected only:

- the "best" curve: the maximum reduction of keypoints we had while having approximately the same performance
- the "worst" curve: the maximum loss of performance we had before filtering all the keypoints

It is also interesting to mention that comparing the two visual attention systems showed that percentage of keypoint reduction was higher for Itti+Zhang than PP+Zhang. This can explain why the performance for Itti system was worse than for Perreira system.

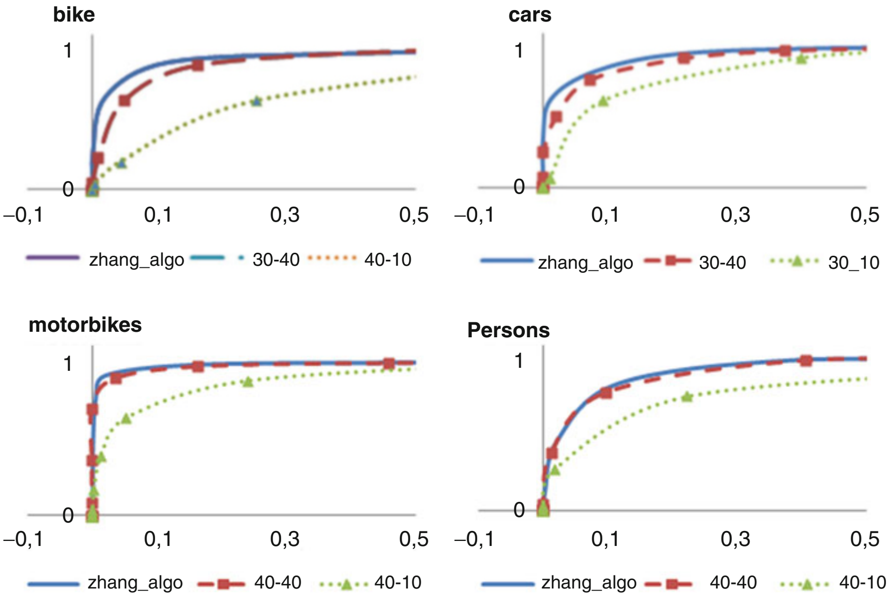


Fig. 19.10 ROC curve with and without our filter approach for the different classes-S_1

19.3 More on the System Architecture and Results

19.3.1 Bottom-Up Attention Model

As mentioned in Sect. 19.2.1, we chose to test our hypothesis on two visual attention models which share the same image-extracted low-level features. The first model is Itti et al.'s basic model [15]. This model was proposed in 1998 and served as basis for later models and standard benchmark for comparison. The second attention model is Perreira Da Silva model [39], a new real-time computational model which allows modeling the temporal evolution of visual focus of attention and its validation. The first part of this model is inspired from Itti classical model in which three conspicuity maps, representing the three main human perceptual channels, are extracted. In the second part, Perreira proposed a competitive system prey/predator with the following features:

- The system is comprised of three types of preys and one type of predators.
- These three types of preys represent the spatial distribution of curiosity generated by our three conspicuity maps(intensity, color, orientation).
- The predator represents the interest generated by the consumption of curiosity (preys) associated with different conspicuity maps.
- The global maximum of the predator's maps (interest) represents the focus of attention at time t.

Perreira Da Silva et al. [39] show that despite the nondeterministic behavior of prey/predator equations, the system exhibits interesting properties of stability, reproducibility, and reactiveness while allowing a fast and efficient exploration of the scene.

19.3.2 Object Recognition

As mentioned in Sect. 19.2.2, we chose Zhang algorithm as reference for our development. This algorithm can be divided in three parts:

1. Sparse image representation: this part extracts a set of SIFT keypoints $\mathrm{K}_{\mathrm{Zhang}}(I)$ from an image $I(x,y)$ which was provided as input. It consists of two steps:
   - Interest point detectors: Zhang uses two complementary local region detectors to extract *interesting* image structures: Harris-Laplace detector [45], dedicated to corner-like region, and Laplacian detector [46], dedicated to blob-like regions. These two detectors have been designed to be scale invariant.
   - Local descriptor: To compute appearance-based descriptor on the extracted patches, Zhang used the SIFT descriptor [8]. It computes descriptors less sensitive to scale variations and invariant to illumination changes.
2. Bag-of-features representation: Zhang builds a visual vocabulary by clustering the descriptors from the training set. Each image is represented as a histogram of visual words drawn from the vocabulary. He randomly selects a set of descriptors for each class extracted from the training set, and he clusters these features using k-means to create 1000-element vocabulary. Finally, each feature in an image is assigned to the closest word, and a histogram that measures the frequency of each word in an image is built for each image.
3. Classification: Zhang uses a nonlinear SVM in which the decision function has the following form:

$$g(x) = \sum \alpha_i y_i k(x_i, x) - b \tag{19.1}$$

   with $k(x_i, x)$ the kernel function value for the training sample x_i and the test sample x. α_i is the learned weighted coefficient for the training sample x_i and b is the learned threshold. Finally, to compute the efficiency of the algorithm, SVM score has been considered as a confidence measure for a class.

19.3.3 Proposed Architecture

As mentioned before, attentive CBIR is a combination of attentional systems and CBIR algorithms. We chose Zhang algorithm as the reference of our development. Analyzing the different steps of the algorithms, it can be noticed that the first step consists in using the interest point detectors. According to [12], not all of those

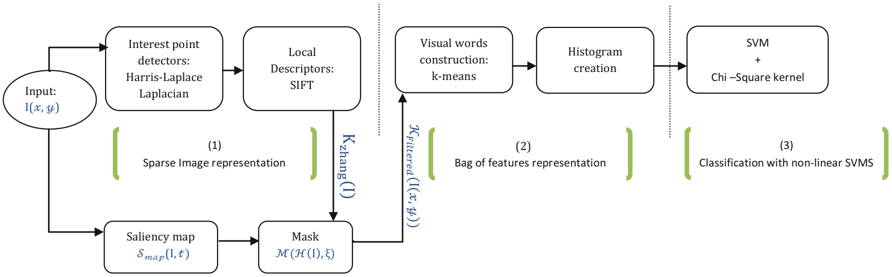

Input:
$I(x,y)$
Interest point detectors: Harris-Laplace Laplacian
Local Descriptors: SIFT
Visual words construction: k-means
Histogram creation
SVM + Chi –Square kernel
(1) Sparse Image representation
(2) Bag of features representation
(3) Classification with non-linear SVMS
$K_{\text{Zhang}}(I)$
$\mathcal{K}_{Filtered}(I(x,y))$
Saliency map $\mathcal{S}_{map}(I,t)$
Mask $\mathcal{M}(\mathcal{H}(I),\xi)$

Fig. 19.11 Architecture of our framework: the saliency module comes as a filter before creating the bag of words

points are useful to categorize the image. On the contrary, we assume the idea that nonrelevant “noisy” information can also be detected. Thus, the idea is that attentional system can be used to select among all the keypoints only those which are the most salient. Given the selection of salient keypoints, the rest of Zhang algorithm could stay unchanged for a CBIR application (see Fig. 19.11). In the following, we will refer to Perreira Da Silva system to explain our approach. The same concept can be applied to other computational attention systems.

Practically speaking, the process we propose consists in providing both Zhang and Perreira Da Silva systems the same image $I(x, y)$. After step 2 of Zhang’s framework, a first set of $\mathrm{K}_{\mathrm{Zhang}}(I)$ of keypoints is obtained. In parallel, for the same image $I(x, y)$, Perreira Da Silva’s system provides a saliency map $\mathrm{S}_{\mathrm{map}}(\mathrm{I}, \mathrm{t})$ which evolves with time. In order to emphasize the visual regions the systems mainly focuses on, $\mathrm{S}_{\mathrm{map}}(\mathrm{I}, \mathrm{t})$ is integrated along the time axis to get what is usually known as a “heatmap.” Formally, the heatmap can be defined as $\mathrm{H}(I) = \int_0^T \mathcal{S}_{\mathrm{map}}(\mathrm{I}, \mathrm{t}).dt$, with T the integration window.

To take advantage of the saliency map within the context of Zhang’s framework, the idea is to generate a mask $\mathrm{M}(\mathrm{H}(I), \xi)$ that is used as a filter of the SIFT keypoint set, with the minimum level of saliency considered in the image. Formally, the generated mask could be defined as:

$$\mathrm{M}(\mathrm{H}(I), \xi) = \begin{cases} 1 \ \mathit{if}\ \mathrm{H}(x_h, y_h) > \xi \\ 0 \ \mathit{otherwise} \end{cases} \tag{19.2}$$

The filtering process by itself consists of selecting the subset $\mathrm{K}_{\mathrm{Filtered}}(I)$ of keypoints in $\mathrm{K}_{\mathrm{Zhang}}(I)$ for which the mask $\mathrm{M}(\mathrm{H}(I), \xi)$ is on:

$$\begin{aligned} &\mathrm{K}_{\mathrm{Filtered}}(I(x, y)) = \\ &\left\{ Key_j \in \mathrm{K}_{\mathrm{Zhang}}(I(x_h, y_h)) \mid \mathrm{M}(\mathrm{H}(I), \xi) = 1 \right\} \end{aligned} \tag{19.3}$$

This subset $K_{\text{Filtered}}(I)$ serves as input for the next parts of Zhang algorithm for object recognition. In the following, we will verify if the attentive CBIR can produce a meaningful enhancement in both communities presented in Sects. 19.2.1 and 19.2.2.

19.3.4 More Results

As mentioned in Sect. 19.2.3, experiments have been done to test the utility of attentive CBIR for two communities: visual attention and object recognition. It consists of using the same concept of our approach mentioned in Sect. 19.3.3, where the saliency map was used as mask to detect the most relevant keypoints. Thus, any computational attention system that generates a saliency map can be used in our attentive CBIR methods. In our test, we compare our approach with Zhang's nominal implementation on VOC 2005 database. In this context, we chose Perreira Da Silva as our reference for attentional filter. As the number of keypoints depends on the images, we have chosen to adapt the parameter to the ratio ρ between the number of remaining keypoints $Card(\mathrm{K}_{\text{Filtered}}(I))$ and the number of keypoints in the image $Card(\mathrm{k}_{\text{Zhang}}(I))$.

ρ was varied $\{10\,\%, 20\,\%, 30\,\%, 40\,\%\}$. We were not able to vary ρ over $40\,\%$, since with the minimum value $\xi = 0$ (the whole heatmap is considered), $60\,\%$ of the keypoints were already filtered.

The filtering of the keypoints can be performed independently during learning and for the test process. Therefore, ρ vary on both the learning set (ρ_L) and the training set (ρ_T): our idea was to determine whether more or less keypoints during training or test may affect the effectiveness of our approach. We performed quantitative evaluation of ROC curves by computing the area under curve (AUC) and the equal error rate (EER) following the procedure defined for the challenge in Tables 19.3, 19.4, 19.5, and 19.6. Each table presents the results for each class with, respectively, Zhang's original score as reported in the challenge summary, our implementation of Zhang's algorithm without filtering, and several couples of (ρ_L, ρ_{S_i}) keypoint filtering.

Table 19.3 AUC/EER values for persons class

AUC/EER	S_1			
	Zhang	Reimpl. of Zhang	40 %, 40 %	40 %, 10 %
	0.97/	0.93/	0.92/	0.79/
	0.91	0.87	0.86	0.77
	S_2			
	Zhang	Reimpl. of Zhang	10 %, 30 %	40 %, 10 %
	0.813/	0.67/	0.69/	0.58/
	0.728	0.56	0.62	0.47

Table 19.4 AUC/EER values for cars class

AUC/EER	S_1			
	Zhang	Reimpl. of Zhang	30 %, 40 %	30 %, 10 %
	0.98/	0.95/	0.94/	0.83/
	0.93	0.90	0.87	0.79
	S_2			
	Zhang	Reimpl. of Zhang	30 %, 20 %	10 %, 40 %
	0.802/	0.73/	0.76/	0.61/
	0.720	0.73	0.76	0.44

Table 19.5 AUC/EER values for bikes class

AUC/EER	S_1			
	Zhang	Reimpl. of Zhang	30 %, 40 %	40 %, 10 %
	0.98/	0.94/	0.93/	0.72/
	0.93	0.90	0.86	0.64
	S_2			
	Zhang	Reimpl. of Zhang	10 %, 30 %	40 %, 10 %
	0.813/	0.67/	0.69/	0.58/
	0.728	0.56	0.69	0.44

Table 19.6 AUC/EER values for motorbikes class

AUC/EER	Zhang	Reimpl. of Zhang	40 %, 40 %	40 %, 10 %
	0.99/	0.98/	0.98/	0.89/
	0.96	0.94	0.93	0.83

Table 19.7 Evaluation of computational cost

	Reimpl. of Zhang	40 %	30 %	20 %	10 %
SIFT descriptor	1.52 s	0.6 s	0.36 s	0.27 s	0.15 s
Histogram construction	1.96 s	0.74 s	0.51 s	0.42 s	0.20 s

Observing the results for S_1 shows that reducing about 60 % the number of keypoints does not affect the performance sensibly. For S_2, we obtained a loss in performance for motorbike class. This can be explained that 50 % of the images in $S_{2,M}$ contained two objects defined as classes (motorbike and persons). In addition to that, applying the saliency map on these images had reduced 70 % of points of interest. This loss can be shown in AUC (0.80 as the result of re-implementation of Zhang; 0.40 for the best result we got for (30 %, 10 %)).

The evaluation of the running time of *INRIA-Zhang* is illustrated in Table 19.7. This table reports the average running time obtained by dividing the total running time of each stage by the number of images. All components of our system are implemented in C++ and run on a computer with a 3.06 GHz Intel Core 2 Duo CPU and 4G of RAM. Analyzing the result demonstrated that when we apply our filtering approach using 40 % of the total interest points have given us about 60 % time gain.

19.4 Summary and Conclusion

We show that attentive CBIR can improve the query run time and information quality in object recognition. Therefore, we propose our approach for selecting the most relevant SIFT keypoints according to human perception, using Perreira Da Silva saliency-based region detection system. Testing this approach on VOC 2005 demonstrated that we can maintain approximately the same performance by selecting only 40 % of SIFT keypoints. Based on this result, we propose this approach as a first step to solve problems related to the management of memory and query run time for recognition systems based on SIFT detectors.

Furthermore, our approach can be used as a new kind of evaluation framework for visual attention models [47]. This framework aims to evaluate the ability of visual attention systems to maintain the performance of a CBIR approach. As shown in Fig. 19.12, it is complicated to compare the different systems as their behaviors vary according to the percentage of keypoint reduction. Moreover, some of the state-of-the-art models shown here have different dynamics function of τ, some of them not being able to cover the entire 0 % to 100 % keypoint reduction range. In [47], we show that the use of an eye-tracking ground truth or a CBIR-based ground truth can

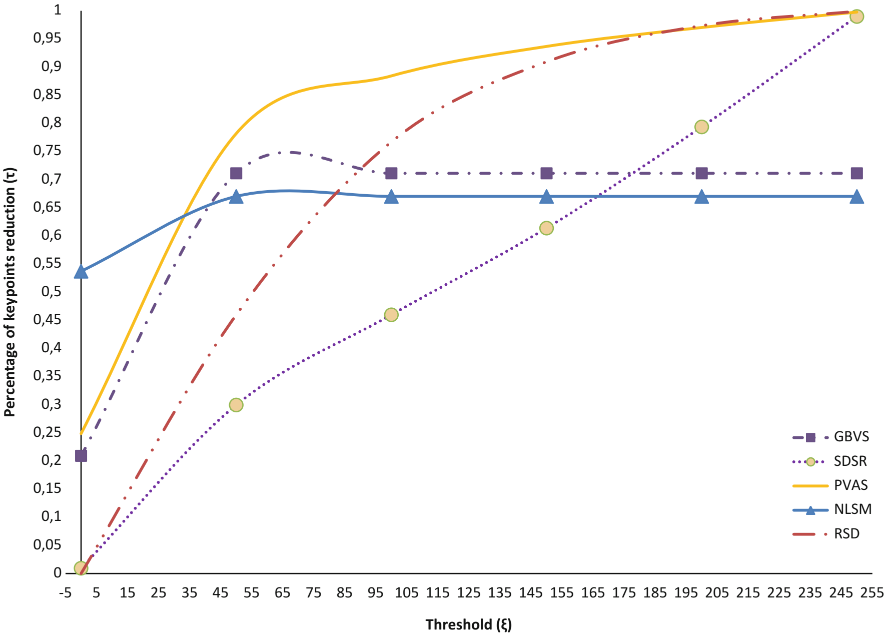


Fig. 19.12 Percentage (τ) of keypoint reduction for several attention models: graph-based visual saliency (GBVS) [48], saliency detection by self-resemblance (SDSR) [49], prey/predator visual attention system (PVAS) [31], nonparametric low-level saliency model (NLSM) [50], and rarity-based saliency detection [51] on VOC 2007. Compared with PVAS that we use here

provide different ranking between the attention models. In other words an attention model which is efficient in predicting eye gaze using a traditional ground truth is not necessarily efficient in improving object recognition. The use of a CBIR-based framework as the one proposed in this chapter is thus very interesting for people who want to choose the best saliency model given their CBIR application.

References

1. Walther, D., Rutishauser, U., Koch, C., & Perona, P. (2005). Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding, 100*(1–2), 41–63.
2. Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.
3. Itti, L., & Koch, C. (2001). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging, 10*, 161–169.
4. Hubel, D. H., & Wiesel, T. N. (2004). *Brain and visual perception: The story of a 25-year collaboration* (Vol. 31). New York: Oxford University Press.
5. Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 136*(12), 97–136.
6. Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207.
7. Tuytelaars, T., & Mikolajczyk, K. (2007). Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision, 3*(3), 177–280.
8. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.
9. Everingham, M., Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2009). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision, 88*(2), 303–338.
10. Awad, D., Courboulay, V., & Revel, A. (2012). Saliency filtering of sift detectors: Application to cbir. *Advanced concepts for intelligent vision systems*, 350 (Vol. 7517 of LNCS, pp. 290–300). Berlin/New York: Springer.
11. Dave, A., Dubey, R., & Ghanem, B. (2012). Do humans fixate on interest points? In *Pattern Recognition (ICPR)* (pp. 2784–2787).
12. Foo, J. J. (2007). Pruning SIFT for scalable near-duplicate image matching. In *Australasian Database Conference*, Ballarat, p. 9.
13. Alhwarin, F., Ristić-Durrant, D., & Gräser, A. (2010). Vf-sift: Very fast sift feature matching. In *Proceedings of the 32Nd DAGM Conference on Pattern Recognition* (pp. 222–231). Berlin/Heidelberg: Springer.
14. Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*(4), 219–227.
15. Itti, L., Koch, C., Niebur, E., & Others (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.
16. Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(5), 802–817.
17. Frintrop, S. (2005). VOCUS: A visual attention system for object detection and goal-directed search, Phd thesis, accepted at the University of Bonn in July 2005 (Lecture notes in artificial intelligence (LNAI), Vol. 3899/2006). Berlin/Heidelberg: Springer. ISBN: 3-540-32759-2.
18. Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America A, 20*(7), 1407–1418.

19. Oliva, A., Torralba, A., Castelhano, M.S., & Henderson, J.M. (2003). *Top-Down Control of Visual Attention in Object Detection*, Proceedings of the IEEE International Conference on Image Processing. Vol. I, pages 253-256; September 14-17, in Barcelona, Spain.
20. Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49*(10), 1295–1306.
21. Gao, D. G. D., Han, S. H. S., & Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(6), 989–1005.
22. Gu, E., Wang, J., & Badler, N. I. (2007). Generating sequence of eye fixations using decision-theoretic attention model. In *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint* (pp. 277–292). Berlin/Heidelberg: Springer.
23. Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. *Advances in Neural Information Processing Systems*, (18), 155–162.
24. Salah, A. A., Alpaydin, E., & Akarun, L. (2002). A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 420–425
25. Rao, R. P. N. (2005). Bayesian inference and attentional modulation in the visual cortex. *NeuroReport, 16*(16), 1843–1848.
26. Liu, Y. J., Luo, X., Xuan, Y. M., Chen, W. F., & Fu, X. L. (2011). Image retargeting quality assessment. *EUROGRAPHICS, 30*(2).
27. Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. *2007 IEEE Conference on Computer Vision and Pattern Recognition, 1*(800), 1–8.
28. Peters, R. J., & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis (pp. 1–8).
29. Kienzle, W., Franz, M. O., Schölkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision, 9*(5), 7.1–15.
30. Ramström, O., & Christensen, H. I. (2002). Visual attention using game theory. In *Proceedings of the Second International Workshop on Biologically Motivated Computer Vision* (BMCV'02), London (pp. 462–471). Springer.
31. Perreira Da Silva, M., Courboulay, V., & Estraillier, P. (2011). Objective validation of a dynamical and plausible computational model of visual attention. In *IEEE European Workshop on Visual Information Processing*, Paris (pp. 223–228).
32. Riche, N., Duvinage, M., Mancas, M., Gosselin, B., & Dutoit, T. (2013). Saliency and human fixations: State-of-the-art and study of comparison metrics. *Proceedings of IEEE 13th International Conference on Computer Vision* (pp. 1153–1160). Sydney, Australia.
33. Perreira Da Silva, M., Courboulay, V., Prigent, A., & Estraillier, P. (2010). Evaluation of preys/predators systems for visual attention simulation. In *VISAPP 2010 – International Conference on Computer Vision Theory and Applications*, Angers (pp. 275–282). INSTICC.
34. Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.
35. Frintrop, S., & Jensfelt, P. (2008). Attentional landmarks and active gaze control for visual slam. *IEEE Transactions on Robotics, 24*(5), 1054–1065.
36. Frintrop, S. (2011). Towards attentive robots. *Paladyn. Journal of Behavioral Robotics, 2*, 64–70. doi:10.2478/s13230-011-0018-4.
37. Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review, 95*(1), 15–48.
38. Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, 1*(2), 202–238.
39. Perreira Da Silva, M., Courboulay, V., Prigent, A., & Estraillier, P. (2010). Evaluation of preys/predators systems for visual attention simulation. In P. Richard & J. Braz (Eds.), *VISAPP 2010 – International Conference on Computer Vision Theory and Applications*, Angers (Vol. 2, pp. 275–282). INSTICC.
40. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision, 88*(2), 303–338.

41. Sivic, J., Russell, B.C, Efros, A. A, Zisserman, A., & Freeman, W. T. (2005). Discovering objects and their location in images. *In Proceedings of IEEE International Conference on Computer Vision ICCV*, Beijing.
42. Kondor, R., & Jebara, T. (2003). A kernel between sets of vectors. In *Machine Learning: Tenth International Conference*, Washington, DC.
43. Laaksonen, J. (2000). PicSOM – content-based image retrieval with self-organizing maps. *Pattern Recognition Letters, 21*(13–14), 1199–1207.
44. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.
45. Mikolajczyk, K. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision, 60*(1), 63–86.
46. Lindeberg, T. (1998). Feature detection with automatic scale selection. *Computer Vision, 30*(2), 96.
47. Awad, D., Mancas, M., Riche, N., Courboulay, V., & Revel, A. (2015). A cbir-based evaluations framework for visual attention models. In *23rd European Signal Processing Conference (EUSIPCO)*, Nice.
48. Koch, C., Harel, J., & Perona, P. (2006). Graph-based visual saliency. In *Proceedings of Advances in neural information processing systems (NIPS)*, Vancouver.
49. Seo, H. J., & Milanfar, P. (2009). Static and spacetime visual saliency detection by self-resemblance. *Journal of Vision, 9*(12), 1–27.
50. Murray, N., Vanrell, M., Otazu, X., & Alejandro Parraga, C. (2011). Saliency estimation using a nonparametric low-level vision model. In *Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs (pp. 433–440).
51. Riche, N., Mancas, M., Gosselin, B., & Dutoit, T. (2012). Rare: A new bottom-up saliency model. In *Proceedings of the IEEE International Conference of Image Processing (ICIP)*, Lake Buena Vista.

Chapter 20
Saliency and Attention for Video Quality Assessment

Dubravko Culibrk

20.1 Introduction

Recent years have witnessed an explosion of visual and multimedia applications across the globe. Digital television and other home entertainment applications; multimedia sharing platforms such as YouTube, Flickr, and Panoramio; social networking applications such as Facebook; mobile multimedia applications; personal multimedia collections; immersive multimedia and virtual reality applications; video tele-conferencing; gaming; and educational multimedia presentations are all examples of multimedia applications that have become an integral part of our lives.

Our increased ability to transmit large amounts of data, through both land and wireless networks, has led to rising increase in the quality of experience that people get and expect from multimedia applications and services. Since the consumers of the multimedia content are ultimately humans, it is vital to be able to efficiently measure and detect the quality of the final product delivered to the users for large amounts of multimedia data. The problem is made more complex by the increasing share of multimedia generated by users, as opposed to the professional media-producing companies. At the time of writing, 300 h of video are uploaded to YouTube (the leading Internet video site) every minute. In fact the amount of video uploaded to YouTube in a month surpasses all the content produced by three

This work supported in part by FP7 QoSTREAM (no. 295220) and COST SoftStat IC0702 projects.

D. Culibrk (✉)
Department of Information Engineering and Computer Science, University of Trento,
Via Sommarive 5, Povo-Trento, Italy
e-mail: dculibrk@uns.ac.rs

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_20

major US networks in 60 years [42]. The Cisco Visual networking index [2] predicts that in 2019, the gigabyte equivalent of all movies ever made will cross Global IP networks every 2 min and that video will account for 80 % of all IP traffic. Global mobile IP traffic will comprise 14 % of total IP traffic in 2019, a sharp rise from 4 % in 2014. User-generated content and mobile applications impose new level of complexity when quality assessment (QA) is concerned, since access to the original (pristine) media content, which most mature methodologies require as a reference [27], is not available in these scenarios.

While aspects of the human visual system (HVS) have been modeled to arrive at an estimate of the perceived level of coding artifacts in video sequences, attention and saliency in videos have only recently begun to be considered as a way to enhance video quality assessment (VQA) [8, 21, 27]. Bottom-up attention can be modeled computationally [12] and has been successfully used in a number of applications such as content-based image retrieval [25], scene classification [28], and vision-based localization [29]. Lately, researchers have started looking into using computational models of (motion) attention to enhance the performance of video coding algorithms [22], address the problem of video skimming [23, 24], and improve VQA [7, 21, 36].

When VQA is concerned, the motivation for taking attention into account lies in the fact that the HVS sensitivity to motion and texture differs significantly between areas of the stimuli focused upon (attended to) and those in peripheral vision [30]. This leads to different sensitivity to coding artifacts in the two regions of the visual field, which has rarely been taken into account in the QA algorithm design.

The rest of this chapter is organized as follows: Sect. 20.2 deals with subjective tests that need to be carry out to establish the quality as perceived by the users. Section 20.3 describes the mechanism through which coding affects the quality. Section 20.4 gives an overview of video quality assessment. Section 20.5 focuses on saliency-based VQA. Finally, Sect. 20.6 holds the concluding remarks.

20.2 Measuring Subjective (Perceptual) Quality

As the consumers of multimedia content are humans, a measure of quality related to viewers' quality of experience (QoE) is needed to establish "ground truth" for any quality assessment methodology.

A subjective quality measure typically used is the mean opinion score (MOS), which is obtained by averaging scores from a number of human observers. The correct procedure for conducting such experiments is described in ITU-R BT.500-10 recommendations[15].

These recommendations encompass a number of different procedures designed for different scenarios. If the reference (pristine) material is available, one might opt for Double Stimulus Impairment Scale (DSIS) method. The assessor is first presented with an unimpaired, reference sequence and then with the same sequence impaired. He is then asked to vote on the second sequence, keeping in mind the first. Voting is done on a 1 to 5 scale, 1 being the lowest score, where perceived impairments are very annoying, and 5 being the highest, where impairments cannot be perceived. Series of sequences with random levels of impairments are presented, and for control purposes, reference sequences are also included in the assessment set, but assessors are not informed about this. The final MOS value for a sequence is the average score over all assessors for that sequence.

If there is no information about the original video, Absolute Category Rating (ACR) can be used. In this method, test clips are presented to assessors one at a time and rated independently on a discrete 9-level scale, ranging from “Bad” to “Excellent.” Naturally, the ratings for each test clip are then averaged over all subjects to obtain a mean opinion score (MOS). It should be noted that the variance of the opinion scores is much higher when ACR is used [6] and the procedure accommodates for this effect by including two sessions. Subjects grade the same sequence twice, but at different points in time (once per session). Thus, intra-subject reliability as well as intersubject variability could be measured. This allows for unreliable observers to be eliminated from the final MOS scores, using a paired t-test.

When it comes to the video material usually used, both to measure MOS and evaluate VQA approaches, a golden standard is the Video Quality Experts Group (VQEG) FRTV Phase 1 database [35]. This data set contains standard definition sequences and is primarily intended for purposes of testing the quality of video codecs. Most authors use VQEG videos, either exclusively or as part of a larger corpora of video sequences to evaluate the effect of different impairments on the perceived quality. An extensive and fairly recent list of VQ resources and databases is provided by Winkler [39]. A number of those databases contain eye-tracking data, which is of particular interest to the saliency and attention VQA researchers.

20.3 Coding Artifacts and Their Relation to Perceptual Quality

The perceived quality depends on the video codec, bit rates required, and the content of video material. User-oriented video quality assessment research is aimed at providing means to monitor the perceptual quality of the service.

Overall degradation in the quality of the sequence is due to the encoder/decoder implementations as part of the transport stream at various bit rates and is a compound effect of different coding artifacts and/or packet losses that occur during transmission. Figure 20.1 shows the effect of coding on details taken from the frames of two VQEG test sequences, for MPEG-2 coder and two bit rates (0.5 and 2 Mbps).

Three types of artifacts are typically considered pertinent to DCT block (e.g., JPEG, MPEG, H.264) coded data: blocking, ringing, and blurring. Blocking appears in all block-based compression techniques due to the coarse quantization of frequency components [37, 38]. It can be observed as surface discontinuity

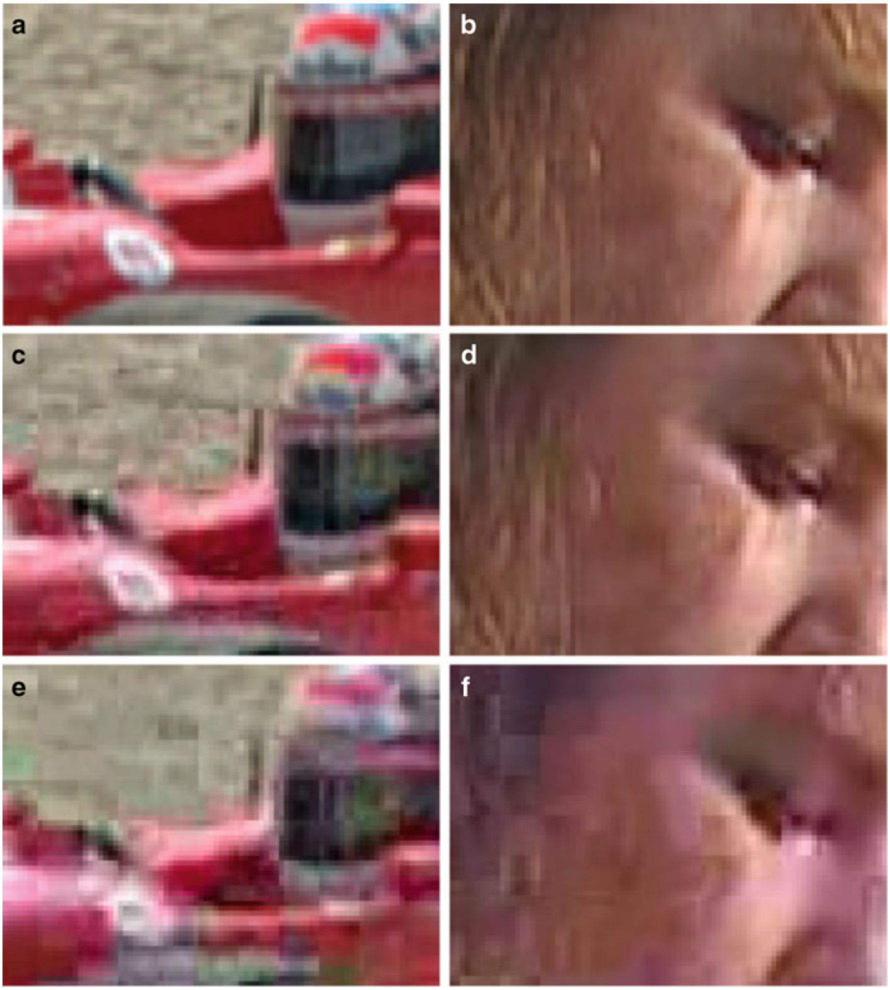


Fig. 20.1 Details of sample frames taken from two VQEG sequences degraded through MPEG-2 coding. (**a**) Sample frame detail (original). (**b**) Sample frame detail (original). (**c**) Detail coded at 2 MB/s. (**d**) Detail coded at 2 MB/s. (**e**) Detail coded at 0.5 MB/s. (**f**) Detail coded at 0.5 MB/s

(edge) at block boundaries. These edges are perceived as abnormal high-frequency components in the spectrum. Ringing is observed as periodic pseudo edges around the original edges [18]. It is due to the improper truncation of the high frequency components. This artifact is also known as the Gibbs phenomenon or Gibbs effect. In the worst case, the edges can be shifted far away from the original edge locations and observed as false edge. Blurring, which appears as edge smoothness or texture blur, is due to the loss of high-frequency components when compared with the original image. Blurring causes the received image to be smoother than the original one [9].

20.4 Video Quality Assessment

Three classes of approaches to image and video quality assessment exist:

- Full reference
- Reduced reference
- No-reference approaches

The distinction is made based on the level of information about the original sequence that is used by the QA methodology. Full-reference approaches require access to the original multimedia content that has not been degraded through coding. As such, they are suitable for use in the classical multimedia production scenario, where the content provider (i.e., a television network) has the access to the pristine signal and the technical capability to store/process it. Thus, the QA methodology is primarily used to optimize the QoE of viewers through the tuning of coding algorithms at the production stage. Seshadrinathan and Bovik provide a very recent survey [27] of full-reference QA approaches. The focus has been placed on such methodologies, as they are most mature and have been commercialized to the greatest extent.

Full-reference quality assessment is a fairly easier problem than the other two classes and has indeed been the focus of most researchers up to this point. Unfortunately, the methodology is quite unsuitable for application in the case of large amounts of user-generated content, since the not-coded versions of images and videos are not available with the widely used production technology (i.e., cameras and camcorders). In addition, if one envisions a scenario where the QA needs to be done on mobile platforms, then there is no question of being able to send uncompressed images and videos to the mobile device and use it as a reference.

Reduced-reference approaches find themselves between the full-reference and no-reference approaches and attempt to use lower-bandwidth information about the original sequence/image. Again, one needs to have access to the pristine content to extract the reference information, and this will be less and less likely in the multimedia landscape of the future.

20.5 Saliency-Based VQA

Saliency-based VQA approaches typically share the same basic structure, illustrated in Fig. 20.2.

Different models and features are used to compute a saliency map for each frame of the sequence. In parallel, conventional quality metrics (designed to measure the degradation in quality due to different coding artifacts) are computed for each frame. The saliency map is then used to weigh the importance of different parts of the sequence/frame in terms of quality and provide a better measure of quality than

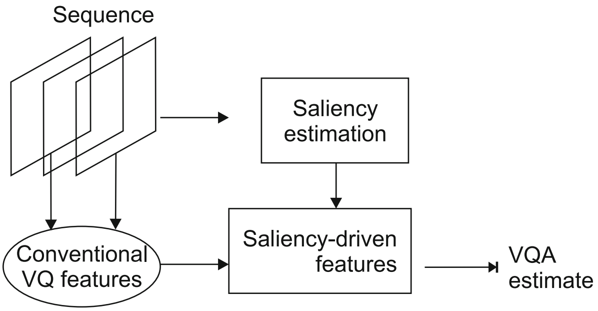


Fig. 20.2 Structure of saliency-based VQA approaches

one that could be estimated without the saliency model. MOS estimates for single frames are then aggregated to arrive at a single estimate per sequence.

Where the approaches differ is in the approach to saliency computation, the quality metrics employed, and the way the information from the quality metrics and the saliency map is fused to arrive at a MOS estimate.

20.5.1 Measuring the Effect of Coding Artifacts on Quality

The origins of research into the influence of coding artifacts on perceptual quality lie in the still-image quality (IQ) domain. Most metrics used in various VQA approaches are derived from IQ metrics.

Even when the reference (not degraded) visual content is available, objective measures of signal degradation such as Peak Signal-to-Noise Ratio (PSNR) are poorly correlated to MOS [43], leading to significant research effort aimed at the design of measures which will allow computers to determine MOS effectively. The measures typically focus on specific coding artifacts and attempt to take into account the effect of the content of the images (video frames). Thus, when perceived blockiness is concerned, most measures are based on the notion that the block-edge-related effects can be masked by high spatial activity in the image itself and that the blockiness cannot be observed in very bright and very dark regions.

While it is not feasible to enumerate all published approaches to the measuring of the different coding effects, several methodologies stand out in terms of their impact on the community and/or are of interest for the discussion in the following sections. For a recent review of the approaches to image and video quality assessment, which attempt to make use of different models of attention deployment, please refer to Engelke et al. [8].

Wang et al. [37] proposed a no-reference approach to quality assessment in JPEG-coded images. Their final measure is derived as a nonlinear combination of a blockiness, local activity, and a so-called zero-crossing measure. The combination is supposed to provide information regarding both blockiness and blurring (via the two latter measures) in JPEG-coded images. Their approach is usually compared against, when no-reference MOS estimation is concerned.

More recently, Babu et al. [34] proposed a blockiness measure for use in VQA, which takes effects along each edge of the block into account separately. They report their measure surpassing the Wang et al. approach in terms of MOS prediction accuracy.

In a recent paper [5], a neural-network approach to MOS estimation was used to evaluate a number of measures (18 in total) of image and video quality in terms of their predictive value, when VQA is concerned. The evaluated measures included still-image quality features designed to describe spatial image activity and contrast [10, 19], ringing effects [18], noise, and blur [17]. Wang et al. proposed measures of quality and Babu et al. blockiness measure. Additional measures were introduced to account for the temporal dynamics of the sequence. Two motion intensity measures were used: (i) global motion intensity, calculated from the global motion field, and (ii) object motion intensity, calculated by subtracting the global motion from the MPEG motion vectors [38].

Feature selection was performed based on training a simple multilayer perceptron (MLP) estimator with each measure as input, separately. The measures were ranked according to their performance and a subset of 5 measures was selected as input for the final estimator, which was an MLP with 7 nodes in the hidden layer. Since the prediction was done on a single frame basis, median filtering was used to arrive at a single estimate for the whole sequence. The approach achieved better results than any measure considered separately. Feature selection indicated that measures proposed by Wang et al. and Babu et al. contributed most to quality estimation.

20.5.2 Saliency, Motion, and Attention

When faced with visual stimuli, the human vision system (HVS) does not process the whole scene in parallel. Part of the visual information sensed by the eyes is discarded in a systematic manner to attend to objects of interest. The most important function of selective visual attention is to direct our gaze rapidly towards objects of interest in our visual environment [13, 26]. The objects that are not of interest are still processed, but with reduced spatial resolution and motion sensitivity [30]. Critical fusion frequency, on the other hand, is higher in the peripheral vision, making the HVS more sensitive to sudden changes in illumination in the not-attended region [1].

It is not possible for the HVS to process an image entirely in parallel. Instead, our brain has the ability to prioritize the order; the potentially most important points are attended to when presented within a new scene. The result is that much of the visual information our eyes sense is discarded. Despite, we are able to quickly gain remarkable insight into a scene.

This type of attention is referred to as attention for perception: the selection of a subset of sensory information for further processing by another part of the information processing system [25, 32].

Current research considers attention deployment as a two-component mechanism [3, 13]. Subjects selectively direct attention to objects in a scene using both bottom-up, image-based saliency cues and top-down, task-dependent cues, an idea that dates back to the nineteenth-century work of William James [16].

Bottom-up processing is driven by the stimulus presented [32]. Some stimuli are intrinsically conspicuous or salient (outliers) in a given context. For example, a red dinner jacket among black tuxedos at a somber state affair, a flickering light in an otherwise static scene, or a street sign against gray pavement automatically and involuntarily attracts attention. Saliency is independent of the nature of the particular task, operates very rapidly, and is primarily determined in a bottom-up manner. If a stimulus is sufficiently salient, it will pop out of a visual scene. This suggests that saliency is computed in a pre-attentive manner across the entire visual field, most probably in terms of hierarchical *center-surround mechanisms*. As for the moving stimuli, they are perceived to be moving only if they are undergoing motion different from their wider surround [26]. The speed of saliency-based form of attention is on the order of 25 to 50 ms per item [13]. The second form of attention is a more deliberate affair and depends on the task at hand, memories, and even past experience [32]. Such intentional deployment of attention has a price, because the amount of time that it takes (200 ms or more) rivals that needed to move the eyes. Thus, certain features in the visual world automatically attract attention and are experienced as "visually salient." Directing attention to other locations or objects requires voluntary "effort." Both mechanisms can operate in parallel.

Significant progress has been made in terms of computational models of bottom-up visual attention [11, 20, 31, 33]. While bottom-up factors that influence attention are well understood [41], the integration of top-down knowledge into these models remains an open problem. Because of this, the fact that bottom-up components of a scene influence our attention before top-down knowledge does [3] and that they can hardly be overridden by top-down goals, applications of visual attention commonly rely on bottom-up models [24, 25, 28, 29].

Full-fledged biologically inspired computational models of attention are primarily designed for still images and are too computationally intensive for real-world video processing applications such as video quality assessment [21] and video skimming [24]. In the case of VQA, this is especially true if a large number of features are calculated based on the output of the visual attention model.

Nevertheless, the complex saliency model of Itti et al. has recently been employed to improve the prediction of packet loss effects [21]. In addition to weighing their conventional quality measures with the computed saliency map, the authors observed that the changes in the saliency map can help when determining perceptual quality. Therefore, they proposed the use of mean squared error between the saliency map computed for the reference frame and the impaired frame and temporal variance of saliency map as features for quality estimation. Using generalized linear models as VQ estimators, they concluded that the results are improved significantly when using saliency-based features (the error improved by 9–15 %, depending on the number of features used).

Ma et al. [24] made an early attempt to design an attention model that integrates motion cues and achieves the performance required in real-time video applications. Ma et al. distinguish motion and static attention parts of their model, since they rely on previously calculated motion vector field to discern the regions of the frame salient due to motion. They propose measures based on motion intensity, spatial and temporal coherence to detect points salient due to motion, and contrast to determine static saliency. It should be noted that the spatial coherency of motion seems to have no bearing on saliency at the lowest levels of attention [26]. Ölveczky et al. [26] report that the driving force of the attention at this level is the difference in the speed of motion between a center and the surrounding region.

Nevertheless, the approach of Ma et al. used general principles of the visual attention in the HVS to drive the design of a lightweight attention model. However, their intended application was video abstraction, and they have not considered using their model to help assess perceptual quality.

More recently, several methodologies for computing saliency due to spatial and temporal cues (motion) were proposed. In 2007, Wang and Li [36] extended their full-reference still image quality assessment methodology by incorporating a speed perception model. They rely on measures of background global and relative motion to create a saliency map describing the perceptual uncertainty for different parts of the frame. Their assumption is that the frames with large global motion which exhibit the largest perceptual uncertainty will not influence the perceived quality much. Based on the perceptual uncertainty map, they derive a weighing map for their quality metric and propose a nonlinear equation to arrive at the MOS estimate.

In [7], the authors proposed the use of a multi-scale background modeling and foreground segmentation approach proposed in [4] as an efficient attention model driven by both motion and static cues, which adheres to the principles reported in [26]. The model employs the principles of multi-scale processing, cross-scale motion consistency, outlier detection, and temporal coherence. The output of the segmentation has been used to derive features describing the salient motion in the frame, as well as to calculate a number of video quality features separately for regions of the frame observed as salient and the rest of the frame. This enabled the evaluation of the influence of the saliency on the predictive ability of the proposed VQ estimators.

The list of measures considered as features for VQA in [7] is shown in Tables 20.1 and 20.2. The measures shown in Table 20.1 represent conventional VQA measures adopted by the authors from previous published work. The measures in Table 20.2 are related to saliency and were proposed in the paper.

As Tables 20.1 and 20.2 show, rather than using the saliency information as weight, the authors considered separately the salient/non-salient regions and the border between the two. Figure 20.3 shows salient motion segmentation masks for sample frames as well as saliency maps detected by the static-saliency approach of Itti and Koch [13].

Table 20.1 Initial list of measures evaluated in [7] with pertinent references

#	Feature	Reference
1	Two field difference	[40]
2	Variance ratio	[17]
3	Blockiness	[34]
4	Ringing	[18]
5	Ringing 2	[18]
6	Global motion vector intensity	[38]
7	Activity	[37]
8	Blocking effect	[37]
9	Zero-crossing rate	[37]
10	Z score	[37]
11	Gradient activity	[19]
12	Edge activity	[19]
13	Contrast	[19]
14	Correlation	[10]
15	Energy	[10]
16	Homogeneity	[10]
17	Variance	[10]
18	Contrast	[10]

Table 20.2 List of saliency-related measures proposed in [7] with pertinent references

#	Feature	Reference
19	Salient reg. count	Proposed
20	Avg. reg. size	Proposed
21	Mean change non-salient	Proposed
22	Change Std.Dev. non-salient	Proposed
23	Mean Change salient	Proposed
24	Change Std.Dev. salient	Proposed
25	Activity non-salient	Modified [37]
26	Blocking effect non-salient	Modified [37]
27	Zero-crossing rate non-salient	Modified [37]
28	Z score non-salient	Modified [37]
29	Activity salient	Modified [37]
30	Blocking effect salient	Modified [37]
31	Zero-crossing rate salient	Modified [37]
32	Z score salient	Modified [37]
33	Blockiness non-salient	Modified [34]
34	Blockiness salient	Modified [34]
35	Blockiness border	Modified [34]

The study concluded that significant improvements in the quality estimation can be achieved when using saliency information. In addition, the study considered the impact of the presence of specific artifacts in the salient/non-salient parts of the sequence on the perceived quality. The authors suggest that the intensity of the

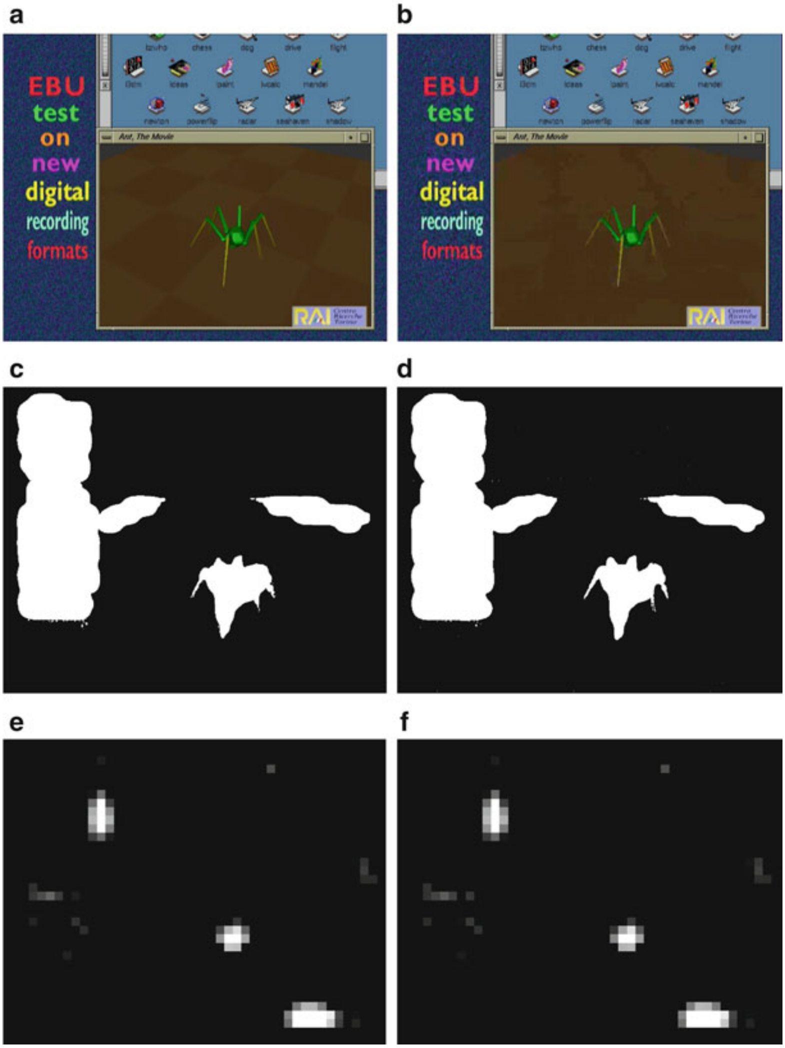


Fig. 20.3 Salient motion detection for the "Ant" VQEG sequence. (**a**) Sample frame coded at 4 MB/s. (**b**) Sample frame coded at 0.5 MB/s. (**c**) Salient motion detected at 4 MB/s. (**d**) Salient motion detected at 0.5 MB/s. (**e**) Static saliency [14] at 4 MB/s. (**f**) Static saliency [14] at 0.5 MB/s

blurring and blocking effects in the salient regions has most bearing on the perceived video quality. On the other hand, temporal changes in the non-salient part of the frame seem to be of more importance when VQA is concerned. This is a significant parting from other methodologies which assume that saliency of a region is directly related with the significance of artifacts present in that region.

Moreover, the results presented in the study show that the root mean square error of the proposed estimator is significantly lower than the mean standard deviation of the opinion scores of the human observers in the subjective tests conducted. That is, the proposed methodology provides more accurate and consistent estimates than those obtained from subjective experiments.

20.6 Conclusion

Recent years have witnessed an explosion of visual and multimedia applications across the globe. This proliferation is accompanied by the shift towards user-generated content in the media production and mobile wireless devices in terms of media production/consumption platforms.

Such developments put an emphasis on the need to produce no-reference VQA approaches that will allow for real-time processing of video, using devices with limited computing power.

Within the VQA research community, saliency-based approaches are gaining interest, as the saliency information is seen as a potent way to increase the accuracy and make the VQA methodologies more general. Early studies presented in this chapter give ground to such optimism.

Methods that allow real-time saliency extraction in video and models of attention that incorporate motion have not achieved the stage of maturity of still-image processing approaches, at the time of writing, making spatial-and-motion-based bottom-up attention models an interesting field of future research.

If one envisions the future in which a mobile device should provide real-time QoE feedback to the broadcasting entity, then we should assume that lightweight VQA approaches will gain significance in the years to come. Visual saliency can help achieve this goal in much the same way it helps optimize the way in which HVS resources are used.

References

1. Brooke, R. (1951). The variation of critical fusion frequency with brightness at various retinal locations. *JOSA, 41*(12), 1010–1016.
2. Cisco. (2015). Cisco visual networking index: Global – 2019 forecast, San Jose.
3. Connor, C., Egeth, H., & Yantis, S. (2004). Visual attention: Bottom-up versus top-down. *Current Biology, 14*(19), R850–R852.
4. Culibrk, D., Crnojevic, V., & Antic, B. (2009). Multiscale background modelling and segmentation. In *Proceedings of the 16th International Conference on Digital Signal Processing*, Chicago, Santorini, Greece (pp. 922–927).
5. Culibrk, D., Kukolj, D., Vasiljevic, P., Pokric, M., & Zlokolica, V. (2009). Feature selection for neural-network based no-reference video quality assessment. In *ICANN* (2). (pp. 633–642).
6. Culibrk, D., Mirkovic, M., Lugonja, P., & Crnojevic, V. (2010). Mining web videos for video quality assessment. In *2010 International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, Paris (pp. 75–80). doi: 10.1109/SOCPAR.2010.5686400, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5686400&isnumber=5685841

7. Culibrk, D., Mirkovic, M., Zlokolica, V., Pokric, M., Crnojevic, V., & Kukolj, D. (2010). Salient motion features for video quality assessment. *IEEE Transactions on Image Processing, 20*, 948–958.
8. Engelke, U., Kaprykowsky, H., Zepernick, H. J., & Ndjiki-Nya, P. (2011). Visual attention in quality assessment. *IEEE Signal Processing Magazine, 28*(6), 50–59.
9. Ferzli, R., & Karam, L. A no-reference objective image sharpness metric based on just-noticeable blur and probability summation. Proceedings of IEEE 2007 International Conference on Image Processing 3, III –445–III –448 (16 2007-Oct 19 2007)
10. Idrissi, N., Martinez, J., & Aboutajdine, D. (2005). *Selecting a discriminant subset of co-occurrence matrix features for texture-based image retrieval*. (pp. 696–703). Advances in visual computing. Berlin/Heidelberg: Springer.
11. Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing, 13*(10), 1304–1318.
12. Itti, L., & Baldi, P. F. (2009). Bayesian surprise attracts human attention. *Vision Research, 49*(10), 1295–1306.
13. Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194–203.
14. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.
15. ITU-R BT.500. (2002). Methodology for the Subjective Assessment of the Quality of Television Pictures. Video Quality Experts Group.
16. James, W. (1950). *The principles of psychology* (Vol. 1). Dover Publications. http://www.worldcat.org/isbn/0486203816
17. Kim, K., & Davis, L. (2004). A fine-structure image/video quality measure using local statistics. In *Proceedings of IEEE 2004 International Conference on Image Processing*, Singapore (Vol. V, pp. 3535–3538).
18. Kirenko, I. (2006). Reduction of coding artifacts using chrominance and luminance spatial analysis. In *International Conference on Consumer Electronics, 2006. ICCE '06. 2006 Digest of Technical Papers*, St. Petersburg, Las Vegas (pp. 209–210).
19. Kusuma, T., Caldera, M., & Zepernick, H. (2004). Utilising objective perceptual image quality metrics for implicit link adaptation. In *Proceedings of IEEE 2004 International Conference on Image Processing*, Singapore (Vol. IV, pp. 2319–2322).
20. Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(5), 802–817
21. Liu, T., Feng, X., Reibman, A., & Wang, Y. (2009). Saliency inspired modeling of packet-loss visibility in decoded videos. In *International Workshop VPQM*, Scottsdale (pp. 1–4).
22. Liu, Z., Yan, H., Shen, L., Wang, Y., & Zhang, Z. (2009). A motion attention model based rate control algorithm for h.264/avc. In *Eighth IEEE/ACIS International Conference on Computer and Information Science*, Shanghai (pp. 568–573).
23. Longfei, Z., Yuanda, C., Gangyi, D., & Yong, W. (2008). A computable visual attention model for video skimming. In *ISM '08: Proceedings of the 2008 Tenth IEEE International Symposium on Multimedia* (pp. 667–672). Washington, DC: IEEE Computer Society.
24. Ma, Y. F., Hua, X. S., Lu, L., & Zhang, H. J. (2005). A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia, 7*(5), 907–919.
25. Marques, O., Mayron, L. M., Borba, G. B., & Gamba, H. R. (2007). An attention-driven model for grouping similar images with image retrieval applications. *EURASIP Journal on Advances in Signal Processing, 2007*, 116
26. Olveczky, B. P., Baccus, S. A., & Meister, M. (2003). Segregation of object and background motion in the retina. *Nature, 423*, 401–408.
27. Seshadrinathan, K., & Bovik, A. (2011). Automatic prediction of perceptual quality of multimedia signals – a survey. *Multimedia Tools and Applications, 51*, 163–186.

28. Siagian, C., & Itti, L. (2007). Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(2), 300–312.
29. Siagian, C., & Itti, L. (2009). Biologically inspired mobile robot vision localization. *IEEE Transactions on Robotics, 25*(4), 861–873.
30. Solomon, J., & Sperling, G. (1995). 1st-and 2nd-order motion and texture resolution in central and peripheral vision. *Vision Research, 35*(1), 59–64.
31. Stentiford, F. W. (2003). An attention based similarity measure with application to content-based information retrieval. In *Proceedings of the Storage and Retrieval for Media Databases Conference, SPIE Electronic Imaging*, Santa Clara
32. Styles, E. A. (2005). *Attention, perception, and memory: An integrated introduction*. New York: Taylor & Francis/Routledge.
33. Tsotsos, J. K., Culhane, S. M., Winky, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence, 78*(1–2), 507–545. http://dx.doi.org/10.1016/0004-3702(95)00025-9
34. Venkatesh Babu, R., Perkis, A., & Hillestad, O. (2008). Evaluation and monitoring of video quality for UMA enabled video streaming systems. *Multimedia Tools and Applications, 37*(2), 211–231.
35. Video Quality Experts Group (VQEG). (2000). Final report from the Video Quality Experts Group on the validation of objective quality metrics for video quality assessment. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx (online).
36. Wang, Z., & Li, Q. (2007). Video quality assessment using a statistical model of human visual speed perception. *JOSA A, 24*(12), B61–B69.
37. Wang, Z., Sheikh, H. R., & Bovik, A. C. (2002). No-reference perceptual quality assessment of jpeg compressed images. In *Proceedings of IEEE 2002 International Conferencing on Image Processing*, Rochester (pp. 477–480).
38. Warwick, G., & Thong, N. (2004). Classification of video sequences in MPEG domain. In *Signal Processing for Telecommunications and Multimedia* (Chapter 6). New York: Springer. http://link.springer.com/book/10.1007%2Fb99846
39. Winkler, S. (2012). Analysis of public image and video databases for quality assessment. *IEEE Journal of Selected Topics in Signal Processing, 6*(6), 616–625.
40. Wolf, S., & Pinson, M. (2002). Ntia report 02-392: Video quality measurement techniques. Technical report, Institute for Telecommunication Sciences. http://www.its.bldrdoc.gov/pub/ntia-rpt/02-392/
41. Wolfe, J. M. (2000). Visual attention. In *Seeing* (pp. 335–386). San Diego, CA: Academic Press.
42. YouTube. (2015). Youtube: Press statistics. http://www.youtube.com/t/press_statistics
43. Zhou Wang, L. L., & Bovik, A. C. (2004). Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication, 19*(2), 121–132.

Chapter 21
Attentive Robots

Simone Frintrop

21.1 Introduction to Attentive Robots

While it is clear that selective attention is a useful concept for human perception, it is less obvious that such a concept is helpful for machines, for example, for autonomous service robots. However, it turns out that the more robotic systems advance, the higher the need for mechanisms such as computational attention. The reason is that robots share many requirements with humans: they operate in the same world as humans and they shall interact with humans. In this chapter, we will outline why computational attention systems are of large interest for machines, and we will give an overview of the state of the art in this field. We will focus on three application areas of computational attention mechanisms in the context of robotics: first, the detection of salient landmarks that robots can use to navigate in their environment; second, the detection of objects, which is important for many robot tasks such as object manipulation or map building; and third, attention as a guiding mechanism for robot action, in which it helps to direct the camera to regions of interest or supports human-robot interaction.

Before we start with these topics, let us clarify in more detail why attentional mechanisms are helpful for robots. First, robots have limited resources to process an overwhelming amount of perceptual input. While humans have their senses which open the world for them, robots have cameras, laser scanners, sonar sensors, etc., and all of them offer more information than can be processed in reasonable time. As humans, robots usually have to react instantaneously; therefore, it is not possible to process all of the sensory input, and decisions have to be made which part of

S. Frintrop (✉)
Department of Computer Science, University of Hamburg, Room R-105 Vogt-Kölln-Str. 30, 22527 Hamburg, Germany
e-mail: frintrop@informatik.uni-hamburg.de

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_21

the sensory input is currently of most interest. This means, on robots as well as in human perception, attention serves to prioritize the processing of the huge amount of sensory input.

Second, robots have physical limitations such as humans. Humans have only one head, two eyes, two arms, and hands, and if they want to act, they have to decide what to do first. Similarly, robots have only one or a few cameras, grippers, etc. and cannot do everything at the same time. Even if they had unlimited processing power, they would have to decide where to drive, which object to grasp, where to direct the camera to, and whether to zoom into an object or better explore the surroundings first.

Third, robots often act in the same world as humans and shall interact with them. Especially for mobile robots that are supposed to support humans in the household, in hospitals, or museums, the interaction with humans is essential. Since these humans are usually no computer experts, it is important that interaction can take place as naturally and intuitively as possible. The more humanlike robots act, the easier it is for a human to understand its intentions and to interact with it. An attention system similar to the human one, which makes the robot pay attention to the same things as humans and to generate a shared focus of attention with the interaction partner, helps to facilitate the communication and interaction. Figure 21.1 visualizes the concept of such an attentive robot.

During the last decade, interest in the robotics community for computational attention models has increased strongly. There are two reasons for this. First, algorithms and computer hardware have advanced enough to compute a focus of attention in real time (e.g., [71]), and methods are robust enough to deal with real-world conditions such as clutter in a scene, noise in the input data, and unexpected conditions. Second, basic capabilities of robots such as localizing themselves in an environment and avoiding obstacles while navigating have reached a quite mature level. Therefore, interest in the community has moved on to more complex tasks and challenges. Currently, one of the largest challenges is to deal with the perceptual input, especially with visual data, and the more complex a system becomes, the more urgent the need for a selection mechanism that decides which part of the input to process next.

The large interest in such capabilities has led to a large number of research projects on cognitive robots worldwide. For example, the European Commission has funded more than 100 projects on cognitive systems since the year 2000. Many of the robots that have been developed in these projects have a computational attention module to focus the processing on relevant parts of the sensory input, e.g., the projects MACS,[1] PACO-PLUS,[2] RobotCub,[3] and GRASP.[4]

[1] http://www.macs-eu.org/

[2] http://www.paco-plus.org/

[3] http://www.robotcub.org/

[4] http://www.csc.kth.se/grasp/

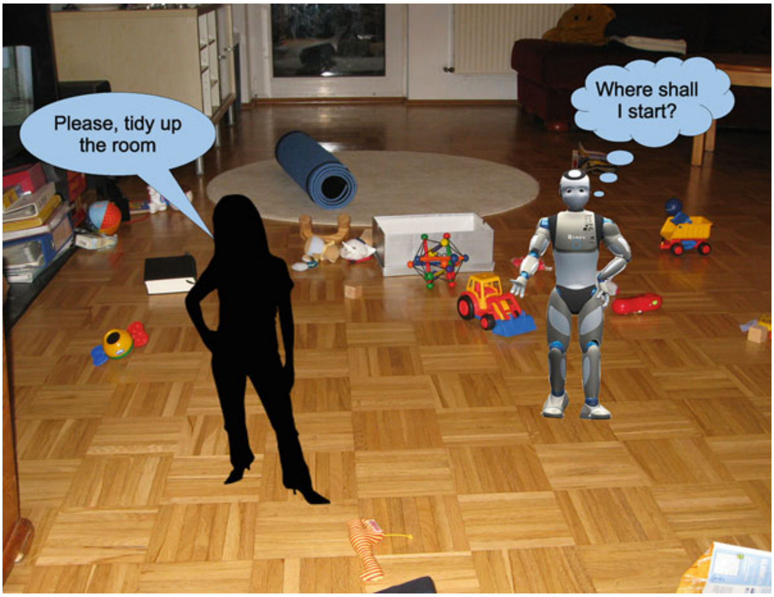


Fig. 21.1 The scene visualizes the concept of an attentive robot: to tidy up the room, the robot has to attend to the human, understand the task, and investigate the scene by attending to the objects on the floor, one object at a time. An attention module endows it with the capability to focus on regions of most potential interest. This enables efficient processing and prioritizes the robot's actions (Reprinted with permission from [18])

In this chapter, we give an overview of the applications of attention systems to autonomous robots.[5] We start in Sect. 21.2 with a classification of applications into three areas, from low-level to high-level techniques. In Sect. 21.3, we present three example applications in more detail.

21.2 State of the Art

In this section, we give an overview over past and current research in the field of attentive robots. The tasks of a robot that involve visual attention might be classified roughly into three categories. The first, most low-level category uses attention to detect salient landmarks that can be used for localization and scene recognition (Sect. 21.2.1). The second, mid-level category considers attention as a front end for object recognition (Sect. 21.2.2). In the third, highest-level category, attention is

[5] Parts of this chapter have already been published in [18].

used in a humanlike way to guide the actions of an autonomous system like a robot, i.e., to guide object manipulation or human-robot interaction (Sect. 21.2.3).

21.2.1 Salient Landmarks

A basic capability for autonomous mobile robots is to determine their position automatically, a technique that is called *self-localization* or often simply *localization*. The robot usually has a map of the environment that has either been provided by humans or has been built automatically in a preceding training phase. Then, the sensor data of the robot is interpreted to determine the robot's current position in the map. Traditionally, this task has been solved with 2D laser scanners that provide depth information of the obstacles around the robot. Depending on the structure of this data, probabilities about the current position are computed which are updated instantaneously as soon as new perceptions are available. During the last decade, interest in localization based on visual data has increased since cameras are low-cost, low-weight, and low-power sensors that provide rich information about the environment. On the other hand, this rich visual data bears many challenges, starting from the problem to obtain acceptable processing times to dealing with illumination changes and noise.

One approach for localization based on visual data is to use visual landmarks. This idea is not new. Humans have used it since many centuries to navigate on land and on sea. They have used natural landmarks such as rocks and trees, and they have built artificial ones, such as lighthouses. When seeing a known landmark, this gives information about the current position, as long as the landmark is unique. If there are ambiguities due to the similar appearance of several landmarks, this can lead to immense problems and many ships sank because of such misinterpretations. The same problem arises for robots. If they base their processing on visual landmarks, it is desired that the landmarks are clearly separable from their surroundings and easy to recognize when coming back to the same position. Here, salient landmarks that are detected with a visual attention system are excellent candidates since salient regions have by definition a high rareness/uniqueness. We have shown that the repeatability of salient image regions is significantly higher than for other standard region detectors [16].

Several studies have used salient landmarks for robot localization. An early project was the ARK project [50] that relied on hand-coded maps with the locations of known static obstacles as well as the locations of natural visual landmarks. Ouerhani et al. tracked salient spots over time and used them as landmarks for robot localization [51]. Siagian and Itti presented an approach for scene classification and global localization based on salient landmarks [62]. Additionally to the landmarks, the authors use the "gist" of the scene, a feature vector which captures the appearance of the scene, to obtain a coarse localization hypothesis.

A variant of the localization problem, the SLAM problem (simultaneous localization and mapping), has been investigated in [20]. In SLAM, the environmental

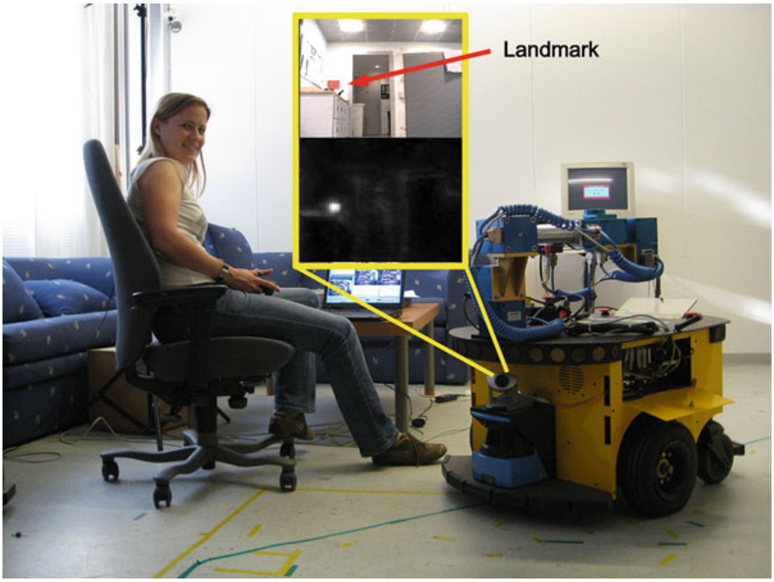


Fig. 21.2 Salient landmark detection on a robot: robot Dumbo corrects its position estimate by redetecting a landmark which it has seen before. Landmark detection is done with the attention system VOCUS. The *yellow rectangle* shows the view of the robot: an image with a landmark and the corresponding saliency map (Reprinted with permission from http://www.iai.uni-bonn.de/~frintrop/research.html)

map is not known in advance and has to be built during localization which makes the problem intrinsically harder. Figure 21.2 illustrates the process of detecting landmarks. The application will be described in detail in Sect. 21.3.1.

21.2.2 Object Detection and Recognition

Object detection and recognition are important tasks for mobile robots that are especially required when a robot is supposed to manipulate objects or interact with humans. The terms object detection, object discovery, object localization, object recognition, and object classification are closely related and often used interchangeably. Let us therefore clarify our understanding of the terms.

Object detection or localization tackles the problem of localizing objects in images. Usually, the object is comparably small in the scene which makes the task challenging. Hereby, it is important to distinguish several cases that we will call *general object detection*, *object instance detection*, and *object class detection* in the following.

In general object detection, also called *object discovery* or *object proposal detection*, nothing is known about the objects that might occur in the scene, and the task is to "find all objects in the scene." It is a pure bottom-up process that does not use preknowledge about a target or the scene. While this task is easily and effortlessly solved by humans, it is a challenging task for machine vision and belongs to the open problems in the field.

Object instance detection describes the task to find a specific, previously specified object, for example, the book "From Human Attention to Computational Attention: A Multidisciplinary Approach." In human vision, this task is usually called visual search and is supported by top-down visual attention.

On the other hand, object class detection means to find all instances of a certain class (any book). This task is another variant of visual search. In both, object instance and object class detection, it depends strongly on the type of object (class) whether the task is easy or hard. For example, a rigid object with plenty of texture is easy to detect, e.g., a specific book. One can use, for example, SIFT-based object detection for this purpose [38]. Object class detection also profits from a clear structure and, as usual in classification, is easy if the intra-class variation is low and the interclass variation is high. An example of an object detector that works very well is face detection that is implemented in almost all current digital cameras. For other types of object classes, the detection can become much harder. This can be seen, for example, on the results of the ImageNet Recognition Challenge [57]. While some classed achieved a recognition rate of 100 % (e.g., tiger), others were classified correctly only in 59 % of the images (e.g., letter opener). If the task was to not only determine whether or not the object is present in the image, but to precisely localize the object, the performance went down to 8 % for difficult objects such as letter openers or nails.

Object instance detection as well as object class detection are often also denoted as object recognition, and object class detection is also referred to as classification. Classification is usually applied to pre-segmented objects, or it uses a sliding window approach, in which subregions of the images are successively investigated by the classifier. Recently, this trend is changing, and many groups apply classifiers to regions of interest supplied by a proposal detection method that delivers a collection of windows that are likely to contain objects (see survey in [30]).

Visual attention can support object detection in several ways. On one hand, bottom-up attention facilitates general object detection in an unknown scene by restricting the processing to promising image regions. Since it does not require any preknowledge about the objects, it is a perfect candidate to detect arbitrary objects. If on the other hand a specific object instance is searched, top-down attention (visual search) can facilitate the task. The advantage of an attentional front end is that it limits the search space and thus reduces computational complexity. Additionally, most recognition methods work best if the object covers a dominant portion of the image; and thus, the recognition performance is improved.

General object detection based on a visual attention system was investigated by Walther and Koch [69]. They use an attention system to obtain saliency maps and generate proto-objects inside this map by thresholding and labeling the resulting blobs. Rudinac et al. [55] have presented a similar approach in a robotic context. The attention module detects object hypotheses which are then tracked over time while they are manipulated by a human caregiver. This process shall imitate infant learning. In [19], we have generated object candidates with a method that combines saliency and segmentation. The approach was extended in [28] to image sequences, in which candidate regions were tracked over time to generate sequence-level candidates. In robotics, it is often necessary to not only locate an object in the 2D image but also in the 3D world. Potapova et al. [52] find objects based on a symmetry-based saliency method that operates on depth data from an RGB-D depth camera. Martín García et al. [40] integrate color and depth data to obtain complementary object candidates. In [39], a spatial inhibition of return mechanism enables to sequentially focus regions of interest without losing track of the regions under camera motion.

While general object detection is a bottom-up process, object instance and object class detection are top-down processes since they include preknowledge about the target. In human vision, such search processes are supported by top-down cues that guide the visual attention system to target-specific regions of interest. While a natural choice for modeling this process computationally seem to be top-down attention systems, most groups use instead bottom-up saliency models as preprocessing for recognition [45, 67, 68]. This is mainly due to the fact that many bottom-up attention systems are freely available, while top-down models are rare. An early approach to combine bottom-up attention with classification was presented in Miau et al. [45]. They combined the attention system iNVT (iLab Neuromorphic Vision Toolkit) [33] with the biologically motivated object recognition system HMAX. The experiments are restricted to recognize simple artificial objects like circles or rectangles. Alternatively, the authors have used a support vector machine to detect pedestrians in natural images. Walther [68] combines his Saliency Toolbox, a Matlab implementation of the iNVT, with an object recognizer based on SIFT features [38] and shows that the recognition results are improved by the attentional front end. Vogel and de Freitas combine the iNVT with a classifier to perform gaze planning in complex scenes [67].

In the abovementioned approaches, the attentional part is separated from the object recognition; both systems work independently. In human perception, these processes are strongly intertwined. Accordingly, Walther and Koch suggest a unifying framework for object recognition and attention [70]. It is based on the HMAX model and modulates the activity by spatial and feature modulation functions which suppress or enhance locations or features due to spatial attention.

A few approaches use instead a top-down attention system as pre-processing for classification, for example, [22, 46]. Here, the top-down part of the attention system VOCUS generates object hypotheses which are verified or falsified by a classifier for the application of ball detection in the robot soccer scenario RoboCup.[6] Xu et al. [72] have used visual bottom-up and top-down attention to detect objects with the Autonomous City Explorer (ACE) robot (cf. Fig. 21.3).

[6] http://www.robocup.org

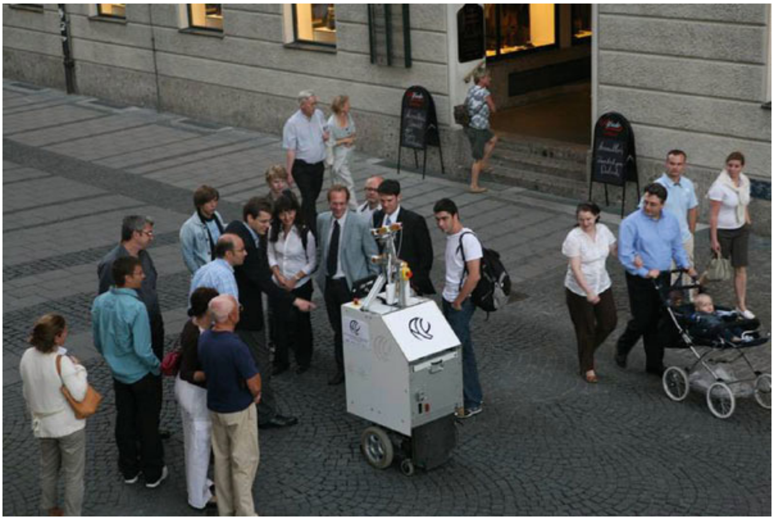


Fig. 21.3 The ACE robot (Autonomous City Explorer) while it is exploring the pedestrian area in Munich (Reprinted with permission from http://www.lsr.ei.tum.de/research/research-areas/robotics/ace-the-autonomous-city-explorer-project/)

In a robotics context, some groups have integrated attentive object detection on real robots. To simulate the different resolutions of the human eye, several groups use two cameras: one wide-angle camera for peripheral vision and one narrow-angle camera for foveal vision. For example, Gould et al. [25] and Meger et al. [42] determine regions of interest with visual attention in a peripheral vision system, focus on these regions with a foveal vision system, and investigate these high-resolution images with an object recognition method. The robot in the latter approach, Curious George, placed first in the robot league of the Semantic Robot Vision Challenge (SRVC) [27] both in 2007 and 2008 and first in the software league for 2009. The SRVC and Curious George will be described in more detail in Sect. 21.3.2.

Some groups have used attentive object detection to support object manipulation on robots or robot arms. One of the earliest works on this topic was presented by Bollmann et al. [6]: a Pioneer1 robot used the neural active vision system (NAVIS) to play dominoes. The group around Tsotsos is working on a smart wheelchair to support disabled children [54, 64]. The wheelchair has a display as easily accessible user interface which shows pictures of places and toys. Once a task like "go to table, point to toy" is selected, the system drives to the selected location and searches for the specified toy, using mechanisms based on a visual attention system. Rasolzadeh et al. [53] use bottom-up and top-down attention to control a KUKA arm for detecting, recognizing, and grasping objects on a table. In [5] and [35] the FOAs from the same attention system were used as seeds for 3D segmentation of objects from stereo data.

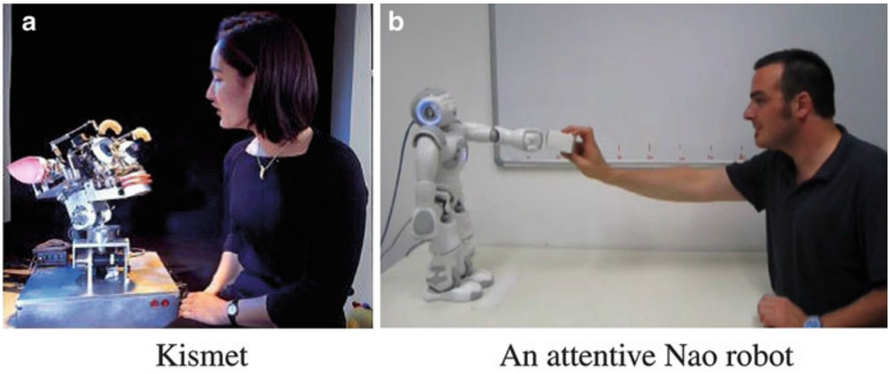
21.2.3 Guiding Robot Action

A robot which has to act in a complex world faces the same problems as a human: it has to decide what to do next. Such decisions include where to go (drive), what to look at, what to grasp, and who to interact with. Thus, even if computational power would allow to find all correspondences, to recognize all objects in an image, and process everything of interest, it would still be necessary to filter out the relevant information to determine the next action [37, 44]. This decision is based first, on the current sensor input and second, on the internal state, for example, the current tasks and goals.

A field in which the decision about the next action is intrinsically based on visual data is active vision, i.e., the problem of where to look next [3]. It deals with controlling "the geometric parameters of the sensory apparatus ... in order to improve the quality of the perceptual results" [2]. Thus, it directs the camera to regions of potential interest as the human visual system directs the gaze, the head, and even the body of a person. Since visual attention triggers this control in humans, it is also an intuitive candidate for the active vision problem on machines.

One of the first active vision systems that integrated visual attention was presented by Clark and Ferrier [10]. They describe how to steer a binocular robotic head with visual attention and perform simple experiments to fixate and track the most salient region in artificial scenes composed of geometric shapes. Bollmann et al. have used the neural active vision system (NAVIS) to steer the pan-tilt unit of a domino-playing Pioneer1 robot [6]. Vijayakumar et al. presented an attention system which is used to guide the gaze of a humanoid robot [66]. The authors consider only one feature, visual flow, which enables the system to attend to moving objects. Each eye is represented by a wide-angle camera for peripheral vision and a narrow-angle camera for foveal vision. In more recent work, the humanoid robot iCub bases its decisions to move eyes and neck on visual and acoustic saliency maps [56] (see also Sect. 21.3.3.2). Robot Dumbo performs active gaze control to support landmark detection for simultaneous robot localization and mapping [20] (see Sect. 21.3.1.2). Additionally, all the object manipulation approaches of the previous section include active vision to focus on the detected objects.

In the future, humans shall interact with robots as naturally and intuitively as possible. Studies in the field of human-robot interaction have shown that humans treat robots like people [14, 49], and the more a robot interacts with people, the more lifelike and intelligent it is perceived and the more excited users are [61]. An essential part for purposefully interacting with humans is to generate a joint focus of attention. A computational attention system similar to the human one can help a robot to focus on the same region as a human. According to this, [7] introduced the social robot Kismet that interacts with humans in a natural and intuitive way. Its gaze is controlled by a visual attention system (see Fig. 21.4a). Schillaci et al. equipped a humanoid Nao robot with an attention mechanism based on optical flow and face detection [61]. The robot is able to interact with humans by looking and pointing at objects (see Fig. 21.4b).



Kismet

An attentive Nao robot

Fig. 21.4 Social robots that interact with people. (**a**) Kismet is one of the earliest social robots. Its gaze is controlled by a visual attention system (Reprinted with permission from [8]). (**b**) The Nao robot is equipped with an attention system that facilitates interaction with a human partner (Reprinted with permission from [61])

For humans, following pointing gestures of other humans is an important ability to jointly focus their attention on objects of interest. Approaches to endow robots with a similar capability were proposed by Heidemann et al. [26] and Schauerte et al. [58, 60]. They analyze the direction of a pointing finger and fuse this top-down information with the bottom-up saliency of objects. A robot that learns visual scene exploration by imitating human gaze shifts is presented by Belardinelli [4]. Nagai developed an action learning model based on spatial and temporal continuity of bottom-up features [48].

Finally, Muhl et al. presented an interesting sociological study in which the interaction of a human with a robot simulation is investigated [47]. A robot face on a screen attends to objects, shown by a human, with help of a visual attention system. If the robot was artificially diverted and directed its gaze away from the object, humans tried to reobtain the robot's attention by waving hands, making noise, or approaching to the robot. This shows that people established a communicative space with the robot and accepted it as a social partner. More about socially interactive robots that use attentional mechanisms can be found in the extensive survey in [13].

21.3 More on Attention Applications for Robots

In this section, we will describe three example application areas for attentive robots. We start with the topic of visual robot localization and mapping (Sect. 21.3.1), continue with attentive object detection (Sect. 21.3.2), and finally present attentive robots that fuse multiple sensor modalities (Sect. 21.3.3). Each subsection starts with a general explanation of the application area and then presents an example system in more detail.

21.3.1 Attentive Visual SLAM

In this section, we introduce the problem of visual SLAM (simultaneous localization and mapping) (Sect. 21.3.1.1) and describe a system that solves this problem with salient landmarks that are detected with a visual attention system (Sect. 21.3.1.2).

21.3.1.1 Introduction to Visual SLAM

SLAM stands for *simultaneous localization and mapping* and describes the problem of automatically building a map of an unknown environment based on sensor data. SLAM is a common and widely investigated problem in robotics, but it can also be applied on systems that do not navigate autonomously such as cars or handheld cameras. A complete survey on SLAM is beyond the scope of this chapter. Instead, we will introduce the main ideas and the key terms. For further reading, we recommend the great tutorial by Durrant-Whyte and Bailey [11].

The SLAM problem is a chicken and egg problem: the robot needs a map to localize itself while on the other hand it requires an accurate pose estimate to build this map. The solution is to successively add new parts to the map while permanently using new sensor data to update existing parts of the map. The process can be compared with a human that explores an unknown area, e.g., a new city. While walking through the streets, she/he successively obtains a clearer picture of the city, of the streets, and their connections. Especially when the streets are narrow and winding, this can be difficult, and one might be surprised when coming to a previously seen location to be not where she/he expected. Based on this new information, the internal picture of the world is updated and corrected. The same is done on a robot. The key idea for this update is that the information about the robot pose and the information about all sensor observations (e.g., landmarks) are correlated. If the position of a single observation is corrected due to better measurements, this can influence the complete map data, e.g., all other observations as well as the robot position itself.

During the SLAM process, the computations take place in two steps: first, the robot moves, which increases the pose uncertainty of robot and landmarks. Then, the robot processes its new sensor data, which decreases the uncertainty. The largest correction of uncertainty, and therefore the most useful one, takes place during so-called *loop closing* situations. When the robot comes back to a region that it had already visited previously, it sees the same observations again and can correct its measurements accordingly. A precondition is of course that the robot realizes that the new measurements belong to the same observations as before. This step is not trivial and belongs to the largest challenges in (visual) SLAM.

A typical example of the accumulated position error that occurs if a robot estimates its position only based on odometry information is shown in Fig. 21.5, left. After driving three rounds on the same path in an office environment, the robot is completely lost and the position error has added up to several meters. On the right of this figure, it is shown how permanent updates with visual SLAM correct the position errors, resulting in an exactly estimated position.

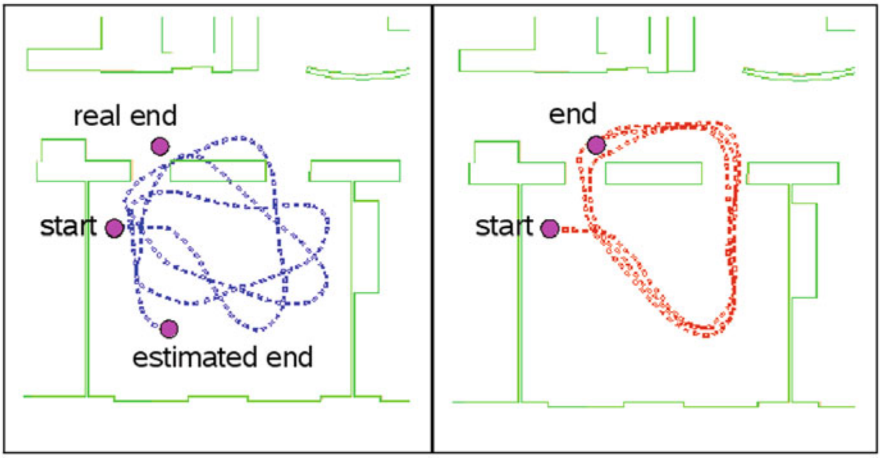


Fig. 21.5 The effect of SLAM: when the robot trajectory was estimated only from odometry (*left*), the position error of the robot accumulates. When the position estimate is updated with a SLAM system (*right*), these errors are corrected (Reprinted with permission from [20])

Traditionally, robots have used distance sensors such as laser range finders to create a map. They are especially well suited since they offer exact information about the distance of obstacles and the layout of buildings. On the other hand, laser scanners are expensive, often heavy, and require much energy. Therefore, other approaches aim to solve the SLAM problem with cameras as sensors. Camera-based SLAM is usually called *visual SLAM*. The main difference in visual SLAM is that first, images provide a huge amount of data which is time-consuming to process completely, and second, the 3D position of image regions is not available instantaneously but has to be estimated from stereo data or by structure from motion. Therefore, visual SLAM methods often extract 2D *features* from the images (e.g., corners or blob-like regions) and estimate their 3D position, resulting in so-called *landmarks*. Since several years, cheap RGB-D sensors are available that offer 3D data directly and several groups have investigated SLAM based on such data (e.g., [12]). This facilitates the 3D localization of landmarks, but the challenges of feature and landmark detection remain mostly the same. Important to note at this point is that a *map* that a robot builds with a landmark-based visual SLAM approach consists only of the positions of landmarks relative to the position of the robot.

Landmark selection and matching belong to the most important issues in visual SLAM. Feature selection is performed with a *detector* and the matching with a *descriptor*. A stable detector is necessary to redetect the same regions in different views of a scene. In applications like visual SLAM with time and memory constraints, it is also favorable to restrict the amount of detected regions. A powerful descriptor on the other hand has to capture the image properties at the detected region of interest and enable a stable matching of two regions with a high detection and low false-detection rate. It has to be able to cope with viewpoint variations as

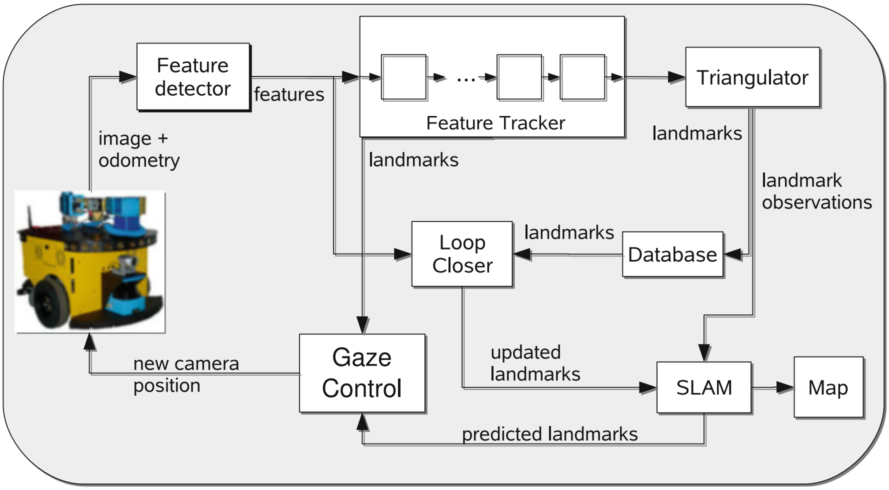


Fig. 21.6 Overview of the active visual SLAM system that estimates a map of the environment from image data and odometry based on salient landmarks (Reprinted with permission from [20])

well as with illumination changes. An overview of feature detection methods can be found in the survey of Tuytelaars and Mikolajczyk [65].

21.3.1.2 An Example of Attentive Visual SLAM: Robot Dumbo

Here, we present an approach for visual SLAM that is based on salient landmarks that are detected with a visual attention system. Such landmarks are especially suitable since they have a high uniqueness and are therefore easy to detect, to track, and to redetect in loop closing situations. The approach is presented in detail in [20].

The robotic platform that we used for our experiments is the robot Dumbo [34]. Dumbo is an ActivMedia PowerBot platform and is visible in Fig. 21.6. The sensor that is used in our application is a Canon VC-C4 pan/tilt/zoom camera that is mounted in the front of the robot at a height of 0.35 m above the ground. Additionally, we use odometry information. The platform possesses also other sensors such as laser scanners, but these are not used for the current application.

The SLAM module on Dumbo is based on an extended Kalman filter. Details about the SLAM architecture can be found in [34]. Here, we focus on the visual front end that detects, tracks, and redetects landmarks and provides their estimated positions to the SLAM module which computes the map. The architecture of the visual front end for SLAM is displayed in Fig. 21.6. When a new frame from the camera is available, it is provided to the *feature detector* which finds salient regions of interest (ROIs) in the images based on a visual attention system. Next, the features are provided to the *feature tracker* which stores the last *n* frames, performs matching of ROIs in these frames, and creates landmarks. A *triangulator* identifies useful landmarks and estimates their position in 3D. Triangulated landmarks are stored in a *database* and a *loop closer* matches current ROIs to database entries to detect if the robot returned to a known position. Finally, a *gaze control module* determines where to direct the camera to, based on the three behaviors: tracking visible landmarks, actively redetecting expected landmarks, and exploring unseen areas. In the following, we will concentrate on the detection and matching of landmarks. The other modules, especially the active camera control, are described in more detail in [20].

Attentional Feature Detection and Matching

An ideal candidate for selecting a few, discriminative regions in an image is a visual attention system. In previous chapters, several computational attention systems have been introduced. Here, we use the attention system VOCUS [15, 17].[7]

VOCUS consists of a bottom-up part which computes saliency purely based on the content of the current image and a top-down part which considers preknowledge and target information to perform visual search. Here, we consider only the bottom-up part of VOCUS; however, top-down search can be used additionally if a target is specified. For the approach presented here, any real-time capable attention system which computes a feature vector for each region of interest could be used.

An overview of VOCUS is shown in Fig. 21.7. The bottom-up part is similar to the Itti-Koch model [33] (differences in [15]). The computations for the features intensity, orientation, and color are performed on three different scales with image pyramids. The feature intensity is computed by *center-surround filters* that approximate the response of retinal ganglion cells in the human visual system. After summing up the scales, this results in two intensity maps. Similarly, four orientation maps ($0^\circ, 45^\circ, 90^\circ, 135^\circ$) are computed by Gabor filters and four color maps (green, blue, red, yellow) which highlight salient regions of the corresponding color.

Before the features are fused, they are weighted according to their *uniqueness*: a feature which occurs seldomly in a scene is assigned a higher saliency than a frequently occurring feature. This mechanism enables to detect outliers in a scene as in human perception. The uniqueness $\mathscr{W}$ of map X is defined as

$$\mathscr{W}(X) = X/\sqrt{m}, \tag{21.1}$$

[7] A more recent reimplementation of VOCUS, called VOCUS2 [24], is available online at http://www.iai.uni-bonn.de/~frintrop/vocus2.html. It achieves state-of-the-art performance on current benchmarks for saliency computation, operates in real-time, and is well suited for real-world images as obtained from a robot.

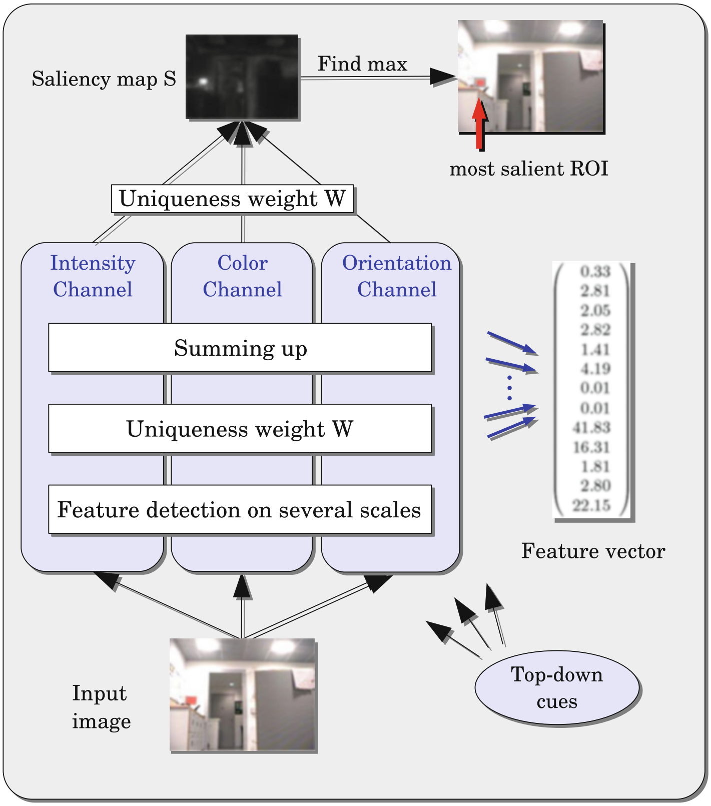


Fig. 21.7 The visual attention system VOCUS detects regions of interest (ROIs) in images based on the features intensity, orientation, and color. For each ROI, it computes a feature vector which describes the contribution of the features to the ROI (Reprinted with permission from [20])

where m is the number of local maxima that exceed a threshold and "/" is here the point-wise division of an image with a scalar. The maps are summed up to three conspicuity maps I (intensity), O (orientation), and C (color) and combined to form the *saliency map*:

$$S = \mathscr{W}(I) + \mathscr{W}(O) + \mathscr{W}(C). \tag{21.2}$$

The feature and conspicuity maps for one example image are displayed in Fig. 21.8.

From the saliency map, the brightest regions are extracted as *regions of interest (ROIs)*. This is done by first determining the maxima in the map and then finding for each maximum a surrounding region with *seeded region growing* [1]. This method finds recursively all neighbors with sufficient saliency. For the sake of efficient storage of the ROIs, we approximate the region by a rectangle. The output of VOCUS for one image is a list of ROIs, each defined by its 2D location, size, and a feature vector that describes its content (see next section).

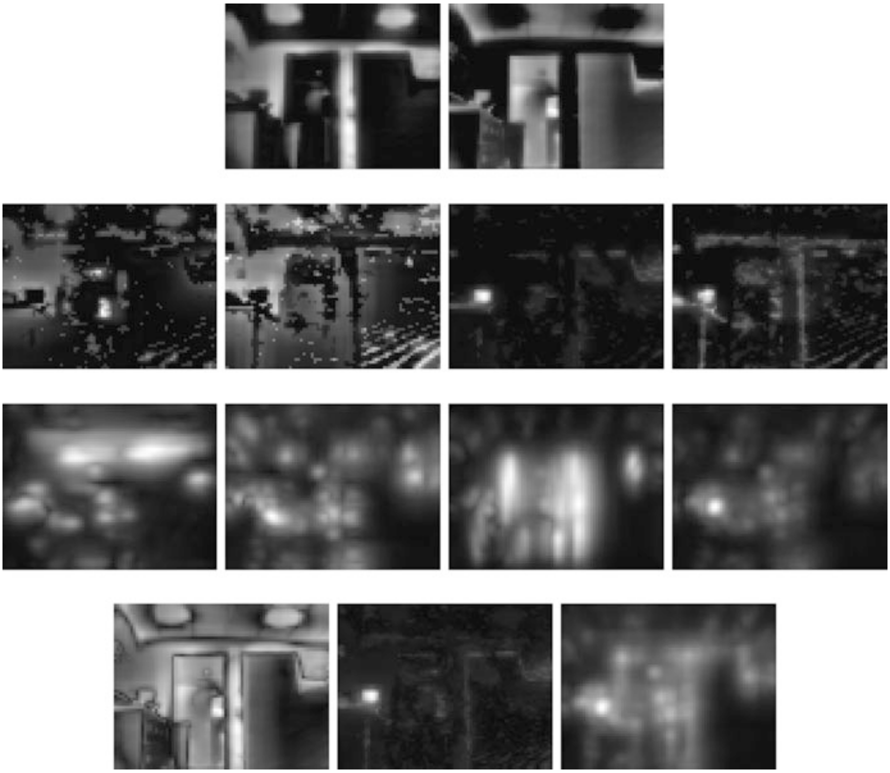


Fig. 21.8 The feature and conspicuity maps for the input image from Fig. 21.7. Top-left to bottom-right: intensity on-off, intensity off-on, color maps *green, blue, red, yellow*, orientation maps 0 °, 45 °, 90 °, 135 ° and conspicuity maps *I*, *C*, *O*. Since the *red region* sticks out as a unique peak in the feature map *red*, this map is weighted strongly by the uniqueness weight function, and the corresponding region becomes the brightest in the saliency map (see Fig. 21.7) (Reprinted with permission from [20])

To compare if two image regions belong to the same part in the world, both regions have a descriptor, which is a vector that describes the appearance of the region and usually its local neighborhood. In this work, we use two kinds of descriptors: first, we determine a simple attentional descriptor for tracking ROIs between consecutive frames. Second, we use the more sophisticated SIFT descriptor to match ROIs in loop closing situations [38].

The *attentional descriptor* can be obtained almost without cost from the ten feature and three conspicuity maps of VOCUS. For each ROI R, a 13-element feature vector $\mathbf{v}$ is determined, which describes how much each feature contributes to the ROI (cf. Fig. 21.7). The value v_i for map X_i is the ratio of the mean saliency in the target region R and the background $B = X_i - R$:

$$v_i = m_R/m_B, \quad \text{with} \quad m_R = \Big(\sum_{p\in R} X_i(p)\Big)/|R| \quad \text{and} \quad m_B = \Big(\sum_{q\in B} X_i(q)\Big)/|B|, \tag{21.3}$$

where $|R|$ and $|R|$ denotes the number of pixels in the regions R and B. This computation does not only consider which features are dominant in the target region but also which features separate the region best from the rest of the image (details in [15]). In the tracker, two vectors $\mathbf{v}$ and $\mathbf{w}$ are matched by calculating the similarity $d(\mathbf{v}, \mathbf{w})$ according to a distance similar to the Euclidean distance [21]. Since the vectors have only 13 elements, matching them is faster than matching SIFT descriptors. It is less powerful, but sufficient, in tracking situations.

The *SIFT descriptor* [38] belongs to the most powerful and widely used descriptors in computer vision and robotics. It is a $4 \times 4 \times 8 = 128$ dimensional descriptor vector which results from placing a 4×4 grid on a point and calculating a gradient histogram for each of the grid cells. Usually, SIFT descriptors are computed at intensity extrema in scale space [38]. Here, we calculate one SIFT descriptor for each ROI. The center of the ROI provides the position, and the size of the ROI determines the size of the descriptor grid. The grid should be larger than the ROI to allow catching information about the surrounding but should also not include too much background and stay within the image borders.[8] The difference d_S of two SIFT descriptors is calculated as the sum of squared differences (SSD) of the descriptor vectors.

Feature Tracking and Loop Closing

The ROIs are matched with previous ones over several frames, and all successfully matched ROIs form a chain of ROIs that all belong to the same item in the world. This item is called a *landmark*; that means, while a feature is a 2D region in an image, a landmark is a 3D part of the real world which is represented here by a collection of 2D views. Next, the triangulator estimates the 3D location of the landmark, and finally, the landmark is stored in the database.

In the loop closing module, it is detected if the robot has returned to an area where it has been before. *Loop closing* is done by matching the ROIs from the current frame to landmarks from the database with the SIFT descriptor as described before. When a match is detected, the coordinates of the matched ROI in the current frame are provided to the SLAM system, to update the coordinates of the corresponding landmark. Some examples of correct matches in loop closing situations are displayed in Fig. 21.9, col. 1–4. False matches occur seldomly with this approach. If they do, the ROIs usually correspond to almost identical objects. An example is shown in Fig. 21.9, right.

[8] We chose a grid size of 1.5 times the maximum of width and height of the ROI.

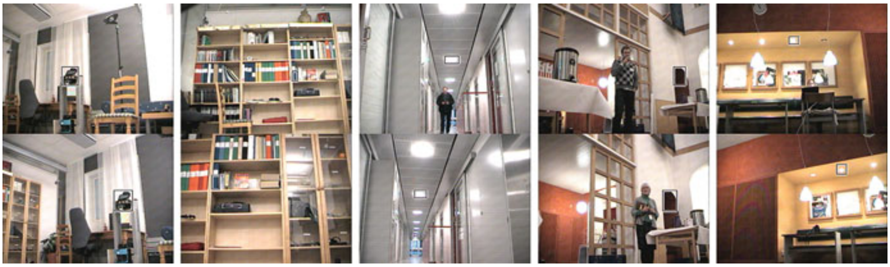


Fig. 21.9 Some examples of matched ROIs, displayed as rectangles. *Top*: current frame. *Bottom*: frame from the database. Col. 1–4 are correct matches; col. 5 shows a false match (Reprinted with permission from [20])

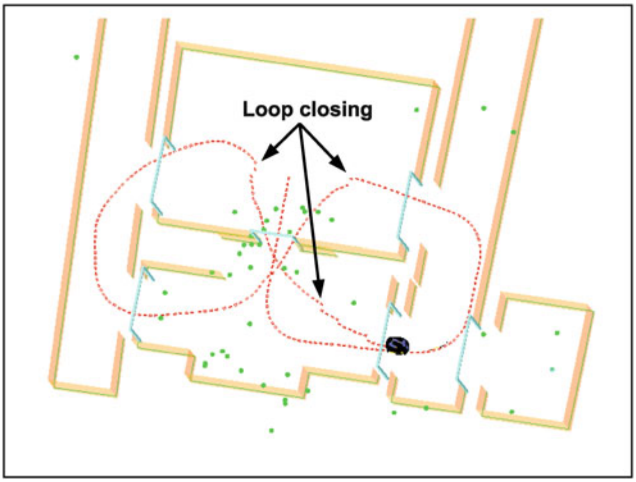


Fig. 21.10 Attentive visual SLAM: The estimated robot trajectory (*red*) and the created map consisting of detected landmarks (*green dots*) (Reprinted with permission from [20]) (See also videos on http://www.informatik.uni-bonn.de/~frintrop/research/aslam.html)

Attentive Visual SLAM

The estimated positions of the landmarks that were detected by the attention system are handed to the SLAM module that computes a map of the environment, consisting of estimated positions of the robot and the detected landmarks. An example of such a map is displayed in Fig. 21.10. Note that the walls are only superimposed onto the map for better visibility; the robot has no knowledge about them. This example was obtained with active camera control, in which the robot actively directed the camera to regions with expected landmarks. This enabled loop closing also in situations in which the current viewpoint of the robot differed strongly from the previous observation of this landmark.

21.3.2 Attentive Object Detection

In this section, we describe attentive object detection in more detail (Sect. 21.3.2.1) and describe as example application the robot system Curious George (Sect. 21.3.2.2).

21.3.2.1 Introduction to Attentive Object Detection

As introduced in Sect. 21.2.2, the broad area of object detection can be subdivided into general object detection, object instance detection, and object class detection. *Attentive object detection* denotes any type of object detection that involves a visual attention system, either bottom-up or top-down.

In the case of attentive general object detection, usually a bottom-up attention system provides object hypotheses, sometimes also called proto-objects [69]. These proto-objects are salient blobs from a bottom-up saliency map. This step is mostly followed by a segmentation step that uses properties such as feature similarity and proximity to obtain a better shape of the proto-objects. Finally, the camera can focus the proto-object and zoom in on it. In object instance and object class detection, an additional step is to recognize the identity of the objects.

Attentive object detection is especially useful in scenarios in which a robot has to find objects in a complex, realistic scenario. In contrast to the task of object recognition in web images, it is not possible here to directly apply object classifiers to all possible subwindows since their number is large (10^6-10^7 windows per image [30]), and real-time computation is usually required.

In attentive object detection, the selection of promising views is performed by a visual attention system. Selecting a view includes determining the direction of the object to center it in the field of view as well as determining a zoom level to obtain an image that includes as much of the object as possible without cutting its borders. To solve this task, most robotic systems use a peripheral-foveal vision system. This includes a peripheral camera with a wide field of view and low resolution to find regions of potential interest and a foveal camera with a high resolution and the ability to zoom.

In a realistic scenario, objects of interest are usually not all visible from one position in the scene. That means, in order to find the objects, the robot has to explore the environment by moving around. This involves many challenges: it has to avoid obstacles, map the environment to maintain a spatial representation of the surrounding, and plan its motions to obtain better views of objects and to discover new ones in unexplored areas. While many mature methods exist for such problems, building a complete integrated system with all of these modules is still a big challenge.

A research competition that has been designed in order to push and evaluate recent developments of recognition systems on autonomous mobile robots is the

pumpkin
orange
red ping pong paddle
white soccer ball
laptop
dinosaur
bottle
toy car
frying pan
book ”I am a Strange Loop” by Douglas Hofstadter
book ”Fugitive from the Cubicle Police”
book ”Photoshop in a Nutshell”
CD ”And Winter Came” by Enya
CD ”The Essential Collection” by Karl Jenkins and Adiemus
DVD ”Hitchhiker’s Guide to the Galaxy” widescreen
game ”Call of Duty 4” box
toy Domo-kun
Lay’s Classic Potato Chips
Peperidge Farm Goldfish Baked Snack Crackers
Peperidge Farm Milano Distinctive Cookies

Fig. 21.11 List of objects to find in the SRVC 2009 challenge [63]

Semantic Robot Vision Challenge (SRVC)[9] [27]. The competition consists of two phases: a training phase, in which the robot receives a text list of objects and uses the web to learn visual classifiers for these objects. The list contains both object categories, such as “bottle,” as well as specific objects, such as a specific CD or book. The list of objects from the 2009 challenge is shown in Fig. 21.11. Thus, object instance detection and object class detection have to be performed. The second phase is the exploration phase, in which the robots have to explore an unknown arena, arranged roughly like a living room, and locate the objects of the list within it.

In the following section, we describe the robot “Curious George” that participates in this challenge and performed successfully in several contests.

21.3.2.2 An Example of Attentive Object Detection: Curious George

Curious George is a robotic platform that was built to perform real-world object recognition in realistic scenarios [42, 43]. Its recognition abilities include instance as well as class detection. A picture of Curious George is shown in Fig. 21.12, left. The robot is an ActivMedia PowerBot, equipped with a peripheral-foveal vision system mounted on a pan-tilt unit. This enables the robot an effective

[9] The challenge took place until 2009; after that, similar challenges have been organized, e.g., the ImageCLEF Robot Vision Challenge: http://www.imageclef.org/2014/robot

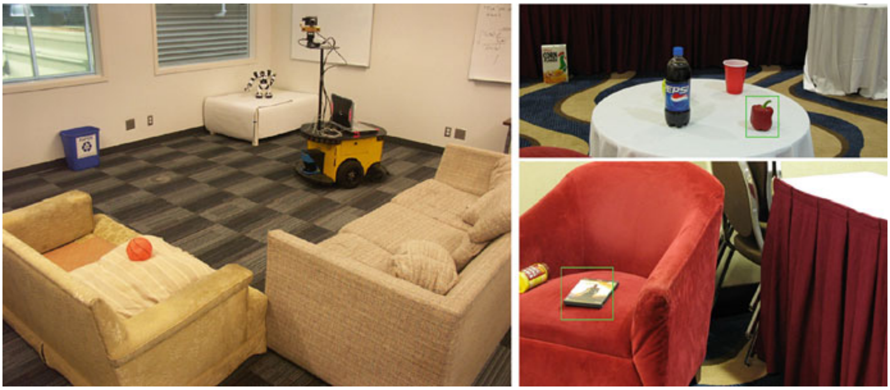


Fig. 21.12 *Left*: the robot Curious George within the room it is exploring. *Right*: two objects detected and successfully recognized by Curious George. (Reprinted with permission from [42])

360° gaze range. The camera for peripheral vision is a Bumblebee color stereo camera with 1024 × 786 resolution and a 60° field-of-view. The foveal camera is a Canon PowerShot G7 camera, with 10 megapixel resolution and 6× optical zoom. Additionally to the camera system, the robot has a laser range finder that is used for mapping the environment. Curious George executes up to 50 independent processes simultaneously to perform all tasks from navigation to object recognition. To enable this, six computation units are used: one on-board processor and five laptops.

The main application area of Curious George has been the Semantic Robot Vision Challenge that was described in the previous section. Figure 21.12, left, shows the scenario in which the robot had to operate during the challenge and Fig. 21.12, right, shows two of the objects that have been successfully identified by it. Curious George has placed first in the SRVC 2007 and 2008 robot league and in the 2009 software league. In the latter, it has recognized 13 out of 20 objects.

The vision system of Curious George consists mainly of an attention system that focuses the regions of interest and a recognition system that analyzes these regions in detail and determines the identity and/or class of the visible objects. Let us have a closer look at these two system parts.

The visual attention system of the robot aims at selecting the interesting views out of the enormous number of actual and possible views that the robot faces during exploration. It identifies potential objects within its peripheral vision system, centers these objects in the camera view, selects an appropriate zoom level, and finally obtains a detailed image using the foveal vision system.

The saliency approach to identify potential objects is based on the Spectral Residual Saliency method by Hou and Zhang [31]. The main idea of this approach is to analyze the frequency spectrum of the image and assign higher saliency to rarely occurring frequencies. It was extended to color, resulting in three feature channels: one intensity channel and two color channels, one for red-green and one for yellow-blue. Instead of the classical winner-takes-all method to determine most

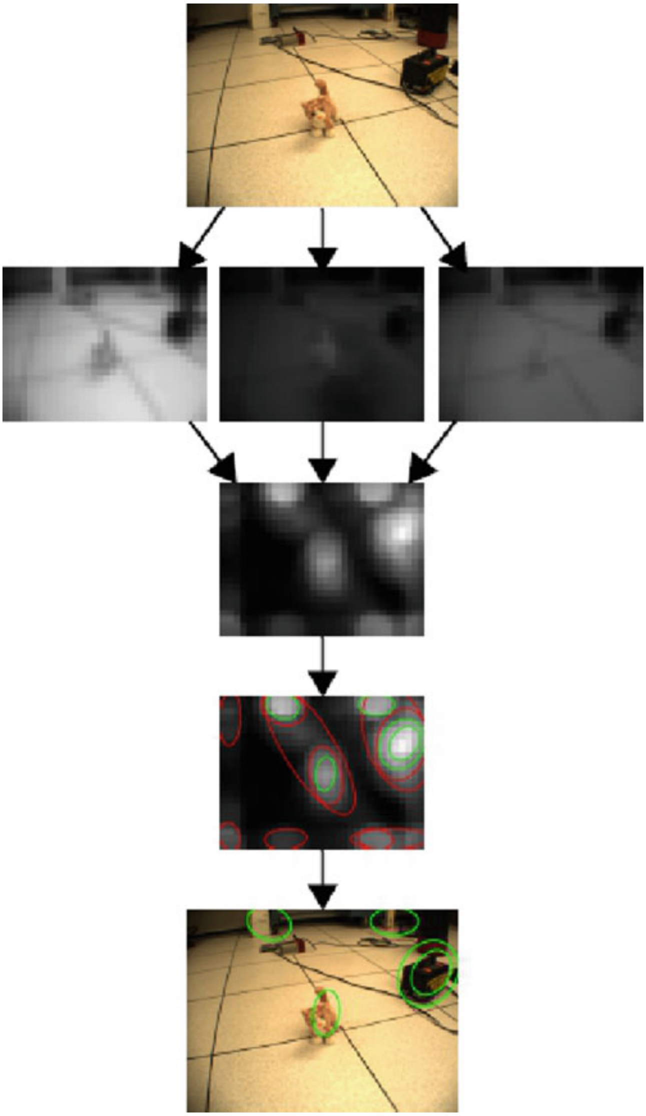


Fig. 21.13 Saliency computation on the robot Curious George. Top to bottom: input image, color opponency channels (intensity, *red-green*, *yellow-blue*), spectral saliency map, detected regions, regions superimposed on input image (Reprinted with permission from [42])

salient regions, maximally stable extremal regions (MSERs) [41] are detected in the saliency map. The region that is finally selected for further processing is the tightest one at which the MSER fits entirely in the image. Figure 21.13 shows the process of saliency computation for an example image.

After a potential object has been detected, the selected image is provided to the object identification system. This contains three different methods: a SIFT-based method that is used to recognize specific objects based on texture, a contour-matching method based on edge detection that recognizes objects based on shape, and a deformable parts model (DPM) classifier that is used for category recognition. Details about these methods and how they are integrated into Curious George can be found in [42].

Additional to the visual components, the robot requires several other capabilities to detect the objects. Since not all objects are visible or recognizable from the starting position, Curious George has to navigate and explore the environment. The map that is required for successful exploration is built with a SLAM approach (cf. Sect. 21.3.1.1) based on laser scanner data and odometry data. The exploration strategy itself is based on the frontier-based exploration strategy by Yamauchi et al. [73].

While a robot like Curious George is still far from the human level of object recognition, there has been large progress during the last years, and the visual understanding by mobile robots has the potential to build useful household robots in the future.

21.3.3 Multimodal Attention

In this section, we describe how saliency from different modalities, especially visual and auditory saliency, can be integrated into a mobile robot. Section 21.3.3.1 starts with a general introduction to multimodal attention on machines and Sect. 21.3.3.2 presents the robot iCub as an example application.

21.3.3.1 Introduction to Multimodal Attention

While most work on attention in psychophysics as well as in computer modeling of attention focuses on visual cues, it is well known that attentional mechanisms also exist for the other senses. The best investigated among the nonvisual attentional cues in human perception is probably auditory attention, known, for example, by the cocktail party effect [9].

It is obvious that not only humans but also robots profit strongly from exploiting perceptual data from different sources. Different sensors enable to extract complementary aspects of the environment and to exploit the advantages of different sensing capabilities. These sensors can be similar (but never equal) to human senses, e.g., cameras that correspond roughly to eyes or microphones that correspond roughly to ears. On the other hand, robots can have sensors that do not correspond directly to human perception but offer additional sensing capabilities, e.g., laser range finders, ultrasonic sensors, or infrared cameras.

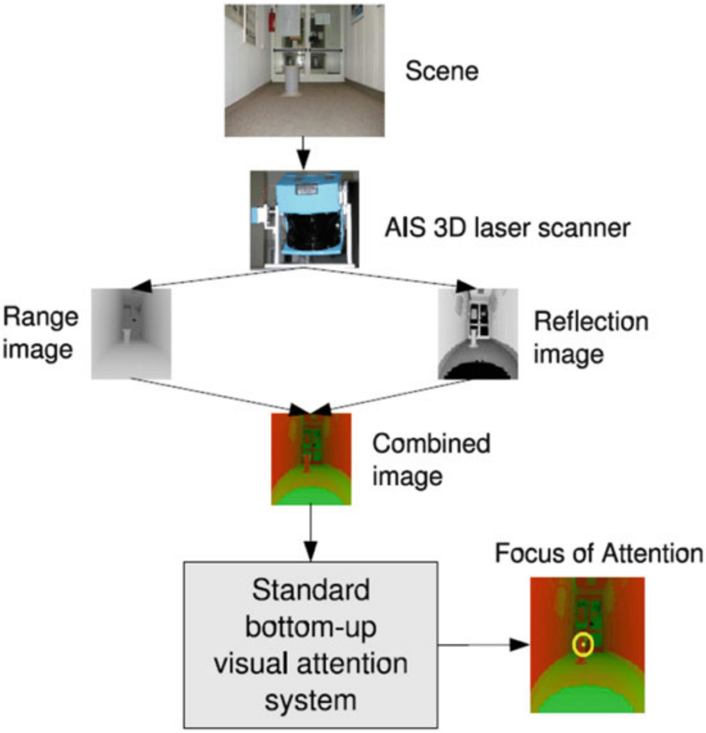


Fig. 21.14 Multimodal attention: Range and reflection data from a 3D laser range finder are visualized as images. For depth visualization, close objects obtain bright intensities while faraway regions are visualized dark. Then, saliencies are computed with a standard visual attention system (Reprinted with permission from [15])

The attentional concept can be transferred easily to this nonhuman perceptual data. The idea is simple: detect parts of the data that differ from their (local) surround. The extraction mechanisms however differ for each sensor. One possibility is to visualize the data and then apply standard visual feature extraction methods on the visualizations. This idea was pursued, for example, in [15, 23], where the depth and reflection data of a laser range finder were visualized as images, their saliencies were computed with visual feature extraction methods according to Itti et al. [33], and the saliencies from the two modalities were finally fused to a single map (see Fig. 21.14). Also the auditory attention of the iCub robot, that will be described in more detail in the following section, visualizes the auditory data before fusing it with the visual cues. An alternative for audiovisual saliency detection is presented by Schauerte et al. who fuse visual bottom-up saliency with a surprise-based auditory saliency module [36, 59].

The same group has also worked on an interesting approach to fuse saliency with higher-level information based on language: They introduce a top-down attention system that performs visual search for objects in spoken human-robot interaction by integrating visual information with linguistic descriptions about the visual appearance of a searched object. Visual attention is hereby guided by integrating spatial information of a bottom-up saliency map with an area of interest obtained from of pointing gestures of the human partner [58]. More about multimodal attention on robots can be found in the extensive survey in [13].

One challenge in multimodal attention is the fusion of data from the different modalities. It is not clear how visual data and auditory data can be combined and even less how data from other sensors fits into the picture. Most approaches ignore the problem to a wide extend and simply apply the same mechanisms that are used to fuse visual information from different feature channels in visual attention models: the saliency maps from the different modalities are summed up or the maximum is taken. Since the influence of data from different modalities depends on the hardware of the robot as well as on the current situation and context, a good solution is certainly to learn the weighting of modality saliencies from experience.

21.3.3.2 An Example of a Humanoid with a Multimodal Attention System: iCub

The iCub is a humanoid robot of the size of a 3.5-year-old child that was developed within the EU project RobotCub.[10] The main purpose of the robot platform was to study cognition. Figure 21.15 shows a picture of iCub; a detailed explanation of iCub's attention system can be found in Ruesch et al. [56].

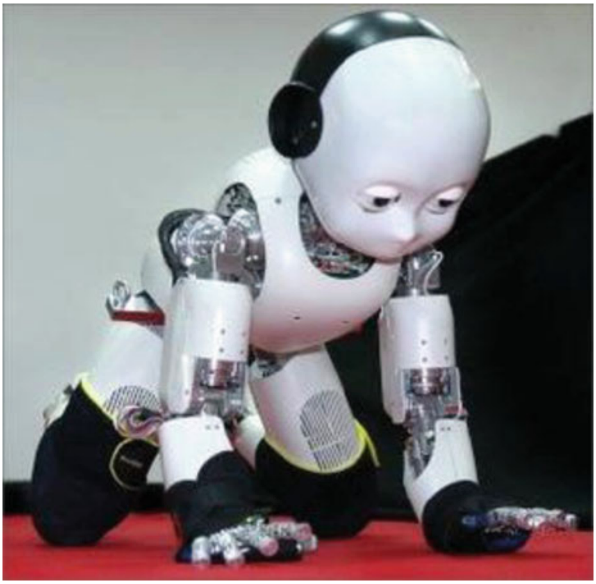


Fig. 21.15 The humanoid robot iCub (Fig. from http://www.robotcub.org/)

[10] http://www.robotcub.org/

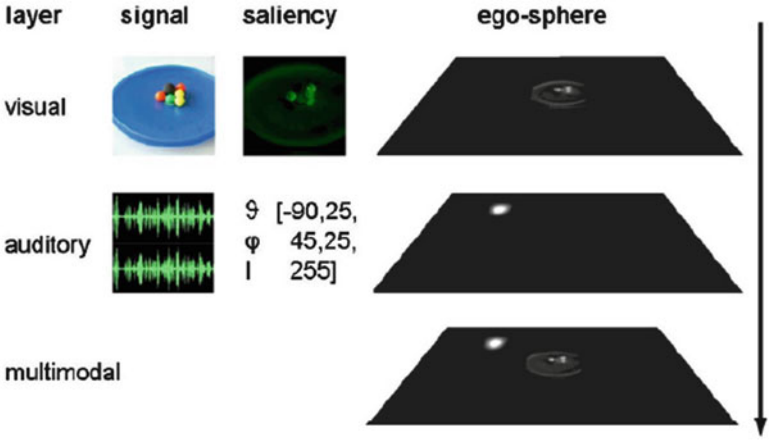


Fig. 21.16 Multimodal fusion of visual and auditory saliency information into a single saliency representation (Reprinted with permission from [56])

The sensor modes used for the iCub system are visual and auditory data. Visual saliency is computed according to the approach of Itti et al. [33], using the feature channels' intensity, color, and orientation. Additionally, a motion feature channel is integrated according to [32]. Auditory saliency is determined by estimating the position of a sound source with interaural spectral differences (ISD) and interaural time difference (ITD). Details about the process can be found in [29]. An example of a visual and an auditory saliency map can be seen in Fig. 21.16.

In iCub, the saliencies from different modes are integrated into a topologically organized ego-sphere. The ego-sphere is a continuous spherical surface with infinite radius that is centered at the robot's head. Saliency maps from different sensor modalities are mapped into the ego-sphere, resulting in a coherent representation of multimodal saliency. While for the iCub system only visual and auditory saliency is computed, in principle, also cues from other sensor modes can be integrated. Figure 21.17, left, shows the concept of the ego-sphere; Fig. 21.17, right, shows a spherical mosaic to illustrate the mapping of data onto it. It was obtained by directly mapping camera images instead of saliency maps onto the ego-sphere.

Saliency information from the different modes is projected onto the sphere by:

- Converting stimulus orientation to head-centered, spherical coordinates
- Projecting stimulus intensity onto rectangular egocentric maps, one per modality
- Aggregating multimodal sensory information

Next, the saliencies from the different sensor modes are integrated into a single saliency representation by taking the maximum over the saliencies of all sensor modes. This process is visualized in Fig. 21.16.

Now, the ego-sphere can be used to control the attention of the robot in order to explore the environment. The approach is simple: first, attend to the most salient

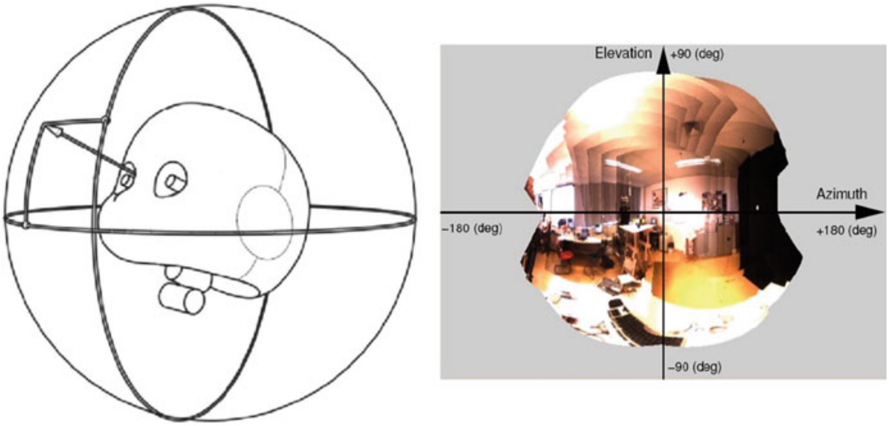


Fig. 21.17 *Left*: the iCub ego-sphere. *Right*: a spherical mosaic, obtained by mapping camera images onto the ego-sphere. (Reprinted with permission from [56])

location on the sphere by moving the neck and eyes; second, inhibit this region in the ego-sphere. Then, repeat this process (details in [56]).

Several videos that show the behavior of iCub are available on the project's web pages[11] and on Youtube.

21.4 Summary and Perspectives

In this chapter , we have introduced attentive robots, that is, robots that are equipped with a computational attention system that guides their focus of attention to special parts of the sensory input. This capability that originates from human perception is also very useful for robots. It helps to decide which parts of the sensory input to process first and which actions to perform, and it enables to establish a joint focus of attention in human-robot interaction. The areas in which attention is useful for a robot can be classified roughly into three areas: salient region detection, object detection and recognition, and guiding robot action. We have illustrated each of these areas by one example application.

Attentive robots are a step into the direction of more humanlike robots which are therefore more intuitive and natural to interact with. Currently, most existing robots are applied for specialized tasks, but the more general a robot system will be, the more urgent the need of an attention system that guides the processing. An example is household robots that obtain orders of a human supervisor and that should be able to act in complex, unknown real-world environments.

[11] http://www.robotcub.org/

Among the biggest challenges of future research are dealing intelligently with the large amount of sensory input, learning to deal autonomously with new situations and new objects, and integrating the data from all the different sensors and modules of a robot.

References

1. Adams, R., & Bischof, L. (1994). Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 16*(6), 641–647.
2. Aloimonos, Y., Weiss, I., & Bandopadhay, A. (1988). Active vision. *International Journal of Computer Vision (IJCV), 1*(4), 333–356.
3. Bajcsy, R. (1985). Active perception vs. passive perception. In *Proceedings of the IEEE Workshop on Computer Vision: Representation and Control*, Bellaire.
4. Belardinelli, A. (2008). Salience features selection: Deriving a model from human evidence. PhD thesis, Sapienza Universita di Roma, Rome.
5. Björkman, M., & Kragic, D. (2010). Active 3D scene segmentation and detection of unknown objects. In *IEEE International Conference on Robotics and Automation (ICRA)*, Anchorage.
6. Bollmann, M., Hoischen, R., Jesikiewicz, M., Justkowski, C., & Mertsching, B. (1999). Playing domino: A case study for an active vision system. In H. Christensen (ed.), *Computer Vision Systems* (pp. 392–411). Berlin: Springer.
7. Breazeal, C. (1999). A context-dependent attention system for a social robot. In *Proceedings of the International Joint Conference on Artifical Intelligence (IJCAI 99)*, Stockholm (pp. 1146–1151).
8. Breazeal, C. (2000). Sociable machines: Expressive social exchange between humans and robots. PhD thesis, Department of Electrical Engineering and Computer Science. MIT.
9. Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America, 25*, 975–979.
10. Clark, J. J., & Ferrier, N. J. (1988). Modal control of an attentive vision system. In *Proceedings of the 2nd International Conference on Computer Vision*, Tampa.
11. Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous localization and mapping (SLAM): Part I the essential algorithms. *Robotics and Automation Magazine, 13*(2), 99–110.
12. Engelhard, N., Endres, F., Hess, J., Sturm, J., & Burgard, W. (2011). Real-time 3D visual SLAM with a hand-held camera. In *Proceedings of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum*, Västerås.
13. Ferreira, J. F., & Dias, J. (2014). Attentional mechanisms for socially interactive robots–a survey. *IEEE Transactions on Autonomous Mental Development, 6*(2), 110–125.
14. Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42*(3–4), 143–166.
15. Frintrop, S. (2006). *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search* (Lecture Notes in Artificial Intelligence (LNAI), Vol. 3899). Berlin/Heidelberg: Springer.
16. Frintrop, S. (2008). The high repeatability of salient regions. In *Proceedings of ECCV workshop "Vision in Action: Efficient Strategies for Cognitive Agents in Complex Environments"*, Marseille.
17. Frintrop, S. (2011). Computational visual attention. In A. A. Salah & T. Gevers (Eds.), *Computer Analysis of Human Behavior* (Advances in Pattern Recognition). London/New York: Springer.
18. Frintrop, S. (2011). Towards attentive robots. *PALADYN Journal of Behavioral Robotics, Springer, 2*(2), 64–70.

19. Frintrop, S., Martín García, G., & Cremers, A. B. (2014). A cognitive approach for object discovery. In *International Conference on Pattern Recognition (ICPR)*, Stockholm.
20. Frintrop, S., & Jensfelt, P. (2008). Attentional landmarks and active gaze control for visual SLAM. *IEEE Transactions on Robotics, Special Issue on Visual SLAM, 24*(5), 1054–1065
21. Frintrop, S., Jensfelt, P., & Christensen, H. (2007). Simultaneous robot localization and mapping based on a visual attention system. In L. Paletta & E. Rome (Eds.), *Attention in Cognitive Systems* (Lecture Notes on Artificial Intelligence (LNAI), Vol. 4840). Berlin/New York: Springer.
22. Frintrop, S., Nüchter, A., Pervölz, K., Surmann, H., Mitri, S., & Hertzberg, J. (2009). Attentive classification. *International Journal of Applied Artificial Intelligence in Engineering Systems, 1*(1), 47–66.
23. Frintrop, S., Rome, E., Nüchter, A., & Surmann, H. (2005). A bimodal laser-based attention system. *Journal of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance in Computer Vision, 100*(1–2), 124–151.
24. Frintrop, S., Werner, T., & Martín García, G. (2015). Traditional saliency reloaded: A good old model in new shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston.
25. Gould, S., Arfvidsson, J., Kaehler, A., Sapp, B., Messner, M., Bradski, G., Baumstarck, P., Chung, S., & Ng, A. Y. (2007). Peripheral-foveal vision for real-time object recognition and tracking in video. In *Proceedings of the 20th International Joint Conference on Artifical intelligence (IJCAI)*, Hyderabad
26. Heidemann, G., Rae, R., Bekel, H., Bax, I., & Ritter, H. (2004). Integrating context-free and context-dependent attentional mechanisms for gestural object reference. *Machine Vision and Applications, 16*(1), 64–73.
27. Helmer, S., Meger, D., Viswanathan, P., McCann, S., Dockrey, M., Fazli, P., Southey, T., Muja, M., Joya, M., Little, J., et al. (2009). Semantic robot vision challenge: Current state and future directions. arXiv preprint arXiv:0908.2656.
28. Horbert, E., Martín García, G., Frintrop, S., & Leibe, B. (2015). Sequence-level object candidates based on saliency for generic object recognition on mobile systems. In *Proceedings of the IEEE International Conference Robotics and Automation (ICRA)*, Seattle.
29. Hörnstein, J., Lopes, M., Santos-Victor, J., & Lacerda, F. (2006). Sound localization for humanoid robots – building audio-motor maps based on the hrtf. In *Proceedings of IROS*, Beijing
30. Hosang, J., Benenson, R., Dollár, P., & Schiele, B. (2015, 7 August). What makes for effective detection proposals? *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, IEEE (*99*). http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7182356.
31. Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Proceedings of CVPR*, Minneapolis.
32. Iida, F. (2003). Biologically inspired visual odometer for navigation of a flying robot. *Robotics and Autonomous Systems, 44*(3–4), 201–208
33. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.
34. Jensfelt, P., Kragic, D., Folkesson, J., & Björkman, M. (2006). A framework for vision based bearing only 3D SLAM. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Orlando.
35. Johnson-Roberson, M., Bohg, J., Björkman, M., & Kragic, D. (2010). Attention based active 3d point cloud segmentation. In *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei.
36. Kühn, B., Schauerte, B., Kroschel, K., & Stiefelhagen, R. (2012). Multimodal saliency-based attention: A lazy robots approach. In *Proceedings of the 25th International Conference on Intelligent Robots and Systems (IROS)*, Vilamoura.

37. Loach, D., Frischen, A., Bruce, N., & Tsotsos, J. K. (2008). An attentional mechanism for selecting appropriate actions afforded by graspable objects. *Psychological Science, 19*(12), 1253–1257
38. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV), 60*(2), 91–110.
39. Martín-García, G., & Frintrop, S. (2013). A computational framework for attentional 3D object detection. In *Proceedings of the Annual Conference of the Cognitive Science Society*, Berlin.
40. Martín-García, G., Potapova, E., Werner, T., Zillich, M., Vincze, M., & Frintrop, S. (2015). Saliency-based object discovery on rgb-d data with a late-fusion approach. In *Proceedings of the IEEE International Conference Robotics and Automation (ICRA)*, Seattle.
41. Matas, J., Chum, O., Urban, M., & Pajdla, T. (2002). Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference*, Cardiff.
42. Meger, D., Forssen, P.-E., Lai, K., Helmer, S., McCann, S., Southey, T., Baumann, M., Little, J. J., Lowe, D. G., & Dow, B. (2008). Curious george: An attentive semantic robot. *Journal Robotics and Autonomous Systems, 56*(6), 503–511.
43. Meger, D., Muja, M., Helmer, S., Gupta, A., Gamroth, C., Hoffman, T., Baumann, M., Southey, T., Fazli, P., Wohlkinger, W., Viswanathan, P., Little, J. J., Lowe, D. G., & Orwell, J. (2010). Curious george: An integrated visual search platform. In *Canadian Conference on Computer and Robot Vision*, Ottawa.
44. Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2000). Intermodal selective attention in monkeys. I: Distribution and timing of effects across visual areas. *Cerebral Cortex, 10*(4), 343–358.
45. Miau, F., Papageorgiou, C., & Itti, L. (2001). Neuromorphic algorithms for computer vision and attention. In *Proceedings of the SPIE 46 Annual International Symposium on Optical Science and Technology*, Bellingham (vol. 4479, p. 12–23).
46. Mitri, S., Frintrop, S., Pervölz, K., Surmann, H., & Nüchter, A. (2005). Robust object detection at regions of interest with an application in ball recognition. In *IEEE Proceedings of the International Conference on Robotics and Automation (ICRA '05)*, Barcelona (pp. 126–131)
47. Muhl, C., Nagai, Y., & Sagerer, G. (2007). On constructing a communicative space in HRI. In J. Hertzberg, M. Beetz, & R. Englert (Eds.), *Proceedings of the 30th German Conference on Artificial Intelligence (KI 2007)*, Osnabrück. Springer.
48. Nagai, Y. (2009). From bottom-up visual attention to robot action learning. In *IEEE 8th International Conference on Development and Learning*, Shanghai.
49. Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103.
50. Nickerson, S. B., Jasiobedzki, P., Wilkes, D., Jenkin, M., Milios, E., Tsotsos, J. K., Jepson, A., & Bains, O. N. (1998). The ARK project: Autonomous mobile robots for known industrial environments. *Robotics and Autonomous Systems, 25*(1–2), 83–104.
51. Ouerhani, N., Bur, A., & Hügli, H. (2005). Visual attention-based robot self-localization. In *Proceedings of European Conference on Mobile Robotics (ECMR 2005)*, Ancona (pp. 8–13).
52. Potapova, E., Varadarajan, K. M., Richtsfeld, A., Zillich, M., & Vincze, M. (2014). Attention-driven object detection and segmentation of cluttered table scenes using 2.5D symmetry. In *IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong
53. Rasolzadeh, B., Björkman, M., Huebner, K., & Kragic, D. (2010). An active vision system for detecting, fixating and manipulating objects in real world. *International Journal of Robotics Research, 29*(2–3), 133–154
54. Rotenstein, A., Andreopoulos, A., Fazl, E., Jacob, D., Robinson, M., Shubina, K., Zhu, Y., & Tsotsos, J. (2007). Towards the dream of intelligent, visually-guided wheelchairs. In *Proceedings of the 2nd International Conference on Technology and Aging*, Toronto
55. Rudinac, M., Kootstra, G., Kragic, D., & Jonker, P. (2012). Learning and recognition of objects inspired by early cognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vilamoura.

56. Ruesch, J., Lopes, M., Bernardino, A., Hörnstein, J., Santos-Victor, J., & Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention: A framework for the humanoid robot iCub. In *Proceedings of the International Conference on Robotics and Automation (ICRA 2008)*, Pasadena
57. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision 115*(3), 211–252.
58. Schauerte, B., & Fink, G. A. (2010). Focusing computational visual attention in multi-modal human-robot interaction. In *Proceedings of the 12th International Conference on Multimodal Interfaces (ICMI)*, Beijing.
59. Schauerte, B., Kühn, B., Kroschel, K., & Stiefelhagen, R. (2011). Multimodal saliency-based attention for object-based scene analysis. In *Proceedings of IROS*, San Francisco.
60. Schauerte, B., Richarz, J., & Fink, G. A. (2010). Saliency-based identification and recognition of pointed-at objects. In *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, Taipei.
61. Schillaci, G., Bodiroža, S., & Hafner, V. V. (2012). Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *International Journal of Social Robotics*.
62. Siagian, C., & Itti, L. (2009). Biologically inspired mobile robot vision localization. *IEEE Transaction on Robotics, 25*(4), 861–873.
63. SRVC: The 2009 Semantic Robot Vision Challenge. http://google-opensource.blogspot.de/2010/01/2009-semantic-robot-vision-challenge.html.
64. Tsotsos, J. K., Verghese, G., Stevenson, S., Black, M., Metaxas, D., Culhane, S. Dickinson, S., Jenkin, M., Jepson, A., Milios, E., Nuflo, F., Ye, Y., & Mann, R. (1998). PLAYBOT: A visually-guided robot to assist physically disabled children in play. *Image and Vision Computing 16, Special Issue on Vision for the Disabled, 16*(4), 275–292.
65. Tuytelaars, T., & Mikolajczyk, K. (2007). Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision, 3*(3), 177–280.
66. Vijayakumar, S., Conradt, J., Shibata, T., & Schaal, S. (2001). Overt visual attention for a humanoid robot. In *Proceedings of the International Conference on Intelligence in Robotics and Autonomous Systems (IROS 2001)*, Hawaii (pp. 2332–2337).
67. Vogel, J., & de Freitas, N. (2008). Target-directed attention: Sequential decision-making for gaze planning. In *Proceedings of ICRA*, Pasadena.
68. Walther, D. (2006). Interactions of visual attention and object recognition: Computational modeling, algorithms, and psychophysics. PhD thesis, California Institute of Technology, Pasadena.
69. Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*(9), 1395–1407.
70. Walther, D., & Koch, C. (2007). Attention in hierarchical models of object recognition. *Computational Neuroscience: Theoretical insights into Brain Funciton, Progress in Brain Research, 165*, 57–78.
71. Xu, T., Pototschnig, T., Kühnlenz, K., & Buss, M. (2009). A high-speed multi-GPU implementation of bottom-up attention using CUDA. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Kobe.
72. Xu, T., Zhang, T., Kühnlenz, K., & Buss, M. (2010). Attentional object detection of an active multi-focal vision system. *International Journal of Humanoid Robotics, 7*(2), 223–243.
73. Yamauchi, B. (1997). A frontier-based approach for autonomous exploration. In *Proceedings of the 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, Monterey (pp. 146–151).

Part IV
Conclusion

Chapter 22
The Future of Attention Models: Information Seeking and Self-awareness

Matei Mancas, Vincent P. Ferrera, and Nicolas Riche

This book contributes to the crucial endeavor of understanding and modeling human attention. It gives an overview of physiological and computer science models, an extensive approach to model validation, as well as new trends and applications of attention models. It also paves the way for further investigations. Some directions for future research are discussed in the next section, in relation to the major contributions summarized above. In the second section, a perspective on issues beyond attention, such as higher level processing and consciousness, is provided. We propose that human attention can be viewed as a suite of computational strategies that are essential for autonomous behavior by agents both natural and artificial. The study of attention should go beyond filtering of sensory data to develop an understanding of how relevant and valuable information is actively gathered by agents who possess an integrated awareness of both their internal goals, needs and abilities and external sources of sustenance or danger. This kind of awareness implies an ability to model both the environment and the self that acts within that environment. Understanding the computational mechanisms underlying active, goal-oriented attention may be a first step toward artificial consciousness.

M. Mancas (✉) • N. Riche
Numediart Institute, University of Mons (UMONS), 31, Bd. Dolez, 7000 Mons, Belgium
e-mail: matei.mancas@umons.ac.be; nicolas.riche@umons.ac.be

V.P. Ferrera
Department of Neuroscience, Columbia University, 1051 Riverside Drive, Unit 87, New York, NY 10032, USA
e-mail: vpf3@cumc.columbia.edu; vincent.ferrera@gmail.com

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5_22

22.1 Perspectives in Attention Modeling

22.1.1 Models

22.1.1.1 Computing Eye Scan Paths from Saliency Maps

Most saliency models take color images as input and produce saliency maps that estimate the probability distribution of the gaze in the image. The static nature of these maps could be an issue for some applications as these models do not predict the temporal sequence of human fixations (also called scan path). What is the order of fixations? How is the image seen dynamically? Some models like FSM [1], DVA [2] or ESAL [3] propose algorithms to predict the scan path from a saliency map. However, this question deserves to be investigated more deeply. Indeed, the dynamic nature of attention was less investigated in the saliency models than in the visibility models (see Chap. 7).

22.1.1.2 Modeling Visual Attention with Learning Algorithms Such as Deep Neural Networks

Recently, with the advent of deep learning in many areas, some general multilayer deep networks have been proposed to detect human fixations (see Chap. 7). Although the neural network methods are very effective, some issues remain to be solved by further research to mix the advantages of those models with more classical handcrafted ones. First, current deep neural networks remain unclear about the nature of the learned representation and about their results. This is not the case for handcrafted models which are often clear and interpretable. In neural networks, it is not easy to distinguish learning of bottom-up factors from top-down factors or even viewer biases like the center bias. The second major issue is the dependency of deep neural network on databases. The database containing eye tracking data is small compared to databases in other fields like object recognition. There are solutions like mouse tracking or webcam-based systems to build medium-sized databases but to the detriment of precision. Moreover, the database has to be representative of each situation because the neural network will be trained on this database.

22.1.1.3 Multimodal Modeling of Attention

One of the future trends will be to aggregate results from attention algorithms on videos, but also on sound or 3D objects. This integration will drastically augment the already numerous engineering applications of attention modeling. Of course those new models will need new ground truth and new validation techniques, but this effort is crucial to boost the attention modeling community and to augment its visibility both in other research communities and in industry. The arrival of those new techniques comes along with an increasing access to video and 3D data using simple devices.

22.1.1.4 Curiosity: Uncertainty Reduction Through Guided Exploration

When humans encounter a novel environment, one of their first priorities is to learn whom, what and where to attend. Random, novelty-driven exploration has to be balanced with hypothesis-driven mechanisms for identifying sources of valuable or meaningful information [4]. This kind of everyday goal-directed information seeking can be called curiosity. Reward circuits in the brain respond not only to the likelihood and amount of reward but also to uncertainty [5]. Quantifying uncertainty is essential for decision-making [6]. Attention can be thought of as a mechanism to seek information that reduces uncertainty through guided exploration. Future studies should focus on the neural and computational mechanisms of attention and decision-making in environments where risk, uncertainty and ambiguity can be controlled. An autonomous agent, such as a robot, should be able to pose questions that are relevant to its current situation and to formulate plans to seek answers to those questions.

22.1.2 Model Validation

22.1.2.1 Exploring Databases with New Classes of Stimuli

Currently, saliency models for still images are very effective on natural images. However, their performance is frequently disappointing on other kinds of stimuli like websites, paintings, etc. Indeed, for such stimuli, there is a significant decrease in performance. To address this issue, more databases like [7] have to be collected on diverse sources (websites, advertising, etc.) to understand what attracts attention when observing stimuli different from classical "natural images."

22.1.2.2 Large-Scale Human Data During Natural Explorations of Videos

Recently, large-scale datasets have become a necessity in several domains of computer vision. In saliency, it is complicated to obtain large-scale datasets due to the nature of the ground truth. Conventional eye tracking studies are time consuming. However, there are alternatives such as mouse tracking [8] or webcam-based eye tracking [9], but they are currently applied only on still image databases.

22.1.2.3 Metrics for Comparing Temporal Sequences of Eye Fixations

As is the case for saliency models, standard metrics to compare the output of attention models and human fixations are static. They do not take into account the temporal sequence of human fixations. Although in [10], the authors proposed some dynamic metrics to compare two scan paths, there are still too few databases and models providing temporal sequences of fixations.

22.1.2.4 Study of Biases Inside Stimulus Categories

As described in [11], there are different classes of stimuli, such as noisy images, indoor pictures, etc. For each type of image, major biases exist either from viewers who are watching static and dynamic scenes or from features of the stimuli. The impact of salient object size in outdoor pictures or different kind of movements in videos is addressed in [12]. These studies show how saliency models manage the size of salient objects or videos from static or moving cameras. They also help to have a better understanding of current model biases and show how to improve them. The recording of new databases with other biases or the addition of new categories inside existing datasets is needed.

22.1.2.5 Other Application-Based Validation Frameworks

A new trend in attention model validation is to have application-based protocols. In Chap. 19, for example, a system is established to compare several attention models for a precise application (CBIR), as the rankings might be different between the existing models depending on the application of interest.

22.1.2.6 Audiovisual Validation Based on Eye Tracking Data

There is no easy way to validate the use of audio information in saliency models because there is no auditory ground truth and the dimensions of audio and video information are different. The idea is thus to find a way to build an audiovisual validation protocol. Eye tracking data acquired with visual and audiovisual conditions should be used. All saliency models for videos could be validated by only one condition inside the validation framework. Some audio features could be used to weight the saliency results on videos and better predict where people attend.

22.1.2.7 Deployment of Attention During the Performance of Natural Tasks

Attention is often studied in situations where subjects are given various cues or reinforcement. In other words, the tasks are structured to guide attention to locations, features, or objects that have been chosen by the experimenter. This begs of the question of how subjects naturally deploy attention when performing everyday tasks such as driving, making a sandwich, or playing sports. Ballard, Hayhoe, and colleagues [13] have recently developed an immersive virtual reality system for recording eye, head, and hand movements when human subjects are performing simple tasks. Such systems allow experimental control over external variables like novelty, reward, and context while imposing minimal constraints on subject behavior. In the future, such systems can be combined with mobile EEG recording to map the brain activity during natural, goal-directed behavior.

22.2 Attention Beyond Information Filtering

22.2.1 Structure, Semantics, and Objects

In Chaps. 1 and 2, we saw that attention is the gate between the outer world (or subconscious in the case of dreams) and consciousness. Part of the attention process occurs before information reaches awareness; it is pre-attentive or reflex and can use a parallel processing strategy to acquire data. The attentive process is conscious and uses a serial discovery strategy based on selected data from the pre-attentive acquisition. Serial deployment of attention is highly dynamic and depends on a number of factors:

- Bottom-up: related to information maximization of the field of view where features are extracted and processed.
- Top-down: depends on the task on hand or on specific object recognition.
- Previous eye fixation location: the dynamical process of vision provides different bottom-up cues depending on the fixation location.

The eye scan path can be very different from one person to another depending on the image content, the initial eye fixations, and the importance of the task. Attention acts as an information filter and prioritization strategy, transforming a huge amount of unstructured information to a serial discovery of the most interesting areas or objects. This eye scan path shows that attention not only discovers important areas but also seeks to find how those areas are related (object-subject relationship). Moreover, attention due to object changes (motion) has a huge impact on the final result and is obtained using a different pathway in the brain than an object-oriented attention. Indeed, visual signals mainly follow two different pathways in the brain, a dorsal and a ventral, one focusing more on object recognition (what?) and the other on space and movement (where?). This distinction between the fundamental question of what (the objects/subjects) and where/how (their interaction) is directly related to the minimal semantics of a sentence composed of a subject/verb/object. In [14], it is hypothesized that children already base their first thoughts (in a language nonspecific way) as a "transfer" (verb) between an "agent" (subject) and a "willing recipient" (object). When children learn a language, they actually try to encode this first conscious representation into a given language, as an expression of their consciousness.

Beyond being a simple information filter, attention gives cues on which areas (subject/object) are interesting in the scene and their changes (verb), leading to the first notion of semantics. It seems that the influence of the attentive system goes beyond the initial role of selecting the information that gains access to awareness, and it also intends to provide it with a structure. Indeed in Chap. 15, it is shown how attention can be related to the notions of "proto-objects" or "objectness" which are the first steps toward the notion of object. This represents a milestone in the understanding of our environment.

22.2.2 *Emotions, Memory, and Actions*

Attention and memorability are heavily interlinked. While the influence of emotions and memory on attention is obvious and this influence is part of the definition of top-down attention, in the other direction (attending toward emotions and memory), things are less clear. Nevertheless, even if the links are not as obvious as one would think, the first step toward memorizing an object may require attention. In Chap. 18, the link between visual attention and image memorability is detailed as an application of attention modeling.

In the brain, a basic structure within the thalamus provides very interesting clues about a possible relationship between attention, memory, and emotions. This is the Papez circuit (Fig. 22.1) which was initially seen as a mechanism for emotions [15].

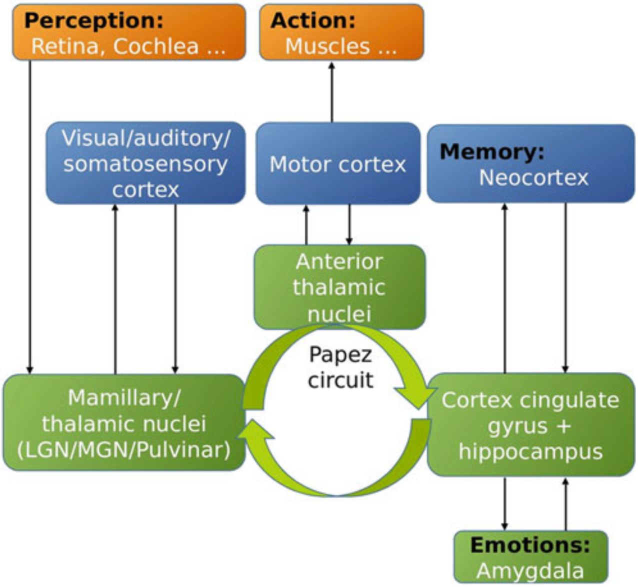


Fig. 22.1 A simplified schematic view of the Papez circuit. In *green* the areas which are directly involved in the circuit. This circuit is linked to both emotions and memorability. Indeed, if impaired, new data will not go into the long-term memory, but older recollection is not affected

The main element of this circuit is the hippocampus which is related to episodic and spatial memory [16].

At the rostral end of the temporal lobe, a collection of nuclei called the amygdala is involved in emotions. The emotions related to the amygdala are mainly negative, but positive emotions also evoke a response in this area [17]. The amygdala also responds to high interest or unusual images which attract attention [18].

At the other side of the Papez circuit, one can find the mamillary body and the anterior thalamic nuclei. The mamillary body relays the output of the hippocampus and amygdala to the anterior thalamus and has an important role in spatial memory [19]. The anterior thalamic nuclei are linked to action and the motor cortex. Other thalamic nuclei that are important in sensory processing and attention are the lateral geniculate nucleus (relays signals from the retina to visual cortex [20]), the medial geniculate nucleus (auditory perception), and the pulvinar which is directly related to attention by modulating or gating sensory signals in relay nuclei [21].

It is very difficult to isolate locations in the brain which are responsible for complex tasks such as attention (see Chap. 4) or memory, but the Papez circuit is of particular interest because in the limbic system, attention, memory, action, and emotions have a close anatomical proximity and are all needed in the process of memory formation. Thus, attention is heavily interconnected with emotions, memory, and action. Indeed the effects of an agent's own body on its environment are highly important in scene understanding, and it also has a crucial impact in the feeling of self-awareness and ownership which are at the basis of consciousness.

22.2.3 Toward Consciousness

In [22] Schmidhuber links the notion of attention to compression progress. It is interesting to note that the concepts of compression and prediction are not only emphasized in [22], but this concept is quite well accepted as a driver of our brain. The idea of matching sensory inputs to stored patterns from memory to predict what will happen in the future is also present in [23] where the interaction of neocortex, hippocampus, and thalamus is central to the model. The brain is analyzed as an encoding problem, complete with error-correction codes.

In [22], the incoming data has regularities and irregularities: each time regularities are found, they can be further compressed and simplified. A better explanation of the environment results in its simplification, and thus its compression. A complex concept which is very well compressed will be considered as subjectively beautiful. For example, a beautiful face will be one close to the average face previously learned, as only few bits are necessary to code the deviation of the new face from the previously learned one. Also symmetries are important regularities which can be compressed and which can be seen as subjectively beautiful. The approach of Schmidhuber implies that a subjectively beautiful woman, for example, has a face with perfect symmetries and very close to the average of faces that the observer has seen (related to the observer family and social class). But beautiful does not mean interesting. In [22], the author states that beauty is interesting only during a given time period, as long as it is new. He thus defines the interest or attention as the derivative of the subjective beauty. A beautiful object attracts attention only if the beauty (compression) progresses. For example, a perfect beautiful woman for a subjective observer will attract attention at the beginning, but this attention rapidly fades. If this woman has a small defect, the need for compression of this small defect will provide more interest to this woman than to a perfect one. It is important to differentiate the subjective beauty due to information compression from the beauty coming from external rewards (linked to emotions and memory). In [22] attention is also linked to consciousness as compression induces the use of symbols or codes summarizing regularities with a high occurrence probability in the information. In the same way, a subject which occurs very frequently in a subjective observation of the world is the observer or agent himself; thus, it is very efficient for him to find a code for himself, and by doing this, the agent becomes self-aware and conscious.

There are several models of consciousness. In those models, attention plays an important role as in [24] where the colliculus, in addition to its role of fusing attention of several modalities (visual and auditory attention) on a unique topographic map, also generates a simulation of the sensory world that corresponds to primary (sensory) consciousness. The interaction between perception and action (eye orientation) is a key to the development of consciousness.

In [25] attention is an emerging property due to subsets of neurons which activate at the expense of other groups related to stimuli which will thus remain unattended. When no stimulation is present, groups of neurons related to a past stimulus can activate and synchronize with other related groups of neurons. In that way, the network pays attention to a stimulus which is actually not present but occurred in the past. This way of thinking about a stimulus is viewed in [25] as the emerging of consciousness, which is like "attention to memory."

In one model of consciousness, attention and consciousness are viewed inseparable: the CODAM model of John Taylor [26]. Taylor developed and refined his model until [27] which is a vibrant and complete legacy of 10 years of research on the notion of consciousness. CODAM stands for "Corollary Discharge of Attention Movement." The importance of attention is signified by its appearance in the name of the model. This model is based on control theory. Modules representing goals integrate top-down attention with a module for the input signal which focuses on bottom-up attention. Two other key modules are used in CODAM. One is the working memory buffer and the other is the corollary discharge, which is the key of the model. This corollary discharge is a copy of the attention movement control signal, and it activates the working memory buffer. The corollary discharge signal is an internal copy of the location-to-be-attended, and it can be interpreted as a signal of the ownership of the about-to-be-experienced content as well as a signal of guarantee of the content-to-be-acquired to be the one that the agent chooses to acquire. This signal of ownership and self-awareness or correct self-identification is of a crucial importance in the building of the "I" and the consciousness. The CODAM model also helps explain pathologies like schizophrenia where a higher weight to the corollary discharge is enough to amplify the self-awareness so the subject is too involved with his own thoughts and cannot cope with the reality of an external world.

22.2.4 The Rise and Fall of Consciousness

CODAM is one of the models where attention is tightly related to consciousness. But is this hypothesis necessary? In [28] the author insists on the clear separation of consciousness and attention. Attention may only select among already conscious concepts. Also attention is a process where not all of the attended regions reach the conscious state: looking is not watching and hearing does not mean listening. Attention can be parallel and serial. During parallel strategies, the gist of a scene or other global details can become conscious without the need of being explicitly attended. However, if attention is considered in both covert and overt aspects, it is hard to find examples where attention and consciousness do not match. Even if attention is NOT consciousness, it is safe to argue that the relationship between the two concepts is very tight and that full consciousness requires some form of attentive mechanism.

Given these considerations, measuring the state of attention can be an indirect measure of the state of consciousness. Meditation studies using EEG, PET, or fMRI imaging have found that (1) there might be an increased attentional control in the frontal brain regions [29, 30] and (2) this occurs at the same time primary visual cortex regions acquiring signals from the outside world decrease [31, 32]. The above studies seem to lead toward the possibility that meditation is an augmented consciousness state, but focused on the self and not on the external world.

A strange case was reported in [33] in which a young woman was in a vegetative state following severe brain injuries sustained in a car accident. While no overt sign of awareness could be observed, an fMRI study suggested that she was able to hear, understand speech, and follow doctor instructions. Indeed, when she was asked to think about playing tennis, for example, her brain acted in a very similar way as a control patient. This is a case showing that despite no visible awareness, consciousness can still be at work.

A final question is how attention and consciousness initially develop and later decline throughout the lifespan of an individual. In [34, 35] evidence of REM sleep as early as around the sixth month of fetal development suggests that even with rudimentary auditory sensing, spatial coding of sound and processing of structure and meaning are already at work. Following the CODAM theory, these early attentional signs suggest that the first sense of ownership begins to be set up before birth. However, the exact mechanisms of attentional development remain unknown.

On the other side of life, during agony, one would expect attention and consciousness to slowly fade until their complete shutdown when vital functions cease. While work relating agony to attention and consciousness is very sparse, some evidence supports the idea that, contrary to the slow fading of consciousness, the approach of death induces massive cerebral activity closer to attentional fireworks than slow shutdown. As [33] shows in the case of a vegetative state, the visible state of a person is not necessarily related to his/her awareness or state of consciousness. In [36, 37] the authors show that in rare cases, patients can recall auditory details during general anesthesia where attention and consciousness are presumably inhibited or absent.

In [38], the authors performed EEG measurements on rats during waking periods, during anesthesia, and after induced cardiac arrest. While awake, the rats showed a normal EEG. During anesthesia, the activity drastically decreased. But, just after the heart stopped, EEG activity exceeded activity levels found during the conscious waking state (especially in the gamma range, which is related to visual attention [39]). Contrary to expectations, attention-related EEG signals were greater than normal for roughly 30 s after the heart stopped beating. In [40], similar results are described, this time demonstrating hyperactivity of the sympathetic system. Sympathetic system arousal generally occurs during stressful situations that require rapid preparation of the body to fight or flee from danger. This hypothesis might explain part of the sudden near-death sympathetic hyperactivity in the brain as a kind of hyperawareness might improve the odds of survival in cases where the animal is able to escape or recover from a near-death situation.

In humans, about 10 % of dying people are conscious during their agony. Within those 10 %, it is estimated that 50–60 % have deathbed visions. A deathbed vision is thus present in most of the deaths where people are conscious, and they consist in visions of dead relatives or friends, religious figures, or a language related to travel [41]. Those deathbed visions are positive visions that comfort the dying and prepare them psychologically for death [42]. The visions are positive, structured, and meaningful, in contrast to the case of more chaotic and mostly auditory illusions associated with mental illnesses, dementia, delirium, or drugs which are mostly negative and with little meaning [43]. Interestingly the near-death visions in humans could be related to the high arousal in visual attention demonstrated in moribund rats. They are also concomitant with people gazing in fixed regions with few eye movements and dilated pupils and reduced social interaction with people nearby. Bottom-up attention seems thus reduced, while all the attention focuses on the top-down component which is related to memory and emotions. The state of consciousness of the dying is closer to that experienced during meditation and focus on the self and not on outside. The higher weight of top-down component might explain part of those visions, but it seems difficult to explain them entirely. Another explanation might be a psychological defense against the idea of ceasing to exist and the shutdown of consciousness. But again, this approach probably only partly explains those visions.

These studies show that the shutdown of consciousness is not necessarily slow, but in the very last moments (when physically possible), people experience a very high degree of attention and consciousness before brain death and maybe even after heart stops like in rats. However, due to the lack of scientific experiments and measurements on dying people, it is very difficult to provide an objective view on the topic.

In this book, we focused on attention from the level of single neurons to visual detection and on through computational modeling of salience and scan paths. Those models are mostly concerned with information filtering, but some applications like image memorability (see Chap. 18) go further than simple information filtering. In this last chapter, we provided some insights on structure and objects, memorability, and the first steps toward consciousness where attention plays a crucial role. We also saw that attention and consciousness are at the very beginning of life (during fetal development) focused inwardly on the self; they then open to others during life and return again to the self during the natural dying process. How exactly attention begins to work in babies and how it acts in the very final moments of life still remains very speculative and reaches the boundaries of current science.

References

1. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 20*(11), 1254–1259.
2. Hou, X., & Zhang, L. (2008). Dynamic visual attention: Searching for coding length increments. *Proceedings of Neural Information Processing Systems (NIPS), 5*, 7.
3. Avraham, T., & Lindenbaum, M. (2010). Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 32*(4), 693–708.
4. Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods, 45*(1), 251–266.
5. Schultz, W., Preuschoff, K., Camerer, C., Hsu, M., Fiorillo, C. D., Tobler, P. N., et al. (2008). Explicit neural signals reflecting reward uncertainty. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 363*(1511), 3801–3811. doi:10.1098/rstb.2008.0152.
6. Grinband, J., Hirsch, J., & Ferrera, V. P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron, 49*(5), 757–763.
7. Yujian, L., & Bo, L. (2007). A normalized levenshtein distance metric. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29*(6), 1091–1095.
8. Borji, A., & Itti, L. (2015). CAT2000: A large scale fixation dataset for boosting saliency research. arXiv preprint arXiv:1505.03581.
9. Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). *Salicon: Saliency in context*. Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1072–1080. Boston, US.
10. Gottlieb, J., Oudeyer, P. Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Science, 17*(11), 585–593. doi:10.1016/j.tics.2013.09.001.
11. Shen, C., & Zhao, Q. (2014). Webpage saliency. In *The European conference on computer vision (ECCV)* (pp. 33–46). Springer, Zurich.
12. Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., & Xiao, J. (2015). Turkergaze: Crowdsourcing saliency with webcam based eye tracking. arXiv preprint arXiv:1504.06755.
13. Hayhoe, M., & Ballard, D. (2014). Modeling task control of eye movements. *Current Biology, 24*(13), R622–R628. doi:10.1016/j.cub.2014.05.020. Review.
14. Goldberg, A. (1994). *Constructions, a construction grammar approach to argument structure*. Chicago: Chicago University Press.

15. Papez, J. W. (1937). A proposed mechanism of emotion. *Archives of Neurology and Psychiatry, 38*(4), 725–743.
16. Burgess, N., Maguire, E. A., & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron, 35*(4), 625–641.
17. Garavan, H., et al. (2001). Amygdala response to both positively and negatively valenced stimuli. *Neuroreport, 12*(12), 2779–2783.
18. Hamann, S. B., et al. (2002). Ecstasy and agony: Activation of the human amygdala in positive and negative emotion. *Psychological Science, 13*(2), 135–141.
19. Schmidhuber, J. (2009). Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Anticipatory Behavior in Adaptive Learning Systems, 5499*, 48–76.
20. O'Connor, D. H., et al. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience, 5*(11), 1203–1209.
21. Arend, I., et al. (2008). 15-the role of the human pulvinar in visual attention and action: Evidence from temporal-order judgment, saccade decision, and antisaccade tasks. *Progress in Brain Research, 171*, 475–483.
22. Vann, S. D. (2010). Re-evaluating the role of the mammillary bodies in memory. *Neuropsychologia, 48*(8), 2316–2327.
23. Hawkins, J., & Blakeslee, S. (2007). *On intelligence*. New York: Macmillan.
24. Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences, 30*(01), 63–81.
25. Izhikevich, E. M. (2006). Polychronization: Computation with spikes. *Neural Computation, 18*(2), 245–282.
26. Taylor, J. G. (2002). Consciousness: Theories of. In M. A. Arbib (Ed.), *Handbook of brain theory and neural computation*. Cambridge, MA: MIT Press.
27. Taylor, J. G. (2013). *Solving the mind-body problem by the CODAM neural model of consciousness?* Dordrecht: Springer.
28. Lamme, V. A. F. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences, 7*(1), 12–18.
29. Shear, J., & Jevning, R. (1999). Pure consciousness: Scientific exploration of meditation techniques. *Journal of Consciousness Studies, 6*(2–3), 189–210.
30. Jevning, R., Wallace, R. K., & Beidebach, M. (1992). The physiology of meditation: A review. A wakeful hypometabolic integrated response. *Neuroscience & Biobehavioral Reviews, 16*(3), 415–424.
31. Herzog, H., et al. (1990). Changed pattern of regional glucose metabolism during yoga meditative relaxation. *Neuropsychobiology, 23*(4), 182–187.
32. Baerentsen, K. B., et al. (2001). Onset of meditation explored with fMRI. *NeuroImage, 13*(6), 297.
33. Owen, A. M., et al. (2006). Detecting awareness in the vegetative state. *Science, 313*(5792), 1402–1402.
34. Mirmiran, M. (1995). The function of fetal/neonatal rapid eye movement sleep. *Behavioural Brain Research, 69*(1), 13–22.
35. Hopson, J. L. (1998). Fetal psychology. *Psychology Today, 31*(5), 44.
36. Osterman, J. E., et al. (2001). Awareness under anesthesia and the development of posttraumatic stress disorder. *General Hospital Psychiatry, 23*(4), 198–204.
37. Sebel, P. S., et al. (2004). The incidence of awareness during anesthesia: A multicenter United States study. *Anesthesia and Analgesia, 99*(3), 833–839.
38. Borjigin, J., et al. (2013). Surge of neurophysiological coherence and connectivity in the dying brain. *Proceedings of the National Academy of Sciences, 110*(35), 14432–14437.
39. Müller, M. M., Gruber, T., & Keil, A. (2000). Modulation of induced gamma band activity in the human EEG by attention and visual information processing. *International Journal of Psychophysiology, 38*(3), 283–299.
40. Fujiwara, A., & Kobata, H. (2015). Paroxysmal sympathetic hyperactivity after near-hanging. *The American Journal of Emergency Medicine, 33*(5), 735-e1.

41. Mazzarino-Willett, A. (2009). Deathbed phenomena: Its role in peaceful death and terminal restlessness. *American Journal of Hospice and Palliative Medicine, 27*(2), 127–133.
42. Brayne, S., Lovelace, H., & Fenwick, P. (2008). End-of-life experiences and the dying process in a Gloucestershire nursing home as reported by nurses and care assistants. *American Journal of Hospice and Palliative Medicine, 25*(3), 195–206.
43. Sondermann, M., & Janzen, C. (2011) Deathbed phenomena: Real or imagined? In *Palliative Care Conference*, *Edmonton, US*.

Index

A

Additional comparative statistical analysis, 251, 252

Amplitude-based metrics, 213, 224, 238

Anterior cingulate, 40, 41, 44, 54

Applications, 3–6, 9, 13, 25, 125–127, 130, 142, 160, 177, 195, 278–279, 308, 331–369, 399, 400, 406, 407, 410, 415, 422–430, 447, 448, 457

Attention, 1–6, 9–11, 21–36, 39–54, 63–83, 87–102, 105–119, 123–138, 141, 178, 191, 227, 271–286, 291, 305, 331, 380, 399–410, 413, 447–457

Attention modeling, 105, 107, 161, 285, 294, 299, 331–333, 347, 360, 369, 448–452

Audio saliency, 112, 113, 296, 301, 359

Audiovisual, 292, 294–301, 341, 357, 436, 450

Awareness, 1–3, 9, 10, 12, 276, 341, 447–457

B

Beauty, 454

Bottom-up, 5, 15, 17, 18, 42, 46, 49, 50, 52, 53, 105–119, 123–138, 141–173, 177–189, 192, 204, 248–250, 274, 278, 279, 281, 283, 284, 286, 291, 310, 311, 319, 320, 325, 327, 331, 347, 350, 359, 363–365, 367, 369, 380, 383, 384, 386, 390–391, 400, 406, 410, 418–420, 422, 431, 436, 437, 448, 451, 454, 456

Brain damage, 39–54

C

Cognitive neuroscience, 9, 13–15

Cognitive psychology, 3, 9, 11–14

Computer science, 5, 9, 13, 15, 17, 18, 36, 105–106, 119, 360, 364, 447

Consciousness, 2, 3, 5, 10, 11, 13, 16, 447, 451, 453–457

Content based image retrieval, 379–396, 400

Covert, 11, 16–18, 21, 42, 124, 282, 283, 291, 455

D

Databases, 115, 161, 191–205, 209, 228, 234–237, 240, 245, 250, 255–257, 266, 267, 285, 297, 306, 364, 401, 448–450

Decision criterion, 65, 68–71

Depth, 26, 87, 111, 112, 114, 272, 305–309, 312–316, 318–320, 324–327, 344, 347, 349, 358, 416, 419, 436

Detectability, 64–67, 72–76, 79, 80, 82, 83, 89, 90, 94, 95, 97–99, 102, 301

Disparity, 306, 312–316, 344

Distribution-based metrics, 212, 213, 215–219, 237, 238

Dynamic saliency models, 201, 340

E

Electroencephalography (EEG), 13, 30–32, 35, 36, 52, 451, 455, 456

Emotions, 16–18, 36, 105, 331, 452–454, 456

Evaluation, 28, 161, 191–205, 209, 210, 227, 228, 230, 237, 239, 240, 245–267, 278, 324, 325, 367, 368, 384, 386, 388, 393–395, 407

M. Mancas et al. (eds.), *From Human Attention to Computational Attention*,
Springer Series in Cognitive and Neural Systems 10,
DOI 10.1007/978-1-4939-3435-5

Experiments, 30, 123, 127, 192, 193, 227, 240, 260, 292, 293, 313, 384, 393, 400, 410, 419, 421, 425, 456
Eye-tracking, 2, 13, 15, 22, 24–27, 29, 115, 118, 192, 193, 195–203, 205, 209, 246, 248, 249, 257–260, 267, 292–294, 296, 297, 346, 361, 362, 364, 368, 395, 401
Eye tracking-based databases, 228, 231, 235

F
Fairly, 96, 227, 240, 401, 403
Fano factor, 90, 91, 96–98
Features, 12, 26, 50, 102, 105, 123, 144, 177, 203, 246, 271, 291, 305, 331, 380, 403, 419, 450
Frontal lobe, 41, 44
Functional magnetic resonance imaging (fMRI), 32–36, 45, 49, 52, 126, 273, 455
Future of attention, 447–457

G
Gain control, 102
Ground-truths, 191

H
Human vision, 123, 133, 138, 305, 315, 327, 405, 418, 419

I
Image, 4, 12, 23, 67, 101, 106, 123–138, 141–173, 177, 191, 210, 227, 245, 276, 293, 305, 331, 379–396, 400, 416, 448
Industrial applications, 356

L
Location-based metrics, 209–212, 219–223, 237

M
Magnetoencephalography (MEG), 13, 32–34, 36
Memory, 3, 11–14, 16–18, 36, 39, 105, 114, 115, 258, 274, 275, 277, 331, 364, 365, 369, 395, 424, 452–454
Metrics, 5, 192, 209–224, 227–229, 232, 235, 237–241, 245, 250–257, 260, 261, 263, 266, 267, 314, 364, 403, 404, 449, 450
Mismatch negativity (MMN), 30–32
Models, 4, 10, 30, 79, 88, 105, 124, 141–173, 177–189, 191–205, 209–224, 227, 245–267, 276, 291–301, 305, 331–369, 381, 400, 414, 447–448
Monkey, 43, 45–48, 72, 87–89, 91–93
Mouse-tracking, 15, 26, 109, 115, 362
Multi-disciplinary, 1, 3, 6, 271, 418
Multimodal, 5, 291–301, 341, 359, 435–439, 448

N
Neurophysiology, 5

O
Object-based attention, 14, 271–286
Object recognition, 5, 18, 115, 117, 129, 271, 272, 274, 275, 278, 281, 285, 331, 358–360, 380, 382, 385–388, 391, 393, 395, 396, 415, 417–420, 433, 448, 451
Object saliency, 280, 366
Overt, 11, 16–18, 21–28, 283, 291, 297, 455

P
Parameters, 5, 63, 80, 89, 90, 96, 101, 125, 127, 128, 144, 171, 188, 213, 227–241, 245, 280, 296, 298, 299, 342, 348, 350, 386, 393, 421
Parietal lobe, 41, 50, 54
Philosophy, 5, 9–10, 18
Point cloud, 112, 308, 313, 317–326, 349
Pooling, 72–76, 82, 100–102, 118, 318, 343, 344, 367
Positron electron tomography (PET) scan, 14, 33–35
Proto-objects, 5, 272, 275–277, 279–281, 283, 284, 348, 358, 384, 418, 431, 452

Q
Quality assessment, 182–183, 343, 399–410

R
Researchers, 4, 6, 21, 202, 229, 238, 386, 400, 401, 403
Review, 4, 5, 36, 44, 141–173, 177–189, 209–223, 228, 272, 273, 291, 312–319, 328, 332, 358, 404
Robotics, 4, 5, 324, 332, 335, 340, 357–360, 413, 414, 419, 420, 423, 429

S
Salience, 39, 44, 46, 49, 50, 53, 64, 82, 83, 102, 129, 134, 278, 282, 293, 312, 313, 327, 328, 457
Salient object detection databases, 110, 142, 160–169, 171, 172, 191, 195–198, 200, 209–212, 231, 232, 240, 265, 281, 324, 351, 384
Sequential sampling, 79
Signal detection, 5, 63–83, 95, 219
State-of-the-art, 173, 186, 189, 210, 228, 231, 235, 237, 241, 245, 257, 260, 265, 327, 355, 426
Static saliency models, 154
Structure, 127, 128, 130–132, 134–138, 142, 152, 153, 169, 186, 272, 277, 305, 313, 316–320, 325, 331, 351–352, 360, 365, 366, 379, 383, 391, 403, 404, 416, 418, 424, 450–452, 455–457
Students, 4, 6, 202
Study, 2, 5, 9, 11, 12, 14, 18, 19, 28, 30, 32, 42, 46, 49, 52, 53, 76, 89, 93, 94, 128, 141, 161, 177, 192, 197, 209, 212, 224, 227–241, 273, 276, 277, 280, 296, 298, 314, 327, 340, 348, 361, 362, 364, 408, 409, 422, 437, 447, 450, 455

T
Temporal lobe, 41, 48, 50, 453
3D data, 305, 308, 312, 313, 319, 328, 331, 369, 424
3D saliency, 5, 111, 112, 305–306, 308–319, 328, 360, 366
Top-down, 17, 18, 41, 42, 46, 50, 53, 105, 107, 109, 114–119, 125, 126, 129–131, 135, 138, 148, 150, 178, 186, 250, 253, 272, 276–278, 283–285, 291, 320, 321, 327, 331, 341, 347, 350, 359, 360, 363, 364, 367, 369, 380, 383, 406, 418–420, 426, 431, 436, 448, 451, 452, 454, 456

V
Video(s), 4, 13, 23, 89, 110, 177–189, 191, 227, 257, 273, 291–301, 306, 331, 384, 399–410, 430, 448
Visual attention, 13, 22, 53, 87–102, 112, 123–138, 141, 182, 191, 227, 271, 291, 305, 341, 383, 405, 415, 448
Visual attention modeling, 161, 294
Visual neglect, 50–53
Visual saliency
  assessment, 227–241
  modeling, 305–328